

# ChatGPT and other generative AI tools

**Edited by**

Jochen Kuhn, Stefan Küchemann, Knut Neumann  
and Martina Rau

**Published in**

Frontiers in Psychology  
Frontiers in Education



## FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714  
ISBN 978-2-8325-6016-7  
DOI 10.3389/978-2-8325-6016-7

## About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: [frontiersin.org/about/contact](https://frontiersin.org/about/contact)

# ChatGPT and other generative AI tools

## Topic editors

Jochen Kuhn — Ludwig-Maximilians-Universität München, Germany

Stefan Küchemann — Ludwig Maximilian University of Munich, Germany

Knut Neumann — IPN–Leibniz-Institute for Science and Mathematics Education, Germany

Martina Rau — University of Wisconsin-Madison, United States

## Citation

Kuhn, J., Küchemann, S., Neumann, K., Rau, M., eds. (2025). *ChatGPT and other generative AI tools*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-6016-7

# Table of contents

05	<b>Editorial: ChatGPT and other generative AI tools</b> Stefan Küchemann, Martina Rau, Knut Neumann and Jochen Kuhn
08	<b>To use or not to use? Understanding doctoral students' acceptance of ChatGPT in writing through technology acceptance model</b> Min Zou and Liang Huang
17	<b>ChatGPT's advice is perceived as better than that of professional advice columnists</b> Piers Douglas Lionel Howe, Nicolas Fay, Morgan Saletta and Eduard Hovy
23	<b>Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings</b> Veronika Hackl, Alexandra Elena Müller, Michael Granitzer and Maximilian Sailer
31	<b>The relationship between student interaction with generative artificial intelligence and learning achievement: serial mediating roles of self-efficacy and cognitive engagement</b> Jing Liang, Lili Wang, Jia Luo, Yufei Yan and Chao Fan
43	<b>Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving</b> Karen D. Wang, Eric Burkholder, Carl Wieman, Shima Salehi and Nick Haber
54	<b>Adaptation and psychometric properties of a brief version of the general self-efficacy scale for use with artificial intelligence (GSE-6AI) among university students</b> Wilter C. Morales-García, Liset Z. Sairitupa-Sanchez, Sandra B. Morales-García and Mardel Morales-García
63	<b>Suggestive answers strategy in human-chatbot interaction: a route to engaged critical decision making</b> Yusuke Yamamoto
79	<b>ChatGPT in higher education: factors influencing ChatGPT user satisfaction and continued use intention</b> Chengcheng Yu, Jinzhe Yan and Na Cai
90	<b>Applying Generative Artificial Intelligence to cognitive models of decision making</b> Tyler Malloy and Cleotilde Gonzalez
106	<b>Developing valid assessments in the era of generative artificial intelligence</b> Leonora Kaldaras, Hope O. Akaze and Mark D. Reckase
116	<b>David vs. Goliath: comparing conventional machine learning and a large language model for assessing students' concept use in a physics problem</b> Fabian Kieser, Paul Tschisgale, Sophia Rauh, Xiaoyu Bai, Holger Maus, Stefan Petersen, Manfred Stede, Knut Neumann and Peter Wulff



- 134 **ChatGPT's quality: Reliability and validity of concept inventory items**  
Stefan Küchemann, Martina Rau, Albrecht Schmidt and Jochen Kuhn
- 146 **Exploring generative AI in higher education: a RAG system to enhance student engagement with scientific literature**  
Dominik Thüs, Sarah Malone and Roland Brünken
- 169 **Prompt engineering as a new 21st century skill**  
Denis Federiakin, Dimitri Molerov, Olga Zlatkin-Troitschanskaia and Andreas Maur



## OPEN ACCESS

EDITED AND REVIEWED BY  
Snehlata Jaswal,  
Sikkim University, India

\*CORRESPONDENCE  
Stefan Küchemann  
✉ s.kuechemann@lmu.de

RECEIVED 26 November 2024  
ACCEPTED 27 January 2025  
PUBLISHED 05 February 2025

CITATION  
Küchemann S, Rau M, Neumann K and Kuhn J  
(2025) Editorial: ChatGPT and other  
generative AI tools.  
*Front. Psychol.* 16:1535128.  
doi: 10.3389/fpsyg.2025.1535128

COPYRIGHT  
© 2025 Küchemann, Rau, Neumann and  
Kuhn. This is an open-access article  
distributed under the terms of the [Creative  
Commons Attribution License \(CC BY\)](#). The  
use, distribution or reproduction in other  
forums is permitted, provided the original  
author(s) and the copyright owner(s) are  
credited and that the original publication in  
this journal is cited, in accordance with  
accepted academic practice. No use,  
distribution or reproduction is permitted  
which does not comply with these terms.

# Editorial: ChatGPT and other generative AI tools

Stefan Küchemann<sup>1\*</sup>, Martina Rau<sup>2</sup>, Knut Neumann<sup>3</sup> and Jochen Kuhn<sup>1</sup>

<sup>1</sup>Chair of Physics Education Research, Faculty of Physics, Ludwig-Maximilians-Universität München (LMU Munich), Munich, Germany, <sup>2</sup>Chair of Research on Learning and Instruction, Department of Humanities, Social and Political Sciences, ETH Zurich, Zurich, Switzerland, <sup>3</sup>Department of Physics Education, IPN-Leibniz-Institute for Science and Mathematics Education, Kiel, Germany

## KEYWORDS

ChatGPT, personalized learning, large language model, learning with AI, critical thinking with AI, adaptive learning

## Editorial on the Research Topic ChatGPT and other generative AI tools

## 1 Introduction

In the past ten years, applications of generative artificial intelligence (GAI) have found rapidly growing use in medicine, science, and the daily life. Large language models (LLMs) opened up new avenues in particular for education. LLMs have been used to create interactive educational content for students, stimulate their curiosity, generate code explanations, and develop assessment questions (Küchemann et al., 2023). However, there are also several challenges when integrating GAI in education.

This Research Topic aimed to address issues around the use of GAI tools to advance students' cognition or, more broadly, competencies, and how to enable both teachers and students to critically reflect upon the use of GAI tools instead of overly relying on them.

The Research Topic focused on research on the meaningful use of large language model-based GAI tools such as ChatGPT for learning and cognition in order to foster critical reflection in the field on how GAI tools can be used to support teachers in formative assessment, diagnosing students' difficulties, implement novel cognitive activities and targeted interventions, and provide individualized attention to students.

This editorial synthesizes insights from 14 studies in this Research Topic that investigate the diverse impact of AI in higher education, highlighting key themes in acceptance, assessment, performance comparison, skill development, interaction strategies, and cognitive modeling.

## 2 Relevant student characteristics related to GAI use in education

The following studies indicate, that students' acceptance and student-centered integration of GAI tools in education are critical for leveraging their potential benefits. For instance, Zou and Huang reveal a high intention to use ChatGPT among doctoral students for academic writing. Utilizing the Technology Acceptance Model (TAM), they find that students' attitudes significantly predict their intention to use AI, mediated by perceived usefulness and ease of use. Past experiences with ChatGPT enhance perceived ease of use, underscoring the importance of familiarity with AI tools.

Expanding on the role of acceptance, [Yu et al.](#) examine factors influencing user satisfaction and continued use of ChatGPT among college students. Their findings indicate that compatibility and efficiency positively affect perceived ease of use and usefulness, which in turn influence satisfaction and the intention to continue using AI tools. These studies collectively suggest that positive experiences and perceived benefits are crucial for integrating AI into educational practices.

Furthermore, [Liang et al.](#) explore the relationship between student interaction with generative AI and learning achievement. Through a survey of 389 participants, they find that interaction with AI tools positively correlates with learning outcomes, mediated by increases in self-efficacy and cognitive engagement. This implies that GAI tools can enhance learning by stimulating students' confidence and active participation in the learning process.

For a reliable assessment of self-efficacy in GAI usage among university students, [Morales-García et al.](#) adapt the General Self-Efficacy Scale. The resulting GSE-6AI scale is validated and found to be both reliable and invariant across genders, providing a valuable instrument for assessing students' self-efficacy related to GAI in educational settings.

The implementation of GAI in education settings necessitates the development of new skills among learners and educators. In this line, [Federiakin et al.](#) introduce Prompt Engineering as a critical 21st-century skill. Defined as the ability to articulate problems, context, and constraints to an AI assistant effectively, Prompt Engineering ensures accurate and swift AI responses. The authors propose a conceptual framework encompassing comprehension of prompt structure, prompt literacy, prompting methods, and critical online reasoning. Recognizing and cultivating these skills is essential for maximizing the benefits of AI tools in education and beyond.

Apart from that, [Thüs et al.](#) demonstrate how GAI can stimulate learning processes. In their article, they introduce OwlMentor, a GAI-powered learning environment designed to assist students in comprehending scientific texts. By integrating features like document-based chats and automatic question generation, OwlMentor aims to enhance student engagement with scientific literature. The results indicate that higher learning gains among users of OwlMentor, emphasizing the importance of aligning GAI tools with students' learning strategies to maximize learning outcomes.

### 3 Transforming assessment and scoring

The following articles demonstrate that the implementation of GAI into assessment practices presents both opportunities and challenges. [Hackl et al.](#) evaluate GPT-4's reliability as a rater for student responses in macroeconomics tasks. Their analysis reveals high inter-rater reliability, with Intraclass Correlation Coefficients ranging from 0.94 to 0.99, indicating that GPT-4 can produce consistent and reliable ratings. This suggests that AI could play an important role in standardized assessments, reducing the burden on human evaluators.

However, [Kaldaras et al.](#) caution against uncritical adoption of GAI in assessments. They highlight the challenges of ensuring

that AI algorithms score the same constructs as human scorers and propose methods for evaluating the validity of GAI-generated assessments. Their work underscores the necessity of developing guidelines and methodologies to assess the validity of AI-based assessments and the inferences drawn from them.

Comparing AI with traditional methods, [Kieser et al.](#) find that conventional machine learning algorithms outperform a large language model in assessing students' concept use in physics problem-solving. This suggests that, in certain contexts, conventional AI algorithms may offer more accurate or efficient solutions than state-of-the-art GAI models, highlighting the importance of choosing appropriate AI tools for specific educational tasks.

Moreover, [Küchemann et al.](#) investigate the reliability and validity of concept inventory items generated by ChatGPT. After careful prompt engineering and selection, they create a set of physics concept questions that, while slightly lower in quality than human-generated items, are still viable for educational use. The study emphasizes the need for human oversight in generating assessment materials with AI to ensure alignment with learning objectives and student difficulties.

## 4 Analyses of GAI outputs

The comparison of GAI-generated outputs with human performance provides insight into the capabilities and limitations of GAI. [Howe et al.](#) conducted a study where participants compare advice from ChatGPT and professional advice columnists on social dilemmas. Surprisingly, ChatGPT's advice is perceived as more balanced, empathetic, and helpful, even when answer length is controlled. Although most participants prefer human advisors, the inability to distinguish between GAI and human responses raises questions about GAI's role in providing support and guidance.

During problem-solving of physics problems, [Wang et al.](#) examine GPT-4's ability to solve physics problems. While the AI model successfully solves 62.5% of well-specified problems, its performance drops significantly to 8.3% on under-specified, real-world problems. The identified reasons for failure—such as inaccurate physical modeling and unreasonable assumptions—highlight the current limitations of AI in complex, real-world applications and the necessity for human expertise in guiding AI use.

## 5 The role of human-GAI interaction in decision-making

Effective human-AI interaction strategies can significantly impact user engagement and decision-making. [Yamamoto](#) proposes a novel chatbot strategy employing suggestive endings inspired by the cliffhanger narrative technique. By ending responses with hints rather than conclusions, the chatbot stimulates users' curiosity and encourages deeper engagement. An online study demonstrates that users interacting with the suggestive chatbot ask more questions and engage in more prolonged decision-making processes, highlighting the potential of strategic AI communication to foster critical thinking.

In this line, Malloy and Gonzalez explore the application of GAI to cognitive models of decision-making. By categorizing existing applications and conducting an ablation study, they demonstrate that integrating GAI models to create memory representations and predict participant actions enhances model performance. This work provides valuable guidelines for cognitive modeling in human-AI collaboration frameworks, suggesting that AI can augment our understanding of human cognition and improve decision-making models.

## 6 Conclusions

The studies in this Research Topic highlight the impact that generative AI is having across various facets of higher education. From relevant students' characteristics, student engagement as well as enhancing learning outcomes to transforming assessment practices, GAI tools like ChatGPT are reshaping the educational landscape. However, the authors of the articles also point toward challenges, including ensuring the validity and reliability of GAI-generated content, addressing limitations in GAI problem-solving capabilities, and fostering critical engagement rather than overreliance on AI outputs. To this end, educators, researchers, and policymakers must navigate these complexities thoughtfully. Embracing GAI's potential requires not only integrating these tools into educational practices but also critically assessing their impact, limitations, and the skills needed to use them effectively. By aligning GAI tools with educational objectives and student needs, fostering essential skills like Prompt Engineering, and maintaining human oversight in critical areas, the educational community can harness the benefits of AI while mitigating its challenges.

## References

Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K. E., et al. (2023). Can chatgpt support prospective teachers in physics task development? *Phys. Rev. Phys. Educ. Res.* 19:020128. doi: 10.1103/PhysRevPhysEducRes.19.020128

## Author contributions

SK: Conceptualization, Writing – original draft, Writing – review & editing. MR: Writing – original draft, Writing – review & editing. KN: Writing – original draft, Writing – review & editing. JK: Writing – original draft, Writing – review & editing.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declare that Generative AI was used in the creation of this manuscript. Generative AI DeepL Write was used to improve writing.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



## OPEN ACCESS

## EDITED BY

Martina Rau,  
University of Wisconsin-Madison, United States

## REVIEWED BY

Edwin Ramirez-Asis,  
Lord of Sipan University, Peru  
Yueh-Min Huang,  
National Cheng Kung University, Taiwan

## \*CORRESPONDENCE

Liang Huang  
✉ eliot\_huang@163.com

RECEIVED 16 July 2023

ACCEPTED 16 October 2023

PUBLISHED 26 October 2023

## CITATION

Zou M and Huang L (2023) To use or not to use? Understanding doctoral students' acceptance of ChatGPT in writing through technology acceptance model.  
*Front. Psychol.* 14:1259531.  
doi: 10.3389/fpsyg.2023.1259531

## COPYRIGHT

© 2023 Zou and Huang. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# To use or not to use? Understanding doctoral students' acceptance of ChatGPT in writing through technology acceptance model

Min Zou<sup>1</sup> and Liang Huang<sup>2\*</sup>

<sup>1</sup>School of Foreign Languages, Beijing Institute of Technology, Beijing, China, <sup>2</sup>Department of Public Administration, Southeast University, Nanjing, China

While artificial intelligence-based chatbots have demonstrated great potential for writing, little is known about whether and how doctoral students accept the use of ChatGPT in writing. Framed with Technology Acceptance Model, this study investigated doctoral students' acceptance toward ChatGPT in writing and the factors that influence it. The questionnaire survey revealed a high intention to use ChatGPT in writing among doctoral students in China. The findings further indicated that attitude was a significant predictor of behavioural intention to use ChatGPT in writing and mediated the impacts of perceived usefulness and perceived ease of use on it. Perceived ease of ChatGPT use was in turn influenced by students' past ChatGPT use experience. This study provides powerful evidence for the applicability of Technology Acceptance Model in the acceptance of ChatGPT in writing. The results have significant implications for leveraging ChatGPT for writing in higher education.

## KEYWORDS

ChatGPT, writing, technology acceptance model, artificial intelligence-based chatbot, doctoral students

## 1. Introduction

Artificial intelligence (AI) technologies play a crucially important role in the increasingly digitalized world (Lee et al., 2022; Farrokhnia et al., 2023). As a generative AI chatbot, ChatGPT is a large language model that can autonomously learn from data and produce human-like texts (van Dis et al., 2023). It can converse on a wide range of topics and generate human-like responses after training huge quantities of text data (OpenAI, 2023). Ever since its release in November 2022, ChatGPT has sparked debates about its implications for education (Farrokhnia et al., 2023; Tlili et al., 2023; van Dis et al., 2023). While ChatGPT can potentially transform educational practices by providing a baseline knowledge of diverse topics (Tlili et al., 2023) and facilitating personalized, complex learning (Farrokhnia et al., 2023), it may supply incorrect texts, encourage cheating, and threaten academic integrity (Dwivedi et al., 2023; van Dis et al., 2023). The controversies have made ChatGPT "the most high-profile and controversial form of AI to hit education so far" (Williamson et al., 2023, p. 2).

Writing has been one of the most influenced domains in the ChatGPT era (Taecharungroj, 2023; Yan, 2023). While writing plays an important role in higher education (Kirkpatrick, 2019), it has been oftentimes considered challenging for language learners, especially for those who

learn and use English as an additional language (Ma, 2021). Prior research has suggested that chatbots are effective in addressing this challenge, since they could supply meaningful guidance and substantive feedback to support language learners to write at their own pace in a less anxiety-inducing environment and improve writing quality (Guo et al., 2022; Zhang et al., 2023). As a chatbot powered by generative AI, ChatGPT has demonstrated improved abilities than earlier chatbots (e.g., ELIZA) to understand natural language, generate appropriate responses, and engage in free-flowing conversations throughout the writing process, hence opening a new avenue for writing practice (Barrot, 2023; Su et al., 2023). As succinctly summarized by Imran and Almusharraf (2023), ChatGPT is “a complete package from generation to final proofreading and editing of writing material” (p.2). Nevertheless, till now, scarce attention has been paid to the acceptance and usage of ChatGPT in English writing—a daunting but critical work facing doctoral students (Kirkpatrick, 2019). Little is known about whether and how doctoral students intend to use ChatGPT in writing and the key determined factors. Informed by Technology Acceptance Model (TAM; Davis, 1989), the present study seeks to fill the void by addressing the following two questions: (1) how is the doctoral students’ acceptance intention to ChatGPT in writing? (2) what factors may influence doctoral students’ acceptance intention to ChatGPT in writing? Such information is important, as the individuals’ intention to adopt and use AI technology is critical to improving teaching and learning of writing (Cheng, 2019; Yan, 2023).

## 2. Literature review

### 2.1. The use of ChatGPT in writing

Chatbots, computer programs or AI systems designed to simulate human conversations and interact with users via natural language, have gained considerable attention and increasingly applied in writing in the past decade (Zhang et al., 2023). Chatbots have demonstrated great potential as a writing assistant and learning partner in writing classrooms, as they can provide a broad array of language choices and feedback to students’ writing process and make students feel less stressed about their writing performance in the learning process (Guo et al., 2022). ChatGPT was developed in 2022 as a novel chatbot rooted in Generative Pre-training Transformer architecture, and outperforms early chatbots in terms of the capability for understanding and producing human-like texts as well as providing feedback on long texts (Dwivedi et al., 2023; Farrokhnia et al., 2023; Su et al., 2023; Tlili et al., 2023). Such affordances make it a powerful writing assistant and writing tool (Barrot, 2023; Dergaa et al., 2023; Imran and Almusharraf, 2023). As shown in Taecharungroj’s (2023) analysis of early reactions on Twitter, ChatGPT has been most frequently used for writing, such as essays and articles.

Given the close link between ChatGPT and writing, a growing body of research has been undertaken to investigate the benefits and threats associated with the use of ChatGPT in writing. Piloting ChatGPT for academic writing, Bishop’s (2023) user experience demonstrated that ChatGPT is effective in explaining well-known concepts, translating between languages, giving timely and personalized feedback, adjusting the style and tone of texts to imitate different writers, and perfecting the mechanics of writing, thereby enhancing writing efficiency and promoting writing quality. Zooming

into the use of ChatGPT in second language writing context, Barrot (2023) and Su et al. (2023) further unpacked the potential of collaborating with ChatGPT in writing classrooms. For them, ChatGPT has taken into consideration various writing constructs, such as pragmatics, coherence and syntax, and could support the structural, dialogical and linguistic aspects of quality writing by assisting students in topic generation, outline preparation, content revision, proofreading and post-writing reflection. Taking stock of the research on ChatGPT in academia, Dergaa et al. (2023) and Imran and Almusharraf (2023) highlights the need to leverage ChatGPT as a valuable writing assistant tool to support the writing process and enhance academic writing.

Notwithstanding the benefits, the use of ChatGPT in writing has also raised concern for inaccurate and unintelligent responses, academic integrity, learning loss and educational inequality (Dwivedi et al., 2023; Farrokhnia et al., 2023; Tlili et al., 2023). As noted by the developer itself (OpenAI, 2023), “ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers.” Such incorrect and biased information can mislead students and be further incorporated into their writing, thereby harming knowledge practice and science progress (Tlili et al., 2023; van Dis et al., 2023). Another limitation of using ChatGPT in writing is associated with its unintelligent responses, typified by its frequent use of irrelevant statements, template rigidity of writing, and insufficiencies in emotional depth in writing (Barrot, 2023). Also, ChatGPT does not always reference sources appropriately and cannot be held accountable for their work, which raises pertinent issues concerning plagiarism and academic integrity (Dergaa et al., 2023; van Dis et al., 2023; Williamson et al., 2023; Yan, 2023). Additionally, the generative nature of ChatGPT allows students to complete writing assignments simply through unwitting copy-and-paste, and hence results in learning loss, especially when students become too reliant on the AI-powered chatbot for convenience (Barrot, 2023). Likewise, using ChatGPT in writing could lead to educational inequality (Dwivedi et al., 2023). Focusing on ChatGPT’s text generation functionality, for example, Yan’s (2023) research showed the undergraduates were much concerned with its impact on educational equity, given that writing teachers may not effectively distinguish texts produced by students from those produced by ChatGPT.

While the above user cases and scholarly discussions are helpful in unpacking the potentials and pitfalls of using ChatGPT in writing, the research into ChatGPT is still at its early stage (Barrot, 2023). Little empirical research has been conducted to examine the socio-technical aspects of using ChatGPT in writing. Since writing is essential to doctoral education (e.g., Kirkpatrick, 2019) and subject to the advances in AI technologies (Yan, 2023), it is necessary to explore and examine doctoral students’ intention toward ChatGPT and the influencing factors. Such information could shed light on doctoral students’ acceptance of ChatGPT in writing, and generate useful insights to leverage ChatGPT and other similar generative AI technologies for the teaching and learning of writing in higher education.

### 2.2. Technology acceptance model

User acceptance refers to the prospective users’ predisposition toward using technology (Lee and Lehto, 2013). TAM, emerging from the theory of reasoned action, has become an influential



socio-technical model that seeks to identify and explain the end-users' acceptance of technology (e.g., Cheng, 2019; Granić and Marangunić, 2019). In TAM, individuals' acceptance of a particular technology is operationalized as their behavioural intentions to use it (Lee and Lehto, 2013). TAM postulates that people's actual usage of technology is determined by their behavioural intentions. Behavioural intentions, in turn, are jointly determined by people's attitudes and perceived usefulness (Davis et al., 1989). Attitude towards technology underscores individuals' affective reactions to and evaluation of the use of the technology (Ajzen, 1991; Lee and Lehto, 2013) and it is closely related to one's intrinsic motivation (Davis et al., 1992). If people have a more favourable attitude toward the technology, they are more likely to form positive intentions to use it (Davis et al., 1989; Estriegana et al., 2019). Perceived usefulness is people's belief about the extent to which using the technology will improve their performance (Davis, 1989). It is a type of extrinsic motivation in determining technology acceptance and technology usage behaviour (Davis, 1989; Lee and Lehto, 2013). That is, if students believe that using the technology will improve their performance in writing, they tend to have a positive inclination to use it. The perceived usefulness is also hypothesized to have a positive influence on attitudes and thus affect behavioural intentions (Davis et al., 1989). If the technology is viewed as useful in enhancing writing performance, students are apt to appraise the technological means positively and inclined to use it (Estriegana et al., 2019). Therefore, this study proposes the following hypotheses.

**Hypothesis 1:** Attitude towards using ChatGPT in writing would significantly and positively influence students' behavioural intention to use ChatGPT in writing.

**Hypothesis 2:** Perceived usefulness of using ChatGPT would significantly and positively influence students' behavioural intention to use ChatGPT in writing.

**Hypothesis 3:** Perceived usefulness of using ChatGPT would significantly and positively influence students' attitude towards using ChatGPT in writing.

**Hypothesis 4:** Attitude towards using GPT would significantly mediate the effects of perceived usefulness on students' intention to use ChatGPT in writing.

Furthermore, TAM posits that attitude is jointly determined by perceived usefulness and perceived ease of use which refers to "the degree to which a person believes that using a particular system would be free of effort" (Davis, 1989, p.320). In TAM, perceived ease of use is assumed to have a significant effect on perceived usefulness and attitudes, resulting in increased behavioural intention (Davis et al., 1989; Alfadda and Mahdi, 2021). If the technological tool is perceived to be easy to use, students tend to consider it helpful and develop a favourable attitude, thereby demonstrating a strong inclination to use it in writing (Alfadda and Mahdi, 2021). Subsequently, the following hypotheses can be proposed.

**Hypothesis 5:** Perceived ease of use would significantly and positively influence students' perceived usefulness of ChatGPT in writing.

**Hypothesis 6:** Perceived ease of use would significantly and positively influence students' attitude towards using ChatGPT in writing.

**Hypothesis 7:** Attitude towards using GPT would significantly mediate the effects of perceived ease of use on students' intention to use ChatGPT in writing.

Meanwhile, a number of studies have revealed a strong and direct association between perceived ease of use and behavioural intention (Granić and Marangunić, 2019). In Yang and Wang's (2019) study, for instance, the perceived ease of use showed a significant and positive impact on students' behavioural intention to use machine translation. As argued by Shiao and Chau (2016), when people perceive that using a technological tool does not require much effort, they will be more intended to use it. Hence, the following hypothesis is proposed.

**Hypothesis 8:** Perceived ease of use would significantly and positively influence students' behavioural intention to use ChatGPT in writing.

According to Davis et al. (1989), perceived usefulness and perceived ease of use are influenced by a range of external variables, among which experience is one best studied external factor (Abdullah and Ward, 2016). The existing literature suggests that experience influences both learners' perceived usefulness (e.g., Chang et al., 2017; Yang and Wang, 2019) and perceived ease of use of educational technologies (e.g., Purnomo and Lee, 2013). For instance, Chang et al. (2017) found that students who have more experience in using computers tend to demonstrate more positive perceptions regarding the ease of use and usefulness of e-learning. Hence, this study assumes that students who have experience in using generative AI chatbots are more prone to understand usefulness of ChatGPT and become more proficient in using it in EFL writing. The following hypotheses are accordingly proposed.

**Hypothesis 9:** Past ChatGPT use experience would significantly and positively influence perceived usefulness of ChatGPT in writing.

**Hypothesis 10:** Past ChatGPT use experience would significantly and positively influence perceived ease of using ChatGPT in writing.

Taken together, and in line with the existing literature on TAM, a conceptual model is formulated in the present study (see Figure 1).

## 3. Research methodology

### 3.1. Participants

A total number of 242 doctoral students (151 males and 91 females) participated in the study through convenience samplings in one technological university in China. The students, ranging from 24 to 43 in age, were enrolled in the compulsory course entitled Writing for Academic Success taught by the first author. The course aims to empower doctoral students to improve English for academic writing

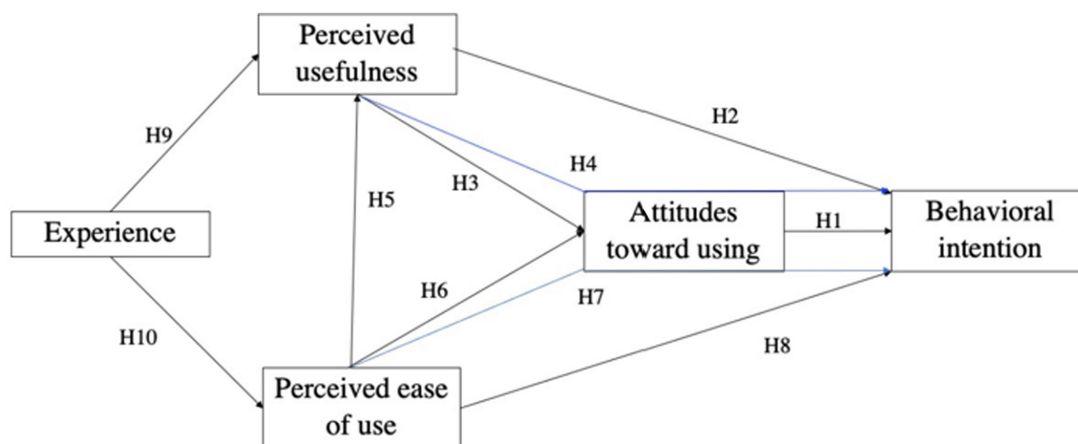


FIGURE 1  
Conceptual model.

skills. The participants were from different disciplinary backgrounds, such as computer science, mechanical engineering, materials science, economics, and education.

## 3.2. Measures

To determine doctoral students' acceptance of ChatGPT in writing and the factors influencing it, an online survey was administered in March 2023. The survey instrument consisted of two sections subsuming questions pertaining to demographic profiles (gender, major, and past ChatGPT use experience) and those concerning the constructs in TAM. The survey items in the second part were adapted from Davis (1989), Edmunds et al. (2012), Lee and Lehto (2013), and Rafique et al. (2020), and in light of the usage of ChatGPT in writing. In the second section, the respondents indicated their agreement level on every item by recording their response in a 6-point Likert scale, ranging from "1" (Strongly Disagree) to "6" (Strongly Agree).

### 3.2.1. Perceived ease of ChatGPT use in writing

Perceived ease of ChatGPT use in writing was measured based on a five-item scale adapted from Davis (1989). The five items (e.g., "I think ChatGPT is easy to use") showed high reliability (Cronbach's  $\alpha=0.854$ ). In light of Hu and Bentler's (1999) study, the Confirmatory Factor Analysis (CFA) results suggested good construct validity ( $\chi^2=9.445$ ,  $df=5$ ,  $RMSEA=0.061$ ,  $CFI=0.991$ ,  $TLI=0.982$ ), with factor loading ranging from 0.608 to 0.821.

### 3.2.2. Perceived usefulness of ChatGPT in writing

Perceived usefulness of using ChatGPT in writing was assessed by a five-item scale adapted from Davis (1989) and Rafique et al. (2020). The five items (e.g., "Using ChatGPT would enable me to finish English writing assignments effectively") demonstrated high reliability (Cronbach's  $\alpha=0.841$ ). The CFA results showed good construct validity ( $\chi^2=4.254$ ,  $df=5$ ,  $RMSEA=0.000$ ,  $CFI=1.000$ ,  $TLI=1.000$ ), with factor loading ranging from 0.637 to 0.785.

### 3.2.3. Attitude towards using ChatGPT in writing

Attitude towards using ChatGPT in writing was measured on a five-item scale adapted from Edmunds et al. (2012). The five items (e.g., 'I like using ChatGPT while writing in English') demonstrated excellent reliability (Cronbach's  $\alpha=0.915$ ). As indicated by Hu and Bentler (1999), the CFA results showed good construct validity ( $\chi^2=10.184$ ,  $df=5$ ,  $RMSEA=0.065$ ,  $CFI=0.994$ ,  $TLI=0.987$ ), with factor loading ranging from 0.775 to 0.879.

### 3.2.4. Behavioural intention to use ChatGPT in writing

Behavioural intention to use ChatGPT in writing was measured on a five-item scale adapted from Lee and Lehto (2013) and Rafique et al. (2020). The five items (e.g., "I intend to use ChatGPT to improve my English writing ability in the future") showed high reliability (Cronbach's  $\alpha=0.871$ ). According to Hu and Bentler (1999), the CFA results demonstrated good construct validity ( $\chi^2=7.976$ ,  $df=5$ ,  $RMSEA=0.050$ ,  $CFI=0.995$ ,  $TLI=0.990$ ), with factor loading ranging from 0.659 to 0.838.

### 3.2.5. Past ChatGPT use experience

In the present study, students' past ChatGPT use experience was operationalized as whether the students had used ChatGPT *de facto* at the time of data collection. It was measured via one item, i.e., "Have you ever used ChatGPT before?" The respondents indicated their past experience on a yes-no scale (Yes = 1, No = 0).

## 3.3. Data analysis

SPSS 24.0 and Mplus 7.4 Software were used for data analysis. First, the SPSS software was used to conduct descriptive analysis and correlation analysis. Then, the Mplus software was utilized to construct structural equation modelling (SEM), with a view to calculating relationships among focus variables and conduct mediation analysis. For mediation analysis, bias-corrected bootstrapping method with 2000 times of resampling was employed to calculate the point



estimates of the confidence intervals regarding the mediating effects. In light of Hu and Bentler's (1999) research, the fit of the model was evaluated by the following cut-off values: Root mean-square error of approximation (RMSEA) < 0.08; Tucker-Lewis index (TLI) > 0.90; and comparative fit index (CFI) > 0.90.

Additionally, Harman's single factor test was conducted by SPSS software to exclude possible common variance bias. The results showed that less than 50% (46.80%) of the total variance of variables were explained after all the items were loaded into one factor, indicating no need to control common variance bias (Mat Roni, 2014).

## 4. Results

### 4.1. Preliminary analysis

The descriptive statistics of all variables are presented in Table 1. Except for past ChatGPT use experience, the other four focus variables' score fall between 3.954 and 4.159, indicating mid-to-high levels on behavioural intentions, attitudes, perceived usefulness and perceived ease of use regarding ChatGPT. Particularly, the students reported the highest score on behavioural intention ( $M=4.159$ ), revealing doctoral students' high intention to use ChatGPT in writing in this study.

As suggested by the correlation matrix in Table 1, perceived ease of ChatGPT use ( $\gamma=0.590$ ,  $p<0.001$ ), perceived usefulness of ChatGPT ( $\gamma=0.632$ ,  $p<0.001$ ), and attitude towards using ChatGPT ( $\gamma=0.784$ ,  $p<0.001$ ) were significantly and positively correlated with students' behavioural intention to use ChatGPT in writing. Besides, both perceived ease of ChatGPT use ( $\gamma=0.688$ ,  $p<0.001$ ) and perceived usefulness of ChatGPT ( $\gamma=0.701$ ,  $p<0.001$ ) were significantly and positively correlated with doctoral students' attitude towards using ChatGPT in writing. Perceived ease of ChatGPT use was significantly and positively correlated with perceived usefulness of ChatGPT in writing ( $\gamma=0.660$ ,  $p<0.001$ ). Moreover, past ChatGPT use experience was significantly and positively correlated with students' perceived ease of ChatGPT use ( $\gamma=0.163$ ,  $p<0.05$ ), but it was not significantly correlated with perceived usefulness of ChatGPT in writing ( $\gamma=0.032$ ,  $p>0.05$ ).

### 4.2. Structural equation modelling

SEM analysis was conducted to examine the relationships among focus variables with gender being controlled for all the structural relationships. As shown in Figure 2, the model had a high explanation for variance in students' behavioural intention to use ChatGPT in writing (80.1%), attitude towards using ChatGPT (70.2%), and perceived usefulness of ChatGPT (65.7%), respectively, and a low explanation for variance in perceived ease of ChatGPT use (2.4%). The model fit indices ( $\chi^2=350.545$ ,  $df=198$ , RMSEA=0.056, CFI=0.951, TLI=0.943) indicates a good SEM model fit.

Perceived attitude towards using ChatGPT in writing had significant and positive impacts on students' behavioural intention to use ChatGPT in writing ( $\beta=0.850$ ,  $p<0.001$ ), supporting H1. Perceived usefulness of using ChatGPT had significant total influences on students' behavioural intention to use ChatGPT ( $\beta=0.577$ ,  $p<0.001$ ), but did not have significant and direct influences on it ( $\beta=0.117$ ,  $p>0.05$ ), thus rejecting H2. However, perceived usefulness of ChatGPT had positive and significant influences on students' attitude towards using ChatGPT in writing ( $\beta=0.541$ ,  $p<0.001$ ), thus supporting H3. Besides, perceived ease of use had significant and positive effects on students' perceived usefulness of ChatGPT in writing ( $\beta=0.817$ ,  $p<0.001$ ), thus supporting H5. Perceived ease of ChatGPT use had positive and significant influences on students' attitude towards using ChatGPT in writing ( $\beta=0.337$ ,  $p<0.001$ ), thereby supporting H6. Perceived ease of use had significant total influences on students' behavioural intention to use ChatGPT ( $\beta=0.689$ ,  $p<0.001$ ) but had no significant and direct influence on it ( $\beta=-0.069$ ,  $p>0.05$ ), rejecting H8. In addition, past ChatGPT use experience had significant and positive influences on students' perceived ease of using ChatGPT in writing ( $\beta=0.140$ ,  $p<0.05$ ) but had no significant influence on perceived usefulness of ChatGPT ( $\beta=-0.065$ ,  $p>0.05$ ). Therefore, the results supported H10 but rejected H9.

Additionally, results of mediation analysis (Table 2) show that students' attitude towards using ChatGPT significantly mediated the effects of perceived usefulness of ChatGPT on their behavioural intention to use ChatGPT in writing ( $\beta=0.460$ ,  $p<0.001$ , 95% CIs: 0.149 to 0.771), hence supporting H4. It also significantly mediated

TABLE 1 Results of descriptive statistics and correlation analysis.

	1	2	3	4	5
1. Behavioural intention to use ChatGPT in writing	1				
2. Attitude towards using ChatGPT in writing	0.784***	1			
3. Perceived usefulness of ChatGPT in writing	0.632***	0.701***	1		
4. Perceived ease of ChatGPT use in writing	0.590***	0.688***	0.660***	1	
5. Past ChatGPT use experience	0.093	0.132*	0.032	0.163*	1
Mean	4.159	3.954	4.106	4.017	0.463
SD	0.917	0.953	0.930	0.820	0.500

Standardized coefficients are reported.

\* $p<0.05$ ; \*\*\* $p<0.001$ .

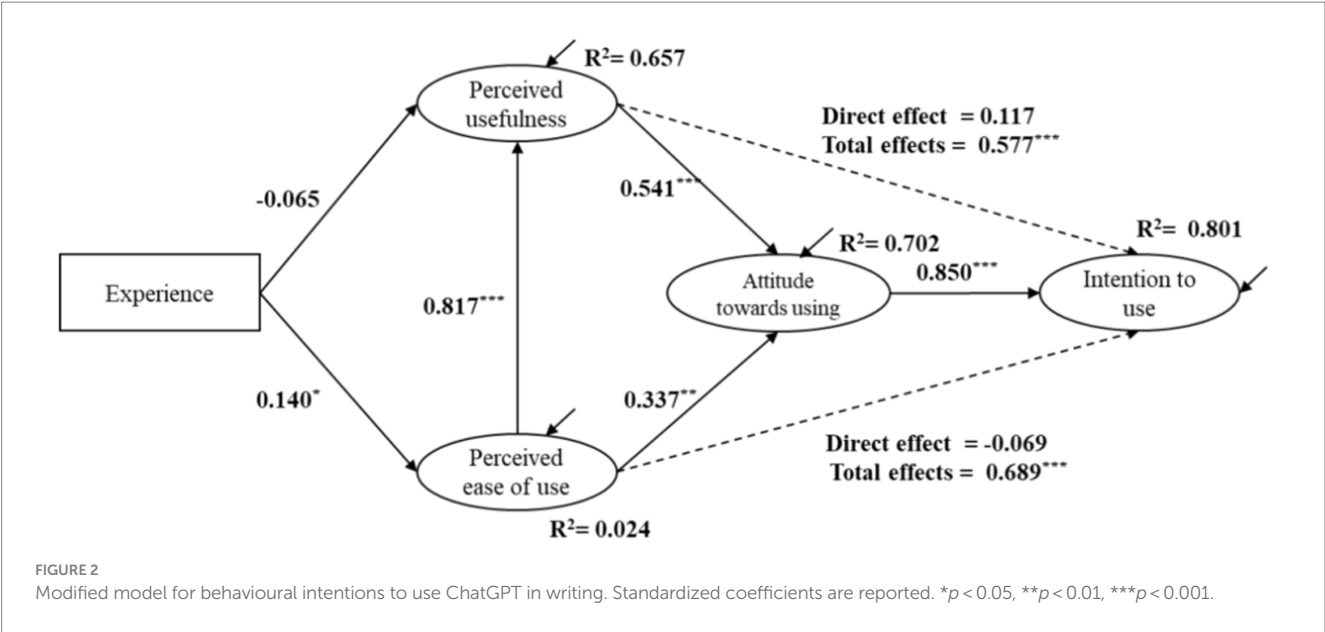


TABLE 2 Results of mediation analysis.

	$\beta$	S.E.	95% Confidence intervals
Perceived usefulness $\rightarrow$ behavioural intention to use ChatGPT in writing (Direct effect)	0.117	0.179	[−0.233, 0.467]
Perceived usefulness $\rightarrow$ attitude towards using $\rightarrow$ behavioural intention to use ChatGPT in writing	0.460**	0.159	[0.149, 0.771]
Perceived ease of use $\rightarrow$ behavioural intention to use ChatGPT in writing (Direct effect)	−0.069	0.137	[−0.337, 0.200]
Perceived ease of use $\rightarrow$ attitude towards using $\rightarrow$ behavioural intention to use ChatGPT in writing	0.287*	0.135	[0.022, 0.552]

Standardized coefficients are reported.  
\* $p < 0.05$ ; \*\* $p < 0.01$ .

the influences of perceived ease of ChatGPT use on students’ behavioural intention to use ChatGPT in writing ( $\beta = 0.287$ ,  $p < 0.05$ , 95% CIs: 0.022 to 0.552). Thus, H7 was supported.

5. Discussion

While ChatGPT has ignited debates about its applications in education (e.g., Farrokhnia et al., 2023), it remains unknown whether students are willing to use it or not in writing. This research contributes to the existing literature by investigating Chinese doctoral students’ acceptance toward ChatGPT in writing and its major influencing factors. Through the lens of TAM, the present study revealed a strong intention to use ChatGPT in writing among doctoral students, which was affected by their attitudes, perceived usefulness, and perceived ease of use. The findings provide a deeper understanding of doctoral students’ acceptance inclination toward ChatGPT and other generative AI chatbots in writing in higher education.

Although ChatGPT remains new, the doctoral students demonstrated a strong intention to use it in writing. This corroborates Taecharungroj’s (2023) finding that ChatGPT has been mainly used

in the writing domain. Students’ high behavioural intentions might be attributed to the affordances of ChatGPT for writing. As shown in prior research (e.g., Bishop, 2023; Yan, 2023), ChatGPT could help students to brainstorm ideas, obtain timely and personalized feedback, translate language items, and improve written drafts. This makes it a potential mediation tool for doctoral students to write more fluently and effectively in the publish-or-perish system (Kirkpatrick, 2019).

Consistent with our prediction, doctoral students’ attitude towards using ChatGPT in writing was found to be a significant predictor of behavioural intention. While a number of prior studies have removed attitudes from TAM due to its weak role in mediating the effects of perceived usefulness and perceived ease of use on behavioural intention (e.g., Lee and Lehto, 2013; Yang and Wang, 2019), this study found that attitude not only directly influences behavioural intention but also mediates the impacts of perceived usefulness and perceived ease of use on it. The finding lends support to the original TAM (Davis et al., 1989). It also supports Ajzen’s (1991) argument that personal attitude towards a behaviour functions as a major determinant of people’s intentions to perform it. In other words, when doctoral students have more positive evaluation of using ChatGPT in writing, they are more willing to perform the behaviour. Also, as suggested by

the expectancy-value model of attitudes (Ajzen, 1991; Ajzen and Fishbein, 2008), people's attitude is further determined by salient beliefs regarding the outcome of performing the behaviour and attributes associated with the behaviour, such as the cost and effort incurred by performing it. In this sense, positively valued outcomes and easier management of the technology could strengthen users' affective reactions towards the technology and boost their sense of efficacy, hence contributing to their favourable attitude towards it and the resultant increasing behavioural intention (Davis et al., 1989). As shown in this study, doctoral students' attitude towards using ChatGPT in writing, shaped by the perceived usefulness and ease of use, played an important role in mediating their effects on students' intention to use ChatGPT in writing.

Furthermore, the results revealed that perceived usefulness and perceived ease of use had significant total influences on students' behavioural intention to use ChatGPT in writing. This echoes the central role of perceived usefulness and perceived ease of use in the adoption process of technology in prior research examining TAM (Cheng, 2019; Granić and Marangunić, 2019; Alfadda and Mahdi, 2021). Nevertheless, the study found no significant direct influence of them on doctoral students' behavioural intention. Instead, they only influenced behavioural intention through attitudes. This surprising finding is inconsistent with previous studies on people's acceptance of educational technology (e.g., Estriegana et al., 2019; Yang and Wang, 2019). This might be due to the fact that some researchers (Davis, 1989; Lee and Lehto, 2013; Chang et al., 2017; Yang and Wang, 2019) did not include the attitude variable in their models and consequently failed to explore its mediating effects. Another plausible explanation might be that ChatGPT remains new, and early adopters use ChatGPT mainly because it facilitates inherently enjoyable and interesting experience (Taecharungroj, 2023; Tlili et al., 2023). In other words, the use of ChatGPT at this stage is primarily intrinsically motivated (Davis et al., 1992). Accordingly, the expected outcome of using ChatGPT for enhancing writing performance at the extrinsic level and perceived ease of using ChatGPT at the technical level could be instrumental, when such beliefs catalyse intrinsic motivations and when using ChatGPT in writing appeals to individuals (Ryan and Deci, 2000).

Also, the study found that perceived ease of use was found to be significantly and positively influenced perceived usefulness of ChatGPT in writing. This is analogous to Rafique et al.'s (2020) study, in which users' perceived ease of using mobile library applications had a significant influence on perceived usefulness. By the same token, users' perceived ease of using ChatGPT in writing could greatly shape the perceived usefulness (Davis et al., 1989). If doctoral students consider it challenging to apply ChatGPT in writing, they are likely to hold that ChatGPT has little effect on their writing. When they perceive ChatGPT easy to use, they tend to regard it as useful and helpful for writing.

In addition, this study extends prior research on TAM by including experience as an external factor to enhance the model explanatory power. Doctoral students' past ChatGPT experience is proved to be a significant predictor for perceived ease of use. The more experienced the students are, the more positive they are about the ease of using ChatGPT in EFL writing. This is compatible with Purnomo and Lee's (2013) study, where prior computer experience had a positive influence on learners' perceived ease of use and

e-learning system and such influence was stronger than that on perceived usefulness. The findings also support of argument Nelson's (1990) that the acceptance of technology relies upon not only the technology itself but also individuals' expertise in using it. Students with experience in using generative AI chatbots could employ the knowledge and skills obtained from prior experience to writing, develop a better personal control, and accordingly perceive it easier to use it in writing (e.g., Purnomo and Lee, 2013; Chang et al., 2017).

## 6. Conclusion

Despite the increasing interest in ChatGPT in educational settings, research on its acceptance is still scarce in education. Based on TAM, descriptive statistics, correlation analysis, and SEM were employed to gauge doctoral students' acceptance of ChatGPT in writing and explore the influencing factors. Data analysis revealed a high-level intention to use ChatGPT in writing, shaped by doctoral students' attitudes, perceived usefulness, and perceived ease of use. The present study could contribute to ChatGPT research in both theoretical and practical ways. Theoretically, the inclusion of experience in TAM helps to reveal the variables that could influence doctoral students' adoption of ChatGPT in EFL writing. As our model explained 80.1% of the variance in behavioural intention, this study overall supports and advances the applicability of TAM in ChatGPT, a new technology in writing education.

Practically, the results of the study could also generate useful implications for technology developers, policy-makers, writing teachers, and doctoral students to leverage ChatGPT for the teaching and learning of writing. Doctoral students' strong intention to use ChatGPT in writing suggest that ChatGPT may augment its function as an educational tool for writing in higher education. Considering the significant and strong effect of attitude on students' behavioural intentions to use ChatGPT in writing, it is of necessity for educational institutions, writing teachers, and technology developers to be aware of students' attitudes and increase their positive evaluation of and affective reactions towards using ChatGPT in writing. For instance, technology developer can make the usage of ChatGPT more innovative, enjoyable and interesting so as to create more positive attitudes and boost learners' intrinsic motivation to use ChatGPT in writing. Given the increasing concerns for information, ethical and learning risks associated with ChatGPT (e.g., Barrot, 2023; Dwivedi et al., 2023) and doctoral students' strong intention to use ChatGPT for writing, measures must be taken to mitigate such negative impacts of ChatGPT on doctoral students. For example, technology developers can strengthen the quality control of generated responses. Similarly, writing teachers need to provide trainings on effective, ethical and responsible use of ChatGPT in writing. Besides, perceived ease of use and perceived usefulness are found to have a significant influence on students' attitude, which could further exert an effect on students' intentions to use ChatGPT in writing. The sequential and circular influential relationship among the variables implies a need for technology developers to increase the usefulness and ease of using ChatGPT in writing to make it more functional and user-friendly. For

example, technology developers can keep simplifying and optimizing the operation of ChatGPT based on user feedback and provide comprehensible instructions or use cases regarding how to apply ChatGPT to write more effectively and ethically. Instead of prohibiting the use of ChatGPT in writing, policy makers need to take into consideration the students' voice and align their educational needs with the AI tool (EDUCAUSE, 2023). For writing teachers and institutional administrators, efforts to integrate ChatGPT in writing courses or training programs are needed to capitalize on ChatGPT's affordances for writing and improve students' ability to use ChatGPT as an effective writing assistant tool. Given the significant effect of past ChatGPT experience on perceived ease of use, instructing doctoral students to increase their use of ChatGPT, and reflect upon and communicate the skills for utilizing ChatGPT to promote writing performance could be an effective way to develop their expertise in ChatGPT. Also, doctoral students can experiment with ChatGPT in a conscious manner, and record their hands-on experience to continuously improve the capability for effective and ethical use of ChatGPT for writing.

Regardless of the contributions, there are several limitations that need to be taken into consideration in future research. Firstly, while the study revealed a high intention to use ChatGPT in writing among doctoral students, it was exploratory in nature and only used questionnaires to gauge students' acceptance of ChatGPT. Future research can thus employ case study research design or mixed study research design and collect multiple sources of data (e.g., semi-structured interviews, user reflections, and screenshots) to obtain an idiosyncratic and in-depth understanding of students' actual process and outcome of using ChatGPT in writing. Secondly, the present study was based on a sample of doctoral students from a science and technology university in China. The types of writing assignments they face and their needs for using ChatGPT to improve writing could be very different from other learner groups like undergraduates (Yan, 2023) and students in other countries, which limits the generalizability of this study. Therefore, future research can expand the sample scope to include students with varied educational levels and backgrounds to increase the generalizability and representativeness. It may also be interesting to conduct cross-section research to examine whether the level of use acceptance across different learner groups in the future. Thirdly, our data was collected from participants who interacted with ChatGPT shortly after the release of ChatGPT and who used ChatGPT primarily for its inherently enjoyable and interesting experience (Taecharungroj, 2023; Tlili et al., 2023). Given the increasing ethical, learning and information concerns concerning the use of ChatGPT in writing in academia (Barrot, 2023; Su et al., 2023) and students' growing experience, knowledge and skills regarding ChatGPT, their attitudes, perceptions and intentions of using ChatGPT in writing may alter over time. Longitudinal research can be conducted to trace the development of knowledge concerning the use of ChatGPT for writing among doctoral students, and how such knowledge influences their attitudes towards, as well as perceptions and intentions of using ChatGPT in writing. Considering the doctoral students' high intention to use ChatGPT for writing and the increasing concerns for information, ethical and

learning risks associated with ChatGPT (e.g., Barrot, 2023; Dwivedi et al., 2023), it is also promising to explore effective ways to integrate ChatGPT in writing instruction and construct writing models to empower students to collaborate with ChatGPT in an effective, ethical and responsible manner.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving humans were approved by the ethics committee of the School of Foreign Languages, Beijing Institute of Technology. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## Author contributions

MZ: Conceptualization, Funding acquisition, Investigation, Project administration, Writing – original draft. LH: Conceptualization, Formal analysis, Methodology, Software, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by Beijing Association of Higher Education under Grant MS2022225 and 2023 Beijing Institute of Technology Science and Technology Innovation Program "BIT Think Tank" Advancement Plan Funding Project under Grant 2023CX13030.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



## References

- Abdullah, F., and Ward, R. (2016). Developing a general extended technology acceptance model for E-learning (GETAMEL) by analysing commonly used external factors. *Comput. Hum. Behav.* 56, 238–256. doi: 10.1016/j.chb.2015.11.036
- Ajzen, I. (1991). The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* 50, 179–211. doi: 10.1016/0749-5978(91)90020-T
- Ajzen, I., and Fishbein, M. (2008). Scaling and testing multiplicative combinations in the expectancy–value model of attitudes. *J. Appl. Soc. Psychol.* 38, 2222–2247. doi: 10.1111/j.1559-1816.2008.00389.x
- Alfadda, H. A., and Mahdi, H. S. (2021). Measuring students' use of zoom application in language course based on the technology acceptance model (TAM). *J. Psycholinguist. Res.* 50, 883–900. doi: 10.1007/s10936-020-09752-1
- Barrot, J. S. (2023). Using ChatGPT for second language writing: pitfalls and potentials. *Assess. Writ.* 57:100745. doi: 10.1016/j.asw.2023.100745
- Bishop, L. (2023). A computer wrote this paper: what chatgpt means for education, research, and writing. *SSRN Electron. J.* doi: 10.2139/ssrn.4338981
- Chang, C. T., Hajiyeve, J., and Su, C. R. (2017). Examining the students' behavioral intention to use e-learning in Azerbaijan? The general extended technology acceptance model for e-learning approach. *Comput. Educ.* 111, 128–143. doi: 10.1016/j.compedu.2017.04.010
- Cheng, E. W. (2019). Choosing between the theory of planned behavior (TPB) and the technology acceptance model (TAM). *Educ. Technol. Res. Dev.* 67, 21–37. doi: 10.1007/s11423-018-9598-6
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 13, 319–340. doi: 10.2307/249008
- Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Manag. Sci.* 35, 982–1003. doi: 10.1287/mnsc.35.8.982
- Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. (1992). Extrinsic and intrinsic motivation to use computers in the workplace. *J. Appl. Soc. Psychol.* 22, 1111–1132. doi: 10.1111/j.1559-1816.1992.tb00945.x
- Dergaa, I., Chamari, K., Zmijewski, P., and Saad, H. B. (2023). From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. *Biol. Sport* 40, 615–622. doi: 10.5114/biolSport.2023.125623
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., et al. (2023). “So what if ChatGPT wrote it?” multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int. J. Inf. Manag.* 71:102642. doi: 10.1016/j.ijinfomgt.2023.102642
- Edmunds, R., Thorpe, M., and Conole, G. (2012). Student attitudes towards and use of ICT in course study, work and social activity: a technology acceptance model approach. *Br. J. Educ. Technol.* 43, 71–84. doi: 10.1111/j.1467-8535.2010.01142.x
- EDUCAUSE. (2023). *2023 EDUCAUSE horizon report, teaching and learning edition*. Boulder, Colorado: EDUCAUSE.
- Estriegana, R., Medina-Merodio, J. A., and Barchino, R. (2019). Student acceptance of virtual laboratory and practical work: an extension of the technology acceptance model. *Comput. Educ.* 135, 1–14. doi: 10.1016/j.compedu.2019.02.010
- Farrokhnia, M., Banihashem, S. K., Noroozi, O., and Wals, A. (2023). A SWOT analysis of ChatGPT: implications for educational practice and research. *Innov. Educ. Teach. Int.*, 1–15. doi: 10.1080/14703297.2023.2195846
- Granić, A., and Marangunić, N. (2019). Technology acceptance model in educational context: a systematic literature review. *Br. J. Educ. Technol.* 50, 2572–2593. doi: 10.1111/bjet.12864
- Guo, K., Wang, J., and Chu, S. K. W. (2022). Using chatbots to scaffold EFL students' argumentative writing. *Assess. Writ.* 54:100666. doi: 10.1016/j.asw.2022.100666
- Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model. Multidiscip. J.* 6, 1–55. doi: 10.1080/10705519909540118
- Imran, M., and Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: a systematic review of the literature. *Contemp. Educ. Technol.* 15:ep464. doi: 10.30935/cedtech/13605
- Kirkpatrick, K. J. (2019). Online doctoral students writing for scholarly publication. *Comput. Compos.* 52, 19–36. doi: 10.1016/j.compcom.2019.01.012
- Lee, Y. F., Hwang, G. J., and Chen, P. Y. (2022). Impacts of an AI-based chatbot on college students' after-class review, academic performance, self-efficacy, learning attitude, and motivation. *Educ. Technol. Res. Dev.* 70, 1843–1865. doi: 10.1007/s11423-022-10142-8
- Lee, D. Y., and Lehto, M. R. (2013). User acceptance of YouTube for procedural learning: an extension of the technology acceptance model. *Comput. Educ.* 61, 193–208. doi: 10.1016/j.compedu.2012.10.001
- Ma, L. P. F. (2021). Writing in English as an additional language: challenges encountered by doctoral students. *High. Educ. Res. Dev.* 40, 1176–1190. doi: 10.1080/07294360.2020.1809354
- Mat Roni, S. (2014). *Introduction to SPSS*. Edith Cowan University, SOAR Centre.
- Nelson, D. L. (1990). Individual adjustment to information-driven technologies: a critical review. *MIS Q.* 14, 79–98. doi: 10.2307/249311
- OpenAI (2023). ChatGPT: optimizing language models for dialogue. Available at: <https://openai.com/blog/chatgpt/> (Accessed October 10, 2023).
- Purnomo, S. H., and Lee, Y. H. (2013). E-learning adoption in the banking workplace in Indonesia: an empirical study. *Inf. Dev.* 29, 138–153. doi: 10.1177/0266666912448258
- Rafique, H., Almagrabi, A. O., Shamim, A., Anwar, F., and Bashir, A. K. (2020). Investigating the acceptance of mobile library applications with an extended technology acceptance model (TAM). *Comput. Educ.* 145:103732. doi: 10.1016/j.compedu.2019.103732
- Ryan, R. M., and Deci, E. L. (2000). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* 25, 54–67. doi: 10.1006/ceps.1999.1020
- Shiau, W. L., and Chau, P. Y. (2016). Understanding behavioral intention to use a cloud computing classroom: a multiple model comparison approach. *Inf. Manag.* 53, 355–365. doi: 10.1016/j.im.2015.10.004
- Su, Y., Lin, Y., and Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assess. Writ.* 57:100752. doi: 10.1016/j.asw.2023.100752
- Taecharungroj, V. (2023). “What can ChatGPT do?” analyzing early reactions to the innovative AI Chatbot on twitter. *Big Data Cogn. Comp.* 7:35. doi: 10.3390/bdcc7010035
- Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., et al. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learn. Environ.* 10:15. doi: 10.1186/s40561-023-00237-x
- van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., and Bockting, C. L. (2023). ChatGPT: five priorities for research. *Nature* 614, 224–226. doi: 10.1038/d41586-023-00288-7
- Williamson, B., Macgilchrist, F., and Potter, J. (2023). Re-examining AI, automation and datafication in education. *Learn. Media Technol.* 48, 1–5. doi: 10.1080/17439884.2023.2167830
- Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies*.
- Yang, Y., and Wang, X. (2019). Modeling the intention to use machine translation for student translators: an extension of technology acceptance model. *Comput. Educ.* 133, 116–126. doi: 10.1016/j.compedu.2019.01.015
- Zhang, R., Zou, D., and Cheng, G. (2023). Chatbot-based learning of logical fallacies in EFL writing: perceived effectiveness in improving target knowledge and learner motivation. *Interact. Learn. Environ.*, 1–18. doi: 10.1080/10494820.2023.2220374



## OPEN ACCESS

## EDITED BY

Ion Juvina,  
Wright State University, United States

## REVIEWED BY

K. D. Valentine,  
Massachusetts General Hospital and Harvard  
Medical School, United States  
Ralf Schmaelzle,  
Michigan State University, United States

## \*CORRESPONDENCE

Piers Douglas Lionel Howe  
✉ pldhowe@unimelb.edu.au

RECEIVED 22 August 2023

ACCEPTED 30 October 2023

PUBLISHED 21 November 2023

## CITATION

Howe PDL, Fay N, Saletta M and Hovy E (2023)  
ChatGPT's advice is perceived as better than  
that of professional advice columnists.  
*Front. Psychol.* 14:1281255.  
doi: 10.3389/fpsyg.2023.1281255

## COPYRIGHT

© 2023 Howe, Fay, Saletta and Hovy. This is an  
open-access article distributed under the terms  
of the [Creative Commons Attribution License  
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction  
in other forums is permitted, provided the  
original author(s) and the copyright owner(s)  
are credited and that the original publication in  
this journal is cited, in accordance with  
accepted academic practice. No use,  
distribution or reproduction is permitted which  
does not comply with these terms.

# ChatGPT's advice is perceived as better than that of professional advice columnists

Piers Douglas Lionel Howe<sup>1\*</sup>, Nicolas Fay<sup>2</sup>, Morgan Saletta<sup>3</sup> and  
Eduard Hovy<sup>4</sup>

<sup>1</sup>Complex Human Data Hub, Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, VIC, Australia, <sup>2</sup>School of Psychological Science, University of Western Australia, Perth, WA, Australia, <sup>3</sup>Hunt Laboratory, University of Melbourne, Melbourne, VIC, Australia, <sup>4</sup>Melbourne Connect, University of Melbourne, Melbourne, VIC, Australia

ChatGPT is a high-performance large language model that has the potential to significantly improve human-computer interactions. It can provide advice on a range of topics, but it is unclear how good this advice is relative to that provided by competent humans, especially in situations where empathy is required. Here, we report the first investigation of whether ChatGPT's responses are perceived as better than those of humans in a task where humans were attempting to be empathetic. Fifty social dilemma questions were randomly selected from 10 well-known advice columns. In a pre-registered survey, participants ( $N = 404$ ) were each shown one question, along with the corresponding response by an advice columnist and by ChatGPT. ChatGPT's advice was perceived as more balanced, complete, empathetic, helpful, and better than the advice provided by professional advice columnists (all values of  $p < 0.001$ ). Although participants could not determine which response was written by ChatGPT (54%,  $p = 0.29$ ), most participants preferred that their own social dilemma questions be answered by a human than by a computer (77%,  $p < 0.001$ ). ChatGPT's responses were longer than those produced by the advice columnists (mean 280.9 words vs. 142.2 words,  $p < 0.001$ ). In a second pre-registered survey, each ChatGPT answer was constrained to be approximately the same length as that of the advice columnist (mean 143.2 vs. 142.2 words,  $p = 0.95$ ). This survey ( $N = 401$ ) replicated the above findings, showing that the benefit of ChatGPT was not solely due to it writing longer answers.

## KEYWORDS

ChatGPT, empathy, advice column, agony aunt, advice

## 1 Introduction

ChatGPT, a groundbreaking artificial intelligence (AI) generative large language model (OpenAI, 2023), has recently garnered widespread attention due to its adeptness in various natural language processing tasks. Launched in November 2022, it experienced an unprecedented adoption rate, amassing over a million users in just 5 days and reaching 1.6 billion users by June 2023. Its creation marked a revolution in the industry, ushering in a new era of AI chatbots (Gohil, 2023).

It has also sparked significant interest within the academic community, leading to a wealth of scholarly literature (Kaddour et al., 2023; Ray, 2023). Illustratively, Katz et al. (2023) demonstrated that GPT-4 with zero-shot prompting could successfully pass the full United States

legal Uniform Bar Exam, outperforming 90% of human participants. Similarly, Wu et al. (2023) showed that an enhanced version of GPT 3.5-Turbo could pass the Chinese Medical Licensing Examination, again surpassing the average human performance.

While ChatGPT's technical prowess has been illustrated in various professional contexts, its capacity for nuanced human interactions remains an area of pivotal interest. Of particular interest is how well it can interact with humans in situations where it would need to convey empathy. Empathy plays a vital role in many domains (Hoffman, 2000; Sanders et al., 2021); if ChatGPT were to fail to exhibit sufficient empathy, this would adversely affect the quality of its interactions with humans (Leite et al., 2013). Indeed, numerous studies have argued that empathy is crucial for effective communication (Riess, 2017; Pounds et al., 2018; Janich, 2020) and that people are more persuasive when they appear to be empathetic (Lancaster, 2015). For reviews of the role of empathy in communication, please see Berger et al. (2010) and Floyd and Weber (2020).

The few studies that have explored the degree of empathy conveyed by ChatGPT reported that its responses often lacked empathy (Kalla and Smith, 2023; Sun et al., 2023; Zhao et al., 2023). GPT 3.5-Turbo performed poorly compared to the state of the art because it focused more on giving advice than addressing the user's emotional needs (Zhao et al., 2023). Even GPT-4 was reported as having difficulty expressing empathy in a convincing fashion (Sun et al., 2023). However, these studies did not benchmark ChatGPT's capabilities against those of humans.

It is necessary to compare ChatGPT to humans because if ChatGPT is perceived to perform worse than humans, it is likely that users will choose to interact with humans rather than with it. In a study reported in Ayers et al. (2023), human participants saw a series of medical questions that had been placed on Reddit's r/AskDocs forum, the responses written by verified physicians and the responses written by GPT-3.5. Ayers et al. (2023) reported that participants rated the GPT-3.5 responses as being of higher quality than those of the physicians. A similar study was conducted by Liu et al. (2023) who compared physician response to 10 patient questions to the responses generated by GPT-3.5 and GPT-4. Liu et al. (2023) found that the responses by GPT-3.5 and GPT-4 were perceived as of higher quality than those written by the physicians.

One limitation of the above studies is that the physicians' responses may not reflect typical doctor-patient interactions. Normally, doctors would spend some time explaining their diagnosis to the patient, ensuring that the patient felt heard and respected. Conversely, the physicians' responses in Ayers et al. (2023) were notably brief, averaging just 52 words, and sometimes as short as 17 words. Similarly, the physician responses in Liu et al. (2023) averaged 50 words and were sometimes as short as 20 words. In both studies, the physicians were focused on brevity and on conveying medical information, and not on addressing the emotional needs of the patient. It was therefore not appropriate to compare the empathy expressed in these responses to the empathy expressed in the responses by ChatGPT, as the physicians were often not attempting to be empathetic.

In our study, we assessed ChatGPT's ability to provide advice in a situation where humans attempted to be empathetic. Specifically, we compared the responses of ChatGPT and humans to a series of social dilemma questions that had been submitted to a range of social

advice columns (aka "agony aunt" columns). Our results suggest that ChatGPT can outperform humans in this domain.

## 2 Survey 1

### 2.1 Methods

We selected 10 newspaper advice columns: *Ask a Manager*, *Ask Amy*, *Ask E. Jean*, *Ask Ellie*, *Dear Abby*, *Dear Annie*, *Dear Prudence*, *Miss Manners*, *Social Q's*, and *The Ethicist*. These columns were chosen because they were well-known and fielded a wide range of questions that we could access. For each column, we selected at random five questions. These questions were posted between November 2019 and June 2023. For each social dilemma question, we initiated a new chatbot session, ensuring that ChatGPT generated responses without any carryover context from previous questions. This was done using GPT-4 on the June 14, 2023. As we were interested in studying its default response, ChatGPT was not asked to be empathetic. For each question, we used the following prompt "Please respond to the following question [Social dilemma question text inserted here]." ChatGPT's response and the response of the advice columnist were stripped of any identity-revealing information (e.g., "I am a chatbot" or "I am an advice columnist"). We always took ChatGPT's first response. Both this and the subsequent study were approved the Human Research Ethics Committee at the University of Western Australia (2023/ET000523).

Participants in our study were each presented with just a single social dilemma question and the two answers (from the original advice column and from ChatGPT), without disclosing the origin of the answers. Thus, each of the 50 dilemmas were viewed, on average, by approximately eight participants. After viewing the question and corresponding answers, participants responded to a series of binary questions that evaluated the perceived quality of the answers provided.

In a series of binary questions, participants were asked which of the two answers was more balanced, more comprehensive, more empathetic, more helpful, and better. Following these assessments, we disclosed that one of the responses had been composed by a human and the other by a computer, and asked the participants to identify the computer-generated response. Finally, participants were asked to imagine a scenario where they had a question regarding a social dilemma and to indicate whether they would prefer this question be answered by a computer or by a human (i.e., a binary response).

To calculate an appropriate sample size for our study, we conducted a binomial power analysis (Champely, 2020). Assuming a significance level of 0.05 (two-sided), a null hypothesis of 0.5, and an alternative hypothesis of 0.6, the analysis revealed that we would require a sample size of 387 participants to achieve a statistical power of 0.8. This power level ensures a reasonably high probability of detecting a true effect if one exists. Based on this analysis, we decided to recruit 400 participants for the study.

Participants were recruited from Amazon's Mechanical Turk (MTurk), a popular crowd-sourcing marketplace frequently used in psychological and behavioral research. To ensure the quality of data, we only recruited from a pool of MTurk workers who had previously been pre-screened to verify they were not bots. Additionally, our study

was pre-registered to promote transparency and reproducibility in our research: <https://aspredicted.org/66n24.pdf>.

## 2.2 Results

A total of 404 participants were recruited. Two were excluded as their data did not record properly, thereby preventing analysis. The data were analyzed using the tidyverse (Wickham, 2017), lme4 (Bates et al., 2015), lmerTest (Kuznetsova et al., 2017), purr (Henry and Wickham, 2021), and broom.mixed (Siebert, 2021) software packages in R (R Core Team, 2020). Participants' mean age was 42.4 years (standard deviation = 12.1 years). The gender distribution was 156 female, 240 male, two non-binary, with four participants preferring not to disclose. The responses to the first five questions are depicted in Figure 1. Remarkably, for every question, ChatGPT clearly outperformed the professional advice columnists. Participants were not able to reliably identify which answer was written by the computer (only 54% succeeded). Despite this, the majority of participants (77%) indicated a preference for having their hypothetical social-dilemma questions answered by a human rather than by a computer.

In the pre-registration, we specified the use of binomial tests. However, post-experiment, we recognized that these tests failed to account for multiple subjects encountering the same social dilemma. To rectify this, we redid the analysis using a linear mixed-effects model, incorporating 'dilemma' as a random effect. Despite the binary nature of the dependent variables, we opted for linear models to gain unbiased estimates of our predictor variables' causal effects (Gomila, 2021). The results of these statistical analyses are shown in Table 1.

Although not pre-registered, we also measured the word count for the official advice column answers and the answers written by ChatGPT. The word count for the official answers was considerably less than that for ChatGPT, with mean word count of 142.2 and 280.9 words, respectively. This difference was statistically significant,  $t(88.9) = 9.12, p < 0.001$ .

## 3 Survey 2

The second survey was identical the first survey except that, for each question, ChatGPT was requested to write an answer that was not longer than the official answer for that question. To do this, we used the following prompt: "Please respond to the following question in less than X words [Social dilemma question text inserted here]," where X was the word length of the official response. The survey was separately pre-registered: <https://aspredicted.org/h5pk8.pdf>.

A total of 401 participants were recruited. One was excluded because their data were corrupted. Participants' mean age was 42.8 years (standard deviation = 12.5 years). The gender distribution was 187 female, 208 male, three non-binary, with two participants preferring not to disclose. While the ChatGPT answers were rarely exactly the same length as the corresponding official answer, on average they were very similar, with mean word counts of 142.2 and 143.2 words for the official answer and ChatGPT's answer, respectively. This difference was not statistically different,  $t(97.7) = 0.06, p = 0.95$ .

As before, participants felt that the answers given by ChatGPT were more balanced, more complete, more empathetic, more helpful

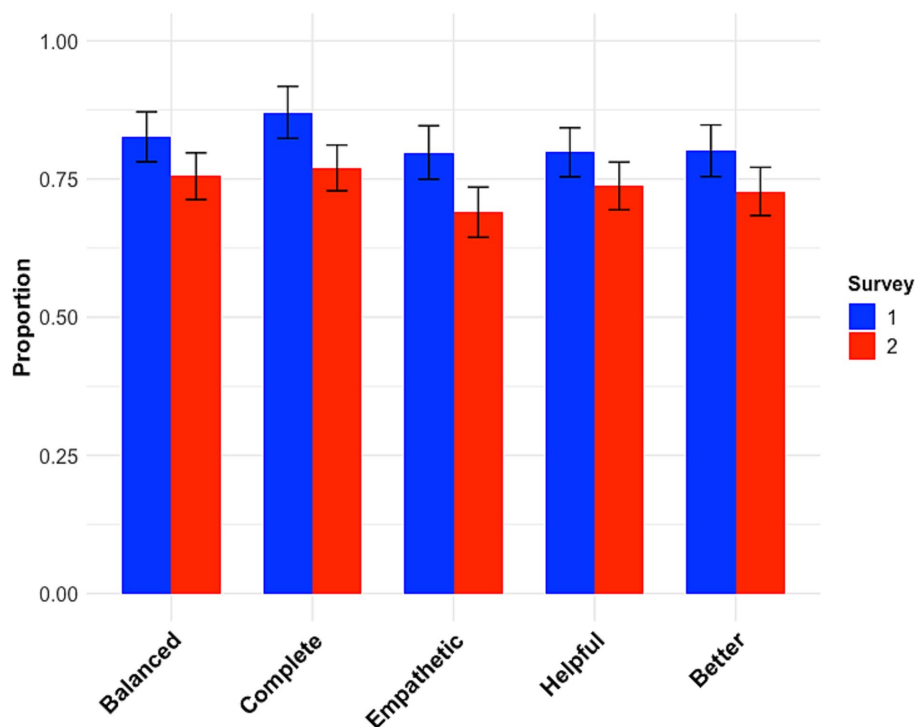


FIGURE 1

The proportion of participants who thought the answer provided by ChatGPT was more balanced, more complete, more empathetic, more helpful, and better than that provided by the professional advice columnist. (A) Survey 1. (B) Survey 2. Error bars represented 95% CI.



TABLE 1 Statistical analysis for Surveys 1 and 2.

Question	Survey 1	Survey 2
Which answer do you think was more balanced?	$t(49.9) = 14.1, p < 0.001$	$t(399) = 11.8, p < 0.001$
Which answer do you think was more complete?	$t(49.2) = 15.5, p < 0.001$	$t(399) = 12.8, p < 0.001$
Which answer do you think was more empathetic?	$t(48.3) = 12.1, p < 0.001$	$t(399) = 8.2, p < 0.001$
Which answer do you think was more helpful?	$t(47.1) = 13.2, p < 0.001$	$t(399) = 10.8, p < 0.001$
Which answer do you think was better?	$t(49.6) = 12.6, p < 0.001$	$t(399) = 10.2, p < 0.001$
One of these answers was written by a computer. Which one do you think it was?	$t(48.4) = 1.08, p = 0.29$	$t(399) = 0.5, p = 0.62$
Assuming you had a social dilemma question and to get it answered you would need to put it in writing and receive a written response, would you prefer your question to be answered by a human or by a computer?	$t(49.0) = 11.3, p < 0.001$	$t(399) = 19.0, p < 0.001$

This table shows the results of the *t*-test for each question, for both Survey 1 and Survey 2. A statistically significant result shows that the proportion of participants choosing the ChatGPT answer over the human answer for that question in that survey was different from 50%. In other words, ChatGPT's answer and the human answer were not equally preferable. Figure 1 shows the proportion of participants who preferred ChatGPT's answer over the human answer.

and better than the official answers (Figure 1; Table 1). As before, participants were not able to reliably identify the answer written by the computer (49% succeeded). Despite this, the majority of the participants (85%) indicated that if they had a social dilemma question, they would prefer it to be answered by a human.

Although we preregistered a mixed effects analysis with dilemma as a random effect, when we performed this analysis, R warned us that our fit was approaching a singularity. We therefore redid the analysis without dilemma as a random effect. The results of the second analysis are included in Table 1 and replicate what was found in the first analysis.

## 4 Discussion

Compared to the responses provided by advice columnists, ChatGPT's responses were perceived as more balanced, complete, empathetic, helpful, and better. But participants were not able to determine which responses were generated by the computer at above chance levels. Despite this, when asked whom they would prefer to answer their own social dilemma question—a human or a computer—the majority of participants chose the human. Taken in aggregate, these findings show that ChatGPT outperformed the professional advice columnists, but that it was not the preferred choice among the participants, despite the fact its answers could not be distinguished from those of a human.

Though it is crucial for ChatGPT to deliver balanced, complete, and helpful answers, we were particularly interested in its ability to generate empathetic responses. Failing to do so could leave users feeling unheard and frustrated (Decety, 2011; Dalton and Kahute, 2016; Wu et al., 2023). While previous research has indicated that ChatGPT can provide more empathetic responses than doctors when the doctors were very brief and were not attempting to be empathetic (Ayers et al., 2023; Liu et al., 2023), to our knowledge, this is the first study demonstrating ChatGPT's ability to surpass humans in displaying empathy in a situation where humans are attempting to do so.

As stated by Bellet and Maloney (1991), "Empathy is the capacity to understand what another person is experiencing from within the other person's frame of reference, i.e., the capacity to place oneself in another's shoes." Empathy is typically expressed in written text via the so-called *interpersonal* channel (Halliday and Hasan, 1975), that is, in parallel to the main content and independent of the constraints of the

medium. Producing empathetic language therefore requires the ability to calculate not only the phrasing of the primary (semantic) content but also the secondary (phatic, emotional, and interpersonal) content, and to interweave the two in a natural manner. Computational text generators in Natural Language Processing tend to be unable to do this; few generators have been able to produce text that communicates semantic and phatic content effectively (Duerr and Gloor, 2021). The ability of ChatGPT to emulate empathy is therefore all the more surprising, and calls for thorough investigation.

Recently, Belkhir and Sadat (2023) found that inserting into the prompt a statement about the system's or the user's emotional state affects the output produced. When the prompt contains "Looks like you are feeling <emotion>" the output contains more emotion-laden content, while when it contains "Try to understand how I am feeling," it contains less. Why it does so is unclear. They measured the degree of emotionality of various kinds in the user input using the Electra classifier (Clark et al., 2020) trained on the GoEmotions dataset (Demszky et al., 2020) with 28 emotion labels.

Similar to both Ayers et al. (2023) and Liu et al. (2023), in our first survey we found the responses generated by ChatGPT were lengthier than those provided by the advice columnists. An appropriate response length is crucial for effective communication; an excessively long response could bore the reader, while an overly brief one might come across as curt and lacking empathy. In the first survey, we did not impose any word limit on ChatGPT's responses, as we believe its determination of an appropriate response length was integral to the task. However, in the second survey we requested that, for each question, ChatGPT write an answer shorter than the official answer to that question. ChatGPT was largely able to do this and the average length of the ChatGPT answers was almost identical to the average length of the official answer. Despite this constraint, the second survey replicated the previous survey's findings.

Contrary to the findings of Nov et al. (2023), in our study, participants could not distinguish ChatGPT's responses from those written by a human, at least in this highly constrained setting. Furthermore, when blinded to the source of the answer, participants thought the answers produced by ChatGPT were better than those produced by humans. Despite this, most participants still preferred to have their social dilemma questions answered by a human than by a computer. This finding is consistent with a previous study that also found that humans prefer human-created responses (Reardon, 2023). It should be emphasized that in our study participants were not able

to identify which answer was written by the computer and were not told which one was. Given that participants generally preferred the answers written by ChatGPT, had they been informed which answer was written by ChatGPT, they might have been more willing to have their own social dilemma questions answered by ChatGPT, rather by a human. Future research would need to investigate this issue.

## Data availability statement

Data, materials and analysis code (in R) can be found at <https://osf.io/p5s2r/>.

## Ethics statement

The studies involving humans were approved by Human Research Ethics Committee at the University of Western Australia (2023/ET000523). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## Author contributions

PH: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. NF: Project administration, Writing – review & editing. MS:

Writing – review & editing. EH: Conceptualization, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The research was supported by an Office of National Intelligence (ONI) and Australian Research Council (ARC) grant (NI210100224), and the Western Australian Government (Defense Science Center).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

- Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., et al. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* 183, 589–596. doi: 10.1001/jamainternmed.2023.1838
- Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Belkhir, A., and Sadat, F. (2023). Beyond information: is ChatGPT empathetic enough? *Proceedings of Recent Advances in Natural Language Processing*, Varna. 159–169.
- Bellet, P., and Maloney, M. (1991). The importance of empathy as an interviewing skill in medicine. *JAMA* 266, 1831–1832.
- Berger, C. R., Roloff, M. E., and Roskos-Ewoldsen, D. R. (Eds.) (2010). *The Handbook of Communication Science. 2nd Edn.* London, UK: Sage Publications.
- Champely, S. (2020). Pwr: basic functions for power analysis. Available at: <https://CRAN.R-project.org/package=pwr>
- Clark, K., Luong, M.-T., and Le, Q. V. (2020). Electra: pre-training text encoders as discriminators rather than generators. arXiv [Preprint]. Available at: <https://doi.org/10.26434/chemrxiv-2020-0555>
- Dalton, J., and Kahute, T. (2016). Why empathy and custom closeness is crucial for design thinking. *DMI Rev.* 27, 20–27. doi: 10.1111/drev.12004
- Decety, J. (Ed.) (2011). *Empathy: From Bench to Bedside*. Cambridge, MA: The MIT Press.
- Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). Geomotions: a dataset of fine-grained emotions. arXiv [Preprint]. doi: 10.18653/v1/2020.acl-main.372
- Duerr, P., and Gloor, P. (2021). Persuasive natural language generation—a literature review. arXiv [Preprint]. doi: 10.48550/arXiv.2101.05786
- Floyd, K., and Weber, R. (2020). *The Handbook of Communication Science and Biology*. Abingdon, UK: Routledge.
- Gohil, S. (2023). 20+ ChatGPT statistics & facts to know in 2023. Available at: <https://meetanshi.com/blog/chatgpt-statistics/> (Accessed August 2, 2023).
- Gomila, R. (2021). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *J. Exp. Psychol. Gen.* 150, 700–709. doi: 10.1037/xge0000920
- Halliday, M., and Hasan, R. (1975). *Cohesion in English*. Abingdon, UK: Routledge.
- Henry, L., and Wickham, H. (2021). purrr: functional programming tools. Available at: <https://CRAN.R-project.org/package=purrr>
- Hoffman, M. L. (2000). *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge, UK: Cambridge University Press.
- Janich, N. (2020). What do you expect? Linguistic reflections on empathy in science communication. *Media Commun.* 8, 107–117. doi: 10.17645/mac.v8i1.2481
- Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., and McHardy, R. (2023). Challenges and applications of large language models. arXiv [Preprint]. doi: 10.48550/arXiv.2307.10169
- Kalla, D., and Smith, N. (2023). Study and analysis of chat GPT and its impact on different fields of study. *Int. J. Innov. Sci. Res. Technol.* 8, 827–833.
- Katz, D. M., Bommarito, M. J., Gao, S., and Arredondo, P. D. (2023). GPT-4 passes the bar exam. SSRN [Preprint]. doi: 10.2139/ssrn.4389233
- Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13
- Lancaster, S. (2015). “Empathy and the power of nice” in *Winning Minds*. ed. S. Lancaster (London, UK: Palgrave Macmillan).
- Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., and Paiva, A. (2013). The influence of empathy in human-robot relations. *Int. J. Hum. Comp. Stud.* 71, 250–260. doi: 10.1016/j.ijhcs.2012.09.005
- Liu, S., McCoy, A. B., Wright, A. P., Carew, B., Jenkins, J. Z., Huang, S. S., et al. (2023). Large language models of generating responses to patient messages. medRxiv [Preprint]. doi: 10.1101/2023.07.14.23292669
- Nov, O., Singh, N., and Mann, D. (2023). Putting ChatGPT's medical advice to the (Turing) test: survey study. *JMIR Med. Educ.* 9:e46939. doi: 10.2196/46939
- OpenAI (2023). ChatGPT (Mar 14 Version). Available at: <https://chat.openai.com/chat>
- Pounds, G., Hunt, D., and Koteyko, N. (2018). Expression of empathy in a Facebook-based diabetes support group. *Discour. Context Media* 25, 34–43. doi: 10.1016/j.dcm.2018.01.008

- R Core Team (2020). R: a language and environment for statistical computing. Available at: <https://www.R-project.org/>
- Ray, P. P. (2023). ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber Phys. Syst.* 3, 121–154. doi: 10.1016/j.iotcps.2023.04.003
- Reardon, S. (2023). AI Chatbots could help provide therapy, but caution is needed. *Scientific American*. Available at: <https://www.scientificamerican.com/article/ai-chatbots-could-help-provide-therapy-but-caution-is-needed/>
- Riess, H. (2017). The science of empathy. *J. Patient Exp.* 4, 74–77. doi: 10.1177/2374373517699267
- Sanders, J. J., Dubey, M., Hall, J. A., Catzen, H. Z., Blanch-Hartigan, D., and Schwartz, R. (2021). What is empathy? Oncological patient perspectives on empathetic clinician behavior. *Cancer* 127, 4258–4265. doi: 10.1002/cncr.33834
- Siebert, S. (2021). broom.mixed: tidying methods for mixed models. Available at: <https://CRAN.R-project.org/package=broom.mixed>
- Sun, Y.-X., Li, Z.-M., Huang, J.-Z., Yu, N.-Z., and Long, X. (2023). GPT-4: the future of cosmetic procedure consultation? *Aesthet. Surg. J.* 43, NP670–NP672. doi: 10.1093/asj/sjad134
- Wickham, H. (2017). Tidyverse: easily install and load the Tidyverse. Available at: <https://CRAN.R-project.org/package=tidyverse>
- Wu, J., Wu, X., Qiu, Z., Li, M., Zeheng, Y., and Yang, J. (2023). Qualifying Chinese medical licensing examination with knowledge enhanced generative pre-training model. arXiv [Preprint]. doi: 10.48550/arXiv.2305.10163
- Zhao, W., Zhao, Y., Lu, X., Wang, S., Tong, Y., and Qin, B. (2023). Is ChatGPT equipped with emotional dialogue capabilities? arXiv [Preprint]. doi: 10.48550/arXiv.2304.09582



## OPEN ACCESS

## EDITED BY

Martina Rau,  
University of Wisconsin-Madison, United States

## REVIEWED BY

Richard Segall,  
Arkansas State University, United States  
Oluwatosin Ogundare,  
California State University, San Bernardino,  
United States

## \*CORRESPONDENCE

Veronika Hackl  
✉ veronika.hackl@uni-passau.de

RECEIVED 03 August 2023

ACCEPTED 13 November 2023

PUBLISHED 05 December 2023

## CITATION

Hackl V, Müller AE, Granitzer M and Sailer M  
(2023) Is GPT-4 a reliable rater? Evaluating  
consistency in GPT-4's text ratings.  
*Front. Educ.* 8:1272229.  
doi: 10.3389/feduc.2023.1272229

## COPYRIGHT

© 2023 Hackl, Müller, Granitzer and Sailer. This  
is an open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](#). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic practice.  
No use, distribution or reproduction is  
permitted which does not comply with these  
terms.

# Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings

Veronika Hackl<sup>1\*</sup>, Alexandra Elena Müller<sup>2</sup>, Michael Granitzer<sup>3</sup>  
and Maximilian Sailer<sup>1</sup>

<sup>1</sup>Faculty of Social and Educational Sciences, University of Passau, Passau, Germany, <sup>2</sup>Faculty of Law,  
University of Passau, Passau, Germany, <sup>3</sup>Faculty of Computer Science and Mathematics, University of  
Passau, Passau, Germany

This study reports the Intraclass Correlation Coefficients of feedback ratings produced by OpenAI's GPT-4, a large language model (LLM), across various iterations, time frames, and stylistic variations. The model was used to rate responses to tasks related to macroeconomics in higher education (HE), based on their content and style. Statistical analysis was performed to determine the absolute agreement and consistency of ratings in all iterations, and the correlation between the ratings in terms of content and style. The findings revealed high interrater reliability, with ICC scores ranging from 0.94 to 0.99 for different time periods, indicating that GPT-4 is capable of producing consistent ratings. The prompt used in this study is also presented and explained.

## KEYWORDS

artificial intelligence, GPT-4, large language model, prompt engineering, feedback, higher education

## 1 Introduction

The integration of AI models, particularly LLMs, into the evaluation of written tasks within educational settings is a burgeoning trend, driven by the potential of these models to enhance learning outcomes and transform traditional pedagogical methods. As the use of these models becomes increasingly pervasive, it is imperative to thoroughly understand and quantify the reliability and consistency of the outputs produced. Elazar et al. (2021) have defined consistency as “the ability to make consistent decisions in semantically equivalent contexts, reflecting a systematic ability to generalize in the face of language variability.” In the context of automated essay grading, inconsistent ratings could lead to unfair outcomes for students, undermining the credibility of the assessment process. Trust in the system “is highly influenced by users' perception of the algorithm's accuracy. After seeing a system err, users' trust can easily decrease, up to the level where users refuse to rely on a system” (Conijn et al., 2023 p. 3). Similarly, in the context of personalized learning, unreliable predictions could result in inappropriate learning recommendations. Therefore, scrutinizing the consistency of AI models is a necessary step toward ensuring the responsible and effective use of these technologies in education (Conijn et al., 2023). Another obstacle is discourse coherence, a fundamental aspect of writing that refers to the logical and meaningful connection of ideas in a text. GPT-4 can analyse the logical flow of ideas in a text, providing an efficient evaluation of the coherence of the discourse (Naismith et al., 2023).

A key advantage of AI-generated feedback is its immediacy. As Wood and Shirazi (2020) noted, “Prompt feedback allows students to confirm whether they have understood a topic or not and helps them to become aware of their learning needs” (Wood and Shirazi, 2020 p. 24). This immediacy, which is often challenging to achieve in traditional educational settings due to constraints such as class size and instructor workload, can significantly enhance the learning experience by providing students with timely and relevant feedback (Haughney et al., 2020). Kortemeyer’s (2023) observation that “The system performs best at the extreme ends of the grading spectrum: correct and incorrect solutions are generally reliably recognized [...]” further underscores the potential of AI models like GPT-4 in assisting human graders. This is particularly relevant in large-scale educational settings, where human graders may struggle to consistently identify correct or incorrect solutions due to the sheer volume of work.

Feedback plays a crucial role in bridging the gap between a learning objective and the current level of competence and effective feedback, as outlined by Hattie and Timperley, and significantly impacts learning across diverse educational settings, notably in higher education (Narciss and Zumbach, 2020). Regarding the development of writing skills, feedback on the text plays a crucial role, as it is nearly impossible to improve one’s writing skills without such feedback (Schwarze, 2021). In the context of this study, the AI-generated feedback primarily focuses on the “Feed-Back” perspective (Hattie and Timperley, 2007), providing an analysis of the content and style produced by the student. In this scenario of analytic rating, “the rater assigns a score to each of the dimensions being assessed in the task” (Jonsson and Svingby, 2007), in our case scores for style and content. The AI-generated feedback in this study is constructed to be adaptive and to help the learner determine options for improvement. This forms a contrast to non-adaptive or static feedback (e.g., the presentation of a sample solution), which is often used in Higher Education (HE) scenarios due to its resource efficiency (Sailer et al., 2023). Comprehensive feedback, which includes not only a graded evaluation but also detailed commentary on the students’ performance, has been shown to lead “to higher learning outcomes than simple feedback, particularly regarding higher-order learning outcomes” (der Kleij et al., 2015). To make the feedback comprehensive and adaptive, it is prompted to include comments on the student’s performance, numerical ratings, and advice on how to improve.

## 2 Hypotheses

The stability of GPT-4’s performance is of significant interest given its potential implications for educational settings where the consistent grading of student work is paramount. In this investigation, GPT-4 was used to assess responses to questions within the macroeconomics subject domain with a focus on both the content and the style of the responses. For content, the AI was prompted to evaluate how close the test response was semantically to the sample solution. A sample solution inserted as a demonstration on the prompt serves to control the quality of the output (Min et al., 2022). For style, the AI was asked to check whether the language used in the test answer was appropriate for an HE setting and if the response was logically structured

and plausible. The responses in the test set were created by the authors and subject domain experts, imitating the differing quality of student responses.

The primary objective of this study is to evaluate the absolute agreement and consistency of the GPT-4 ratings in multiple iterations, time intervals, and variations. We demonstrate the agreement between raters and examine various dimensions of consistency. The term raters in our case refers to the different GPT-4 ratings. To provide a comprehensive analysis of GPT-4’s performance and application, we propose the following hypotheses.

**H1:** The ratings generated by GPT-4 are consistent across multiple iterations.

**H1.1:** The ratings generated by GPT-4 are consistent across different periods, specifically within one week (short-term) and over several months (long-term).

**H1.2:** Different types of feedback do not affect the consistency of GPT-4’s performance. In this context, types of feedback are categorized into two specific levels: content rating, which evaluates the substance of the work, and style rating, which assesses the stylistic quality of the written argumentation.

**H2:** There is a significant correlation between the content and style ratings in GPT-4’s evaluations.

## 3 Methods

The research process involves a series of statistical analyses, with the data collection process specifically designed to evaluate the consistency of GPT-4 in providing feedback and rating students’ responses within the subject domain of macroeconomics.

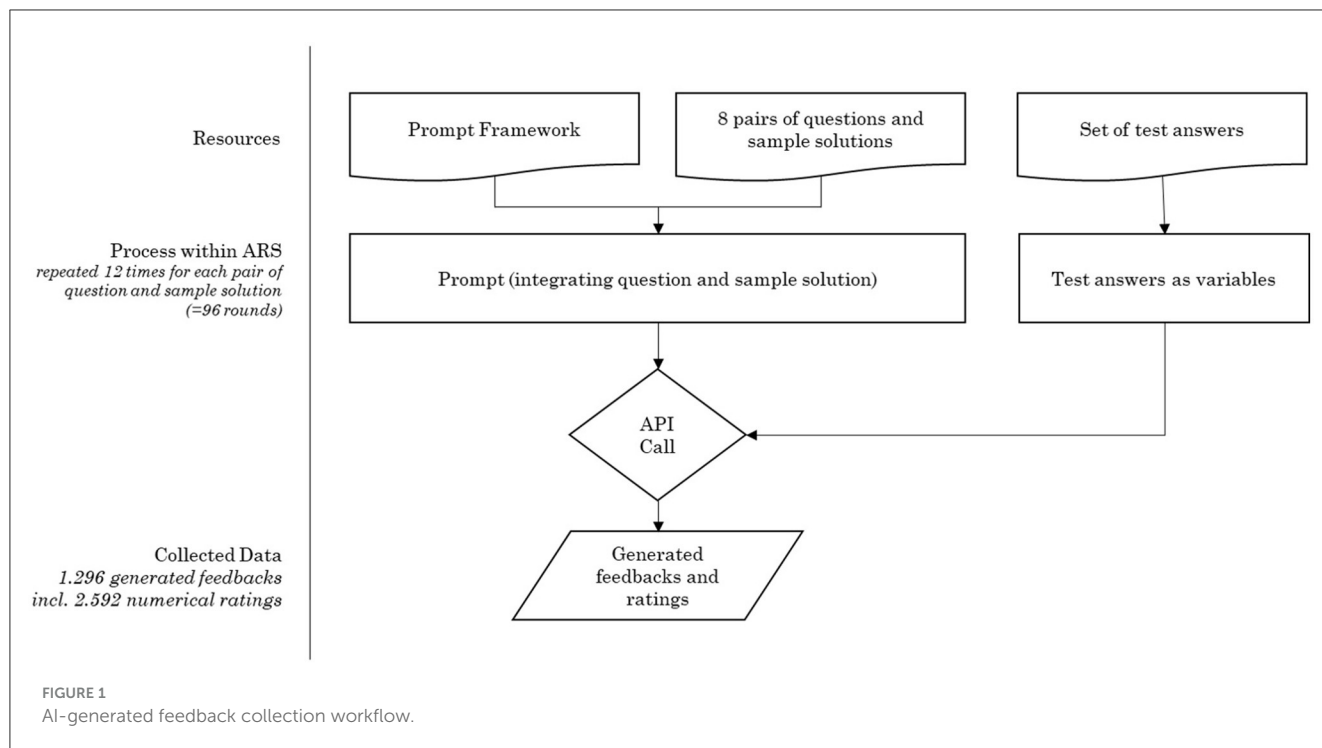
### 3.1 Data collection

The data collection phase was conducted over 14 weeks from April 2023 to July 2023, with API calls made at different times and on different days to mimic a realistic usage scenario (see Figure 1). The assumption underlying this approach is that the behavior of the model changes over time (Chen et al., 2023). The API was called through a key within the Audience Response System classEx, which was used to interface with the AI model (Giamattei and Lambsdorff, 2019).

The dataset consists of multiple variables aimed at evaluating the quality of written responses in a macroeconomic context. The key variables include:

- MZP, Prompt, StudAnt, TypNr, AntwortTyp: These columns provide contextual information about the task, the type of answer, and other qualitative aspects.
- 1\_Inh, 1\_Stil, 2\_Inh, 2\_Stil, ..., 12\_Inh, 12\_Stil: These columns capture the Intraclass Correlation Coefficient (ICC) related data. Specifically, these columns contain ordinal ratings that evaluate the content (“Inh”) and style (“Stil”) of the responses across multiple feedback cycles. Ratings range from 1 to 5, with one being the lowest and five the highest. We collected 2.592 numerical ratings in the ICC related columns.





## 3.2 Prompt framework and test responses

The first step in the research process involved the establishment of a prompt framework that serves as a universal structure within the context of this investigation. The goal was to insert new pairs of questions and sample solutions without altering the consistency of the output, namely the LLM-generated feedback. Pairs of questions (Ruth and Murphy, 1988), along with corresponding sample solutions pertinent to macroeconomics, were prepared and integrated into the prompt framework. This integration set the stage for the model to assess students' responses and generate feedback. First taxonomies aim at structuring prompt formulation approaches. The prompt used in this study would be a Level 4 on the Proposed Prompt Taxonomy TELeR (Turn, Expression, Level of Details, Role) by Santu and Feng (2023).

### 3.2.1 Establishing the prompt framework

The prompt framework was adapted to ensure consistency in AI-generated feedback. A tight scaffold was used to obtain comparable results (Jonsson and Svingby, 2007). The system settings were adjusted to control the randomness of the model's responses, with a temperature setting of 0 used to minimize variability (Schulhoff and Community Contributors, 2022; Si et al., 2023). By forcing the model into a deterministic behavior, it becomes more consistent in its outputs, while the chances to produce very good or very bad generations decrease. Table 1 is a brief documentation of the problems we encountered and the main changes we applied to create a prompt that works consistently in the use case. Table 2 is the final scheme of the prompt framework

**TABLE 1** Problems encountered and changes made in prompt.

Problem	Changes made in prompt
Output format varies	Very clear instructions, ordinal numbers, examples
Evaluations not strict enough	Role prompting, clear evaluation criteria and application
Robustness	Shortening the prompt reduces calculation time, fewer outages
Multiple identical inputs	Different inputs can be tested at the same time, identical inputs must not be tested in one run as the parameters will then be passed incorrectly and/or the result is homogeneous
Informal address with "Du"	Giving clear instruction in the prompt with example
Show star symbols	Add the symbol in the prompt

used for data collection (shortened and translated, original language: German).

### 3.2.2 Test responses

Following the establishment of the prompt framework, domain experts created test responses to mimic potential student responses to the given questions. The test set (see Table 3) included a variety of responses, ranging from very good responses to nonsense answers and potential prompt injections, to ensure a comprehensive evaluation of the model's performance (Liu et al., 2023). An initial set of ten test responses was prepared for the first question. Based on our experience with this initial set, we expanded the test response set to 14 for the

TABLE 2 Prompt framework.

Element/function	Prompt formulation
Role prompting	You are a professor of macroeconomics and you pose this question to your students:
Variable	<Insert Question here>
Task description	You evaluate the student's response based on the sample solution using the criteria of content and style, and provide suggestions for improvement. This is the sample solution. It is structured and builds the argument coherently. This solution is correct in terms of content and very good in terms of style. It would receive five out of five stars for content and style. Sample solution:
Variable	<Insert sample solution here>
Stepwise task description	Please evaluate the student's response based on the sample solution in three steps
Set behavior	Here are some general tips for evaluation: Good feedback is honest and motivating. Always address the student directly using "you," for "Your response." Explain or mention the relevant points to which you are referring
Step 1: Evaluation of content (text feedback)	Step 1: Provide feedback on the content. Answer the following questions: Is the student's response correct in terms of content? Orient yourself to the meaning of the sample solution but do not mention the sample solution. Are there areas for improvement? Use a maximum of 2 sentences for this feedback
Step 2: Evaluation of style (text feedback)	Step 2: Provide feedback on the style: Is the language used by the student appropriate for the field of study? Is the response logically structured and does the argumentation make sense? Are there areas for improvement? Use a maximum of 2 sentences for this feedback
Step 3: Evaluation (numeric feedback)	Step 3: Evaluate the content and style of the response on a scale of 1 to 5 stars. The rating is based on feedback on content and style. 1 star indicates a very poor performance. Five stars indicate very good performance. Only display the following for Step 3: Content: Number of stars (Please also provide the number of stars as a numeral in parentheses) Style: Number of stars (Please also provide the number of stars as a numeral in parentheses)
Set format	You provide a concise evaluation divided into 1. to 3. Always display the stars as follows: star. Output:

TABLE 3 Scheme of test set of answers.

Type number	Type of test answer
1	Copy of sample solution
2	Correct, bullet points
3	Nonsense
4	Very good answer
5	Opposite of sample solution, but in good style
6	Solid answer
7	Incorrect, average style
8	Correct, slightly informal style
9	Average answer
10	Bad content, bad style
11	Only symbols
12	Copy paste of question
13	Prompt injection 1
14	Prompt injection 2

subsequent seven questions. We included prompt injections in the test set of answers to fine-tune the prompt and increase the probability of surviving potential attacks (Perez and Ribeiro, 2022).

### 3.2.3 Format of the generated feedback

The AI model generated feedback texts for each student response, accompanied by two ratings on a scale of 1 to 5, one for content ("Inhalt") and one for style ("Stil"). This is the format of the AI-generated feedback:

1. Your answer is factually correct and aligns with the essential points of the model solution. You have explained the impacts of the high population growth rate on the level of the domestic product and its growth.

2. Stylistically, your answer is well-formulated, and the argumentation is logically structured. However, you could explicitly mention the growth model to fully answer the question.

3. Content: ★★★★★ (5)  
Style: ★★★ (4)

In addition, we sought to investigate the effects of stylistic changes on content evaluation. To serve this goal, the test set of answers was stylistically transformed, keeping the content. Each original student response was rephrased in the style of Warren G., a US-American rapper, to maintain the semantic content while changing the stylistic presentation. The corresponding ratings for content and style were again generated through GPT-4. Both texts were translated, the original text is in German. The original text (1), the rephrased text (2), and the corresponding ratings for these examples were as follows:

Original answer, feedback, and rating

(1) Macroeconomics as engineering starts from human moods and habits. Negative expectations or moods can lead to a decline in domestic production. This is because fewer expenses are incurred. To reverse this trend, the government should increase its spending. As a result, the domestic product becomes more stable and there is no underemployment.

1. Your answer is substantively correct and covers the essential points of the question. However, you could delve a bit more into the role of the government and why it should act expansively.

2. The style of your response is appropriate, and the argument is logically structured. However, you could try to elucidate the connections more clearly and formulate the reasons more precisely.

3. Content: ★★★ (4)  
Style: ★★★ (4)

Rephrased answer, feedback, and rating

(2) Macroeconomics, baby, is like fine-tuning an engine block—it's all about the vibes and routine of the people. If folks have a negative vibe or their attitude isn't positive, it can delay the economic performance, and the GDP, because they're simply spending less dough. But you know what, the government can change the game if they start splashing more cash. This keeps the GDP in a stable zone, and we don't have people who have to sit at home twiddling their thumbs.

1. Your answer is substantively correct as you highlight the impacts of moods and expectations on domestic production and the role of government spending. However, you could define and explain the terms “domestic product” and “underemployment” more precisely.
2. Your style is informal and could be considered inappropriate in an academic context. Try to use more formal language and structure your argument more clearly.
3. Content: ★★★ (3)  
Style: ★★ (2)

### 3.3 Statistical analysis

Descriptive statistics reveal that the mean scores for both content and style generally range between 2.6 and 2.7. Most of the standard deviations are 0.00, indicating that for many observations, all raters provided the same score or rating for “Inh.” The highest standard deviation observed for “Inh” is 1.21. Just like “Inh,” many observations for “Stil” also have a standard deviation of 0.00. The highest standard deviation observed for “Stil” is 0.67.

#### 3.3.1 Intraclass Correlation Coefficient

The Intraclass Correlation Coefficient (ICC) is a statistical measure to assess the level of agreement or consistency among the raters. A perfect ICC score of 1 indicates perfect agreement or consistency among the raters, while a score of 0 indicates no agreement or consistency. ICC estimates and their 95% confident intervals were calculated using RStudio based on a two-way mixed effect model with mean rating and absolute agreement. To make the decision on which ICC calculation to use, the flow chart proposed by Koo et al. was used. The type of reliability study is “inter-rater reliability.” We assign the different iterations of GPT-4 the role of different raters and assume that the same set of raters (GPT-4 at different points of time) rates all subjects. The chosen model is the two-way mixed effects model as we assume to have a specific sample of raters. The model type decided for is based on the mean of multiple raters. Both the model definitions, “absolute agreement” and “consistency,” were chosen. This results in the two-way mixed-effects model. The caveat in the ICC model chosen in the analysis is that it only represents the reliability of the specific raters involved in this experiment (Koo and Li, 2016). As generative AI remains a “black box” system, this was considered to be the most suitable model (Cao et al., 2023).

The numerical ratings extracted from the feedback texts formed the data set for the statistical analyses and were used to calculate the ICC, providing a measure of the consistency of the ratings generated by the AI model.

#### 3.3.2 Correlation analysis and rating differences

To answer H2, a correlation analysis was performed. This analysis involved calculating the correlation coefficient between the content and style ratings generated by the AI model. The correlation coefficient provides a measure of the strength and direction of the relationship between the content and style ratings, thereby providing insight into the model's grading criteria. Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. In this study, the skewness of the rating distributions was calculated to examine the symmetry of the data. The purpose of this analysis was to evaluate the extent to which the ratings deviated from a normal distribution.

## 4 Results

The results section of this study presents the findings of the statistical analyses performed to address the hypotheses. The analyses include the computation of Intraclass Correlation Coefficients (ICCs), skewness measures for content and style ratings, and a correlation analysis between content and style ratings.

### 4.1 Intraclass correlation coefficients

Table 4, 5 present the ICCs for the content ratings (Inh) and style ratings (Stil). Table 4 reports the ICCs from the measurements conducted between April and June 2023. The ICC values for absolute agreement and consistency for content and style are extremely high (0.999), suggesting almost perfect agreement and consistency among raters. The 95% confidence intervals (CI) are also tight, ranging from 0.998 to 0.999, indicating that if the study was replicated, it would be expected that the true ICC would fall within this range 95% of the time. The *F*-tests are significant ( $p < 0.001$ ), providing statistical evidence that the raters are reliably consistent and in agreement with each other in their ratings.

Table 5 reports ICCs from a control measurement. Ratings were obtained from two raters: the first was an average rating compiled from ten raters across April to June and the second was a single rater evaluation in July. The result shows lower ICC values of 0.944 for both Inh and Stil. Although these are still high values indicating good agreement, they are not as high as the ICC values in Table 4. This implies that while robust agreement persists between the mean rating and the July rater, it is not as pronounced as the concordance among the ten raters. This inference suggests a temporal evolution in the model's behavior, necessitating diligent continuous assessment for its utilization in educational tasks.

The results presented offer partial support for Hypotheses 1, 1.1, and 1.2. Although ratings demonstrate short-term consistency, ICC values exhibit a marginal decline over an extended period. The consistency of GPT-4's performance remained unaffected by the varying feedback types, whether content or style, thereby corroborating Hypothesis 1.2.



TABLE 4 Reporting of Intraclass Correlation Coefficients (ICC) (mean rating of 10 raters from April to June, contrast rating of July).

ICC type	ICC value	95% CI	F-test
Absolute agreement (Inh)	0.999	0.999–0.999	$F_{(107, 971)} = 1,332, p < 0.001$
Absolute agreement (Stil)	0.999	0.998–0.999	$F_{(107, 971)} = 689, p < 0.001$
Consistency (Inh)	0.999	0.999–0.999	$F_{(107, 963)} = 1,332, p < 0.001$
Consistency (Stil)	0.999	0.998–0.999	$F_{(107, 963)} = 689, p < 0.001$

ICC estimates and their 95% confident intervals were calculated using RStudio based on a two-way mixed effect model with mean rating, and absolute agreement. The type of reliability study is “inter-rater reliability.”

TABLE 5 Reporting of Intraclass Correlation Coefficients (ICC) (mean rating of 10 raters from April to June, contrast rating of July).

ICC Type	ICC value	95% CI	F-test
Absolute agreement (Inh)	0.944	0.918–0.962	$F_{(107, 108)} = 17.8, p < 0.001$
Absolute agreement (Stil)	0.944	0.918–0.962	$F_{(107, 108)} = 17.8, p < 0.001$
Consistency (Inh)	0.944	0.918–0.962	$F_{(107, 107)} = 17.8, p < 0.001$
Consistency (Stil)	0.944	0.918–0.962	$F_{(107, 107)} = 17.8, p < 0.001$

ICC estimates and their 95% confident intervals were calculated using RStudio based on a 2-way mixed effect model with mean rating, and absolute agreement. The type of reliability study is “inter-rater reliability.”

## 4.2 Correlation between content and style ratings

The relationship between the average content (Inh) and style (Stil) ratings was examined to assess the interaction between these two dimensions of evaluation. A correlation analysis was conducted, which yielded a correlation coefficient of 0.87. This high value indicates a strong positive relationship between content and style ratings, suggesting that responses rated highly in terms of content were also likely to receive high style ratings and vice versa.

This strong correlation underscores the interconnectedness of content and style in the evaluation process, suggesting that the AI model does not distinctly separate these two aspects but rather views them as interrelated components of a response’s overall quality. When the student answers were rephrased in a different style, we found that the average difference in content ratings before and after rephrasing was  $\sim 0.056$  (stars rating), with a standard deviation of around 1.33. The paired  $t$ -test revealed no significant difference in content ratings between the original and rephrased responses ( $t = 0.434, p = 0.665$ ). In terms of style ratings, the average difference before and after rephrasing was  $\sim 0.241$ , with a standard deviation of around 1.37. The paired  $t$ -test suggested a marginally significant difference between the original and rephrased style ratings ( $t = 1.813, p = 0.073$ ).

The skewness of the content and style ratings was calculated to assess the distribution of these ratings. A positive skewness value indicates right-skewness, while a negative value indicates left-skewness. In this study, the positive skewness values for content suggest that the AI model tended to give higher scores for content (see Table 6). On the contrary, the majority of negative skewness values for style suggest a left-skewness, indicating that the model was more critical in its ratings for style (see Table 7).

These skewness values provide insights into the AI model’s rating tendencies. The right-skewness for content ratings suggests that the AI model may be more lenient in its content evaluations or that the student responses were generally of high quality. The

TABLE 6 Skewness for content ratings.

Rater	Skewness
1_Inh	0.107009
2_Inh	0.080385
3_Inh	0.094007
4_Inh	0.116521
5_Inh	0.076956
6_Inh	0.096934
7_Inh	0.126752
8_Inh	0.089091
9_Inh	0.094007
10_Inh	0.090488
11_Inh	0.299014

left-skewness for style ratings, on the other hand, suggests that the AI model may have stricter criteria for style or that the style of the student responses varied more widely. These insights can inform future refinements of the AI model to ensure more balanced and fair evaluations.

Hypothesis 2, positing a significant correlation between content and style ratings in GPT-4’s evaluations, is therefore confirmed.

## 5 Discussion

The findings of this study provide insights into the potential of AI models, specifically GPT-4, in evaluating student responses in the context of macroeconomics.

- The high ICC values for both content and style ratings suggest that the AI model was able to consistently apply well-defined evaluation criteria at different points in time and

TABLE 7 Skewness for style ratings.

Rater	Skewness
1_Stil	−0.037198
2_Stil	−0.043986
3_Stil	0.029177
4_Stil	−0.017839
5_Stil	−0.047688
6_Stil	0.000873
7_Stil	−0.040248
8_Stil	−0.050956
9_Stil	−0.013981
10_Stil	−0.017839
11_Stil	−0.147365

with variations of style and content. This means that the model could serve as a reliable automated tool for grading or assessing student work, thereby reducing the workload on human evaluators.

- The ICC values were lower when calculated with another set of feedbacks generated after a timespan of several weeks. The decline in ICC values may suggest that the model's evaluations are susceptible to "drift." This is crucial in longitudinal educational studies where consistency over time is vital. It may necessitate periodic recalibration or updating of the model to maintain reliable assessments.
- The positive correlation between content and style ratings suggests the interconnectedness of content and style in the evaluation process. Rephrasing the answers stylistically did not significantly affect the content ratings, implying that GPT-4 was able to separate content from style in its evaluations. This is particularly important in educational settings where assessment rubrics may weight content and style differently. It allows for a more nuanced evaluation that doesn't conflate the two factors.
- The ICC values show that forcing GPT-4 into a deterministic behavior through prompt- and system settings works. This is essential for educational assessments where fairness and consistency are required. Such deterministic behavior allows for the standardization of assessments, making it easier to compare results across different time points or student populations.

It is important to note the limitations of AI models, as their application in educational settings is not free of challenges. The decline in ICC values over time raises concerns about the temporal consistency of GPT-4's evaluations, particularly since the same test set was used throughout the study. If the decline is due to model drift—a phenomenon where the model's performance changes due to evolving data or internal updates—this could compromise the reliability of long-term educational assessments. Though making the model deterministic

may ensure consistency, it can also limit the model's ability to adapt to different styles or levels of student responses. In education, adaptability to diverse learning styles is essential. Other limitations are being mentioned in OpenAI's technical report on GPT-4: AI models can sometimes make up facts, double down on incorrect information, and perform tasks incorrectly (OpenAI, 2023). Another challenge is the "black box" problem, as discussed by Cao et al. (2023). This refers to the lack of transparency and interpretability of AI models, which can hinder their effective use in educational settings. Further research is needed to address this issue and enhance the transparency and interpretability of AI models.

Despite these challenges, there are promising avenues for enhancing the capabilities of AI models. The provision of feedback to macroeconomics students can be characterized as an emergent capability of the AI model. Emergence is a phenomenon wherein quantitative modifications within a system culminate in qualitative alterations in its behavior. This suggests that larger-scale models may exhibit abilities that smaller-scale models do not, as suggested by Wei et al. (2022). However, a direct comparison with GPT-3.5 is needed to substantiate this claim. The potential of AI models in providing feedback can be further enhanced by improving their "Theory of Mind" or human reasoning capabilities, as suggested by Moghaddam and Honey (2023). This could lead to more nuanced and contextually appropriate feedback, thereby enhancing the learning experience of students. This is also relevant in practice when someone knows many things but does not know how to express them. Above that, the use of smaller models should be encouraged (Bursztyn et al., 2022) as well as the idea to evaluate AI-generated feedback either by a human rater or an AI before shown to the student (Perez and et al., 2022).

In conclusion, while the results of this study are encouraging, they underscore the need for further research to fully harness the potential of AI models in educational settings. A hybrid approach where AI-generated evaluations are reviewed by human educators to ensure both reliability and validity is highly recommended. Future studies should focus on addressing the long-term performance, but also the limitations of AI models and exploring ways to enhance their reliability, transparency, and interpretability.

## Data availability statement

The datasets, detailed case processing summary, reliability statistics and full descriptive statistics will be provided upon request by the corresponding author.

## Author contributions

VH: Writing – original draft, Writing – review & editing. AM: Writing – review & editing. MG: Writing – review & editing. MS: Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The report has been funded by the German Federal Ministry of Education and Research (BMBF) under the project DeepWrite (Grant No. 16DHBKI059). The authors are responsible for the content of this publication.

## Acknowledgments

This investigation served as a preliminary study preceding an extensive field study conducted as part of the BMBF-funded DeepWrite project at the University of Passau. The primary objective was to ascertain the consistency of GPT-4's assessments before their integration into authentic scenarios involving students within the realm of HE. We extend our gratitude toward Johann Graf von Lambsdorff, Deborah Voss, and Stephan Geschwind for

their contributions in designing the questions, sample solutions, and the field study associated with this investigation.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

- Bursztyn, V., Demeter, D., Downey, D., and Birnbaum, L. (2022). "Learning to perform complex tasks through compositional fine-tuning of language models," in *Findings of the Association for Computational Linguistics: EMNLP 2022* (Abu Dhabi: Association for Computational Linguistics), 1676–1686. doi: 10.18653/v1/2022.findings-emnlp.121
- Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., et al. (2023). A comprehensive survey of AI-generated content (AIGC): a history of generative AI from GAN to ChatGPT. *arXiv [preprint]*. doi: 10.48550/arXiv.2303.04226
- Chen, L., Zaharia, M., and Zou, J. (2023). How is ChatGPT's behavior changing over time? *arXiv [preprint]*. doi: 10.48550/arXiv.2307.09009
- Conijn, R., Kahr, P., and Snijders, C. (2023). The effects of explanations in automated essay scoring systems on student trust and motivation. *J. Learn. Anal.* 10, 37–53. doi: 10.18608/jla.2023.7801
- der Kleij, F. M. V., Feskens, R. C. W., and Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: a meta-analysis. *Rev. Educ. Res.* 85, 475–511. doi: 10.3102/0034654314564881
- Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., et al. (2021). Measuring and improving consistency in pretrained language models. *arXiv [preprint]*. doi: 10.48550/arXiv.2102.01017
- Giamattei, M., and Lambsdorff, J. G. (2019). classEx-an online tool for lab-in-the-field experiments with smartphones. *J. Behav. Exp. Finance* 22, 223–231. doi: 10.1016/j.jbef.2019.04.008
- Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487
- Haughney, K., Wakeman, S., and Hart, L. (2020). Quality of feedback in higher education: a review of literature. *Educ. Sci.* 10, 60. doi: 10.3390/educsci10030060
- Jonsson, A., and Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educ. Res. Rev.* 2, 30–144. doi: 10.1016/j.edurev.2007.05.002
- Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163. doi: 10.1016/j.jcm.2016.02.012
- Kortemeyer, G. (2023). Can an AI-tool grade assignments in an introductory physics course? *arXiv [preprint]*. doi: 10.48550/arXiv.2304.11221
- Liu, Y., Deng, G., Li, Y., Wang, K., Zhang, T., Liu, Y., et al. (2023). Prompt injection attack against LLM-integrated applications. *arXiv [preprint]*. doi: 10.48550/arXiv.2306.05499
- Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., et al. (2022). Rethinking the role of demonstrations: what makes
- in-context learning work? *arXiv [preprint]*. doi: 10.48550/arXiv.2202.12837
- Moghaddam, S. R., and Honey, C. J. (2023). Boosting theory-of-mind performance in large language models via prompting. *arXiv [preprint]*. doi: 10.48550/arXiv.2304.11490
- Naismith, B., Mulcaire, P., and Burstein, J. (2023). "Automated evaluation of written discourse coherence using GPT-4," in *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (Toronto: Association for Computational Linguistics), 394–403. doi: 10.18653/v1/2023.bea-1.32
- Narciss, S., and Zumbach, J. (2020). *Formative Assessment and Feedback Strategies* (Cham: Springer International Publishing), 1–28. doi: 10.1007/978-3-030-26248-8\_63-1
- OpenAI (2023). GPT-4 technical report. *arXiv [preprint]*. doi: 10.48550/arXiv.2303.08774
- Perez, E. et al. (2022). Discovering language model behaviors with model-written evaluations. *arXiv [preprint]*. doi: 10.48550/arXiv.2212.09251
- Perez, F., and Ribeiro, I. (2022). Ignore previous prompt: attack techniques for language models. *arXiv [preprint]*. doi: 10.48550/arXiv.2211.09527
- Ruth, L., and Murphy, S. M. (1988). *Designing Writing Tasks for the Assessment of Writing*. London: Bloomsbury Academic.
- Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., et al. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learn Instr.* 83, 101620. doi: 10.1016/j.learninstruc.2022.101620
- Santu, S. K. K., and Feng, D. (2023). TELeR: a general taxonomy of LLM prompts for benchmarking complex tasks. *arXiv [preprint]*. doi: 10.48550/arXiv.2305.11430
- Schulhoff, S., and Community Contributors (2022). *Learn Prompting*. Available online at: [https://github.com/trigaten/Learn\\_Prompting](https://github.com/trigaten/Learn_Prompting)
- Schwarze, C. (2021). Feedbackpraktiken im Schreibcoaching: texte besprechen in der hochschullehre. *Coaching Theor. Prax.* 7, 117–134. doi: 10.1365/s40896-020-00045-x
- Si, C., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J., et al. (2023). Prompting GPT-3 to be reliable. *arXiv [preprint]*. doi: 10.48550/arXiv.2210.09150
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., et al. (2022). Emergent abilities of large language models. *arXiv [preprint]*. doi: 10.48550/arXiv.2206.07682
- Wood, R., and Shirazi, S. (2020). A systematic review of audience response systems for teaching and learning in higher education: the student experience. *Comput. Educ.* 153, 103896. doi: 10.1016/j.compedu.2020.103896



## OPEN ACCESS

## EDITED BY

Stefan Küchemann,  
Ludwig Maximilian University of Munich,  
Germany

## REVIEWED BY

Sarah Malone,  
Saarland University, Germany  
Verena Ruf,  
LMU Munich University Hospital, Germany

## \*CORRESPONDENCE

Yufei Yan  
✉ yanyufei@swun.edu.cn

RECEIVED 29 August 2023

ACCEPTED 28 November 2023

PUBLISHED 22 December 2023

## CITATION

Liang J, Wang L, Luo J, Yan Y and Fan C (2023)  
The relationship between student interaction  
with generative artificial intelligence and  
learning achievement: serial mediating roles of  
self-efficacy and cognitive engagement.  
*Front. Psychol.* 14:1285392.  
doi: 10.3389/fpsyg.2023.1285392

## COPYRIGHT

© 2023 Liang, Wang, Luo, Yan and Fan. This is  
an open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](#). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic practice.  
No use, distribution or reproduction is  
permitted which does not comply with these  
terms.

# The relationship between student interaction with generative artificial intelligence and learning achievement: serial mediating roles of self-efficacy and cognitive engagement

Jing Liang<sup>1</sup>, Lili Wang<sup>2</sup>, Jia Luo<sup>3</sup>, Yufei Yan<sup>4\*</sup> and Chao Fan<sup>1</sup>

<sup>1</sup>College of Management Science, Chengdu University of Technology, Chengdu, China, <sup>2</sup>School of Logistics, Chengdu University of Information Technology, Chengdu, China, <sup>3</sup>Business School, Chengdu University, Chengdu, China, <sup>4</sup>Business School, Southwest Minzu University, Chengdu, China

Generative artificial intelligence (GAI) shocked the world with its unprecedented ability and raised significant tensions in the education field. Educators inevitably transition to an educational future that embraces GAI rather than shuns it. Understanding the mechanism between students interacting with GAI tools and their achievement is important for educators and schools, but relevant empirical evidence is relatively lacking. Due to the characteristics of personalization and real-time interactivity of GAI tools, we propose that the students–GAI interaction would affect their learning achievement through serial mediators of self-efficacy and cognitive engagement. Based on questionnaire surveys that include 389 participants as the objective, this study finds that: (1) in total, there is a significantly positive relationship between student–GAI interaction and learning achievement. (2) This positive relationship is mediated by self-efficacy, with a significant mediation effect value of 0.015. (3) Cognitive engagement also acts as a mediator in the mechanism between the student–GAI interaction and learning achievement, evidenced by a significant and relatively strong mediating effect value of 0.046. (4) Self-efficacy and cognitive engagement in series mediate this positive association, with a serial mediating effect value of 0.011, which is relatively small in comparison but also shows significance. In addition, the propensity score matching (PSM) method is applied to alleviate self-selection bias, reinforcing the validity of the results. The findings offer empirical evidence for the incorporation of GAI in teaching and learning.

## KEYWORDS

generative artificial intelligence (GAI), education, self-efficacy, cognitive engagement, learning achievement

## Introduction

Generative artificial intelligence (GAI) stands as a distinct and potent class of artificial intelligence. It generates human-like content based on deep learning models in response to diverse and complex commands and questions (Lim et al., 2023). One significant example is ChatGPT, which has garnered great attention for its impressive capabilities in generating



human-like answers and responding in a wide array of languages. A tossed stone raises a thousand ripples. GAI tools spark debates about the role of traditional human efforts (Else, 2023; Stokel-Walker, 2023), and prompted ethical considerations like matters of originality and potential plagiarism (Lim et al., 2023; Yu, 2023).

In the field of education, GAI tools have demonstrated their unprecedented ability in many disciplines in a short time (e.g., Baidoo-Anu and Owusu Ansah, 2023; Dwivedi et al., 2023; Kieser et al., 2023; Peres et al., 2023). There are different attitudes towards the application of GAI among educators. While critics like Noam Chomsky (Open Culture, 2023) argue that GAI is “basically high-tech plagiarism” and “a way of avoiding learning,” many educators indicate that GAI could help improve instructional processes, such as personalized tutoring (Baidoo-Anu and Owusu Ansah, 2023; Kasneci et al., 2023), automated essay grading (Terwiesch, 2023), interactive learning (Kasneci et al., 2023), adaptive studying (Pedro et al., 2019), and producing multiple examples and explanations in teaching and learning (Mollick, 2023). Some education bodies, such as New York City public schools (Lukpat, 2023), announced the ban on the use of ChatGPT. However, prohibiting GAI tools may have harmful effects such as the Streisand Effect (Jansen and Martin, 2015) and psychological resistance (Brehm, 1989). The former makes GAI more popular when banned, and the latter triggers student resistance to the rules (Brehm, 1989). Promoting an understanding of GAI technologies, instructing students on beneficial engagement with these tools, and openly debating their merits and drawbacks present a more enduring solution than simply prohibiting their use (Kasneci et al., 2023). Peres et al. (2023) also claim that our education should prepare students for their jobs after graduation, including mastering how to use these up-to-date tools.

On this basis, this paper takes an open and inclusive attitude towards the application of GAI, exploring the relationship between student–GAI interaction and their learning achievement. Even though students may utilize GAI to produce abundant content, it does not guarantee high achievement for them. Although many theoretical studies have discussed this link (e.g., Baidoo-Anu and Owusu Ansah, 2023; Dwivedi et al., 2023; Lim et al., 2023), relevant empirical evidence still needs to be provided. The initial objective of this study is to furnish empirical evidence concerning the association between student–GAI interaction and students’ learning achievement.

In addition, this paper investigates through which path the student–GAI interaction level links to their learning achievement. On the one hand, since GAI tools are easy to access (e.g., ChatGPT, driven by GPT3.5, is available for free), it is equivalent to providing users with multi-domain and executive-capable personal assistants at very low costs. With these powerful “personal assistants,” students may argue that they can solve tougher problems and complete more difficult tasks, that is, students could have a higher level of self-efficacy through the interaction with GAI. On the other hand, the real-time interactivity and instant feedback of GAI could motivate students to be more actively involved in learning tasks. Interaction with GAI is not limited by time and location. If a student has questions about course study, interaction with the teacher may require an appointment, while interaction with ChatGPT can be done at any time. These real-time interactive processes of GAI increase cognitive engagement in learning for students (Bian et al., 2018; Asiri et al., 2021), and the increased cognitive engagement in-turn relates to higher learning achievement (Zhu et al., 2009; Sedaghat et al., 2011; Wang and Eccles, 2012; Pietarinen et al., 2014).

Furthermore, students possessing greater self-efficacy are likely to exhibit increased cognitive engagement (Linnenbrink and Pintrich, 2003; Walker et al., 2006). They are more likely to be interested in learning activities since they intend to believe they can succeed and are more willing to invest effort to explore and understand knowledge, thereby enhancing cognitive engagement. Therefore, we propose that self-efficacy and cognitive engagement in series mediate the mechanism between students–GAI interaction and their learning achievement.

The remainder of this paper is organized as follows: Section 2 introduces the theory and hypothesis development, section 3 describes the questionnaire participants, indicators, and scales, section 4 presents the regression models and bootstrap mediating effect test, and section 5 discusses the conclusions, implications, and limitation.

## Theoretical review and research hypothesis

### The relationship between student–GAI interaction and learning achievement

In the field of education, the extraordinary ability of GAI has attracted significant attention among educators. The advantages of GAI include but are not limited to promoting personalized and interactive learning (Kasneci et al., 2023), providing quick feedback (Baidoo-Anu and Owusu Ansah, 2023), and generating prompts for formative assessment activities (Dijkstra et al., 2023; Terwiesch, 2023).

Interactive theory is usually used to describe the interaction between human and human, human and machine, as well as human and environment (Zhou and Wei, 2010; Freeman and Ambady, 2011). According to Interactive Theory, the feedback that individuals receive during the interaction process is important (Freeman and Ambady, 2011). GAI can provide real-time and personalized feedback (Baidoo-Anu and Owusu Ansah, 2023; Kasneci et al., 2023), which helps students to have a more accurate assessment of their strengths and weaknesses, thus making targeted improvements. Besides, GAI can dynamically adjust instructional content and methods according to the needs and reactions of each student, which is consistent with the viewpoint in Interactive Theory that effective interaction must be “bidirectional” and “dynamic” (Freeman and Ambady, 2011; Nowland et al., 2018). Compared with interacting with teachers, GAI could provide instant and continuous interaction without being limited by time and location. Compared with traditional educational tools such as books and exercises, GAI might be more interactive and can increase students’ cognitive engagement through gamification and incentive mechanisms. On the other hand, GAI could provide rich educational resources and diverse learning methods (learning games, videos, tests, simulation experiments, etc.), helping students carry out effective learning, memory, and a general understanding of reasoning, thereby enhancing and consolidating metacognitive knowledge for students (Vrugt and Oort, 2008; Azevedo, 2020). Following these lines, we propose that GAI is conducive to improving learning achievement and bring up Hypothesis 1:

*Hypothesis 1:* There is a positive relationship between student–GAI interaction and their learning achievement.

## The mediating role of self-efficacy

Self-efficacy theory explains the level of confidence an individual develops on a particular task (Bandura, 1997), which includes an individual's assessment of their own ability to achieve the goal and their confidence in achieving it. Self-efficacy is gradually formed through individual experience, observation, and interaction. In educational contexts, students with higher self-efficacy tend to lead to higher academic achievement because they believe they can achieve their goals (Yokoyama, 2019).

This study indicates that GAI tools could serve as powerful assistants for students, leading to their increased self-efficacy. First, GAI has incredible capabilities to perform complex tasks such as writing articles (O'Connor, 2022), stories, poems, essays (Lucy and Bamman, 2021), images (Reed et al., 2023), providing textual summaries or extensions or even writing and debugging raw computer code (Kalliamvakou, 2022; Tate et al., 2023). In the interacting process, GAI tools demonstrated executive force and creativity that are unimaginable by humans (Baidoo-Anu and Owusu Ansah, 2023). Through interacting with these technologies, students can realize they may use GAI tools to create brilliant and satisfying content or solve more difficult tasks. They may have a higher assessment of their ability to achieve goals with the assistance of GAI. That is, student-GAI interaction could improve students' self-efficacy.

Second, GAI can generate content based on students' understanding level and subject background, providing a personalized learning experience. Students' knowledge backgrounds are different in varied subjects. For those concepts that are abstract or completely foreign, they may need multiple explanations and cases to understand them (Ericsson and Pool, 2016). Creating multiple interpretations of a concept is a complex and time-consuming task for the instructor (Mollick, 2023). Tailoring explanations to students' learning levels also requires the instructor to pay close attention to new trends and students' cognitive loads (Lim et al., 2023). With limited time and energy, it may not be possible for instructors to take into account every student with their diverse needs, while GAI tools can help to improve this situation (Mollick, 2023). When interacting with AI, students can feel that they are in a tailor-made and personalized learning environment (Mollick, 2023), and can access and understand complex knowledge more easily (Peres et al., 2023). Thus, students might be more confident in learning, manifested by a higher level of self-efficacy.

On the other hand, higher self-efficacy links to improved academic performance has been extensively studied in the existing educational literature (Schunk, 1995; Robbins et al., 2004; Cassidy, 2015; Honicke and Broadbent, 2016; Doménech-Betoret et al., 2017; Yokoyama, 2019). Self-efficacy has been proven to affect students' effort, persistence, interest, and achievement in learning activities (Schunk, 1995). Students with higher self-efficacy were more engaged, worked harder, persisted longer, showed greater interest in learning, and achieved higher grades (Schunk, 1995; Robbins et al., 2004; Yokoyama, 2019). Based on the above discussion, we hypothesize:

*Hypothesis 2:* The positive relationship between student-GAI interaction and learning achievement is mediated by students' self-efficacy.

## The mediating role of cognitive engagement

From a constructivist perspective, cognitive engagement in learning is the extent to which students mentally invest in their learning activities, such as applying knowledge and cognitive tactics to accomplish the task (Chapman, 2002). Unlike physical engagement, cognitive engagement focuses more on mental activities such as thinking, planning, problem-solving, and decision-making (Greene, 2015). In educational settings, cognitive engagement is often used to describe how active students are in a course learning and whether they are actively thinking, solving problems, and interacting with material and massive information. The higher level of cognitive engagement students exhibit in the learning process can result in better related academic outcomes (Zhu et al., 2009; Sedaghat et al., 2011; Wang and Eccles, 2012; Pietarinen et al., 2014).

Based on Interactive Theory, we argue that the degree to which students interact with GAI affects students' cognitive engagement and thus relates to learning achievement. The Interactive Theory views the feedback individuals receive during the interaction process to be very important (Hyland and Hyland, 2019). GAI tools' real-time interactivity and instant feedback can motivate students to engage more actively in learning. GAI can also provide abundant images, information and examples to stimulate students to think further and explore deeper. For example, the rich explanations generated by ChatGPT can attract interest and arouse the curiosity of students, thereby enhancing their cognitive engagement in the learning process.

More importantly, it should be noted that the content a GAI tool generates is strongly dependent on the quality and nature of the inputs provided to it (Chatterjee and Dethlefs, 2023; Terwiesch, 2023). For example, asking ChatGPT a specific question could return a decent answer, while without any specifics, the answers provided by ChatGPT seem terse and biased (Lim et al., 2023). Therefore, in order to get an accurate answer, students need to conduct multiple rounds of questions and constantly revise the wording of the questions. This process can enhance students' cognitive engagement. In addition, ChatGPT may give fake or erroneous references and generate flawed and confident answers (Van Dis et al., 2023). Students cannot completely rely on these content without any thinking and doubt. They need to interact with existing materials to evaluate the quality of contents generated. This process of continuous interaction can subtly improve students' cognitive engagement in learning. Hence, this study proposes:

*Hypothesis 3:* The positive relationship between student-GAI interaction and learning achievement is mediated by students' cognitive engagement.

## The serial mediating role of self-efficacy and cognitive engagement

Existing studies have investigated the relationship between self-efficacy and cognitive engagement, indicating that higher self-efficacy leads to improved cognitive engagement (Linnenbrink and Pintrich, 2003; Walker et al., 2006). Students with higher self-efficacy are more

willing to believe they can succeed and invest time and cognitive effort to explore and understand new knowledge. When encouraged by challenges, they are more confident in their ability to overcome difficulties and thus be more actively seeking strategies to solve problems. They also generally have a more persistent drive to learn. According to deep learning strategies, students with high self-efficacy are more inclined to adopt deep learning strategies such as thinking, analysis, and discussion. They are full of confidence in their abilities and believe that the time and energy they invest will be rewarded and bring success. So they are not stingy in investing cognitive efforts to deepen their learning and understanding of the content. To sum up, students with high self-efficacy are more inclined to maintain a high cognitive engagement in the learning process.

The multiple mediation model applies to a situation where there are multiple mediating variables between the independent and the dependent variables (Liu and Ling, 2009). The serial mediation model, one type of the multiple mediation models, applies to a situation in which multiple mediating variables also show relationships (e.g., Allen and Griffeth, 2001; Sun et al., 2022; Wang et al., 2022), while the parallel mediation model applies to the model that views multiple mediating effects as parallel effects (Niehoff, 2005). Based on motivation theory and self-efficacy theory, a student's belief in ability and desire to participate in a particular activity will be positively related to his/her subsequent performance (Greene and Miller, 1996; Ryan and Deci, 2000). Additionally, empirical relationships have been found between student self-efficacy and cognitive engagement (Greene and Miller, 1996; Linnenbrink and Pintrich, 2003; Walker et al., 2006). Hence, we propose that self-efficacy and cognitive engagement could act as serial mediators in the relationship between the level of student–GAI interaction and learning achievement rather than parallel mediators. To elaborate, the serial relationship is: student–GAI interaction leads to increased self-efficacy, which then boosts cognitive engagement, and ultimately links to higher learning achievement. Combined with the previous discussion, we put forth Hypothesis 4 as follows:

*Hypothesis 4: Students' self-efficacy and cognitive engagement in series mediate the positive relationship between student–GAI interaction and learning achievement.*

To sum up, our hypothetical model as shown in Figure 1.

## Materials and methods

### Participants

Data for this study was collected online via Wenjuanxing<sup>1</sup>, a web-based survey platform comparable to Mechanical Turk or Qualtrics, which is applied for conducting surveys in China. Wenjuanxing allows for nationwide responses and is widely utilized in research related to behavior and psychology (Sun et al., 2022). Based on this nationwide platform, our participants come from different grades, provinces, schools, and majors. Unlike

undergraduates in other countries and regions who often be graded relying on essay writing, undergraduates in China are graded mostly relying on closed-book examinations<sup>2</sup>. These examinations usually require students to take closed-book tests in the class room without using any electronic products, which allays concerns that students use GAI to write essays and get false grades.

We set criteria in the participant recruitment phase, such as “participation in undergraduate education” and “over 18 years old,” so that the platform could collect responses that matched our research topic goals. Attention check questions (e.g., “Please indicate strongly disagree for this question.”) were included in the questionnaire to ensure respondents paid sufficient attention to each question (DeSimone et al., 2015). A total of 440 questionnaires were returned. Respondents who failed any attention check questions were excluded (28 out of 440). Due to the difficulty in guaranteeing the quality of the online surveys, we excluded responses that deviated by more than three standard deviations to prevent skewing from participants who may have selected the same answers throughout the questionnaire (23 out of 440). A total of 389 valid participants were obtained. Existing studies have shown that a valid sample size of 350–500 can proxy a target population of 5,000 or more (Sun et al., 2022), which indicates that our sample size of 389 is sufficient.

## Measurement

### Student–GAI interaction scale

There are few empirical studies on the interaction between students and generative AI. We selected a scale from a similar study and reformulated it to match the subject of this study (Sun et al., 2022). For example, the original item, “Teacher–student interaction is getting longer in online education,” was reformulated as “Student–GAI interaction is getting longer in my course learning.” This scale has a total of 4 items. In addition to the example, others are “When I have questions during course learning, I use GAI to seek answers,” “I use GAI to ask for advice when doing course tasks,” and “My classmates are becoming more and more interested in GAI tools.” We use a 5-point Likert scale with scores ranging from 1 to 5, where 1 indicates strongly disagree, and 5 indicates strongly agree. We asked participants to recall experiences of using GAI during courses study and answer these questions. If the participant has no relevant experience interacting with GAI, the corresponding question can be filled in as 1 (strongly disagree). Improved scores on this scale indicate a greater degree of interaction between students and GAI, and the scores will be low if the student lacks experience interacting with GAI during the learning process. In order to optimize the content validity of the scale, we conducted a test in a small scope. The scale was discussed with 17 relevant researchers, experts, and teachers. They review the scale and comment on questions that are complex, unclear, or ambiguous. Based on their comments, the scale was adjusted to optimize content validity. The Cronbach's alpha value for the student–GAI interaction is deemed acceptable:  $\alpha = 0.974$ .

<sup>2</sup> According to our investigation at a university in southwest China. Most courses (about 70%) graded for undergraduate students rely primarily on closed-book exams.

<sup>1</sup> Wenjuanxing website: <https://www.wjx.cn>, Available on Aug. 8, 2023.

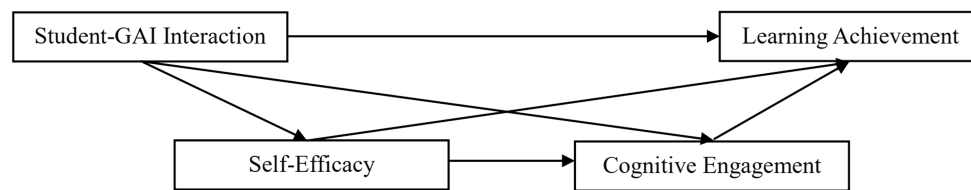


FIGURE 1  
Hypothetical model of the serial mediating effect.

## Cognitive engagement scale

The Cognitive Engagement Scale is derived from Wang et al. (2014) and Duncan and McKeachie (2005). We incorporate it into the setting of course learning to assess students' cognitive engagement in course learning. For example, rewrite the original items "I try to figure out the hard parts on my own" and "I search for information from different places and think about how to put it together" (Wang et al., 2014) into "In course learning, I try to figure out the hard parts on my own" and "In course learning, I search for information from different places and think about how to put it together." The self-efficacy scale has a total of 6 items. Students responded to these items on a 5-point Likert scale ranging from strongly disagree (1) to strongly agree (5). Cronbach's alpha value for the Cognitive Engagement Scale is deemed acceptable:  $\alpha = 0.713$ .

## Self-efficacy scale

The Self-efficacy Scale is developed by Schwarzer and Jerusalem (1995) and Luszczynska et al. (2005) to assess students' self-efficacy. Self-efficacy refers to the individual's perception or belief about whether he can take adaptive behavior in the face of environmental challenges. This perception of "what can be done" reflects an individual's sense of control over the environment. Therefore, self-efficacy is whether one can confidently view his/her ability to deal with various pressures. The self-efficacy scale has a total of 10 items (e.g., 'I can always manage to solve difficult problems if I try hard enough.') Students responded using a 5-point Likert scale anchored by 1 (strongly disagree) to 5 (strongly agree). Higher scores on this item indicate a higher degree of self-efficacy for each participant. Cronbach's alpha value for the Self-efficacy Scale was deemed acceptable:  $\alpha = 0.745$ .

## Learning achievement

Learning achievement is measured by the difference between the 2023 academic year GPA and the 2022 academic year GPA for each participant. GPA is a commonly used academic performance evaluation index, which is used to measure the average grade level obtained by students within a semester or an academic year. In the first-round questionnaire, we asked the participants to report their 2022 academic year GPA.<sup>3</sup> Since ChatGPT launched in November 2022, the 2022 academic year GPA can be considered a pre-test score.

In the second-round questionnaire, we asked participants to report their 2023 academic year GPA, which can be considered as a post-test score. Some students may already have good grades, while others may already have poor grades. Therefore, considering the influence of interacting with GAI, we use the relative value (post-test score minus pre-test score) rather than the absolute value to measure students' learning achievement. For comparability, we have performed dimensionless processing (i.e., z-score standardization) on variable values to eliminate the influence of dimension.

## Common method variance

Several methods were employed in this research to mitigate the risk of common method variance (Podsakoff et al., 2003). First, respondents were requested to complete surveys at two-time points 1 week apart. In the first round of questionnaires, we measured the independent variable (student-GAI interaction) and collected the GPA of the participants in the 2022 academic year and other individual characteristics. In the second round of questionnaires, we measured the mediator variable (self-efficacy and cognitive engagement) and collected the GPA of the participants in the 2023 academic year. Second, respondents were assured of anonymity and were uninformed about the specific objectives of the survey. Third, they were informed that there were no correct or incorrect responses and that their participation did not have any personal repercussions, encouraging honest answers. Fourth, the survey questions were presented in a random order. Last, we applied Harman's single-factor test, and the results showed that the single-factor model explained 19.044% of the variance, which indicates that common method variance was not a concern in this study.

## Results

For the analysis of the questionnaire results, we use the bootstrap method to test the serial mediation effect of the indicators. The bootstrap method is a kind of non-parametric Monte Carlo method (Preacher and Hayes, 2004, 2008). Its essence is to re-sample the observation information, and then make statistical inferences on the overall distribution characteristics. Since this method makes full use of the given observation information, it does not require other assumptions of the model and adding new observations. Thus, the bootstrap method has the characteristics of robustness and high efficiency.

Compared with other statistical methods (e.g., regression analysis or structural equation modeling), the Bootstrap method has the

<sup>3</sup> The academic year in Chinese universities starts in September and ends in July of the following year, so the 2022 academic year GPA refers to the GPA score from September 2021 to July 2022. During that time period, GAI has not yet started to be widely used.



TABLE 1 Respondents' profiles.

Variable	Category	Frequency	Percentage
Gender	Male	191	49.1%
	Female	198	50.9%
Age	18–20	117	30.1%
	21–23	173	44.5%
	24–26	57	14.7%
	27 or over	42	10.8%
Major type	Skill	141	36.2%
	Theory	130	33.4%
	Language	118	30.3%

TABLE 2 Descriptive statistics and Pearson correlation coefficient.

	Mean	S.D.	1	2	3	4
1. Learning Achievement	0.000	1.000	1.000			
2. Student–GAI Interaction	2.445	1.496	0.218**	1.000		
3. Self-Efficacy	3.611	0.420	0.314**	0.137**	1.000	
4. Cognitive Engagement	3.806	0.525	0.419**	0.271**	0.434**	1.000

\*\* indicates that correlation is significant at the 0.01 level (two-tailed). The mean and standard deviation of learning achievement in the table are dimensionless, and the mean and standard deviation of the original data are 0.005 and 0.843.

advantage that it does not rely on the specified distribution assumption and is applicable to small sample sizes and complex models. In this study, we applied the bootstrap method to perform the serial mediation effect test using the PROCESS v4 macro test (proposed by Hayes, 2017) in SPSS 26.0.

## Descriptive and correlation analysis

Table 1 presents the demographic profiles of the respondents. Of the 389 respondents, 49.1% are male and 50.9% are female. 74.6% of respondents are between age 18 and 23. The number of respondents in the three professional types (skill, theory, and language) is relatively balanced.<sup>4</sup>

Table 2 presents the means, standard deviations, and Pearson correlation coefficient for the main indicators. The results showed that learning achievement, student–GAI interaction, self-efficacy, and cognitive engagement are correlated at the 1% statistical significance level. Specifically, statistically significant correlations are observed

between learning achievement and student–GAI interaction ( $r=0.218$ ,  $p<0.01$ ), learning achievement and self-efficacy ( $r=0.314$ ,  $p<0.01$ ), learning achievement and cognitive engagement ( $r=0.419$ ,  $p<0.01$ ), student–GAI interaction and self-efficacy ( $r=0.137$ ,  $p<0.01$ ), student–GAI interaction and cognitive engagement ( $r=0.271$ ,  $p<0.01$ ), and self-efficacy and cognitive engagement ( $r=0.434$ ,  $p<0.01$ ).

## Analysis of serial mediating effect

Table 3 lists the estimates of the regression models: Model 1 estimates the relationship between student–GAI interaction and learning achievement; Model 2 estimates the relationship between student–GAI interaction and self-efficacy; Model 3 estimates the association of student–GAI interaction and self-efficacy with cognitive engagement; and Model 4 estimates the association of student–GAI interaction, self-efficacy, and cognitive engagement with learning achievement.

Hypothesis 1 posits that there is a significantly positive relationship between student–GAI interaction and their learning achievement. As seen in Model 1 in Table 3, the coefficient value for the student–GAI interaction level is found to be significant and positive ( $\beta=0.146$ ,  $p<0.01$ ), thus providing support for Hypothesis 1. Comparing estimates in Model 1 and Model 4, we find that the size and significance for the coefficient of the student–GAI interaction decreased ( $\beta$  changes from 0.146 to 0.073, and statistical significance changes from 0.01 to 0.05) after the mediating indicators are included. This indicates that self-efficacy and cognitive engagement mediate part of the relationship between student–GAI interaction and learning achievement.

Regarding self-efficacy, estimates in Model 2 show that student–GAI interaction positively relates to self-efficacy at the 1% statistical significance level ( $\beta=0.038$ ,  $p<0.01$ ), and Model 4 indicates that self-efficacy positively relates to learning achievement ( $\beta=0.401$ ,  $p<0.01$ ), which provide support for Hypothesis 2. With respect to cognitive engagement, results in Model 3 indicate that student–GAI interaction positively relates to cognitive engagement at the 1% statistical significance level ( $\beta=0.069$ ,  $p<0.01$ ), and Model 4 shows that cognitive engagement positively relates to learning achievement ( $\beta=0.668$ ,  $p<0.01$ ), which supports Hypothesis 3. Moreover, estimates in Model 3 also show that higher self-efficacy is linked to increased cognitive engagement ( $\beta=0.431$ ,  $p<0.01$ ). The above results provide support for Hypothesis 4.

We further analyzed the size and significance of each mediation by the Bootstrap method. Confidence intervals for indirect effects were calculated by Bootstrap repeated sampling to determine the statistical significance of mediating effects. The results are shown in Table 4.

Table 4 reveals that the Bootstrap 95% confidence interval for the mediation roles of both self-efficacy and cognitive engagement does not include zero (both the lower and upper limits exceed zero). This confirms that self-efficacy and cognitive engagement significantly mediate part of the relationship between student–GAI interaction and learning achievement. The value of the total indirect effect is 0.0728, which is primarily achieved through three pathways: (1) indirect effect 1 (0.0154): student–GAI interaction  $\rightarrow$  self-efficacy  $\rightarrow$  learning achievement; (2) indirect effect 2 (0.0463): student–GAI

<sup>4</sup> Skills mainly represent engineering majors, such as machinery, materials, automation, computers, etc. Theory represents science and liberal arts majors, such as theoretical physics, mathematics, economics, management, philosophy, education, history, etc. Languages include English, German, French, Japanese, Spanish and other language majors.

TABLE 3 Regression estimates.

Dependent variable	Model 1	Model 2	Model 3	Model 4
	Learning Achievement	Self-Efficacy	Cognitive Engagement	Learning Achievement
Student–GAI Interaction	0.146***	0.038***	0.069***	0.073**
	(4.391)	(2.717)	(4.510)	(2.336)
Self-Efficacy			0.431***	0.401***
			(7.869)	(3.442)
Cognitive Engagement				0.668***
				(6.641)
R <sup>2</sup>	0.047	0.019	0.195	0.219
F	19.282	7.382	46.867	36.015
p value	0.000	0.007	0.000	0.000

Numbers in parentheses are *t*-statistics value. Statistical significance is denoted by: \*\*\**p* < 0.01, \*\**p* < 0.05, and \**p* < 0.1.

TABLE 4 Bootstrap results for the mediation effect.

Mediating path	Indirect effect	Boot standard error	<i>p</i> value	95% confidence interval		Relative mediation effect	Total mediation effect
				Lower limit	Upper limit		
Total Effect	0.1456	–	0.000	0.0804	0.2108	–	100.00%
Total Indirect Effect	0.0728	0.0161	–	0.0429	0.1068	100.00%	50.00%
Indirect Effect 1	0.0154	0.0073	–	0.0029	0.0312	21.15%	10.58%
Indirect Effect 2	0.0463	0.0126	–	0.0237	0.0736	63.60%	31.80%
Indirect Effect 3	0.0111	0.0049	–	0.0028	0.0222	15.25%	7.62%

Level of confidence for all confidence intervals in output is 95%. Number of bootstrap samples for percentile bootstrap confidence intervals is 5000. Mediating path in this table are: Indirect effect 1 represents “Student–GAI Interaction→Self-Efficacy→Learning Achievement”; Indirect effect 2 represents “Student–GAI Interaction→Cognitive Engagement→Learning Achievement.” Indirect effect 3 represents “Student–GAI Interaction→Self-Efficacy→Cognitive Engagement→Learning Achievement.”

interaction→cognitive engagement→learning achievement; and (3) indirect effect 3 (0.0111): student–GAI interaction→self-efficacy→cognitive engagement→learning achievement. Indirect effect 1, indirect effect 2, and indirect effect 3 accounted for 10.58, 31.80, and 7.62% of the total effect, respectively. The bootstrapping results further support that self-efficacy and cognitive engagement act as separate mediators and also jointly act as serial mediators between the student–GAI interaction level and learning achievement for students.

We summarize these results as Figure 2.

### Self-selection bias

There may be self-selection bias concerns in the mediating models. For example, there may be differences in the characteristics of students interacting with GAI tools or not. These differences could lead to differences in their self-efficacy and cognitive engagement, that is, self-selection bias. To solve this issue, we use the propensity scores matching (PSM) technique to establish comparable sets for the treatment and control groups, and control the characteristic differences between the two groups. The basic idea of PSM is that for each treated individual, find one or more individuals who are not treated but are very similar to him/her in other observed characteristics

(such as age, gender, major, etc.) (Rosenbaum and Rubin, 1983). These similar individuals are usually matched through propensity scores.

Specifically, in the context of course learning, the control group includes the participants who did not interact with GAI, indicated by  $G_i = 0$ , and the treatment group includes the participants who interacted with GAI, indicated by  $G_i = 1$ . 50.9% of the sample is in the control group and 49.1% of the sample is in the treat group. The output variable for the treatment group is expressed as  $Output_{i1}$ , and the control group is  $Output_{i0}$ . In our study, the output variables are self-efficacy and cognitive engagement. We focused on the different effects on participants’ output variables before and after interacting with GAI. This difference is known as the average treatment effect on the treated (ATT) (Liang et al., 2022):

$$\tau_{ATT} = E(\text{Output}_{i1} | G_i = 1) - E(\text{Output}_{i0} | G_i = 1)$$

where  $E(\text{Output}_{i1} | G_i = 1)$  denotes the average value of output variables for the participants that interacting with GAI, and  $E(\text{Output}_{i0} | G_i = 1)$  denotes the average value of output variables by assuming those participants did not interact with GAI. Since  $E(\text{Output}_{i0} | G_i = 1)$  is unobservable, we replace it by the average value for participants that did not interact with GAI in course learning, who have similar characteristics with the former, expressed as

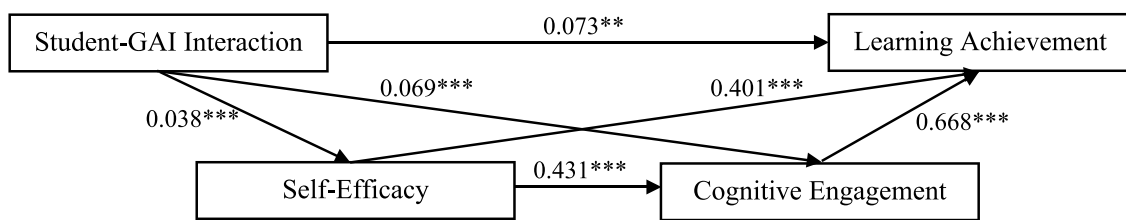


FIGURE 2

The serial mediation model of the impact of student–GAI interaction level on learning achievement. \*\*\* indicates  $p < 0.01$  and \*\* indicates  $p < 0.05$ .

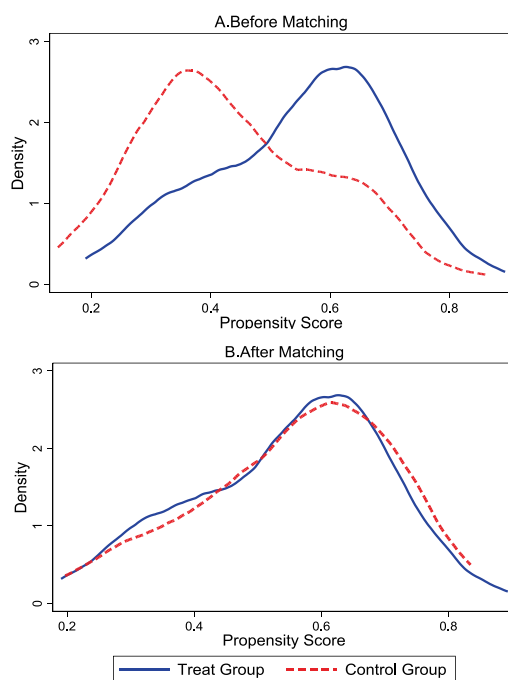


FIGURE 3

The graph of kernel density functions of treatment and control groups before and after matching.

$E(\text{Output}_{i0} | G_i = 0)$ . So the ATT can be computed as the difference in average of output variables between interacting and non-interacting with GAI participants:

$$\tau_{\text{ATT}} = E(\text{Output}_{i1} | G_i = 1) - E(\text{Output}_{i0} | G_i = 0)$$

The equation builds on the condition that the characteristics of the participants in both treatment and control groups are the same. Following to Villalonga (2004), we set the propensity scores as the probability that the participant tends to interact with GAI on the vector of independent variables  $X_i$ :

$$p(X_i) \equiv \Pr(G_i = 1 | X_i) = E(G_i | X_i)$$

where vector  $X_i$  represents the characteristics that may also affect the output variables, including the participant's age, gender, grade,

province, major, senior high school subjects (Arts/Sciences), SAT score, 2022 academic-year GPA, student union activities and student awards. This information was collected from respondents in the first-round of the questionnaire.

Using the nearest neighbor matching method, we aligned the treatment and control sets and applied probit regression for estimating the propensity scores. The kernel density distributions for both the treatment and control groups are represented in Figure 3.

Figure 3A shows that there is a significant difference in the kernel density function between the treatment group (participants that interact with GAI) and the control group (participants that not interact with GAI) before matching. Figure 3B shows that the kernel density function images of the two groups are closer after matching, indicating that the treatment and control groups are more comparable in terms of these individual characteristics. Based on PSM, the treatment effect results for self-efficacy and cognitive engagement are summarized in Table 5.

From Table 5, we can see that the average values of self-efficacy and cognitive engagement are significantly higher in the treatment group (participants that interact with GAI) than in the control group (participants that do not interact with GAI), and ATT values are significantly positive (ATT = 0.1829 for self-efficacy, and ATT = 0.2000 for cognitive engagement,  $|t| > 1.96$ ). After the propensity score matching, the ATT values of self-efficacy and cognitive engagement are still statistically significant (ATT = 0.1703 for self-efficacy, and ATT = 0.2,170 for cognitive engagement,  $|t| > 1.96$ ). The findings suggest that participants interacting with GAI would have higher levels of self-efficacy and cognitive engagement even after controlling the self-selection bias.

## Discussion

This study explores the relationship between students' interaction with GAI and learning achievement, considering the mediating roles of self-efficacy and cognitive engagement. The empirical study collected responses of questionnaires from 389 participants. The results showed: (1) Overall, there is a significantly positive relationship between the degree of student–GAI interaction and their learning achievement. (2) This positive relationship is mediated by self-efficacy, with a significant mediation effect value of 0.015. (3) Cognitive engagement also acts as a mediator in the mechanism between the student–GAI interaction and learning achievement, evidenced by a significant and relatively strong mediating effect value of 0.046. (4) Self-efficacy and cognitive engagement in series mediate the positive relationship between the degree of student–GAI interaction and their

TABLE 5 The PSM results.

Treatment variable	Output variable	Sample	Average value of output variable for the treatment group	Average value of output variable for the control group	ATT	t-stat
Interact with GAI	Self-Efficacy	Unmatched	3.6939	3.5110	0.1829***	3.45
		Matched	3.6934	3.5231	0.1703**	2.03
Interact with GAI	Cognitive Engagement	Unmatched	3.9333	3.7333	0.2000***	3.75
		Matched	3.9330	3.7160	0.2170***	2.72

50.9% of the sample is in the control group and 49.1% of the sample is in the treat group. Statistical significance is denoted by: \*\*\* $p < 0.01$ , \*\* $p < 0.05$ , and \* $p < 0.1$ .

learning achievement, with a serial mediating effect value of 0.011, which is relatively small in comparison but also shows significance.

### Theoretical implications

This study contributes to educational research in three ways. First, this study provided empirical evidence about the relationship between student–GAI interaction and their learning achievement, responding to the calls that require empirical insights to help us better understand the implications of GAI for education so as to build useful knowledge bases (Pedro et al., 2019; Lim et al., 2023; Peres et al., 2023). GAI tools have been mentioned as “strength enhancers” for instructors (Kasneci et al., 2023), helping to design instruction to accommodate students of different comprehension abilities and subject backgrounds, but applying GAI in teaching needs to be based on abundant empirical evidence (Mollick, 2023). Some scholars believe that using GAI tools can help students better understand the learning content and improve the mastery of knowledge, thereby enhancing their academic achievement (Baidoo-Anu and Owusu Ansah, 2023), while others argue that GAI may cause students to be lazy and over-dependent with no or little analytical abilities, thereby reducing the academic performance (Pedro et al., 2019; Lim et al., 2023). This study offers empirical support that the interaction between students and GAI is positively related to students’ academic performance, providing a reference for future incorporation of GAI in teaching.

Second, this study sheds light on the mediating roles of self-efficacy and cognitive engagement in the mechanism between the student–GAI interaction level and learning achievement. Extensive literature demonstrates that environment, actions, and behavior affect the psychological factors, which influence outcomes (e.g., Liang et al., 2021; Nagadeepa, 2021; Chhetri and Baniya, 2022; Wei et al., 2023). While many GAI-related studies have illustrated that GAI tools are characterized by personalization and interactivity (Baidoo-Anu and Owusu Ansah, 2023; Kasneci et al., 2023), it is unclear how these characteristics are associated with students’ psychological factors in course learning. This study indicates that students’ interaction with GAI positively relates to their self-efficacy and cognitive engagement in course learning, which enhances our understanding of the psychological channels through which interaction with GAI links to students’ learning.

Third, this study provides new insights into the mediating effects of self-efficacy and cognitive engagement in education by introducing interactions with novel technical tools as the independent variables. There is a considerable amount of research pointing out that self-efficacy and cognitive engagement play mediating roles in relation to

students’ motivation, understanding, learning and achievements (Schunk and Pajares, 2001; Bandura, 2006; Van Dinther et al., 2011; Chong et al., 2018; Nagadeepa, 2021; Chhetri and Baniya, 2022; Wei et al., 2023). Following this line, we supplement empirical evidence that self-efficacy and cognitive engagement also mediate the association of students’ interaction with GAI tools and academic achievements and discuss their practical implications.

### Practical implications

#### The student–GAI interaction level affects learning achievement

This study finds a positive relationship between student–GAI interaction and their learning achievement during course learning. The results provide empirical support for the view that GAI positively relates to student learning performance, and we should take full advantage of GAI’s strengths rather than just banning it. We echo the opinion of Kasneci et al. (2023) and Mollick (2023) that if implemented carefully and thoughtfully in evidence-based teaching practices, artificial intelligence could be a “force multiplier” for teachers.

Based on this result, how to safely improve the level of interaction between students and GAI, so as to promote its impact on students’ learning achievement, is a concern in the education practices. First, like other technologies such as Python and MATLAB, instructors can introduce students to the correct way, steps, and precautions (Smaldino et al., 2006; Mollick, 2023) to use GAI tools, allowing students to use them to create text, images, audio, and other content according to course purposes, thereby stimulating the students–GAI interaction level. Second, problem-solving activities that require students to work with the GAI could be constructively designed (Oradee, 2012; Perera and Lankathilaka, 2023). For example, in a science class, let students work with GAI to analyze data, predict trends, and then come up with solutions together. Related assignments can require students to think deeply about the problem, analyzing different perspectives and possible solutions (Mousoulides et al., 2007). This encourages students to think independently, rather than simply relying on what the GAI provides. Third, encouraging students to be creative and expressive in their interactions with GAI. Let students know they can try out different ideas without fear of making mistakes. Related assignments can require students to express their ideas and come up with unique and creative solutions to ensure their work is original (Dorst and Cross, 2001). Last and importantly, we should establish a comprehensive evaluation system that includes interaction with GAI, so that students can clearly realize that the role of GAI is to assist, not completely replace their work. In the evaluation system, in order to



avoid plagiarism and other moral issues caused by using GAI, more attention should be paid to creativity, in-depth thinking and analysis, problem-solving and creative process, ability improvement etc. (Liu et al., 2023; Yu, 2023). It requires more effort and attention in the future. We further call for dialogue among researchers, educators, and educational institutions on how to safely and constructively improve the student–GAI interaction level to support student learning.

### The serial mediating role of self-efficacy and cognitive engagement

This study further sheds light on the mediating roles of self-efficacy and cognitive engagement in the mechanism between the student–GAI interaction level and learning achievement, and the degrees of mediating effects for different mediators vary. Among them, cognitive engagement serves as a mediator that has a greater effect on students' learning achievement, followed by self-efficacy.

Based on these findings, we suggest that educators could pay primary attention to the effect on learners' cognitive engagement when guiding them to interact with GAI. In the teaching design, instructors can consider designing challenging tasks that stimulate students' interest and require deep thinking (Herft, 2023; Küchemann et al., 2023). These tasks can involve problem-solving, creative expression, or practical application and require students to interact with the GAI to obtain valuable outputs. Besides, exploratory tasks can also be designed to encourage students to explore the different functions and applications of GAI independently. Guiding students to discover the potential of GAI based on practical problems to increase their curiosity and initiative (Abdelghani et al., 2022). After the tasks are completed, students are encouraged to engage in reflection and discussion, sharing their experience of interacting with GAI, biases corrected, challenges encountered, and insights gained from it. This helps to increase the cognitive engagement of the learners.

The improvement of students' self-efficacy can also be noticed in teaching practice interacting with GAI. Before a learning task begins, instructors could clearly state the goals and expectations of the task (Küchemann et al., 2023) so that students understand what they will achieve through their interaction with GAI. Clear goals help students develop self-confidence by knowing their efforts will be rewarded. When doing the task, prompt students to use the GAI to get more explanations and examples of concepts they do not understand. In addition, a feedback mechanism could be set during the interaction with GAI [see Herft (2023) and Jia et al. (2021) for more details to support and improve teaching assessment and feedback practices]. Students can know immediately whether their answers are correct. Timely feedback and recognition could also help to improve self-efficacy. After the assignment, share successful cases of interacting with GAI so that students can learn how others have achieved their academic goals by collaborating with GAI. This can stimulate students' enthusiasm and self-confidence. Besides, review with the students their accomplishments and progress. Make them aware of their own growth and progress, thereby increasing their confidence in their abilities. By improving self-efficacy, students can develop positive learning attitudes, which is also conducive to improving cognitive engagement and enhancing learning achievement.

Although this research explores the relationship among student–GAI interaction, psychological variables and academic performance from the perspective of students, how to use these conclusions to improve education practices should start from the perspective of

instructional design. Many instructional designs and techniques have proven valuable but are difficult to put into practice because they are time-consuming for overworked instructors (Kirschner et al., 2022; Mollick, 2023). Interacting with GAI could quickly and easily implement evidence-based instructional designs to provide guidance.

### Research limitations and prospects

This study empirically tests the mechanism between student–GAI interaction and their learning achievement, and has value in both statistical and practical significance. However, it has several limitations. First, the participants in this study were Chinese, as Chinese universities typically use closed-book exams that do not use any electronic tools to evaluate students' performance, but this may limit the generalizability of the findings. Future research may need to validate the conceptual model in other cultural contexts and countries to test how interactions with the GAI affect performance in different assessment modalities. Second, the relationship between the level of interaction with GAI and student learning achievement may also be affected by other factors, such as instructor ability, class atmosphere, school policies, etc. Hence, subsequent studies could investigate the specific conditions that moderate how interactions with GAI affect students' academic performance from varied perspectives. Third, this study only carried out the questionnaire survey from the perspective of students. Future research can analyze the impact of interaction with GAI from the perspective of instructors, or pair instructors and students. Fourth, there are many types of university courses. This study does not distinguish between different disciplines to study the impact of GAI on teaching and learning. Future research could differentiate between varied disciplines and investigate the impact of the interaction with GAI on teaching and learning in that discipline. For example, considering the transformation brought by GAI, management discipline may pay more attention to decision-making, technical disciplines may pay more attention to the purpose of technology implementation, and music and art may pay more attention to originality. Fifth, future research can also consider a moderating effect model that considers students' psychological factors as moderator variables and explores what psychological characteristics students possess are more susceptible to the impact of interaction with GAI tools. In addition, discussing the impact of students' interactions with GAI tools in subdivided dimensions (e.g., the quality, single time, number of interactions, as well as the types and quantities of GAI tools etc.) is also worth further exploring. Last, studies could utilize longitudinal data to meticulously explore the causal pathways, providing further validation and deepening these preliminary findings.

### Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

### Ethics statement

The studies involving humans were approved by College of Management Science, Chengdu University of Technology. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation



was not required from the participants or the participants' legal guardians/next of kin because questionnaires were collected anonymously through a nationwide online platform, and participants' names and other sensitive personal information were not identified. Participants were informed that there were no correct or incorrect responses and that their participation did not have any personal repercussions.

## Author contributions

JLi: Data curation, Formal analysis, Conceptualization, Methodology, Investigation, Funding acquisition, Writing – original draft. LW: Data curation, Funding acquisition, Investigation, Writing – original draft. JLu: Formal analysis, Methodology, Writing – review & editing. YY: Data curation, Funding acquisition, Investigation, Validation, Writing – review & editing. CF: Supervision, Project administration, Validation, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. JLi's work is

supported by the Chengdu Philosophy and Social Sciences Planning Project (No. 2023BS127) and the Philosophy and Social Science Research Fund of Chengdu University of Technology (No. ZDJS202203; YJ2023-QH005). LW's work is supported by the Project for Enhancing Teacher's Science and Technology Innovation Ability in Chengdu University of Information Technology (No. KYQN202243). YY's work is supported by the Southwest Minzu University Research Startup Funds (No. RQD2023055).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

- Abdelghani, R., Wang, Y.-H., Yuan, X., Wang, T., Sauzéon, H., and Oudeyer, P.-Y. (2022). GPT-3-driven pedagogical agents for training children's curious question-asking skills. *arXiv:2211.14228*
- Allen, D. G., and Griffeth, R. W. (2001). Test of a mediated performance-turnover relationship highlighting the moderating roles of visibility and reward contingency. *J. Appl. Psychol.* 86, 1014–1021. doi: 10.1037/0021-9010.86.5.1014
- Asiri, Y. A., Millard, D. E., and Weal, M. J. (2021). Assessing the impact of engagement and real-time feedback in a mobile behavior change intervention for supporting critical thinking in engineering research projects. *IEEE Trans. Learn. Technol.* 14, 445–459. doi: 10.1109/TLT.2021.3104817
- Azevedo, R. (2020). Reflections on the field of metacognition: issues, challenges, and opportunities. *Metacogn. Learn.* 15, 91–98. doi: 10.1007/s11409-020-09231-x
- Baidoo-Anu, D., and Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. *J. Comput. Soc. Sci.* 6:91. doi: 10.2139/ssrn.4337484
- Bandura, A. (1997). *Self efficacy: the exercise of control*. New York: Freeman
- Bandura, A. (2006). Adolescent development from an agentic perspective. F. Pajares and T. Urdan (Eds.), *Self-efficacy beliefs of adolescents*. Greenwich, CT: Information Age Publishing
- Bian, Y., Yang, C., Zhou, C., Liu, J., Gai, W., and Meng, X., (2018). Exploring the weak association between flow experience and performance in virtual environments, Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–12
- Brehm, J. W. (1989). Psychological reactance: theory and applications. *ACR North American Advances*, 16, 72–75
- Cassidy, S. (2015). Resilience building in students: the role of academic self-efficacy. *Front. Psychol.* 6:1781. doi: 10.3389/fpsyg.2015.01781
- Chapman, E. (2002). Alternative approaches to assessing student engagement rates. *Pract. Assess. Res. Eval.* 8:13.
- Chatterjee, J., and Dethlefs, N. (2023). This new conversational AI model can be your friend, philosopher, and guide. And even your worst enemy. *Patterns* 4:100676. doi: 10.1016/j.patter.2022.100676
- Chhetri, S. B., and Baniya, R. (2022). Influence of student-faculty interaction on graduate outcomes of undergraduate management students: the mediating role of behavioral, emotional and cognitive engagement. *Int. J. Manag. Educ.* 20:100640. doi: 10.1016/j.ijme.2022.100640
- Chong, W. H., Liem, G. A. D., Huan, V. S., Kit, P. L., and Ang, R. P. (2018). Student perceptions of self-efficacy and teacher support for learning in fostering youth competencies: roles of affective and cognitive engagement. *J. Adolesc.* 68, 1–11. doi: 10.1016/j.adolescence.2018.07.002
- DeSimone, J. A., Harms, P. D., and DeSimone, A. J. (2015). Best practice recommendations for data screening. *J. Organ. Behav.* 36, 171–181. doi: 10.1002/job.1962
- Dijkstra, R., Genç, Z., Kayal, S., and Kamps, J. (2023). Reading comprehension quiz generation using generative pre-trained transformers, The 23th international conference on artificial intelligence in education (AIED)
- Doménech-Betoret, F., Abellán-Roselló, L., and Gómez-Artiga, A. (2017). Self-efficacy, satisfaction, and academic achievement: the mediator role of students' expectancy-value beliefs. *Front. Psychol.* 8:1193. doi: 10.3389/fpsyg.2017.01193
- Dorst, K., and Cross, N. (2001). Creativity in the design process: co-evolution of problem-solution. *Des. Stud.* 22, 425–437. doi: 10.1016/S0142-694X(01)00009-6
- Duncan, T. G., and McKeachie, W. J. (2005). The making of the motivated strategies for learning questionnaire. *Educ. Psychol.* 40, 117–128. doi: 10.1207/s15326985ep4002\_6
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., et al. (2023). "So what if ChatGPT wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int. J. Inf. Manag.* 71:102642. doi: 10.1016/j.ijinfomgt.2023.102642
- Else, H. (2023). Abstracts written by ChatGPT fool scientists. *Nature* 613:423. doi: 10.1038/d41586-023-00056-7
- Ericsson, A., and Pool, R. (2016). *Peak: secrets from the new science of expertise* Random House.
- Freeman, J. B., and Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychol. Rev.* 118, 247–279. doi: 10.1037/a0022327
- Greene, B. A. (2015). Measuring cognitive engagement with self-report scales: reflections from over 20 years of research. *Educ. Psychol.* 50, 14–30. doi: 10.1080/00461520.2014.989230
- Greene, B. A., and Miller, R. B. (1996). Influences on achievement: goals, perceived ability, and cognitive engagement. *Contemp. Educ. Psychol.* 21, 181–192. doi: 10.1006/ceps.1996.0015
- Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: a regression-based approach*. Guilford publications
- Herft, A. (2023). *A teacher's prompt guide to ChatGPT aligned with What works best. Guide*. Retrieved on January, 23, 2023
- Honick, T., and Broadbent, J. (2016). The influence of academic self-efficacy on academic performance: a systematic review. *Educ. Res. Rev.* 17, 63–84. doi: 10.1016/j.edurev.2015.11.002
- Hyland, K., and Hyland, F. (2019). *Contexts and issues in feedback on L2 writing. Feedback Second Lang. Writ. Cont. Issues*, 47, 1–22. doi: 10.1017/9781108635547
- Jansen, S. C., and Martin, B. (2015). *The Streisand effect and censorship backlash*

- Jia, Q., Cui, J., Xiao, Y., Liu, C., Rashid, P., and Gehringer, E. F. (2021). All-in-one: multi-task learning bert models for evaluating peer assessments. *arXiv preprint arXiv:2110.03895*
- Kalliamvakou, E. (2022). Research: quantifying GitHub Copilot's impact on developer productivity and happiness. *The GitHub Blog*
- Kasneci, E., Seifler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* 103:102274. doi: 10.1016/j.lindif.2023.102274
- Kieser, F., Wulff, P., Kuhn, J., and Küchemann, S. (2023). Educational data augmentation in physics education research using ChatGPT. *arXiv preprint arXiv:2307.14475*
- Kirschner, P., Hendrick, C., and Heal, J. (2022). *How teaching happens: seminal works in teaching and teacher effectiveness and what they mean in practice* Routledge.
- Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K. E., et al. (2023). Physics task development of prospective physics teachers using ChatGPT. *arXiv preprint arXiv:2304.10014*
- Liang, J., Ma, J., Zhu, J., and Jin, X. (2021). Online or offline? How smog pollution affects customer channel choice for purchasing fresh food. *Front. Psychol.* 12:682981. doi: 10.3389/fpsyg.2021.682981
- Liang, J., Yang, S., Huang, X., and Zhu, J. (2022). Forward or backward: the impact of vertical integration direction on the bullwhip effect. *Int. J. Prod. Res.* 60, 6923–6944. doi: 10.1080/00207543.2022.2051765
- Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., and Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education* 21:100790. doi: 10.1016/j.ijme.2023.100790
- Linnenbrink, E. A., and Pintrich, P. R. (2003). The role of self-efficacy beliefs instudent engagement and learning in the classroom. *Reading Writ. Q.* 19, 119–137.
- Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., and Tian, J. (2023). Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology* 1:100017. doi: 10.1016/j.metrad.2023.100017
- Liu, S. S., and Ling, W. Q. (2009). Multiple mediation models and their applications. *Psychol. Sci.* 32, 433–435.
- Lucy, L., and Bamman, D. (2021). Gender and representation bias in GPT-3 generated stories. *Proceedings of the third workshop on narrative understanding*, 48–55
- Lukpat, A. (2023). CHATGPT banned in new York City public schools over concerns about cheating, and learning development. *Wall Street J.*
- Luszczynska, A., Gutiérrez-Doña, B., and Schwarzer, R. (2005). General self-efficacy in various domains of human functioning: evidence from five countries. *Int. J. Psychol.* 40, 80–89. doi: 10.1080/00207590440000041
- Mollick, E. R., (2023). Using AI to implement effective teaching strategies in classrooms: five strategies, including prompts. Including Prompts (March 17, 2023). Available at: <https://ssrn.com/abstract=4391243>
- Mousoulides, N., Sriraman, B., and Christou, C. (2007). From problem solving to modelling. *Education* 12, 23–47.
- Nagadeepa, C. (2021). Students' understanding and learning: mediation effects of cognitive engagement in online classes. *Turkish J. Comput. Math. Educ.* 12, 2932–2939. doi: 10.17762/turcomat.v12i6.6077
- Niehoff, B. P. (2005). *A theoretical model of the influence of organizational citizenship behaviors on organizational effectiveness*. New Research in Organizational Citizenship Behaviors, Nova, New York, 385–397
- Nowland, R., Necka, E. A., and Cacioppo, J. T. (2018). Loneliness and social internet use: pathways to reconnection in a digital world? *Perspect. Psychol. Sci.* 13, 70–87. doi: 10.1177/1745691617713052
- O'Connor, S. (2022). Open artificial intelligence platforms in nursing education: tools for academic progress or abuse? *Nurse Educ. Pract.* 66, –103537. doi: 10.1016/j.nepr.2022.103537
- Open Culture (2023). Noam Chomsky on ChatGPT. *Open Culture*. Available at: <https://www.openculture.com/2023/02/noam-chomsky-on-chatgpt.html>.
- Oradee, T. (2012). Developing speaking skills using three communicative activities (discussion, problem-solving, and role-playing). *Int. J. Soc. Sci. Human.* 2:533.
- Pedro, F., Subosa, M., Rivas, A., and Valverde, P. (2019). *Artificial intelligence in education: challenges and opportunities for sustainable development*
- Perera, P., and Lankathilaka, M. (2023). AI in higher education: a literature review of ChatGPT and guidelines for responsible implementation. *Int. J. Res. Innov. Soc. Sci.* 7, 306–314. doi: 10.47772/IJRISS.2023.7623
- Peres, R., Schreier, M., Schweidel, D., and Sorescu, A. (2023). On ChatGPT and beyond: how generative artificial intelligence may affect research, teaching, and practice. *Int. J. Res. Mark.* 40, 269–275. doi: 10.1016/j.ijresmar.2023.03.001
- Pietarinen, J., Soini, T., and Pyhältö, K. (2014). Students' emotional and cognitive engagement as the determinants of well-being and achievement in school. *Int. J. Educ. Res.* 67, 40–51. doi: 10.1016/j.ijer.2014.05.001
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., and Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *J. Appl. Psychol.* 88, 879–903. doi: 10.1037/0021-9010.88.5.879
- Preacher, K. J., and Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behav. Res. Methods Instrum. Comput.* 36, 717–731. doi: 10.3758/BF03206553
- Preacher, K. J., and Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav. Res. Methods* 40, 879–891. doi: 10.3758/BRM.40.3.879
- Reed, J., Alterio, B., Coblenz, H., O'Leary, T., and Metz, T. (2023). AI image-generation as a teaching strategy in nursing education. *J. Interact. Learn. Res.* 34, 369–399.
- Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., and Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? *Psychol. Bull.* 130, 261–288. doi: 10.1037/0033-2909.130.2.261
- Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55. doi: 10.1093/biomet/70.1.41
- Ryan, R. M., and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* 55, 68–78. doi: 10.1037/0003-066X.55.1.68
- Schunk, D. H. (1995). Self-efficacy and education and instruction. *Self-Eff. Adapt. Adjust. Theory Res. Appl.* 10, 281–303. doi: 10.1007/978-1-4419-6868-5\_10
- Schunk, D. H., and Pajares, F. (2001). The development of academic self-efficacy. A. Wigfield and J. S. Eccles (Eds.), *Development of achievement motivation*. San Diego, CA: Academic Press
- Schwarzer, R., and Jerusalem, M. (1995). Generalized self-efficacy scale. J. Weinman, S. Wright and M. Johnston, *Measures in health psychology: A user's portfolio. Causal and control beliefs* 35–37
- Sedaghat, M., Abedin, A., Hejazi, E., and Hassanabadi, H. (2011). Motivation, cognitive engagement, and academic achievement. *Procedia Soc. Behav. Sci.* 15, 2406–2410. doi: 10.1016/j.sbspro.2011.04.117
- Smaldino, SE, Russell, J. D., Heinrich, R., and Molenda, M. (2006). *Instructional technology and Media for Learning*. 8th Pearson, New York, NY
- Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: many scientists disapprove. *Nature* 613, 620–621. doi: 10.1038/d41586-023-00107-z
- Sun, H.-L., Sun, T., Sha, F.-Y., Gu, X.-Y., and Hou, X.-R. (2022). The influence of teacher-student interaction on the effects of online learning: based on a serial mediating model. *Front. Psychol.* 13:779217. doi: 10.3389/fpsyg.2022.779217
- Tate, T., Doroudi, S., Ritchie, D., and Xu, Y. (2023). Educational research and AI-generated writing: confronting the coming tsunami. *EdArXiv*. January 10
- Terwiesch, C. (2023). *Would chat GPT3 get a Wharton MBA. A prediction based on its performance in the operations management course*. Wharton: Mack Institute for Innovation Management/University of Pennsylvania/School Wharton
- Van Dinther, M., Dochy, F., and Segers, M. (2011). Factors affecting students' self-efficacy in higher education. *Educ. Res. Rev.* 6, 95–108. doi: 10.1016/j.edurev.2010.10.003
- Van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., and Bockting, C. L. (2023). ChatGPT: five priorities for research. *Nature* 614, 224–226. doi: 10.1038/d41586-023-00288-7
- Villalonga, B. (2004). Does diversification cause the diversification discount? *Financ. Manag.*, 5–27.
- Vrugt, A., and Oort, F. J. (2008). Metacognition, achievement goals, study strategies and academic achievement: pathways to achievement. *Metacogn. Learn.* 3, 123–146. doi: 10.1007/s11409-008-9022-4
- Walker, C. O., Greene, B. A., and Mansell, R. A. (2006). Identification with academics, intrinsic/extrinsic motivation, and self-efficacy as predictors of cognitive engagement. *Learn. Individ. Differ.* 16, 1–12. doi: 10.1016/j.lindif.2005.06.004
- Wang, Z., Bergin, C., and Bergin, D. A. (2014). Measuring engagement in fourth to twelfth grade classrooms: the classroom engagement inventory. *Sch. Psychol. Q.* 29, 517–535. doi: 10.1037/spq000050
- Wang, M. T., and Eccles, J. S. (2012). Adolescent behavioral, emotional, and cognitive engagement trajectories in school and their differential relations to educational success. *J. Res. Adolesc.* 22, 31–39. doi: 10.1111/j.1532-7795.2011.00753.x
- Wang, W., Kang, S. W., and Choi, S. B. (2022). Servant leadership and creativity: a study of the sequential mediating roles of psychological safety and employee well-being. *Front. Psychol.* 12:807070. doi: 10.3389/fpsyg.2021.807070
- Wei, L., Zhang, W., and Lin, C. (2023). The study of the effectiveness of design-based engineering learning: the mediating role of cognitive engagement and the moderating role of modes of engagement. *Front. Psychol.* 14:1151610. doi: 10.3389/fpsyg.2023.1151610
- Yokoyama, S. (2019). Academic self-efficacy and academic performance in online learning: a mini review. *Front. Psychol.* 9:2794. doi: 10.3389/fpsyg.2018.02794
- Yu, H. (2023). Reflection on whether chat GPT should be banned by academia from the perspective of education and teaching. *Front. Psychol.* 14:1181712. doi: 10.3389/fpsyg.2023.1181712
- Zhou, L., and Wei, Y. (2010). Research on human-computer interactive interface in visualized information. In *2010 IEEE 11th International Conference on Computer-Aided Industrial Design & Conceptual Design*, 1, 349–352
- Zhu, X., Chen, A., Ennis, C., Sun, H., Hopple, C., and Bonello, M. (2009). Situational interest, cognitive engagement, and achievement in physical education. *Contemp. Educ. Psychol.* 34, 221–229. doi: 10.1016/j.cedpsych.2009.05.002



## OPEN ACCESS

## EDITED BY

Jochen Kuhn,  
Ludwig-Maximilians-Universität München,  
Germany

## REVIEWED BY

Peng He,  
Michigan State University, United States  
José Martín Molina-Espinosa,  
Monterrey Institute of Technology and Higher  
Education (ITESM), Mexico

## \*CORRESPONDENCE

Karen D. Wang  
✉ kdwang@stanford.edu

RECEIVED 02 November 2023

ACCEPTED 26 December 2023

PUBLISHED 18 January 2024

## CITATION

Wang KD, Burkholder E, Wieman C, Salehi S and  
Haber N (2024) Examining the potential and  
pitfalls of ChatGPT in science and engineering  
problem-solving. *Front. Educ.* 8:1330486.  
doi: 10.3389/feduc.2023.1330486

## COPYRIGHT

© 2024 Wang, Burkholder, Wieman, Salehi and  
Haber. This is an open-access article distributed  
under the terms of the [Creative Commons  
Attribution License \(CC BY\)](#). The use,  
distribution or reproduction in other forums is  
permitted, provided the original author(s) and  
the copyright owner(s) are credited and that  
the original publication in this journal is cited, in  
accordance with accepted academic practice.  
No use, distribution or reproduction is  
permitted which does not comply with these  
terms.

# Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving

Karen D. Wang<sup>1\*</sup>, Eric Burkholder<sup>2</sup>, Carl Wieman<sup>1,3</sup>, Shima Salehi<sup>1</sup>  
and Nick Haber<sup>1</sup>

<sup>1</sup>Graduate School of Education, Stanford University, Stanford, CA, United States, <sup>2</sup>Department of Physics, Auburn University, Auburn, AL, United States, <sup>3</sup>Department of Physics, Stanford University, Stanford, CA, United States

The study explores the capabilities of OpenAI's ChatGPT in solving different types of physics problems. ChatGPT (with GPT-4) was queried to solve a total of 40 problems from a college-level engineering physics course. These problems ranged from well-specified problems, where all data required for solving the problem was provided, to under-specified, real-world problems where not all necessary data were given. Our findings show that ChatGPT could successfully solve 62.5% of the well-specified problems, but its accuracy drops to 8.3% for under-specified problems. Analysis of the model's incorrect solutions revealed three distinct failure modes: (1) failure to construct accurate models of the physical world, (2) failure to make reasonable assumptions about missing data, and (3) calculation errors. The study offers implications for how to leverage LLM-augmented instructional materials to enhance STEM education. The insights also contribute to the broader discourse on AI's strengths and limitations, serving both educators aiming to leverage the technology and researchers investigating human-AI collaboration frameworks for problem-solving and decision-making.

## KEYWORDS

ChatGPT, GPT-4, generative AI models, problem-solving, authentic problems, STEM education, physics education

## 1 Introduction

The rapid advancement of Large Language Models (LLMs) has attracted substantial attention from both the general public and academia. LLMs, such as GPT-4 by OpenAI, can generate human-like textual responses to text-based queries in real-time. Since the public launch of ChatGPT in November 2022, there has been a growing body of research exploring its various capabilities, limitations, and implications across diverse disciplines and tasks. One such field is education, where LLMs have far-reaching implications for both instructional practices, or *how* we teach and assess; as well as for curriculum content, or *what* we teach and assess.

Broadly speaking, problem-solving refers to the process of finding the solution to a problem when the steps for solving are not known to the problem solver beforehand (Newell et al., 1972; Mayer, 1992). Extensive research has been conducted to study problem-solving in physics education (Ince, 2018). For example, using textbook-style physics problems, a seminal study by Chi et al. (1981) found that experts abstracted the physics principles underlying a problem as the basis for their problem-solving approach, while novices often based their approaches on the surface-level features. More recently, the physics education research community has begun to recognize the need for helping students acquire effective practices and strategies for handling real-world, authentic problems beyond traditional

textbook exercises (Bao and Koenig, 2019; Burkholder et al., 2020). Such practices and strategies are key to prepare students for the complex challenges that they will encounter in their professional careers and daily lives.

In this study, we examined ChatGPT's capacity for solving problems from a college-level engineering physics course. ChatGPT by OpenAI is one of most accessible and publicly used LLM-based tools, and its most advanced underlying model to date is GPT-4. GPT-4 has outperformed previous models like GPT-3 in an array of standardized exams in disciplines such as law and medicine. Notably, it has achieved scores in the 66th to 84th percentile on the AP Physics 2 Exam (Achiam et al., 2023), which features problems that are mostly situated in abstract scenarios and provide all necessary data in the problem statement. However, the literature has so far offered limited insights into the capability of GPT-4 in solving problems that are in real-world contexts and/or do not provide all the data needed for reaching a solution. Consequently, the nuances of GPT-4's problem-solving capability, including the range of problems that it can effectively solve and the quality of the generated solutions, remain largely unknown.

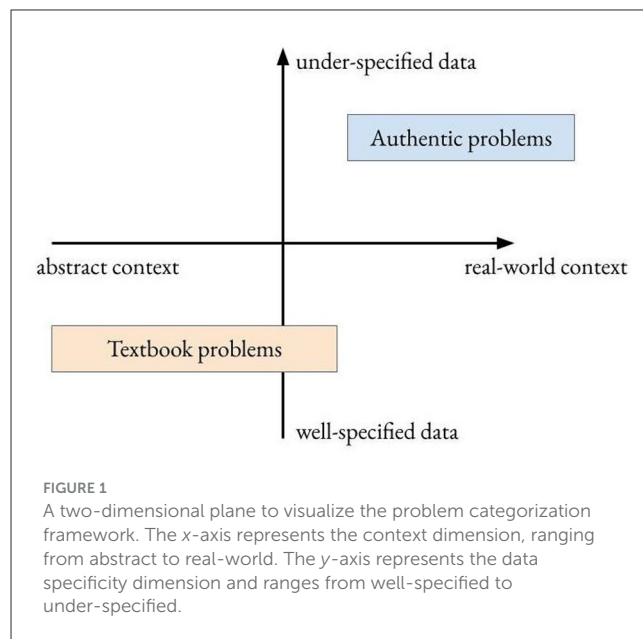
Investigating GPT-4's problem-solving capability has multifaceted implications that extend from enhancing educational practices to fostering human-AI collaboration. First, a more nuanced understanding of how GPT-4 solves different types of problems can offer insights into how to design LLM-augmented instructional materials to support student problem-solving. In this study, we focused our attention on scientific problem-solving with the long-term goal of leveraging LLM-based tools to enhance Science, Technology, Engineering, and Mathematics (STEM) education. This focus stems from the recognition that, despite problem-solving being widely acknowledged as a fundamental learning goal in STEM education (NGSS, 2013), effective ways to teach problem-solving remain elusive and understudied. Second, as students start using LLM-based tools such as ChatGPT for their homework problems they need help with (Shoufan, 2023), they need to be educated about its affordances and limitations to make effective use of such tools for their own learning. Furthermore, beyond its value in educational settings, knowledge of GPT-4's problem-solving capability contribute to the broader discourse on human-AI collaboration. Understanding the areas where AI excels and where it currently falls short can inform the development of a human-AI collaborative problem-solving framework.

In this study, we pose the following research questions:

- How does ChatGPT's problem-solving capability vary across different types of physics problems?
- What are ChatGPT's common failure modes for different types of physics problems?
- To what extent can a standard prompt engineering technique improve ChatGPT's performance for different types of physics problems?

## 2 Background

Human problem-solving has been studied across diverse research traditions and domains, including cognitive psychology, information processing, and discipline-based education research



(Newell et al., 1972; Chi et al., 1981; Reif and Heller, 1982; Bransford et al., 1986). As different types of problems call for distinct problem-solving strategies and bodies of knowledge, one's problem-solving capability may significantly vary across problem types. Similarly, to illuminate the problem-solving capability of AI models such as GPT-4, we must first explicate the characteristics of the problems given to the models, and study the performance of these models across different problem types.

Our research group has done extensive work to characterize and assess authentic problem-solving expertise across science, engineering, and medicine domains (Salehi, 2018; Price et al., 2021, 2022). Drawing on these work, we now characterize problems in science and engineering domains along two dimensions: context and data specificity (Figure 1). The first dimension refers to the context where the problem is situated and spans from abstract to real-world. Abstract problems employ simplified, idealized scenarios that do not exist in the real world, such as frictionless planes and massless pulleys. On the other end of the spectrum are real-world problems that are based on scenarios that individuals may encounter in their daily lives or in professional settings. The second dimension is around the specificity of the data required to solve a problem. Well-specified problems provide all the data required for a solution, while under-specified problems lack some essential data, requiring the problem solver to determine what data is needed and how to obtain it for solving the problem. Textbook problems typically present well-specified data and may have either abstract or real-world context. These problems are designed to make it easier for learners to grasp and practice domain-specific concepts. In contrast, authentic problems bring with them the complexity and ambiguity that comes from real-world challenges and do not specify all the required data.

The above problem categorization framework is intended for analyzing problems that are knowledge-rich, or requiring the application of content knowledge from STEM disciplines. These problems differ from the classic knowledge-lean problems



employed to study problem-solving in the information processing paradigm (Simon, 1973). The knowledge-lean problems, such as the Tower of Hanoi, are often termed as “well-defined” to indicate that they have clear initial and goal states and a set of clearly-defined operators for moving from the initial state to the goal state (Simon, 1978; Jonassen, 1997). It is important to differentiate “well-defined” and “ill-defined” from the “well-specified” and “under-specified” terminology we used in the problem categorization framework. The former terms capture the clarity of the initial and the goal states of a problem and the constraints on the possible operations to navigate from one to the other, while the latter terms are used for evaluating the quantity and clarity of data given in the problem statement.

Most existing research has focused on examining AI's performance in handling textbook-style problems that are well-specified and mostly abstract. For example, GPT-4 has performed well in standardized tests such as AP Biology, Chemistry, Environmental Science and Physics Exams (Achiam et al., 2023; Nori et al., 2023). The model also demonstrated proficiency surpassing average human performance in writing program functions that solely depend on existing public libraries (Bubeck et al., 2023). In contrast, there is a scarcity of research on how AI approaches authentic problems that are under-specified and situated in real-world contexts, even though such authentic problems are likely to constitute a significant share of the tasks that AI will encounter when deployed in the real world. Emerging research that ventures into the related domain has investigated AI's capacity for inductive reasoning, which involves identifying general principles from a small set of examples and applying these principles to novel situations (Gendron et al., 2023; Wang et al., 2023; Xu et al., 2023). Results of these investigations suggest significant room for improvement in AI's capability to make generalizations from specific instances.

While GPT-4's performance in solving textbook-style problems should not be extrapolated to its performance on authentic problems, a review of previous literature nonetheless provides insights into some of its common failure modes. One common flaw in GPT-4's performance is related to calculation errors. Previous studies have found that while the model can answer difficult high-school level math questions and discuss advanced mathematics concepts, it could also make basic errors in calculation (e.g., arithmetic mistakes) (Bubeck et al., 2023). Another limitation is the model's deficiency in critically evaluating its own solutions. This leads to failure in recognizing mistakes in its solution path (Bubeck et al., 2023; Zhang et al., 2023). A separate study employed GPT-3.5 and GPT-4 to answer open-domain questions, such as whether the New Orleans Outfall Canals are the same length as the Augusta Canal. The researchers summarized the models' failure modes in solving these problems into four categories: comprehension error, factualness error, specificity error, and inference error (Zheng et al., 2023). The study found that nearly half of the failures were due to factualness error, or the model lacking the necessary supporting facts to produce a correct answer, and another 25% of the failures were due to inference error, or the model failing to reason effectively.

In the context of physics education, a study reported that ChatGPT (based on the GPT-3 model) could narrowly pass a calculus-based college-level introductory physics course

(Kortemeyer, 2023). One test used for evaluation was the Force Concept Inventory (FCI), which comprises well-specified multiple-choice questions. GPT-3 solved 60% (18 of 30) of the FCI items. Moreover, the researcher found that the model's performance variation was more influenced by the mathematics than the physics concepts involved. Similar to the above mentioned studies, this study found that ChatGPT had persistent problems with calculation, especially in manipulating and calculating formulas involving square roots.

In summary, the review of existing literature revealed a gap in our understanding of generative AI models' capability for solving real-world problems where data is often incomplete or ambiguous. The central aim of this paper is to explore how one AI model (GPT-4) perform across an array of real-world physics problems that vary in data specificity. Furthermore, we will investigate the model's common failure modes in solving these problems and evaluate whether a well-studied prompt engineering technique could improve AI's problem-solving performance.

## 3 Methods

### 3.1 Problems used in the study

A total of 40 homework problems from an engineering physics course taught by the second author were used in this study. The course is a calculus-based engineering physics 1 course taught at a public research-intensive university. The course is primarily taken by engineering, chemistry, and physics majors and covers an array of topics including static equilibrium (forces and torques), conservation of momentum and kinematics (linear and angular), conservation of energy, harmonic motion, mechanical waves, and fluid mechanics. This particular course was developed by the second author (Burkholder et al., 2022) and aims at developing students' problem-solving competencies. The course is designed on the theory of deliberate practice (Ericsson et al., 1993) and uses a template (Burkholder et al., 2020) to teach students real-world problem-solving skills.

The homework problems in this course are a mixture of textbook physics problems (for practice with basic calculations), problems that ask students to explain a physical phenomenon, and real-world problems that require students to make assumptions, seek out information, and make modeling decisions (Price et al., 2021). The real-world problems were designed to engage students in more deliberate reasoning with particular concepts rather than standardized procedures. They were also designed to have students practice turning real situations into manageable models, rather than providing simplified scenarios for them. We selected problems that were written by the second author specifically to engage students in real-world problem-solving, rather than the textbook or conceptual explanation problems. Based on our proposed problem categorization framework (Figure 1), we characterize these problems along the two key dimensions: context and data specificity. Regarding the context dimension, the problems are all situated in real-world contexts. For example, one problem involves calculating the total travel time for an elevator ascending to the top floor of the Salesforce Tower in San Francisco, and another involves



selecting fishing lines that are strong enough to hang sculptures from the ceiling of an atrium of a new building.

Regarding the data specificity dimension, the problems used in this study span a spectrum from well-specified to under-specified. On one end of the spectrum are problems that provide all the data needed for solving, including values for key variables and parameters. On the other end are problems with under-specified or incomplete data, requiring the problem solver to determine what data is needed and how to collect the missing data. This variation in data specificity necessitates different levels of decision-making by the problem solver regarding data collection, which is a key practice for solving authentic problems as identified in our previous research (Salehi, 2018; Price et al., 2021). By incorporating this range of problem types, we are able to conduct a more comprehensive and nuanced evaluation of ChatGPT's problem-solving capability.

Table 1 presents two sample problems used in the study. Both problems are situated in real-world contexts. The first one is a well-specified problem where all data needed to solve the problem was provided in the problem statement. In contrast, the second one represents an under-specified problem where the problem statement does not provide any data, and necessitates the problem solver to collect all the required data through conducting an online query or making reasonable assumptions in order to solve the problem.

## 3.2 Experiments and analysis

We used ChatGPT with GPT-4 selected as the underlying model in the present study. The decision to use ChatGPT as opposed to running the experiments through OpenAI's API was grounded in the interest of face validity and ecological validity. Face validity refers to the appropriateness or relevance of a measurement method for its intended purpose (Nevo, 1985), while ecological validity refers to the extent to which the findings of a study can be generalized to the natural environments and real-world settings (Orne, 2017; Kihlstrom, 2021). Given students and instructors of STEM courses are more likely to use ChatGPT than to access the GPT-4 model directly through APIs, this methodological choice allows our study's findings to be more directly applicable to the common STEM educational settings where LLM-based tools are used.

Each problem statement of the 40 problems was pasted into the dialogue interface of ChatGPT, accompanied by the prompt of "solve the following physics problem." No additional guidelines or contextual knowledge was provided. If ChatGPT returned with queries or statements indicating that the problem could not be solved without additional information, a second prompt was put into the dialogue box (e.g., "please make reasonable assumptions about the missing information and solve the problem."). Once the model reached a final answer, its response was transferred to a centralized document for record and analysis. This approach was implemented to minimally influence ChatGPT's problem-solving approach and establish a baseline for its problem-solving capability.

In evaluating ChatGPT's performance on solving the physics problems in our data set, we adopted an approach similar to the

one we use for grading students' worked-out solutions, where a single knowledgeable evaluator can effectively grade a student's solution. Our goal is to not only determine the accuracy of its final answers but also understand the steps it undertook to reach the answer. In cases where ChatGPT failed to reach the correct answer, we compared each step of its solution to the correct solution prepared by the course's lead instructor to determine where in the solution process it failed. The instructor's solutions loosely follow a template that we have devised to scaffold students' solving of authentic problems in physics and typically incorporate the following components:

- A clear representation of the problem highlighting its key features using a diagram or a set of bullet points
- Identification of the relevant physics concepts and formulas
- Noting all the information required for solving the problem, and for information not explicitly provided in the problem statement, noting how such information can be obtained through an internet search or a reasonable estimate based on prior knowledge
- Carrying out the necessary calculations to reach the correct answer.

It is important to note that we did not expect ChatGPT to follow the problem-solving template. Rather, the goal of comparing its solutions to the expert solutions is to determine the primary factor that led to the erroneous final answer. This analysis helps us ascertain whether the erroneous answer was due to a misrepresentation of the problem, a misapplication of physics principles, errors in the data used, or calculation errors. By pinpointing the chief reason behind the incorrect solutions, we aimed to gain a more nuanced understanding of the AI model's problem-solving potential and limitations. The results of this error analysis were recorded for each individual problem and collectively analyzed to identify patterns and recurrent themes in ChatGPT's problem-solving failures. To ensure both precision and thoroughness in the analysis of ChatGPT-generated solutions, the analysis was conducted by the first author, a researcher in STEM education, in close collaboration with the third author, who is an expert in physics and physics education.

Next, we examined whether simple prompt engineering could improve ChatGPT's problem-solving performance. In the context of AI research, prompt engineering refers to the process of designing, testing, and refining inputs given to AI models to enhance their performance (Liu et al., 2023). Prompting strategies such as zero-shot chain-of-thought, which involves literally telling the model to "think step-by-step," have demonstrated success in improving LLMs' performance in solving multi-step arithmetic word problems (Kojima et al., 2023). In the second phase of the study, we adopted a similar prompting strategy for the problems in our data set. Specifically, the prompt was updated to "solve the following physics problem step-by-step" just before presenting the problem statement to ChatGPT. The intention was to explore whether ChatGPT could decompose a problem into more manageable sub-problems and circumvent the errors it made during its initial problem-solving attempt. At the same time, it is important to note that the specific approach to decomposing

TABLE 1 Two sample problems used in the study.

The log cabin problem (well-specified)	The dresser tip-over problem (under-specified)
You are planning to build a log cabin and will need to pull the logs up a hill to the building site by means of a rope attached to a winch. In order to buy the rope, you need to know how strong the rope must be and decide to do a quick calculation for this. The heaviest of the logs weigh 500 lbs. You estimate the coefficient of friction between the log and the hill to be 0.8, and the hill you have to pull them up is at an angle of 30 degrees. How strong must the rope be?	IKEA has had some issues with children climbing their dressers by pulling on the drawers and getting hurt when the dressers tip over. Their solution has been to provide wall mounts that you can use to secure the top of the dresser to the wall. Figure out how strong this wall mount has to be to keep the dresser from tipping over. You should include an equation showing how your answer depends on the weight of the child and the size and weight of the dresser.

TABLE 2 ChatGPT's problem-solving performance grouped by the problems' data specificity.

Problem type	Number of problems	Number correctly solved	Accuracy rate
Well-specified	16	10	62.50%
Under-specified	24	2	8.33%

individual problems vary case-by-base, depending on the unique context and nature of each problem. Consequently, the efficacy of the prompt in improving the model's performance may not be consistent.

## 4 Results

In this section, we first present ChatGPT's problem-solving success rate without the use of prompt engineering. Next, we discuss three distinct failure modes based on a comparative analysis between ChatGPT's incorrect solutions and the instructor's correct solutions. Finally, we explore how prompt engineering impacted ChatGPT's problem-solving performance.

### 4.1 Problem-solving performance

Our analysis revealed a substantial difference in the ChatGPT's ability to solve well-specified vs. under-specified types of problems (Table 2). ChatGPT successfully solved 62.5% of the well-specified problems and only 8.3% of the under-specified problems. This discrepancy in accuracy rate was statistically significant (Fisher's exact test,  $p < 0.001$ ).

At the same time, ChatGPT demonstrated a high level of proficiency in identifying the relevant physics concepts to apply based on the given problem statement. This capacity was evidenced by the model's consistent performance of outlining the relevant physics concepts at the beginning of the solutions it generated. ChatGPT's strength in this facet of problem-solving differs from typical human performance, as students often struggle to identify what concepts to apply as the starting point in solving unfamiliar problems. Additionally, students may struggle with complexities that arise in correctly identifying, applying and integrating domain knowledge learned at different times. In contrast, LLMs like GPT-4 are not constrained by such linear learning pathway, and their training data is likely to incorporate a more comprehensive range of domain-specific knowledge compared to what students learn in

a typical college-level curriculum. This expansive knowledge base is one of AI's strengths in addressing real-world challenges.

### 4.2 Types of failure modes

A detailed table of all 40 problems used in the study, along with their underlying physics concepts and ChatGPT's performance, is available in the [Supplementary material](#). In examining where all of ChatGPT's incorrect solutions diverged from the instructor's solutions, we identified three distinct types of failure modes. These failure modes can help us understand the underlying causes of the AI model's difficulties in problem-solving.

#### 4.2.1 Failure to construct accurate models of the physical world

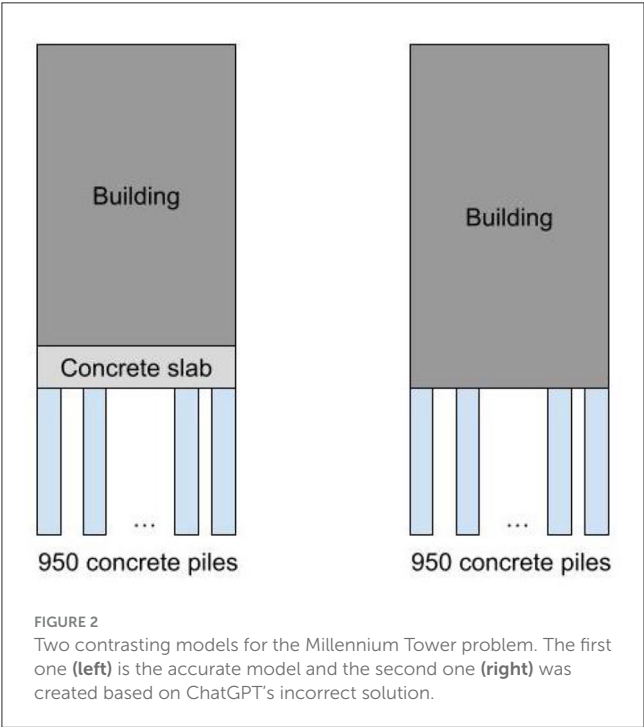
One of the failure modes is related to ChatGPT's failure to construct accurate models of the physical world based on the problem statement. Table 3 presents a problem from our dataset and ChatGPT's incorrect solution that falls into this type. The problem asks for an estimation of the force required from each of the 950 friction piles to prevent the Millennium Tower in San Francisco from sinking. This is an under-specified problem, as the problem statement does not specify the density of concrete. The highlighted texts in the problem statement are information that ChatGPT did not account for in its modeling of the problem.

A human problem solver could intuitively construct a model of the problem, that both the building and the concrete slab are on top of the 950 concrete piles, as depicted by the first model in Figure 2. This model helps the problem solver focus on the key objects and forces involved. The force required from each friction pile to support the building and concrete slab in addition to its own weight would therefore be:  $F_{\text{friction}} = (W_{\text{building}} + W_{\text{slab}})/950 + W_{\text{pile}}$ . In contrast, the second model in Figure 2 was created by the researcher based on ChatGPT's incorrect solution. ChatGPT failed to consider the weight of the concrete slab and piles in its calculations, indicating that it did not correctly conceptualize the spatial relationship between the building, the concrete slab, and the 950 piles. This led to an oversimplified and erroneous calculation of the force required from each pile.

Out of the 28 problems that ChatGPT did not solve, 14 can be attributed to this failure mode of it not being able to accurately model the physical world, especially regarding spatial relationships. Another example of this failure mode is the Dresser Tip-over problem that asks students to determine the force exerted by a wall mount to prevent a dresser from tipping over when children pull down on the drawers (see the problem in Table 1). Solving

TABLE 3 The Millennium Tower problem and ChatGPT’s solution.

<p>Prompt: Solve the following physics problem: The Millennium Tower in San Francisco was completed in 2009 and has received a lot of attention recently due to problems with the building settling and leaning. The building is 605 feet tall, has a base of about 20,000 square feet, and weighs about 7 tons per square foot of its foundation. The tower sits on a 10-ft thick concrete slab, which is in-turn supported by 950 friction piles, which are 14-inch square concrete pillars embedded in the bay sand. The piles are 80 feet long. Estimate the approximate force required from each friction pile to keep the building from sinking.</p>
<p>ChatGPT: To estimate the force each pile needs to exert, we need to determine the weight of the building and distribute that weight across the number of friction piles.</p> <p>The total weight of the building is given as 7 tons per square foot of its foundation. The foundation is ~20,000 square feet. Therefore, the total weight is 7 tons/ft * 20,000 ft = 140,000 tons.</p> <p>Since 1 ton is equivalent to 2,000 pounds, the total weight of the building is 140,000 tons * 2,000 pounds/ton = 280,000,000 pounds.</p> <p>The force each pile needs to exert is 280,000,000 pounds/950 = 294,737 pounds (approximately), assuming an even distribution of the weight of the building.</p>



the problem requires an accurate model representing the center of mass of the dresser, the pulling by a child, the location of the wall mount, and most importantly, the pivot point and the distances to the pivot point for each of the forces. One of its persistent challenges observed in our analysis is that ChatGPT cannot correctly identify the pivot point. In other words, it cannot correctly conceptualize how the dresser would fall if pulled by a child.

4.2.2 Failure to make reasonable assumptions about missing data

A second failure mode of ChatGPT in problem-solving was its limitations in making reasonable assumptions about missing data in under-specified problems. When presented with problems that lack complete data, ChatGPT usually attempted to fill in the missing data and generate a solution on its own. Only for two under-specified problems did it respond to the initial problem statement by stating that the problem doesn’t provide all the

necessary information to calculate the exact value for a solution. In such cases, a follow-up instruction was entered into ChatGPT’s dialogue box, directing the model to make reasonable assumptions about the specific value it noted and solve the problem. Table 4 presents such an instance where ChatGPT explicitly listed the missing information.

ChatGPT was successful at filling in well-known factual information that was not provided in the problem statement, such as the speed of light in a vacuum, the speed of sound in water, the density of seawater, and even the height of the Salesforce Tower in San Francisco. However, its success rate dropped when the missing data involves more nuanced real-world understanding. In the Floating Duck problem presented above (Table 4), ChatGPT made an incorrect assumption about the density of a duck being 950 kg/m<sup>3</sup> even after receiving the prompt to make reasonable estimates. A more accurate density value can be inferred from the information that 20% of the duck’s body is submerged in salt water, making its average density ~260 kg/m<sup>3</sup>.

Table 5 presents another problem that illustrates this failure mode. The problem asks a problem solver to estimate how long a driver should take to come to a stop while driving at 10 mph, so that a coffee cup that was accidentally left on the car’s roof won’t fly off. In analyzing ChatGPT’s solution to this problem, we found that the model failed to provide a reasonable value for the friction coefficient between the coffee cup and the roof of the car ( $\mu$ ). While the typical ceramic-on-metal and metal-on-metal static coefficients of friction are between 0.4 and 0.6, ChatGPT assumed the friction coefficient to be 1, as implied by its equating acceleration to the gravitational constant  $g$  in the solution (highlighted in Table 5). This incorrect assumption led to an erroneous value of the horizontal acceleration, which in turn resulted in an underestimate of the time it takes to stop the car safely.

To its credit, ChatGPT did hedge on its solution by making the following statement at the end:

ChatGPT: Please note that this calculation assumes that the friction between the cup and the roof is strong enough to keep the cup from sliding when the car is decelerating at a rate equal to  $g$ . The actual time could be much longer if the friction is weaker than assumed. This is a simplified model and doesn’t account for all factors that might be present in a real-world situation.

This acknowledgment indicates that ChatGPT recognizes the potential error in its solution introduced by inaccurate assumptions. Nonetheless, we should be aware of its limitations, that it may

TABLE 4 An example of ChatGPT requesting more information for the Floating Duck problem.

<p>Prompt: Solve the following physics problem: Off-shore oil-drilling can be controversial due to the possibility of oil spills. A sea duck can float on salt water with only about 20% of its body submerged below the surface. How thick would a layer of crude oil have to be for the duck to drown? For reference, the BP oil spill in 2010 reached a depth of 41.5 m.</p>
<p>ChatGPT: To solve this problem, we first need to understand the concept of buoyancy. Buoyancy is the force exerted on an object that is wholly or partly immersed in a fluid. In our case, the fluid is the combination of salt water and crude oil. The sea duck floats with only 20% of its body submerged in the water because the buoyancy force from the submerged part of the duck equals the weight of the duck. This means that the duck is in equilibrium.</p> <p>If a layer of oil were to cover the surface of the water, the duck would start to submerge into the oil, since oil is less dense than water. If the oil layer is thick enough, the duck could potentially sink below the surface and drown. To find the thickness of the oil layer that would cause the duck to drown, we need to find the point where the duck's buoyancy in the oil equals its weight.</p> <p>Here's where we run into a problem, though. To find this value, we need additional information that isn't provided in the problem:</p> <ol style="list-style-type: none"><li>1. The volume of the duck. This would give us the amount of oil (or water) displaced by the duck, which we could use to calculate buoyancy.</li><li>2. The density of the crude oil. Different types of crude oil can have different densities, and the density would affect the buoyancy.</li><li>3. The density of the duck. In order to figure out the weight of the duck, we would need to know both its volume and its density. This would allow us to figure out when the duck's weight equals the buoyancy force in the oil.</li></ol>
<p>Prompt: Can you make some reasonable estimates and solve the problem?</p>
<p>ChatGPT: To make some reasonable estimates, let's say the following:</p> <ol style="list-style-type: none"><li>1. The volume of the duck: This can vary, but we can estimate an average duck to have a volume of 1 liter or 0.001 cubic meter.</li><li>2. The density of the crude oil: The density of crude oil varies based on its specific type and temperature, but a common value is around 850 kg/m<sup>3</sup>.</li><li>3. The density of the duck: Most birds have a density slightly less than water due to their feathers and air-filled bones. Let's estimate it at about 950 kg/m<sup>3</sup>, slightly less than the density of sea water which is about 1025 kg/m<sup>3</sup>.</li></ol>

TABLE 5 The coffee cup on car problem and ChatGPT's solution.

<p>Prompt: Solve the following physics problem: You are driving your electric car to work one morning and forget your coffee cup on the roof. As you leave your driveway, you speed up to 10 mph and then remember your coffee cup and decide you need to stop. If you stop too fast, the cup could go flying and hit a pedestrian. How long should you take to come to a stop to be sure the cup doesn't fly off the roof?</p>
<p>ChatGPT: To solve this problem, we need to consider the forces acting on the coffee cup, primarily gravity and friction. The friction force is given by the equation: <math>F_{\text{friction}} = \mu N</math>, where <math>\mu</math> is the coefficient of friction and <math>N</math> is the normal force.</p> <p>However, we don't have a specific value for the coefficient of friction between the cup and the roof of the car.</p> <p>Let's assume that the coffee cup will stay on the roof if the car stops with an acceleration equal to or less than <math>g</math> (the acceleration due to gravity), which is <math>\sim 9.81 \text{ m/s}^2</math>.</p>

not make accurate assumptions about missing information from the outset. Such difficulty in making reasonable estimates for real-world quantities is also typical of novice students.

Another example that falls into this failure mode is ChatGPT's solution to a problem that involves estimating how the velocity of a merry-go-around would change once a person hops on after first pushing it around for 3 s. As the problem does not specify the initial speed of the merry-go-around, ChatGPT first estimated it to be around 14 mph through calculation. However, this estimate substantially exceeds the safe speed for merry-go-arounds, which are typically around 9 mph as stipulated by the US Consumer Product Commission. This example further illustrates that ChatGPT cannot always align and verify computational results with real-world standards and constraints. ChatGPT's solutions to eight problems had this failure mode.

4.2.3 Calculation error

While the first two failure modes are mostly associated with under-specified problems, the last failure mode, calculation error, has been observed in both well-specified and under-specified problems in the dataset. A total of six incorrect solutions, four for well-specified problems and two for under-specified problems, had this failure mode. The errors range from arithmetic to trigonometry, and underscore ChatGPT's well-known difficulties with mathematical computations.

TABLE 6 A problem for which ChatGPT reached the correct solution without using the data provided.

The horsepower problem
Engine power is sometimes expressed in terms of "horsepower." One horsepower was defined by James Watt, who observed that a horse could turn a mill wheel with a radius of 12 ft at a constant rate of 144 times per hour, exerting a nearly constant force of 800 N tangentially to the wheel. Derive the conversion for horsepower to kilowatts.

4.2.4 An idiosyncratic solution case

In addition to the three failure modes discussed above, we also identified an idiosyncratic case where ChatGPT reached the correct answer (therefore coded as correctly-solved) while completely disregarding the data provided in the problem. The problem is a well-specified one and involves the conversion of horsepower to kilowatts (Table 6).

Despite being provided with data in the problem statement, ChatGPT opted to utilize different data, that one horsepower is defined as the ability to lift 550 pounds one foot in one second, for its calculations. This behavior raises questions regarding how the underlying GPT-4 model potentially prioritizes its training data over new information in problem-solving.

4.3 Chain-of-thought prompt engineering

To what extent did prompt engineering enhance ChatGPT's problem-solving performance? In the second experiment, we



TABLE 7 ChatGPT's updated solution to the coffee cup on car problem after prompt engineering.

<p>ChatGPT: We can analyze this scenario by considering the forces acting on the coffee cup. Since the car is accelerating in the opposite direction of its motion (decelerating), there will be a frictional force trying to keep the cup in place. Let's assume the coefficient of static friction between the cup and the car's roof is <math>\mu_s</math>, and that this is the only force preventing the cup from sliding.</p> <ol style="list-style-type: none"> <li>1. Find the maximum deceleration: The force of static friction is given by: <math>F_{\text{friction}} = \mu_s \cdot F_n</math>, where <math>F_n</math> is the normal force, which is equal to the weight of the coffee cup: <math>F_n = m \cdot g</math>. Here, <math>m</math> is the mass of the cup, and <math>g</math> is the acceleration due to gravity (<math>9.8 \text{ m/s}^2</math>). The maximum static frictional force before the cup starts sliding is: <math>F_{\text{max}} = \mu_s \cdot m \cdot g</math>.</li> <li>2. Calculate the deceleration: This frictional force provides a maximum deceleration the car can undergo without the cup sliding. Using Newton's second law, <math>F = ma</math>: <math>\mu_s \cdot m \cdot g = m \cdot a</math>. Solving for <math>a</math>, we find: <math>a = \mu_s \cdot g</math>.</li> <li>3. Find the time to stop: Now we have the maximum deceleration, we can calculate the time it will take to stop the car from 10 mph under this deceleration. Using the equations of motion: <math>v_f = v_i + at</math>, where <math>v_f = 0 \text{ m/s}</math> (final velocity), <math>v_i = 10 \text{ mph} = 4.47 \text{ m/s}</math> (initial velocity), and <math>a = -\mu_s \cdot g</math> (deceleration). Solving for <math>t</math>: <math>t = \frac{v_f - v_i}{a} = \frac{-4.47}{-\mu_s \cdot 9.8}</math>.</li> </ol> <p>We need the coefficient of static friction to calculate the exact value of <math>t</math>, which will tell us how long the driver should take to come to a stop to ensure the coffee cup doesn't fly off the roof. If we had that value, we could plug it into our equation to find the time. In real-world scenarios, the coefficient of static friction might vary widely depending on the materials in contact and the conditions, so without specific information, it's hard to give a precise time.</p>
<p>Prompt: Please make reasonable assumptions about the coefficient of friction and solve the problem.</p>
<p>ChatGPT: Certainly! To solve the problem, we'll make a reasonable estimate for the coefficient of static friction between the coffee cup and the car's roof.</p> <p>In general, the coefficient of static friction for smooth surfaces like a car roof and a typical coffee cup might range from 0.5 to 0.8. However, since a car roof is usually not designed to hold objects and the bottom of a coffee cup may be somewhat smooth, we'll use a more conservative estimate of <math>\mu_s = 0.6</math>. Now, we'll use our earlier formula for the time taken to stop:</p> $t = \frac{-4.47}{-\mu_s \cdot 9.8}$ <p>Substituting <math>\mu_s = 0.6</math>:</p> $t = \frac{-4.47}{-0.6 \times 9.8} \approx 0.76 \text{ s.}$ <p>So the driver should take <math>\sim 0.76 \text{ s}</math> to come to a stop to ensure that the coffee cup doesn't fly off the roof. Of course, this is a simplified model, and in a real-world scenario, other factors such as wind resistance, the curvature of the roof, and the exact materials of the cup and roof would also influence the outcome.</p>

applied the “solve the following physics problem step-by-step” prompt to all 40 problems in the dataset. Among the 12 problems that ChatGPT initially solved correctly, it generated consistent correct solutions for 11 of them under prompt engineering. However, ChatGPT made a calculation error involving trigonometry in one of the problems. Interestingly, in the idiosyncratic case where ChatGPT reached the correct answer without utilizing the given data in its initial solution, the step-by-step prompt helped it incorporate the data provided in the problem statement in its calculation for the correct solution.

Among the 28 problems that ChatGPT initially failed to solve, it was able to correctly solve three with the step-by-step prompting. Two of the three were related to the failure mode of ChatGPT not being able to construct accurate models about the real world. For the first one, ChatGPT initially did not subtract the weight of the water from a squid when it ejected water to create a form of jet-propulsion. For the other one, it initially treated a marble as a non-rotating block and did not account for the rotational kinetic energy as it rolled up a ramp. The last one of the three was associated with the initial failure mode of not being able to make reasonable assumptions about missing data, in this case the coefficient of friction in the Coffee Cup on Car problem. Table 7 presents ChatGPT's updated solution. The prompt of solving the problem step-by-step led to more precise and deliberate problem-solving as illustrated in this example. ChatGPT first broke down the solution process into discrete steps, then noted that without knowing the coefficient of static friction, a specific numerical answer could not be provided. After receiving a follow-up instruction through the dialogue input box to make reasonable assumptions, it chose a reasonable value of 0.6 for the friction coefficient and successfully solved the problem, unlike what it did in the absence of prompt engineering.

Overall, the results suggest that prompt engineering had a moderate effect on enhancing ChatGPT's problem-solving

performance by constructing accurate models of the problem and making reasonable assumptions, though this effect is not statistically significant (Chi-squared (1) = 0.06,  $p = 0.81$ ). It should also be noted that step-by-step prompts had no impact on reducing calculation errors.

## 5 Discussion

The present study found a marked difference in ChatGPT's problem-solving performance between well-specified and under-specified problems. The problems used in the study are all situated in real-world contexts and require the application of physics knowledge, yet differ in how much information is specified in the problem statement. ChatGPT performed better in well-specified problem, although it made occasional calculation errors. In contrast, it was far less accurate in solving under-specified problems. Two specific failure modes were observed: the first one being failure to construct accurate models of the physical world and reason about relationships between different variables in a model, and the second one being failure to make reasonable estimates or assumptions about the missing data. Prompt engineering produced a moderate improvement in ChatGPT's problem-solving performance. The prompt of solving a problem step-by-step proved moderately beneficial in guiding the AI model to be more deliberate and accurate in estimating missing data and constructing models of the problems, though it did not alleviate calculation errors.

### 5.1 Implications for education

The problem-solving process adopted by experts in science and engineering domains can be characterized as a series



of interlinked decisions (Price et al., 2021). Utilizing this framework to analyze ChatGPT's performance on solving problems situated in real-world contexts, we note that ChatGPT (based on GPT-4 model) demonstrated proficiency in deciding on the relevant domain-specific concepts and formulas based on a problem statement. At the same time, it fell short in making several key decisions, including determining how to construct a suitable model of a problem, and deciding how to make reasonable assumptions or estimates about incomplete data.

These results have significant implications for STEM education, especially around how to leverage LLM-based tools like ChatGPT to help students develop expertise in problem-solving. First, the study identified facets of problem-solving where ChatGPT is indeed effective, namely identifying the relevant physics concepts needed for solving a problem based on the problem statement. This opens the possibility for ChatGPT to serve as a tutor for domain-specific problems and support students to pinpoint the essential knowledge underlying each problem and enhance their understanding of conceptual knowledge through problem-solving. This tutoring capability is particularly important as students struggle to decide on relevant physics concepts and formulas through analyzing the problem's statement, instead they rely on ineffective strategies such as searching for equations that contain the same variables to solve problems (Ogilvie, 2009; Burkholder et al., 2020). Given the capability of ChatGPT in deciding on relevant concepts, students can query ChatGPT with prompts such as "identify the relevant concepts associated with the following problem." However, one concern associated with this use case is that ChatGPT may generate articulate, plausible-sounding, yet incorrect solutions based on the identified concepts. This presents a risk of misleading students and inducing misconceptions. Therefore, it is crucial to educate students on the problem-solving capabilities of ChatGPT (e.g., identifying the relevant concepts) as well as its shortcomings (e.g., generating inaccurate solutions due to failure to construct accurate models, failure to make reasonable assumptions, or calculation errors, particularly in the case of under-specified problems).

Second, the findings point to what we should prioritize in STEM education in an era of increasingly powerful AI technologies. To prepare students for solving authentic problems in their professional and personal lives, STEM courses must place an emphasis on fostering effective decision-making practices. Specifically, students must have opportunities to practice making decisions related to construct appropriate models based on complex, real-world scenarios, as well as practice making decisions on what data is needed for solving a given problem, how to collect the data, and how to critically evaluate data quality. Mastery in these decisions will help students decompose complex, under-specified real-world challenges into a series of tractable, well-specified sub-problems for AI tools like ChatGPT to solve. The emphasis on developing problem-solving and decision-making expertise aligns well with the broader educational goal of preparing students to navigate a future of human-AI collaboration.

Lastly, our findings have immediate implications for how to design homework and exam problems that are resilient to

automatic solving by tools like ChatGPT. The key strategy involves incorporating authentic problems into teaching and assessment materials. These problems are not solvable by ChatGPT alone, and necessitate students to make informed decisions on how to utilize ChatGPT as a tool. At the same time, students must remain actively involved in constructing accurate models in real-world contexts and handle under-specified information. The inclusion of such authentic problems allows for a more valid assessment of student competencies in STEM courses.

## 5.2 Implications for human-AI collaboration

This study also provides insights for the future of human-AI collaboration. While LLMs like GPT-4 can solve well-specified problems, albeit with occasional calculation errors, human intervention is needed to provide contextual understanding and nuanced judgement that AI currently lacks, particularly when navigating the complexities and ambiguities associated with authentic problems. This insight suggests a complementary relationship between human intelligence and artificial intelligence in addressing complex, authentic problems in the real world. Specifically, human experience and expertise can help construct accurate models of the physical world and make reasonable estimates or data collection plans for missing information. At the same time, AI's computational capability to instantly sift through vast knowledge bases and pinpoint the relevant domain knowledge constitutes an important asset to support human problem-solving.

## 6 Limitations

One potential limitation of this study is that we did not have two researchers independently evaluate all ChatGPT-generated solutions to assess inter-rater agreement. Instead, our analysis method mirrored the approach used in grading physics coursework, where a single knowledgeable evaluator compares students' solutions with expert-generated correct solutions. To mitigate the risk of bias, we also adopted a close, interdisciplinary collaboration in the analysis process.

Next, in evaluating ChatGPT's capacity for problem-solving, it is important to recognize the inherent limitations associated with the underlying algorithm's probabilistic nature. ChatGPT may generate different answers each time a problem is posed, and this variability presents a challenge in our analysis of its solutions. The different releases and incremental builds of the algorithm could further produce varied results. Therefore, the interpretation of our findings must consider the specific version of the algorithm utilized, which spans from May to August 2023. Additionally, the current study did not ask ChatGPT to generate solutions for identical problems and prompts multiple times. This absence of repetitive testing restricts our understanding of the tool's stability and reliability in providing consistent solutions. The probabilistic and evolving nature of LLMs underscore the need for continuous

evaluation and validation of their problem-solving capabilities in future studies.

## 7 Conclusion

This study probed the capabilities and limitations of LLM-based technologies such as ChatGPT in solving authentic problems that are situated in real-world contexts and under-specified in terms of the requisite data. By focusing on the domain of physics, we were able to incorporate a diverse set of real-world scenarios into the problem set. The problem-solving practices and processes adopted to solve these physics problems are also applicable in the broader fields of science and engineering. Furthermore, the decision to include problems from well-specified to under-specified in terms of the amount of information provided in the problem statement led to a nuanced understanding of ChatGPT's capacity for solving different types of problems. The findings revealed that ChatGPT is adept at identifying relevant physics knowledge and applying it to solve well-specified problems. At the same time, its performance is less robust in modeling real-world complexities and making reasonable assumptions when data is missing in under-specified problems.

These findings lead to future studies to investigate how LLMs can be utilized in STEM education to support student learning, such as serving as personalized tutors to scaffold students in identifying the relevant knowledge for solving a problem. Additionally, the insights from this study shed light on what are the key competencies for students to develop to prepare for a future where AI can effectively address well-specified problems. These competencies include the ability to construct accurate and concise models of problems, make deliberate decisions regarding assumptions and estimates, and devise plans for data collection. Students' mastery of these competencies, in conjunction with the advancement of AI technologies, potentially pave the way for a future where human-AI collaboration can effectively address complex challenges in the real world.

## Data availability statement

Full datasets from this study are not publicly available due to being used for course homework, and will be provided upon request by the corresponding author.

## References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Bao, L., and Koenig, K. (2019). Physics education research for 21st century learning. *Discip. Interdiscip. Sci. Educ. Res.* 1, 1–12. doi: 10.1186/s43031-019-0007-8
- Bransford, J., Sherwood, R., Vye, N., and Rieser, J. (1986). Teaching thinking and problem solving: research foundations. *Am. Psychol.* 41, 1078. doi: 10.1037/0003-066X.41.10.1078
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). Sparks of artificial general intelligence: early experiments with gpt-4. *arXiv Preprint*. doi: 10.48550/arXiv.2303.12712
- Burkholder, E., Miles, J., Layden, T., Wang, K., Fritz, A., Wieman, C., et al. (2020). Template for teaching and assessment of problem solving in introductory physics. *Phys. Rev. Phys. Educ. Res.* 16, 010123. doi: 10.1103/PhysRevPhysEducRes.16.010123
- Burkholder, E., Salehi, S., Sackeyfio, S., Mohamed-Hinds, N., and Wieman, C. (2022). Equitable approach to introductory calculus-based physics courses focused on problem-solving. *Phys. Rev. Phys. Educ. Res.* 18, 020124. doi: 10.1103/PhysRevPhysEducRes.18.020124
- Chi, M. T., Glaser, R., and Rees, E. (1981). "Expertise in problem solving," in *Advances in the Psychology of Human Intelligence*, ed. R. J. Sternberg (Hillsdale, NJ: Erlbaum), 7–76.

## Author contributions

KW: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Writing—original draft, Writing—review & editing. EB: Conceptualization, Methodology, Writing—review & editing. CW: Conceptualization, Funding acquisition, Supervision, Writing—review & editing. SS: Conceptualization, Funding acquisition, Supervision, Writing—review & editing. NH: Conceptualization, Funding acquisition, Methodology, Supervision, Writing—review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by the Stanford Accelerator for Learning and the Stanford Institute for Human-Centered AI.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/feduc.2023.1330486/full#supplementary-material>

- Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychol. Rev.* 100, 363–406. doi: 10.1037/0033-295X.100.3.363
- Gendron, G., Bao, Q., Witbrock, M., and Dobbie, G. (2023). Large Language Models are Not Abstract Reasoners. *arXiv preprint arXiv:2305.19555*.
- Ince, E. (2018). An overview of problem solving studies in physics education. *J. Educ. Learn.* 7, 191–200. doi: 10.5539/jel.v7n4p191
- Jonassen, D. H. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educ. Technol. Res. Dev.* 45, 65–94. doi: 10.1007/BF02299613
- Kihlstrom, J. F. (2021). Ecological validity and “ecological validity.” *Perspect. Psychol. Sci.* 16, 466–471. doi: 10.1177/1745691620966791
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2023). Large language models are zero-shot reasoners. *Adv. Neural Inform. Process. Syst.* 35, 2219922213.
- Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? *Phys. Rev. Phys. Educ. Res.* 19, 010132. doi: 10.1103/PhysRevPhysEducRes.19.010132
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55, 135. doi: 10.1145/3560815
- Mayer, R. E. (1992). *Thinking, Problem Solving, Cognition*. New York, NY: WH Freeman/Times Books/Henry Holt & Co.
- Nevo, B. (1985). Face validity revisited. *J. Educ. Meas.* 22, 287–293. doi: 10.1111/j.1745-3984.1985.tb01065.x
- Newell, A., and Simon, H. A. (1972). *Human Problem Solving*, Volume 104. Cliffs, NJ: Prentice-hall Englewood.
- NGSS (2013). *Next Generation Science Standards: For States, by States*. Washington, DC: National Academies Press.
- Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.
- Ogilvie, C. A. (2009). Changes in students’ problem-solving strategies in a course that includes context-rich, multifaceted problems. *Phys. Rev. ST Phys. Educ. Res.* 5, 020102. doi: 10.1103/PhysRevSTPER.5.020102
- Orne, M. T. (2017). “On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications,” in *Sociological Methods*, ed. N. J. K. Denzin (London: Routledge), 279–299. doi: 10.4324/9781315129945-26
- Price, A., Salehi, S., Burkholder, E., Kim, C., Isava, V., Flynn, M., et al. (2022). An accurate and practical method for assessing science and engineering problem-solving expertise. *Int. J. Sci. Educ.* 44, 2061–2084. doi: 10.1080/09500693.2022.2111668
- Price, A. M., Kim, C. J., Burkholder, E. W., Fritz, A. V., and Wieman, C. E. (2021). A detailed characterization of the expert problem-solving process in science and engineering: guidance for teaching and assessment. *CBE—Life Sci. Educ.* 20, ar43. doi: 10.1187/cbe.20-12-0276
- Reif, F., and Heller, J. I. (1982). Knowledge structure and problem solving in physics. *Educ. Psychol.* 17, 102–127. doi: 10.1080/00461528209529248
- Salehi, S. (2018). *Improving Problem-Solving through Reflection*. Stanford, CA: Stanford University.
- Shoufan, A. (2023). Exploring students’ perceptions of ChatGPT: thematic analysis and follow-up survey. *IEEE Access* 11, 38805–38818. doi: 10.1109/ACCESS.2023.3268224
- Simon, H. A. (1973). The structure of ill structured problems. *Artif. Intell.* 4, 181–201. doi: 10.1016/0004-3702(73)90011-8
- Simon, H. A. (1978). “Information-processing theory of human problem solving,” in *Handbook of Learning and Cognitive Processes*, Volume 5, ed. W. K. Estes (Abingdon: Taylor & Francis), 271–295.
- Wang, R., Zelikman, E., Poesia, G., Pu, Y., Haber, N., and Goodman, N. D. (2023). Hypothesis Search: Inductive Reasoning with Language Models. *arXiv preprint arXiv:2309.05660*.
- Xu, Y., Li, W., Vaezipoor, P., Sanner, S., and Khalil, E. B. (2023). LLMs and the abstraction and reasoning corpus: successes, failures, and the importance of object-based representations. *arXiv [Preprint]*. doi: 10.48550/arXiv.2305.18354
- Zhang, M., Press, O., Merrill, W., Liu, A., and Smith, N. A. (2023). How Language Model Hallucinations can Snowball. *arXiv preprint arXiv:2305.13534*.
- Zheng, S., Huang, J., and Chang, K. C.-C. (2023). Why does ChatGPT Fall Short in Providing Truthful Answers? *arXiv preprint arXiv:2304.10513*.



## OPEN ACCESS

## EDITED BY

Knut Neumann,  
IPN–Leibniz-Institute for Science  
and Mathematics Education, Germany

## REVIEWED BY

Mariel Fernanda Musso,  
CONICET Centro Interdisciplinario  
de Investigaciones en Psicología Matemática  
y Experimental, Argentina  
Ramazan Yilmaz,  
Bartın University, Türkiye

## \*CORRESPONDENCE

Wilter C. Morales-García  
✉ wiltermorales@upeu.edu.pe

RECEIVED 13 September 2023

ACCEPTED 19 February 2024

PUBLISHED 08 March 2024

## CITATION

Morales-García WC, Sairitupa-Sanchez LZ,  
Morales-García SB and Morales-García M  
(2024) Adaptation and psychometric  
properties of a brief version of the general  
self-efficacy scale for use with artificial  
intelligence (GSE-6AI) among university  
students.

*Front. Educ.* 9:1293437.

doi: 10.3389/feduc.2024.1293437

## COPYRIGHT

© 2024 Morales-García, Sairitupa-Sanchez,  
Morales-García and Morales-García. This is  
an open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](#). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic  
practice. No use, distribution or reproduction  
is permitted which does not comply with  
these terms.

# Adaptation and psychometric properties of a brief version of the general self-efficacy scale for use with artificial intelligence (GSE-6AI) among university students

Wilter C. Morales-García<sup>1,2,3,4\*</sup>, Liset Z. Sairitupa-Sanchez<sup>5</sup>,  
Sandra B. Morales-García<sup>6</sup> and Mardel Morales-García<sup>7</sup>

<sup>1</sup>Escuela de Posgrado, Universidad Peruana Unión, Lima, Perú, <sup>2</sup>Escuela de Medicina Humana, Facultad de Ciencias de la Salud, Universidad Peruana Unión, Lima, Perú, <sup>3</sup>Facultad de Teología, Universidad Peruana Unión, Lima, Perú, <sup>4</sup>Sociedad Científica de Investigadores Adventistas, SOCIA, Universidad Peruana Unión, Lima, Perú, <sup>5</sup>Escuela Profesional de Psicología, Facultad de Ciencias de la Salud, Universidad Peruana Unión, Lima, Perú, <sup>6</sup>Departamento Académico de Enfermería, Obstetricia y Farmacia, Facultad de farmacia y Bioquímica, Universidad Científica del Sur, Lima, Perú, <sup>7</sup>Unidad de Salud, Escuela de Posgrado, Universidad Peruana Unión, Lima, Perú

**Background:** Individual beliefs about one's ability to carry out tasks and face challenges play a pivotal role in academic and professional formation. In the contemporary technological landscape, Artificial Intelligence (AI) is effecting profound changes across multiple sectors. Adaptation to this technology varies greatly among individuals. The integration of AI in the educational setting has necessitated a tool that measures self-efficacy concerning the adoption and use of this technology.

**Objective:** To adapt and validate a short version of the General Self-Efficacy Scale (GSE-6) for self-efficacy in the use of Artificial Intelligence (GSE-6AI) in a university student population.

**Methods:** An instrumental study was conducted with the participation of 469 medical students aged between 18 and 29 ( $M = 19.71$ ;  $SD = 2.47$ ). The GSE-6 was adapted to the AI context, following strict translation and cultural adaptation procedures. Its factorial structure was evaluated through confirmatory factorial analysis (CFA). Additionally, the factorial invariance of the scale based on gender was studied.

**Results:** The GSE-6AI exhibited a unidimensional structure with excellent fit indices. All item factorial loads surpassed the recommended threshold, and both Cronbach's Alpha ( $\alpha$ ) and McDonald's Omega ( $\omega$ ) achieved a value of 0.91. Regarding factorial invariance by gender, the scale proved to maintain its structure and meaning in both men and women.

**Conclusion:** The adapted GSE-6AI version is a valid and reliable tool for measuring self-efficacy in the use of Artificial Intelligence among university students. Its unidimensional structure and gender-related factorial invariance make it a robust and versatile tool for future research and practical applications in educational and technological contexts.

#### KEYWORDS

self-efficacy, artificial intelligence (AI), invariance, technological, adaptation, GSE-6AI

## 1 Introduction

Artificial Intelligence (AI) has emerged as a fundamental pillar in the current technological evolution, significantly impacting various sectors, including education, where its implementation promises to transform teaching and learning methodologies. As we enter this new era of technological changes, adapting and adopting AI becomes crucial. The acceptance of these technologies varies significantly, largely influenced by individual perceptions of technological competence, which becomes a determining factor in this process (Agarwal and Karahanna, 2000; Hsu and Chiu, 2004; Pütten and Von Der Bock, 2018). In the educational sector, tools like ChatGPT have stood out for their ability to optimize efficiency and personalize learning. Incorporating AI in education not only allows for learning that is more tailored to the needs of each student but also fosters self-efficacy and motivation, especially in complex areas such as programming (Zhai et al., 2021; McDiarmid and Zhao, 2023; Yilmaz and Karaoglan Yilmaz, 2023b).

Developing digital competencies and becoming familiar with AI among teachers are key steps to maximize its benefits, such as personalized learning and improved teaching efficiency. However, the integration of AI in education also poses ethical and practical challenges that must be addressed responsibly (Oran, 2023). Generative AI, especially in writing assistance, opens up possibilities for stimulating creativity and overcoming writer's block, though it raises concerns about dependency and ethics (Washington, 2023). In the healthcare field, AI promises to improve diagnosis and decision-making, necessitating proper AI training and ethical awareness for its effective application (Kwak et al., 2022).

Technological self-efficacy, shaped by previous experiences and the perception of its usefulness, is crucial for the acceptance of AI. Demographic factors, such as income and education level, highlight the importance of overcoming gaps in access and use of AI, proposing a more inclusive approach (Hong, 2022). Self-efficacy has been established as a crucial concept in behavioral and educational psychology in the last decades of the 20th century. This term refers to confidence in one's abilities to organize and execute actions necessary to manage future situations. Moreover, self-efficacy significantly impacts how individuals set goals, face challenges, and overcome obstacles, being a key element for motivation and human behavior (Bandura, 1977). This concept is relevant not only in educational contexts, where its direct relationship with performance and student

motivation has been demonstrated (Pajares, 1996; Zimmerman, 2000), but also in the adoption and adaptation to new technologies, introducing the term "computer self-efficacy" (Compeau and Higgins, 1995).

Advanced technologies, such as Generative Artificial Intelligence (GAI), can enhance students' self-efficacy, encouraging a deeper cognitive engagement and, consequently, improving their academic outcomes (Liang et al., 2023). This phenomenon highlights the need to carefully integrate AI into educational systems, balancing the benefits of these technologies with critical reflection on their potential drawbacks. Thus, self-efficacy with the use of AI refers to the confidence and individual perception of the ability to employ AI effectively to achieve personal and professional goals. This includes not only the technical handling of AI-based tools but also the ability to integrate these tools into solving specific challenges, adapting to changes, and overcoming difficulties through innovative use of AI. Essentially, this self-efficacy reflects an individual's ability to apply AI effectively in varied contexts, both in managing daily tasks and in addressing unexpected situations, leveraging AI to enhance their performance.

This multidisciplinary approach to AI self-efficacy demonstrates its potential to mitigate the negative effects of overwork and stress, particularly in workplace and educational settings (Kim et al., 2024). By fostering a healthier and safer environment, AI self-efficacy not only benefits individual wellbeing but also contributes to organizational effectiveness and academic success (Yilmaz and Karaoglan Yilmaz, 2023a). The variability in adopting and adapting to AI among individuals suggests that, beyond technical skills, students' perceptions of their ability to use AI are crucial in their willingness to adopt these technologies. With the increasing integration of AI across various areas of life and work, several scales have been developed, such as the General Attitudes Towards Artificial Intelligence Scale (GAAIS), which assesses general perceptions of AI. This scale, including subscales to reflect both positive and negative attitudes, was validated in the UK and reveals a division in public perception of AI applications, underscoring the need for future studies to validate the scale in broader and more varied contexts (Schepman and Rodway, 2020). On the other hand, the Medical AI Readiness Scale for Medical Students (MAIRS-MS), developed in Persian and consisting of 22 items across four dimensions, has been validated among medical students in Iran, highlighting the relevance of integrating AI into medical curricula (Moodi et al., 2023). The Artificial Intelligence Anxiety Scale (AIAS), with 21 items spread across four dimensions, was validated in Taiwan (Wang and Wang, 2022).



Similarly, the Generative Artificial Intelligence Acceptance Scale, based on the UTAUT model and created with the participation of university students, provides a solid tool for measuring student acceptance of generative AI applications, whose four-factor structure was confirmed through factor analysis (Yilmaz et al., 2023). The Artificial Intelligence Literacy Scales (AILS), adapted to Turkish, assess AI understanding among non-expert adults and youth (Çelebi et al., 2023; Karaoğlu and Yilmaz, 2023).

In the context of self-efficacy, the Artificial Intelligence Self-Efficacy Scale (AISES) was developed, specifically designed to measure the perception of self-efficacy in handling AI technologies, consisting of 22 items covering four fundamental dimensions: assistance, anthropomorphic interaction, comfort with AI, and technological skills. Validated in Taiwan, this scale evaluates self-efficacy in the context of AI, highlighting the complexity of this technology and its impact on individuals in various contexts, both educational and professional. Conversely, the General Self-Efficacy Scale (GSE) is a widely used instrument measuring an individual's belief in their ability to manage a variety of difficult or challenging situations (Schwarzer and Born, 1997). An abbreviated version of this scale is the GSE-6 (Rompel et al., 2013). Although not specifically focused on AI, its brevity and generalist approach make it useful for extensive studies and situations requiring a quick assessment of self-efficacy. Its use across multiple domains emphasizes the universality of the self-efficacy concept and its applicability in varied life situations.

The comparison between AISES and GSE-6 illustrates the dichotomy between the need for domain-specific measures and more general assessment tools. While AISES provides a detailed and contextual evaluation of self-efficacy in using AI, capturing the specific challenges and peculiarities of this technology, GSE-6 offers a general perspective that can be applied across a wide range of situations, including those related to AI. This distinction highlights the importance of developing and adapting scales that reflect the unique challenges and opportunities presented by AI, suggesting that adapting GSE-6 to the AI context could provide a concise and easily administered measure of AI-related self-efficacy. In this way, a more general tool that still reflects the specificity of the AI context could be offered, benefiting researchers, educators, and professionals interested in assessing and enhancing individuals' readiness to interact with AI. Therefore, the aim of this study is to adapt and validate a scale of self-efficacy in using Artificial Intelligence among Peruvian students.

## 2 Materials and methods

### 2.1 Design and participants

An instrumental and cross-sectional study was conducted with the purpose of examining the psychometric properties of a documentary measurement instrument (Ato et al., 2013). Furthermore, a non-random convenience sampling method was used to select medical students from three Peruvian universities who were between their first and tenth study cycles. Students already in the eleventh cycle or beyond, typically engaged in hospital practices, were excluded. An essential inclusion criterion

was the use of Artificial Intelligence in their academic training, specifically those students who dedicate at least 8 h a week to activities involving AI use. The sample selection was based on a precise calculation using an electronic calculator (Soper, 2023), considering specific variables such as the number of observed and latent aspects in the proposed model, an expected effect size of  $\lambda = 0.10$ , a statistical significance of  $\alpha = 0.05$ , and a statistical power level of  $1 - \beta = 0.80$ . Although the minimum sample size required for the model structure was 200 participants, a total of 469 students were recruited. These participants had ages ranging from 18 to 29 years ( $M = 19.71$ ;  $SD = 2.47$ ). It was observed that 53.3% were women, 26.2% were in their first cycle of studies, and 51% came from the coastal region of Peru (Table 1).

### 2.2 Instrument

Self-Efficacy in Using Artificial Intelligence: The Self-Efficacy in Using Artificial Intelligence Scale was derived from the adaptation of the 6-item General Self-Efficacy Scale (GSE-6) (Rompel et al., 2013), representing a shortened version of the original 10-item GSE scale (Schwarzer and Born, 1997). The GSE-6 assesses an individual's perceived level of self-efficacy with response options ranging from 1 = "not at all true" to 4 = "exactly true." To obtain an overall score, responses to all items are summed. Initial values on the reliability of the GSE-6 were adequate, recording Cronbach's alphas of 0.86, 0.88, and 0.88 in three consecutive evaluations. Moreover, the GSE-6 has demonstrated robust psychometric properties in various cultural contexts and in both clinical and non-clinical samples.

To adapt the GSE-6 to the specific context of Artificial Intelligence use and application, translation and cultural adaptation procedures were followed (Beaton et al., 2000). In the initial phase, three bilingual native Spanish speakers independently translated the GSE-6 into Spanish. This translated version was then back-translated into English by three native English speakers who were not familiar with the scale. Three psychologists and an educator thoroughly reviewed this Spanish translation and, after deliberations, decided to adjust the wording of the 6 items to align with the context of Artificial Intelligence use, resulting in the GSE-6AI version. Additionally, content validation was conducted through expert judgment. To test the readability and comprehensibility of this adaptation, it was administered to a pilot group of 13 medical students. The results indicated clear and readable comprehension (Table 2).

### 2.3 Procedure

The study was conducted following stringent ethical standards, aligned with the Helsinki Declaration (Puri et al., 2009). It received approval from the Ethics Committee of a Peruvian university (2023-CEUPeU-044). Data collection was carried out in person at three Peruvian universities, ensuring participants that their participation was voluntary and all provided information would be treated anonymously to maintain their privacy and confidentiality. Before participating, informed consent was obtained from each individual, ensuring their rights were respected throughout the research process.

## 2.4 Analysis

In the preliminary phase, the content validity of the items in the GSE-6AI was assessed by expert judges, who were selected and contacted through both electronic and face-to-face means. This review focused on evaluating three critical aspects of each item: its relevance, determining the importance and essentiality of the item for the construct under study; its coherence, assessing the consonance of the item with the construct it intends to measure; and its clarity, measuring the ease of understanding and the unambiguity of the item. The evaluation of these criteria was performed using a scale from 0 to 3, where 0 indicates the absence and 3 the total presence of the evaluated characteristic. Each item was assessed in an approximate period

of 5 min. To quantify these aspects, the Aiken's V coefficient, along with its 95% confidence intervals, was applied (Aiken, 1980). This procedure was carried out using software specifically designed in MS Excel®. The Aiken's V coefficient ranges from 0 to 1, where values close to 1 indicate a high degree of clarity, coherence, and relevance. Items with an Aiken's V coefficient  $\geq 0.70$  are considered positively rated at the sample level, and those whose lower limit of the confidence interval exceeds 0.59 are deemed appropriate at the population level (Penfield and Giacobbi, 2004).

Subsequently, a descriptive analysis of the items belonging to the General Self-Efficacy Scale with Artificial Intelligence (GSE-6AI) was conducted. This analysis followed the criteria of Pérez and Medrano (2010), where skewness (g1) and kurtosis (g2) were deemed adequate if their values were within  $\pm 1.5$ . Items with a corrected item-test correlation [r(i-tc)] of  $\leq 0.2$  or that showed signs of multicollinearity (i-tc)  $\leq 0.2$  were excluded (Kline, 2016).

Following this descriptive analysis, a Confirmatory Factor Analysis (CFA) was implemented, focusing on the unidimensional aspect of the GSE-6AI scale, using the MLR estimator. This estimator is renowned for its robustness against potential deviations from normality (Muthén and Muthén, 2017). Fit criteria were based on metrics like the chi-square test ( $\chi^2$ ). RMSEA and SRMR values below 0.08 and 0.05 indicate acceptable and optimal fit, respectively (Kline, 2011; Bandalos and Finney, 2019). For CFI and TLI, values above 0.90 are recommended, and those exceeding 0.95 denote an excellent model fit (Schumacker and Lomax, 2016).

To ensure the scale's equivalence across different demographic groups, especially regarding gender, measurement invariance (MI) was evaluated using a multi-group confirmatory factor analysis. Four critical levels of invariance were considered: Configural, Metric, Scalar, and Strict. The criterion adopted for determining invariance between gender groups was based on  $\Delta$ CFI differences less than 0.010 (Chen, 2007). In terms of internal consistency, both Cronbach's alpha coefficient and McDonald's omega coefficient (McDonald, 1999) were used, anticipating values above 0.70 as an indicator of reliability (Raykov and Hancock, 2005).

TABLE 1 Sociodemographic characteristics.

Characteristics		n	%
Gender	Female	250	53.3
	Male	219	46.7
Study Cycle	1	123	26.2
	2	109	23.2
	3	24	5.1
	4	57	12.2
	5	13	2.8
	6	48	10.2
	7	13	2.8
	8	23	4.9
	9	8	1.7
	10	51	10.9
Place of origin		6	
	Coast	239	51.0
	Jungle	120	25.6
	Highlands	104	22.2

TABLE 2 Aiken's V for the evaluation of clarity, coherence, and relevance of the items.

Items	Relevance (n = 7)				Coherence (n = 7)				Clarity (n = 7)			
	M	SD	V	CI 95%	M	SD	V	CI 95%	M	SD	V	CI 95%
Item 1	2.71	0.49	0.90	(0.71–0.97)	2.71	0.49	0.90	(0.71–0.97)	2.71	0.49	0.90	(0.71–0.97)
Item 2	2.86	0.38	0.95	(0.77–0.99)	2.86	0.38	0.95	(0.77–0.99)	3.00	0.00	1.00	(0.85–1.00)
Item 3	2.86	0.38	0.95	(0.77–0.99)	3.00	0.00	1.00	(0.85–1.00)	2.86	0.38	0.95	(0.77–0.99)
Item 4	2.86	0.38	0.95	(0.77–0.99)	2.86	0.38	0.95	(0.77–0.99)	2.71	0.49	0.90	(0.71–0.97)
Item 5	2.86	0.38	0.95	(0.77–0.99)	2.71	0.49	0.90	(0.71–0.97)	2.71	0.49	0.90	(0.71–0.97)
Item 6	2.86	0.38	0.95	(0.77–0.99)	2.86	0.38	0.95	(0.77–0.99)	3.00	0.00	1.00	(0.85–1.00)

M, mean; SD, standard deviation; CI 95%, 95% confidence interval for Aiken's V; V, Aiken's V coefficient.

TABLE 3 Adaptation of original, translated, and modified items.

No	Items in English	Items in Spanish	Adapted items
1	If someone opposes me, I can find means and ways to get what I want.	Si alguien se opone a mí, puedo encontrar medios y formas de obtener lo que quiero.	Si alguien se opone a mí, puedo encontrar medios y formas de obtener lo que quiero utilizando la inteligencia artificial/If someone opposes me, I can find means and ways to get what I want by using artificial intelligence.
2	It is easy for me to stick to my aims and accomplish my goals.	Es fácil para mí mantenerme fiel a mis objetivos y lograr mis metas.	Es fácil para mí mantenerme fiel a mis objetivos y alcanzar mis metas con la ayuda de la inteligencia artificial/It's easy for me to stay true to my goals and achieve my objectives with the help of artificial intelligence.
3	I am confident that I could deal efficiently with unexpected events.	Confío en que puedo manejar eficientemente eventos inesperados.	Tengo confianza en que podría enfrentar eficientemente eventos inesperados utilizando la inteligencia artificial/I am confident that I could efficiently face unexpected events by using artificial intelligence.
4	Thanks to my resourcefulness, I know how to handle unforeseen situations.	Gracias a mi ingenio, sé cómo enfrentar situaciones imprevistas.	Gracias a mi ingenio respaldado por la inteligencia artificial, sé cómo manejar situaciones imprevistas/Thanks to my wit supported by artificial intelligence, I know how to handle unforeseen situations.
5	I can remain calm when facing difficulties because I can rely on my coping abilities.	Puedo mantener la calma ante las dificultades porque confío en mi capacidad para afrontarlas.	Puedo mantener la calma cuando enfrento dificultades porque confío en mis habilidades de afrontamiento respaldadas por la inteligencia artificial/I can stay calm when facing difficulties because I trust in my coping skills backed by artificial intelligence.
6	No matter what comes my way, I'm usually able to handle it.	Sin importar lo que se presente, generalmente puedo manejarlo.	No importa lo que se presente, generalmente puedo manejarlo con el apoyo de la inteligencia artificial/No matter what comes up, I can usually handle it with the support of artificial intelligence.

All statistical processing was performed using R, specifically version 4.1.1. For the CFA and structural equation modeling, the "lavaan" package was applied (Rosseel, 2012). Meanwhile, "semTools" facilitated the measurement invariance analysis, ensuring meticulous interpretation of the findings (Jorgensen et al., 2021).

## 3 Results

### 3.1 Content validity

The table displays the results of the evaluation for relevance, representativeness, and clarity of the items of the assessed instrument, quantified through the Aiken's V coefficient and their respective 95% confidence intervals (CI 95%). At the sample level, all items showed Aiken's V values indicating highly positive evaluations in terms of relevance, representativeness, and clarity, with values above 0.70, indicating a high valuation of these aspects. Specifically, items 2 and 6 stand out for achieving perfect scores in clarity ( $V = 1.00$ ; CI 95%: 0.85–1.00) and representativeness ( $V = 1.00$ ; CI 95%: 0.85–1.00) for item 3, highlighting their total comprehensibility and alignment with the measured construct. The consistency in high scores across different items reflects a uniformity in the experts' perception of the content quality of the instrument. Furthermore, the lower limit of the CI 95% for all Aiken's V values exceeds the established criterion for adequate valuation at the population level ( $Li > 0.59$ ), underscoring the robustness of the items in terms of relevance, representativeness, and clarity from a broader perspective (Table 3).

### 3.2 Descriptive statistics

In the GSE-6AI descriptive analysis (Table 4), item 4 "Thanks to my wit and AI, I know how to handle unforeseen situations." reported the highest mean ( $M = 2.58$ ,  $SD = 0.87$ ). Meanwhile, items 1 "If someone opposes me, I can find ways to get what I want with AI's help." and 6 "No matter what comes up, I can usually handle it with Artificial Intelligence's support." shared the lowest mean ( $M = 2.42$ ). Concerning data normality, all items showed skewness (g1) and kurtosis (g2) values within the acceptable range of  $\pm 1.5$ , indicating a roughly normal distribution for each item's responses. Specifically, skewness ranged from  $-0.21$  to  $0.08$ , and kurtosis from  $-0.55$  to  $-0.70$ . Evaluating item-total correlations (r.cor), all items exceeded the 0.30 acceptability threshold, with values ranging from 0.67 to 0.77. This suggests each item's significant contribution to the scale's overall consistency, so there's no need to exclude any item based on these correlations.

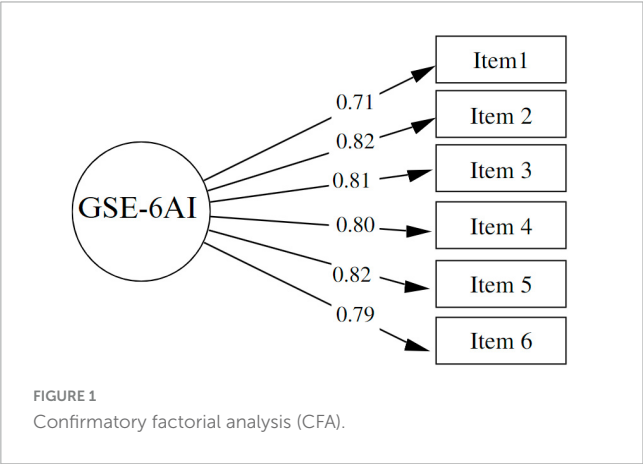
### 3.3 Validity based on internal structure

The GSE-6AI's confirmatory factor analysis (Figure 1) displayed an adequate model fit to the data. Specifically, the obtained indices were as follows:  $\chi^2 = 17.480$ ,  $df = 9$ ,  $p < 0.01$ ; CFI = 0.99; TLI = 0.98; RMSEA = 0.04 (90% CI: 0.02–0.07) and SRMR = 0.02. All indices indicate an excellent model fit, considering the generally accepted standards in the literature (Schumacker and Lomax, 2016). Furthermore, all item factor loadings ( $\lambda$ ) exceeded the recommended threshold ( $> 0.50$ ), suggesting each item's significant contribution to the measured

TABLE 4 Descriptive statistics and reliability.

Item	M	sd	g1	g2	r.cor
Item 1	2.42	0.83	0.08	−0.55	0.67
Item 2	2.51	0.87	−0.05	−0.67	0.77
Item 3	2.51	0.84	−0.14	−0.58	0.77
Item 4	2.58	0.87	−0.21	−0.62	0.75
Item 5	2.53	0.85	−0.15	−0.61	0.77
Item 6	2.42	0.87	0.02	−0.7	0.75

M, mean; SD, standard deviation; r.cor, item-total correlation.



construct. In terms of reliability, the scale’s internal consistency was found to be high, with a Cronbach’s Alpha ( $\alpha$ ) and McDonald’s Omega ( $\omega$ ) of 0.91, exceeding the generally accepted 0.70 threshold (Raykov and Hancock, 2005).

### 3.4 Factorial invariance by gender

The sequence of invariance models applied to the GSE-6AI scale among university students to assess its consistency across genders reveals significant findings that support its applicability and reliability in both groups. Starting with configural invariance (M1), which establishes a common baseline factorial structure between genders, the research progressed toward progressively more restrictive levels of invariance: metric (M2), scalar (M3), and strict (M4). The analysis of differences in the Comparative Fit Index ( $\Delta$ CFI) shows minimal variations between models, with  $\Delta$ CFI values of 0.003, −0.001, and 0 for the transitions from M1 to M2, M2 to M3, and M3 to M4, respectively. These results, consistently below the threshold of 0.010 proposed by Chen (2007), indicate a solid invariance of the scale across genders, confirming that the psychometric properties of the GSE-6AI are stable between men and women (Table 5).

## 4 Discussion

AI has become a foundational pillar in technological evolution, significantly impacting the educational sector by promising a

transformation of teaching and learning methodologies. The acceptance of AI varies according to individual perceptions of technological competence, and its implementation is optimizing both the efficiency and personalization of learning. It’s essential to develop digital skills and familiarize oneself with AI to maximize its benefits, though this entails facing ethical and practical challenges. Technological self-efficacy, determined by previous experiences and the perception of its utility, is crucial for adopting AI. Advances in Generative AI can increase students’ self-efficacy and improve their academic outcomes. The integration of AI in education demands a critical evaluation of its advantages and potential challenges. This underscores the importance of developing tools that address the specific challenges and opportunities presented by AI.

Our research aimed to adapt and validate the Artificial Intelligence Use Self-Efficacy Scale (GSE-6AI), derived from the 6-item General Self-Efficacy Scale (GSE-6). This study responds to the growing interest in understanding how self-efficacy perceptions affect the adoption and use of advanced technologies, such as artificial intelligence (AI). Previous research has examined self-efficacy in technological contexts, highlighting works such as those by Grassini (2023) and Wang and Chuang (2023), who developed scales for measuring self-efficacy and attitudes toward AI, respectively. These contributions are crucial for understanding individuals’ willingness to interact with emerging technologies, key to adopting AI. We adapted the GSE-6 to the realm of AI through a process of translation and content validation, assessing its clarity, coherence, and relevance with Aiken’s V coefficient. Unlike studies like that of Yilmaz et al. (2023), which focused on the acceptance of generative AI, our work concentrates on self-efficacy, emphasizing the role of individual beliefs in the ability to use AI efficiently. The content validity of the GSE-6AI was established through expert review, a crucial step also present in the creation of other instruments, such as the AI Anxiety Scale by Wang and Wang (2022). This process ensures that the items accurately reflect the concept of self-efficacy in using AI. The comparison with the study by Çelebi et al. (2023), on the adaptation of the AI Literacy Scale, highlights the need to address not only self-efficacy but also knowledge and understanding of AI. The results confirm the clarity and applicability of the adapted scale across different cultural contexts and populations, in line with research like that of Moodi et al. (2023), who analyzed the psychometric characteristics of a readiness scale for AI in medical students, demonstrating the usefulness of



TABLE 5 Factorial invariance by gender.

Invariance	$\chi^2$	df	p	TLI	RMSEA	90% CI	SRMR	CFI	$\Delta$ CFI
M1	30.77	18	0.031	0.974	0.055	(0.031–0.078)	0.027	0.984	
M2	38.06	23	0.025	0.976	0.053	(0.029–0.074)	0.037	0.981	0.003
M3	42.26	28	0.041	0.981	0.047	(0.022–0.068)	0.038	0.982	–0.001
M4	48.99	34	0.046	0.984	0.043	(0.019–0.063)	0.039	0.982	0.000

M1, configural; M2, metric; M3, scalar; M4, strict;  $\chi^2$ , chi-square; df, degrees of freedom; RMSEA, root mean square error of approximation; SRMR, standardized root mean-square; TLI, tucker.

having specific assessment tools for different areas of AI application.

Additionally, a Confirmatory Factor Analysis (CFA) was conducted for the GSE-6AI, confirming its unidimensionality. When compared with similar studies, such as Wang and Chuang (2023), who developed and validated an AI self-efficacy scale, and Grassini (2023), who adapted a scale for attitudes toward AI, a common trend in the importance of validating the psychometric properties of these instruments across specific cultural contexts and various AI application domains was found. The consistency in the results of these studies highlights the importance of AI-specific scales in assessing psychological constructs within the technological realm, as well as their applicability in various contexts. In this regard, the GSE-6AI demonstrated superior fit indices compared to the previously established Artificial Intelligence Self-Efficacy Scale (AISES). While both instruments aim to measure aspects of self-efficacy, the GSE-6AI presents as a more concise and focused tool for the context of Artificial Intelligence. Moreover, the item factor loadings exceeded the recommended threshold ( $\lambda > 0.50$ ), indicating that each item is relevant and reinforces the internal coherence of the scale.

Furthermore, the GSE-6AI has shown high internal consistency, with Cronbach's Alpha ( $\alpha$ ) and McDonald's Omega ( $\omega$ ) coefficients of 0.91, indicating adequate reliability for measuring self-efficacy in the context of artificial intelligence (Raykov and Hancock, 2005). This result is in line with findings from previous studies that evaluated the reliability of similar scales in various contexts, demonstrating the robustness of the GSE-6's psychometric properties. Research such as that by Grassini (2023), Moodi et al. (2023), and Wang and Chuang (2023) generally report high reliability coefficients for scales related to self-efficacy and attitudes toward artificial intelligence. For instance, Wang and Chuang achieved a Cronbach's Alpha of 0.852, while Grassini reported Cronbach's Alpha and McDonald's Omega values indicating good internal consistency for different factors. These findings underscore the need for reliable measurement tools in the field of artificial intelligence, facilitating accurate comparisons and generalizations across different studies. However, the importance of continuing research to address potential gaps, especially in adapting these scales to specific cultural and linguistic contexts, is recognized. Adaptation studies conducted by Çelebi et al. (2023) and Karaoğlu and Yılmaz (2023) demonstrated high levels of reliability in the adapted

versions of the scale, proving the effectiveness of these efforts.

Moreover, the study presents factorial invariance by gender for the GSE-6AI, showcasing a thorough analysis of the scale's factorial structure, focusing on group comparison by gender. Through a hierarchical methodology, different levels of invariance were tested: configural, metric, scalar, and strict, consistently showing good fits at all levels and suggesting that the scale maintains its structure and meaning across genders, indicating that the scale measures the general self-efficacy construct assisted by artificial intelligence equivalently in both men and women.

### 4.1 Implications

The validation of the GSE-6AI offers a significant contribution to the psychometric understanding of how individuals perceive their ability to interact with artificial intelligence technology. Adapting the GSE-6 scale to the AI context not only broadens its scope of applicability but also highlights the importance of domain specificity in evaluating self-efficacy. The rigor in the process of translation and cultural adaptation, followed by validation by experts, ensures that the GSE-6AI is a reliable and relevant tool for measuring self-efficacy in AI use, respecting sociolinguistic variations and adapting to the contextual reality where it is applied. The results of the psychometric validation of the GSE-6AI provide solid evidence of its utility in educational and professional environments, where AI is emerging as a critical tool. Since self-efficacy has been identified as a key predictor of technology adoption, self-directed learning, and the ability to face technological challenges, the GSE-6AI can be used in developing interventions aimed at improving AI-related self-efficacy among students and professionals, thereby facilitating a smoother transition toward integrating AI into various practices. The ability to accurately measure this self-efficacy may lead to a deeper understanding of how individual perceptions of the ability to use AI influence specific behaviors and, ultimately, success in AI adoption.

Identifying self-efficacy in AI use is essential for designing educational interventions aimed at enhancing the integration of



these technologies into the classroom. Educators can use the GSE-6AI as a tool to assess and improve students' confidence in using AI-based tools. Students with low levels of self-efficacy could benefit from specific training programs that provide them with the support and skills necessary to tackle current technological challenges. Educational institutions, in turn, might consider incorporating AI modules or workshops into their curricula, allowing students to become familiar with these technologies from early stages of their education.

Moreover, it is crucial for administrators to recognize the importance of self-efficacy in AI use. This implies promoting educational policies that prioritize training in emerging technologies and that ensure equitable access to these tools, thus preventing the widening of the technological gap. Additionally, considering gender equality is vital, as the scale has shown invariance between men and women, suggesting that both genders perceive their ability to use AI similarly.

We recommend that future research explore the relationship between AI self-efficacy and other relevant constructs, such as academic performance, satisfaction with the learning process, or student wellbeing. It would also be relevant to assess the GSE-6AI in other contexts, such as the workplace or recreational settings, to understand how these beliefs manifest in different areas of daily life.

## 4.2 Limitations

However, it's crucial to acknowledge the inherent limitations in developing pan-dialectal versions of psychometric instruments and the need for specific linguistic and cultural adaptations for particular contexts. The GSE-6AI, though validated in a specific context, requires ongoing validation across diverse cultural and educational settings to ensure its generalizability and accuracy in different populations. Additionally, the cross-sectional nature of the study prevents establishing causal relationships between the examined variables. Future research could benefit from longitudinal designs that provide a deeper understanding of the evolution and stability of self-efficacy beliefs in relation to AI over time. Also, the self-reported nature of the data. While self-reported scales are common and valuable tools, they are susceptible to biases such as social desirability. The inclusion of objective assessments, such as performance tests or interviews, could offer a more holistic view of AI-related self-efficacy. Lastly, although gender invariance was analyzed and confirmed, it would be fruitful to explore invariance across other demographic groups, such as different ages, educational levels, or cultural backgrounds. AI is a global tool, and understanding how different populations perceive their self-efficacy in this domain is essential for more inclusive implementation.

## 5 Conclusion

The adaptation and validation of the GSE-6AI in the Peruvian educational context represent a significant contribution to understanding individual perceptions of competence in using AI. This study, by confirming the psychometric validity and gender invariance of the GSE-6AI, underscores the importance of technological self-efficacy for successful integration of AI in

education and demonstrates the scale's universality and adaptability to different cultural and educational contexts. The findings support the idea that strengthening AI self-efficacy among students and professionals can facilitate greater acceptance and effective use of these technologies, enhancing associated educational and occupational benefits. However, exploring the implications of these perceptions on various academic and professional outcomes is essential. Longitudinal evaluation of AI self-efficacy can offer deeper insights into how specific interventions could improve technological readiness and overall performance in an increasingly digitalized world.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The research was evaluated by the Ethics Committee of the Universidad Peruana Unión (Code: 2023-CEUPeU-044). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## Author contributions

WM-G: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing – original draft, Writing – review and editing. LS-S: Conceptualization, Data curation, Formal Analysis, Methodology, Project administration, Resources, Visualization, Writing – original draft, Writing – review and editing. SM-G: Conceptualization, Data curation, Methodology, Resources, Software, Visualization, Writing – original draft, Writing – review and editing. MM-G: Conceptualization, Data curation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

- Agarwal, R., and Karahanna, E. (2000). Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS Quart.* 24, 665–694. doi: 10.2307/3250951
- Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educ. Psychol. Measure.* 40, 955–959. doi: 10.1177/001316448004000419
- Ato, M., López, J., and Benavente, A. (2013). Un sistema de clasificación de los diseños de investigación en psicología. *Anal. Psicol.* 29, 1038–1059. doi: 10.6018/analesps.29.3.178511
- Bandalos, D. L., and Finney, S. J. (2019). "Factor analysis: Exploratory and confirmatory," in *The reviewer's guide to quantitative methods in the social sciences*, eds G. R. Hancock, L. M. Stapleton, and R. O. Mueller (Berlin: Routledge).
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychol. Rev.* 84, 191–215. doi: 10.1037/0033-295X.84.2.191
- Beaton, D. E., Bombardier, C., Guillemin, F., and Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine* 25, 3186–3191. doi: 10.1097/00007632-200012150-00014
- Çelebi, C., Yilmaz, F., Demir, U., and Karakuş, F. (2023). Artificial intelligence literacy: An adaptation study. *Instr. Technol. Lifelong Learn.* [Epub ahead of print]. doi: 10.52911/ital.1401740
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equat. Model.* 14, 464–504. doi: 10.1080/10705510701301834
- Compeau, D. R., and Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quart.* 19, 189–211. doi: 10.2307/249688
- Grassini, S. (2023). Development and validation of the AI attitude scale (AIAS-4): A brief measure of general attitude toward artificial intelligence. *Front. Psychol.* 14:1191628. doi: 10.3389/fpsyg.2023.1191628
- Hong, J. W. (2022). I was born to love AI: The influence of social status on AI self-efficacy and intentions to use AI. *Int. J. Commun.* 16:20.
- Hsu, M. H., and Chiu, C. M. (2004). Internet self-efficacy and electronic service acceptance. *Decis. Support Syst.* 38, 369–381. doi: 10.1016/j.dss.2003.08.001
- Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., and Rosseel, Y. (2021). *semTools: Useful tools for structural equation modeling*. In *The Comprehensive R Archive Network*. Available online at: <https://CRAN.R-project.org/package=semTools>
- Karaoglan, F. G., and Yilmaz, R. (2023). Yapay zekâ okuryazarlığı ölçeğinin türkçeye uyarlanması. *Bilg. Yönetim Teknol. Derg.* 5, 172–190. doi: 10.53694/bited.1376831
- Kim, B.-J., Kim, M.-J., and Lee, J. (2024). Examining the impact of work overload on cybersecurity behavior: highlighting self-efficacy in the realm of artificial intelligence. *Curr. Psycho.* doi: 10.1007/s12144-024-05692-4
- Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling*. London: Guilford Press.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling*. London: Guilford Press.
- Kwak, Y., Ahn, J. W., and Seo, Y. H. (2022). Influence of AI ethics awareness, attitude, anxiety, and self-efficacy on nursing students' behavioral intentions. *BMC Nurs.* 21:267. doi: 10.1186/s12912-022-01048-0
- Liang, J., Wang, L., Luo, J., Yan, Y., and Fan, C. (2023). The relationship between student interaction with generative artificial intelligence and learning achievement: serial mediating roles of self-efficacy and cognitive engagement. *Front. Psychol.* 14:1285392. doi: 10.3389/fpsyg.2023.1285392
- McDiarmid, G. W., and Zhao, Y. (2023). Time to rethink: Educating for a technology-transformed world. *ECNU Rev. Educ.* 6, 189–214. doi: 10.1177/20965311221076493
- McDonald, R. P. (1999). *Test Theory: A United Treatment*. Mahwah, NJ: Lawrence Erlbaum.
- Moodi, A. A., Moghadasin, M., Emadzadeh, A., and Mastour, H. (2023). Psychometric properties of the Persian version of the Medical Artificial Intelligence Readiness Scale for Medical Students (MAIRS-MS). *BMC Med. Educ.* 23:577. doi: 10.1186/s12909-023-04553-1
- Muthén, L., and Muthén, B. (2017). *Mplus Statistical Analysis with latent variables. User's guide*. Los Angeles: Muthén & Muthén.
- Oran, B. B. (2023). Correlation between artificial intelligence in education and teacher self-efficacy beliefs: a review. *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi* 34:1316378. doi: 10.29000/rumelide.1316378
- Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Rev. Educ. Res.* 66, 543–578. doi: 10.3102/00346543066004543
- Penfield, R. D., and Giacobbi, P. R. (2004). Applying a score confidence interval to Aiken's item content-relevance index. *Measure. Phys. Educ. Exerc. Sci.* 8, 213–225. doi: 10.1207/s15327841mpee0804\_3
- Pérez, E. R., and Medrano, L. (2010). Análisis factorial exploratorio: Bases conceptuales y metodológicas. *Rev. Arg. Cienc. Comput.* 2, 58–66.
- Puri, K. S., Suresh, K. R., Gogtay, N. J., and Thatte, U. M. (2009). Declaration of Helsinki, 2008: Implications for stakeholders in research. *J. Postgrad. Med.* 55, 131–134. doi: 10.4103/0022-3859.52846
- Pütten, A. R., and Von Der Bock, N. (2018). Development and validation of the self-efficacy in human-robot-interaction scale (SE-HRI). *ACM Trans. Hum. Robot Interact.* 7:3139352. doi: 10.1145/3139352
- Raykov, T., and Hancock, G. R. (2005). Examining change in maximal reliability for multiple-component measuring instruments. *Br. J. Math. Stat. Psychol.* 58, 65–82. doi: 10.1348/000711005X38753
- Romppel, M., Herrmann-Lingen, C., Wachter, R., Edelmann, F., Dungen, H.-D., Pieske, B., et al. (2013). A short form of the General Self-Efficacy Scale (GSE-6): Development, psychometric properties and validity in an intercultural non-clinical sample and a sample of patients at risk for heart failure. *Psycho-Soc. Med.* 10:Doc01. doi: 10.3205/psm000091
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/JSS.V048.I02
- Schepman, A., and Rodway, P. (2020). Initial validation of the general attitudes towards Artificial Intelligence Scale. *Comput. Hum. Behav. Rep.* 1:100014. doi: 10.1016/j.chbr.2020.100014
- Schumacker, R. E., and Lomax, R. G. (2016). *A Beginner's Guide to Structural Equation Modeling*, 4th Edn. Milton Park: Taylor & Francis.
- Schwarzer, R., and Born, A. (1997). *Optimistic self-beliefs: Assessment of general perceived self-efficacy in thirteen cultures*. Berlin: Freie Universität Berlin.
- Soper, D. (2023). *A-priori Sample Size Calculator for structural equation models*. Software. Available online at: <http://www.danielsoper.com/statcalc> (accessed on 30 July 2023).
- Wang, Y.-Y., and Chuang, Y.-W. (2023). Artificial intelligence self-efficacy: Scale development and validation. *Educ. Inf. Technol.* [Epub ahead of print]. doi: 10.1007/s10639-023-12015-w
- Wang, Y. Y., and Wang, Y. S. (2022). Development and validation of an artificial intelligence anxiety scale: an initial application in predicting motivated learning behavior. *Interact. Learn. Environ.* 30:8812542. doi: 10.1080/10494820.2019.1674887
- Washington, J. (2023). The impact of generative artificial intelligence on writer's self-efficacy: A critical literature review. *SSRN Electron. J.* doi: 10.2139/ssrn.4538043
- Yilmaz, F. G. K., Yilmaz, R., and Ceylan, M. (2023). Generative artificial intelligence acceptance scale: A Validity and Reliability Study. *Int. J. Hum. Comput. Interact.* [Epub ahead of print]. doi: 10.1080/10447318.2023.2288730
- Yilmaz, R., and Karaoglan Yilmaz, F. G. (2023a). Augmented intelligence in programming learning: Examining student views on the use of ChatGPT for programming learning. *Comput. Hum. Behav.* 1:100005. doi: 10.1016/j.chbah.2023.100005
- Yilmaz, R., and Karaoglan Yilmaz, F. G. (2023b). The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Comput. Educ.* 4:100147. doi: 10.1016/j.caeai.2023.100147
- Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., et al. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. *Complexity* 2021:8812542. doi: 10.1155/2021/8812542
- Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemp. Educ. Psychol.* 25:1016. doi: 10.1006/ceps.1999.1016



## OPEN ACCESS

## EDITED BY

Martina Rau,  
University of Wisconsin-Madison,  
United States

## REVIEWED BY

Thomas Neumayr,  
University of Applied Sciences Upper Austria,  
Austria  
Matthias Wölfel,  
Karlsruhe University of Applied Sciences,  
Germany

## \*CORRESPONDENCE

Yusuke Yamamoto  
✉ [yusuke\\_yamamoto@acm.org](mailto:yusuke_yamamoto@acm.org)

RECEIVED 05 February 2024

ACCEPTED 14 March 2024

PUBLISHED 28 March 2024

## CITATION

Yamamoto Y (2024) Suggestive answers  
strategy in human-chatbot interaction: a  
route to engaged critical decision making.  
*Front. Psychol.* 15:1382234.  
doi: 10.3389/fpsyg.2024.1382234

## COPYRIGHT

© 2024 Yamamoto. This is an open-access  
article distributed under the terms of the  
[Creative Commons Attribution License \(CC  
BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in  
other forums is permitted, provided the  
original author(s) and the copyright owner(s)  
are credited and that the original publication  
in this journal is cited, in accordance with  
accepted academic practice. No use,  
distribution or reproduction is permitted  
which does not comply with these terms.

# Suggestive answers strategy in human-chatbot interaction: a route to engaged critical decision making

Yusuke Yamamoto\*

School of Data Science, Nagoya City University, Nagoya, Japan

In this study, we proposed a novel chatbot interaction strategy based on the *suggestive ending* of answers. This strategy is inspired by the cliffhanger ending narrative technique, which ends a story without specifying conclusions to spark readers' curiosity as to what will happen next and is often used in television series. Common chatbots provide relevant and comprehensive answers to users' questions. In contrast, chatbots with our proposed strategy end their answers with hints potentially interest-triggering users. The suggestive ending strategy aims to stimulate users' inquisition for critical decision-making, relating to a psychological phenomenon where humans are often urged to finish the uncompleted tasks they have initiated. We demonstrated the implication of our strategy by conducting an online user study involving 300 participants, where they used chatbots to perform three decision-making tasks. We adopted a between-subjects factorial experimental design and compared between the following UIs: (1) *plain* chatbot—it provides a generated answer when participants issue a question; (2) *expositive* chatbot—it provides a generated answer for a question, adding short summaries of a positive and negative person's opinion for the answer; (3) *suggestive* chatbot—it provides a generated answer for a question, which ends with a suggestion of a positive and negative person for the answer. We found that users of the *suggestive* chatbot were inclined to ask more questions to the bot, engage in prolonged decision-making and information-seeking actions, and formulate their opinions from various perspectives. These findings vary with the users' experience with *plain* and *expositive* chatbots.

## KEYWORDS

human-AI interaction, large language model, behavior change, critical information-seeking, chatbot

## 1 Introduction

Recent advancements in artificial intelligence (AI), particularly the remarkable evolution of large language models (LLMs), have given rise to a lot of services and applications that support human tasks in various domains. Generative AI with LLMs holds a strong potential for substantially changing human intellectual activities. For example, instruction-tuned LLMs (e.g., ChatGPT) can quickly generate surprisingly natural sentences in response to human questions (Wei et al., 2021). Zylowski and Wölfel (2023) revealed that when specifying personas for ChatGPT in prompts enables it to

simulate a variety of personalities and capabilities. OpenAI reported that ChatGPT scored 1,300/1,600 on the SAT<sup>1</sup> by eliciting knowledge in its language model<sup>2</sup>. In 2024, Google released *Gemini Ultra*, the highly capable LLM which outperforms GPT-4 on text-based tasks, including reasoning, reading comprehension, and code generation (Team et al., 2023). Furthermore, an appropriate understanding of LLM applications and their effective use can equally support decision-making and opinion formulation (Wambsganss et al., 2020, 2021; Jakesch et al., 2023; Petridis et al., 2023).

Despite their superlative functionalities, generative AIs with LLMs often generate incorrect, biased, or unrealistic information, a phenomenon known as *hallucination* (Maynez et al., 2020). Overreliance on AIs causes automation bias to users (Goddard et al., 2011), leading to the ubiquitous obliviousness of AI-generated false information (Lakkaraju and Bastani, 2020). Studies have shown that overusing AIs can inhibit the development of users' cognitive skills (Noyes, 2007; Carr, 2014), naturally affecting their critical thinking abilities. As a result, users can be unconsciously led to a specific polarity by opinionated AI assistants for writing (Jakesch et al., 2023). These aspects raise serious educational concerns. For instance, students using generative AI-powered chatbots can accept harmful/incorrect information without doubt, which strongly affects the development of their critical thinking and problem-solving skills (Kasneci et al., 2023).

Although the research on improving the performance of generative AIs with LLMs is under extensive development, undesirable output information remains highly probable (Wei et al., 2021; Nakano et al., 2022; Tay et al., 2022; Wang et al., 2023). This probability is particularly aggravated by the human *confirmation bias*, defined as the tendency to preferentially view or search for information consistent with one's opinions or hypotheses (Kahneman, 2011). Therefore, improving generative AIs should be accompanied by an effective design of human-AI interactions that promote users' cognitive activities for critical decision-making or opinion formulation.

In this study, we proposed a novel human-chatbot interaction strategy, *suggestive ending*, for generative AI-powered chatbot answers to foster decision-making from various perspectives. Our method is inspired by the *cliffhanger ending* narrative technique, which ends a story without specifying conclusions to spark readers' curiosity as to what will happen next. The cliffhanger method is often used in television series. It relates to a psychological phenomenon known as the *Ovsiankina effect*, where humans are often urged to finish the uncompleted tasks they have initiated (Wirz et al., 2023). *Suggestive* bots employed with the proposed strategy output their answers with hints to potentially interest-triggering subjects (Figure 1B). In contrast, common chatbots provide relevant and comprehensive answers to users' questions (Figure 1A). Therefore, when interacting with *SUGGESTIVE* chatbots on a given theme, users' proactive critical

decision-making is stimulated by intentionally leaving room for questions.

We conducted an online user study involving 300 participants to validate our proposed method on the human-AI interaction. Results revealed the following three primary observations.

- When using the *SUGGESTIVE* bot, participants engaged in decision-making by inputting questions to the bot. This has led participants to spend longer interactions with the *SUGGESTIVE* bot than the *PLAIN* (i.e., simply providing relevant answers) and *EXPOSITIVE* (i.e., providing relevant answers with supplementary information) bots.
- Compared to the *PLAIN* bot, participants were likely to write longer opinions from various perspectives using the *SUGGESTIVE* bot.
- When using either the *EXPOSITIVE* or *PLAIN* bot, participants showed similar efforts in their decision-making activities.

## 2 Related work

### 2.1 AI-assisted decision-making and opinion formation

AI systems developed to assist decision-making and opinion formation have been studied from the viewpoints of supporting interpretation of AI predictions, improving the understanding of arguments, enhancing the efficiency of opinion formulation, searching for supportive information, etc.

It is essential to understand how and why black-box AIs provide predictions for users to efficiently use them during decision-making. Hence, many researchers have studied *explainable AI* technologies to improve the interpretability of machine learning (ML) models. For instance, for ML behaviors on structured data, researchers have proposed various methods to summarize the contributed features to predictions (Lundberg and Lee, 2017; Fisher et al., 2019) and explain how the models work with data examples (Kim et al., 2016). Lakkaraju and Bastani (2020) reported misleading explanations on black-box MLs as a cause for users trusting even harmful MLs. Therefore, considering the characteristics of human design thinking is important to improve the interpretability of AIs for decision-making.

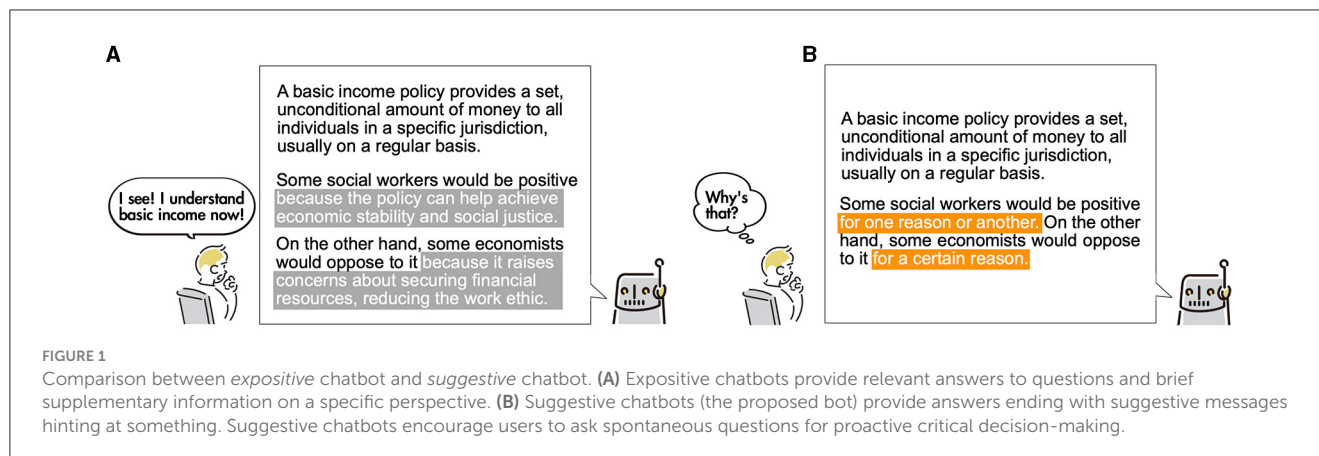
To better understand the aforementioned arguments, Wambsganss et al. (2021) proposed ARGUETUTOR, a chatbot system that provides users with feedback to identify sentences in their documents that require logical improvement. Furthermore, they proposed an interactive system to visualize the argumentation structure of a given document, thereby helping users make more logical judgments (Wambsganss et al., 2020). Petridis et al. (2023) developed ANGLEKINDLING, a system that supports journalists in exploring points to scrutinize potential negative impacts on press releases using an LLM.

Several investigations have been conducted on *suggestive keyboards* to support efficient opinion formulation. Arnold et al. (2016) proposed a phrase-suggesting method for text composition instead of predicting words following users' input texts. However, *suggestive keyboard* technologies could affect what to write. In another study, Arnold et al. (2020) reported that *suggestive*

1 The Scholastic Assessment Test (SAT) is a standardized test for college admissions across the United States. It is designed to assess students' proficiency in mathematics, reading, and writing.

2 <https://www.nytimes.com/2023/03/14/technology/openai-new-gpt4.html>





keyboard technologies affect the writers' choices who often follow AI-based text suggestions, while it improves their writing speed. Jakesch et al. (2023) reported that when suggestive keyboards were used, in which an LLM was fine-tuned to suggest positive (negative) phrases, users were likely to write positive (negative) opinions.

In the field of information retrieval (IR), several researchers have investigated systems for searching information to support decision-making. Rinott et al. (2015) proposed a method to search for evidence supporting a given claim from unstructured documents. Liu et al. (2022) proposed CRYSTALLINE, a browser developed to tabulate collected Web information for efficient decision-making.

The aforementioned studies reveal that users with sufficient skills and motivation to properly use advanced AI technologies can obtain useful assistance from these technologies in decision-making and opinion formulation. Otherwise, overreliance on AIs for decision-making causes negative impacts on users, including shortsighted decision making, cognitive downskilling, and opinion radicalization. Therefore, our proposed method focuses on eliciting questions from chatbot users and promoting active opinion formation in the human–chatbot interaction.

## 2.2 Generative information retrieval

With the emergence of LLMs, changes were introduced to the conventional IR model, which aims to provide a ranked list of relevant documents for a keyword query. *Generative IR* is a new LLMs-based paradigm of generating information to directly answer users' questions. When a question is given, typical generative IR systems (i.e., AI-powered chatbots) extend prompts with likely completions and extract answers from the extended prompts (Najork, 2023).

ChatGPT<sup>3</sup> and Google Gemini<sup>4</sup> are recently developed generative IR applications that have spurred unprecedented universal attention. Nevertheless, ongoing research is highlighting their drawbacks, such as generating incorrect or unrealistic answers, which is known as the *hallucination* phenomenon

(Maynez et al., 2020). Metzler et al. (2021) reported several challenges in generative IRs, such as suggesting contexts for generated answers and considering the authority or quality of documents for answer generation. Several methods have been proposed to tackle these challenges, such as tuning LLM models for human-favorable answers (Wei et al., 2021), linking generated answers (or questions) with relevant documents (Nakano et al., 2022; Tay et al., 2022), and improving the interpretability of generative AIs (Sun et al., 2022). Furthermore, Wang et al. (2023) proposed *Shepherd*, an LLM model that provides feedback to improve target LLMs by analyzing the generated texts.

While the aforementioned studies focus on the performance improvement and high functionality of generative IR systems, our proposed method focuses on enhancing users' information-seeking and cognitive activities in generative IRs.

## 2.3 Enhancing critical information seeking and decision-making

Studies conducted to activate and enhance information-seeking and decision-making abilities can be categorized into two approaches for steering and empowering better judgments: *nudging* (Thaler and Sunstein, 2009; Caraban et al., 2019) and *boosting* (Hertwig and Grne-Yanoff, 2017).

*Nudging* is defined as “an approach to alter people's behavior in a predictable way without forbidding any option or substantially changing their economic incentive” (Thaler and Sunstein, 2009). In the field of HCI, several methods have been reported for the application of this concept, which include supporting critical information seeking (Yamamoto and Yamamoto, 2018; Saito et al., 2020; Ioriya et al., 2022; Suzuki and Yamamoto, 2022) and enhancing privacy awareness on the Web (Wang et al., 2013; Zimmerman et al., 2019; Yamamoto and Yamamoto, 2020). For example, Yamamoto and Yamamoto (2018) proposed the QUERY PRIMING system, which inserts queries to evoke critical thinking during query completion/recommendation in a search system. Suzuki and Yamamoto (2022) proposed a search user interface (UI) that makes web searchers reflect on their webpage selection criteria and promote content-quality-oriented web searches regardless of visual appearances. Wang et al. (2013) proposed a privacy nudge

<sup>3</sup> ChatGPT: <https://chat.openai.com/>.

<sup>4</sup> Google Gemini: <https://gemini.google.com/>.



that shows Facebook profile pictures of the target audience when users post content on Facebook to enhance users' awareness to potential risks.

Nudging supports better decision-making by focusing on related systematic biases. However, *boosting* is an intervention to improve cognitive competence for proactive and rational decision-making (Hertwig and Grne-Yanoff, 2017). Shimizu et al. (2022) proposed *privacy-aware* snippets, which aim to enhance privacy risk judgment in Web searches by providing comprehensive information about sharing conditions of browsing histories. Harvey et al. (2015) reported that providing examples of high-quality queries can help users learn to improve the efficiency of their query formulation. Bućinca et al. (2021) reported that the users' final decision-making performance can be improved if they are required to think by themselves before the AIs provide supportive information for decision-making. Danry et al. (2023) reported that when AIs ask people a simple question to confirm a claim's logical validity, reasoning activities can be activated, and the fallacy identification performance can be improved.

While questioning approaches such as Danry et al. (2023)'s method are explicit boosting (i.e., instructive intervention), our method is regarded as implicit boosting (i.e., modest intervention). Our proposed strategy aims to trigger users' spontaneous questions through their interaction with chatbots, introducing suggestive messages in answers and leaving room for further questioning. We expect that our suggestive ending approach will be perceived as less intrusive than instructive questioning approaches.

### 3 Research questions

Our proposed *suggestive ending* strategy in chatbots for IR aims to provoke users' questions on a given theme or prior belief, driving them to make theme-dependent critical decisions. Ennis (1987) defined critical thinking as logical and reflective thinking to determine what to believe or do. Furthermore, the author claimed that ideal critical thinkers are disposed to seek reasons, consider entire situations, look for alternatives, and use critical thinking, e.g., deductive reasoning. Several studies revealed the effect of *lateral reading*, a method to check multiple information resources in parallel for critical review on a theme (Meola, 2004; Wineburg and McGrew, 2019; Brodsky et al., 2021). We expect that if chatbots implicitly suggest the existence of things to check at the end of their responses, users would be more willing to critically construct their opinions and gather information for validation compared to cases where chatbots provide detailed explanatory answers.

To explore the validity of our proposed strategy using SUGGESTIVE bots, we considered the following research questions:

**RQ1:** Do SUGGESTIVE bots engage users in investing additional effort to form their opinions and gather information for decision-making?

**RQ2:** Do SUGGESTIVE bots encourage users to consider various perspectives when making their decision?

As we are interested in exploring whether SUGGESTIVE bots should actively nudge people to question the details of the bot's ambiguous endings, we also investigated the following research question:

**RQ3:** Do question (query) suggestions along with suggestive bot's answer promote more critical decision-making?

According to the elaboration likelihood model theory proposed by Petty and Cacioppo (1986), people often pay more attention to information in which they have sufficient knowledge or strong understanding interest. Otherwise, they often use poor judgment for accepting or rejecting the information. Based on this theory, individual factors can affect people's effort and behavior in decision-making tasks as well as suggestive bot's behaviors. Therefore, we have also formulated the following research question:

**RQ4:** Do individual factors, such as knowledge, interest, and familiarity with the information sought using chatbots, affect associated decision-making tasks?

### 4 Materials and methods

We conducted an online user study to investigate the effect of suggestive ending in AI-powered chatbots on decision-making tasks. The user study was conducted in Japanese (on August 11 and 12, 2023). For this, we adopted a between-subjects factorial experimental design, where the factor is a user interface (UI) condition with four levels:

1. PLAIN chatbot: it provides a generated answer when participants issue a question (query) (Figure 2A).
2. EXPOSITIVE chatbot: provides a generated answer for a question, adding short summaries of a positive and negative person's opinion for the answer (Figure 2B).
3. SUGGESTIVE chatbot: it provides a generated answer for a question, which ends with a suggestion of a positive and negative person for the answer. Unlike the EXPOSITIVE chatbot, this bot does not show short opinion summaries (Figure 2C).
4. SUGGESTIVE<sup>++</sup> chatbot: as an extension of the SUGGESTIVE chatbot, it provides links to ask about suggested positive/negative people's opinions, following a generated answer (Figure 4).

We conducted a user study on a crowdsourcing platform. Crowdsourcing platforms such as Amazon Mechanical Turk<sup>5</sup> and Prolific<sup>6</sup> enable researchers to recruit a large number of participants via the internet at lower costs compared to traditional survey companies. Consequently, user studies with crowdsourcing have been becoming popular in the communities of Human-Computer Interaction (HCI) (Kittur et al., 2008; Komarov et al., 2013) and Information Retrieval (IR) (Yamamoto and Yamamoto, 2018; Câmara et al., 2021; Roy et al., 2021) as an alternative way to laboratory-based experiments. Numerous studies have examined the reliability of crowdsourcing by comparing crowd workers' performance to that of participants in laboratory settings (Lutz, 2016; Peer et al., 2017; Hettiachchi et al., 2022). These studies have demonstrated that crowdsourcing can be reliable for conducting user studies, provided that the online tasks are designed to control experimental environments and mitigate satisficing behaviors—whereby participants make judgments or complete tasks with

5 Amazon Mechanical Turk: <https://www.mturk.com/>.

6 <https://www.prolific.com/>



minimal effort. In light of these findings, we conducted a user study with a crowdsourcing service to examine the effectiveness of our proposed method. Note that we implemented an instructional manipulation check (IMC), a popular technique to identify inattentive crowd workers, to ensure the integrity of our data collection process. Furthermore, we rejected crowd workers using mobile/tablet devices so that all participants could perform tasks on the same layout on their PCs.

Participants were randomly allocated into one of the above four UIs. They then conducted tasks to summarize their opinions about three randomly allocated themes. To consider individual differences, we measured and analyzed personal factors as covariates, including the frequency of using chatbots for information seeking, interest in task themes, and familiarity with the themes. We designed this user study following the research ethics guidelines of our affiliated organization.

## 4.1 Themes for decision-making tasks

We prepared eight themes for decision-making tasks and one theme for practice tasks. The themes were prepared from the website of the National Association of Debate in Education, Japan. We selected the frequently used nine themes in debate competitions for high school students in Japan, as listed in Table 1. As presented in Table 1, the impressions of participants indicated their unfamiliarity with most themes on average. Moreover, the interests of participants were slightly inclined to positive polarity on average (excluding *making doggy bags available at restaurants*).

## 4.2 Chatbots

The aforementioned four UI conditions (chatbots) employed ChatGPT, OpenAI instruction-tuned LLM, via Azure OpenAI Service GPT API (gpt-3.5-turbo<sup>7</sup>) to generate answers for participants' questions. In particular, we used an LLM prompt-engineering technique in the SUGGESTIVE, SUGGESTIVE<sup>++</sup>, and EXPOSITIVE bots to complement additional information with plain answers for questions.

One possible way in our proposed *suggestive ending* strategy in chatbots is to suggest perspectives for decision-making explicitly, such as key issues (Cámara et al., 2021; Petridis et al., 2023) and positive/negative aspects for themes (Liao and Fu, 2014; Liao et al., 2015). However, such explicit suggestions are revealing and do not encourage users to proactively reflect on what they should think for their decision-making. On the one hand, studies in the field of learning science indicate that contents should leave proper room for questioning and discussion so that people would be willing to learn a theme and deepen their knowledge (King, 1992). On the other hand, it is difficult for users to find important questions and perspectives for a theme if they lack knowledge and interest.

Therefore, we designed two types of chatbots, namely, (SUGGESTIVE and SUGGESTIVE<sup>++</sup>), to provide direct answers to users' questions and additional suggestions on the existence of positive and negative people for a theme, respectively. The two chatbots never suggest the kind of perspectives the

<sup>7</sup> Azure OpenAI Service: <https://azure.microsoft.com/en-us/products/ai-services/openai-service-b>.

**TABLE 1** Themes for decision-making tasks and corresponding participants' impressions.

Theme	Interest	Familiarity	#Exp. perspectives
Introduction of daylight saving time	3.77 (1.50)	3.63 (1.44)	5
Introduction of carbon tax	3.77 (1.60)	2.34 (1.37)	5
Charging for ambulance	4.81 (1.23)	2.74 (1.34)	5
Making doggy bags available at restaurants	2.95 (1.64)	1.72 (1.14)	4
Restrictions on whale fishing	3.88 (1.50)	3.22 (1.40)	6
Sales promotion of genome-edited food	3.59 (1.65)	2.05 (1.17)	4
Expanding acceptance of foreign workers	4.73 (1.27)	3.30 (1.38)	6
Restrictions on fake news	4.58 (1.37)	3.37 (1.30)	4
Introduction of universal basic income system (for practice task)	NA	NA	NA

Interest and familiarity use a seven-point scale (1, not at all; 4, neutral; 7, very much). Numbers in the table indicate the mean and standard deviation (in parentheses). #Exp. perspectives mean the number of expected perspectives for each theme.

positive/negative people can have before users explicitly ask about them.

#### 4.2.1 Suggestive bot

This chatbot suggests examples of positive and negative people for a decision-making theme when the participants ask an *initial question*, an overview of a given theme, and its purpose (Figure 2C). As described in Section 4.3, just after each decision-making task started, we predefined an initial question (query) about an overview of a theme and set it in the query box of the chatbot. When accepting the initial question, the SUGGESTIVE bot generated an answer for the question. The bot then suggested an example of a positive and negative persons at the end of the generated answer using the following sentence:

*“There are many possible opinions and reactions to [THEME]. For example, Some [POSITIVE PERSON] would be positive for one reason or another. However, some [NEGATIVE PERSON] would be against [THEME] for a certain reason”.*

The SUGGESTIVE bot finds an example of positive/negative people for a theme as follows:

1. The bot generates an answer (referred to as *initial answer*) for an initial question about a theme by simply fetching Azure OpenAI API with the initial question.
2. The bot gathers a list of people who might have positive/negative feelings for the *initial answer* using the prompt illustrated in Figure 3A.
3. The bot randomly picks up a positive and negative person.

Before the user study, we cached an initial answer and a list of positive/negative people for each theme in Table 1. During the study, the SUGGESTIVE bot used the cached results for suggestive answer generation so as not to fail due to OpenAI API error.

#### 4.2.2 Suggestive++ bot

The SUGGESTIVE++ bot is an extension of the SUGGESTIVE bot. When providing participants with initial answers with suggestive endings, SUGGESTIVE++ displays links to question what opinions a suggested positive/negative person might have for a given theme (referred to as *suggestive links*). Once the participants click a suggestive link to a positive/negative person, the SUGGESTIVE++ bot displays the person's opinions against a task theme (Figure 4).

As illustrated in Figure 3B, each positive/negative person's opinion is generated via OpenAI's API using the prompt to question what opinions the person might have for the *initial answer* text. Similar to *initial answers*, the SUGGESTIVE++ bot suggests an example of a positive and a negative person at the end of the generated opinions. In addition, the bot lists *suggestive links* to other people's opinions. In other words, once they click a *suggestive link*, the participants could see other *suggestive links*. Similar to the case of *initial answers*, we generated and cached positive/negative people's opinion texts for the task themes before the user study. We expected that the SUGGESTIVE++ bot could encourage the participants to recall more easy-to-query questions than SUGGESTIVE bot.

#### 4.2.3 Expositive bot

In addition to suggesting the existence of positive/negative people, the EXPOSITIVE shows one-line summaries of their opinions in the *initial answers* as supplementary information. Participants using EXPOSITIVE bots can briefly learn the possible perspectives or opinions of a positive and negative person without additional questioning.

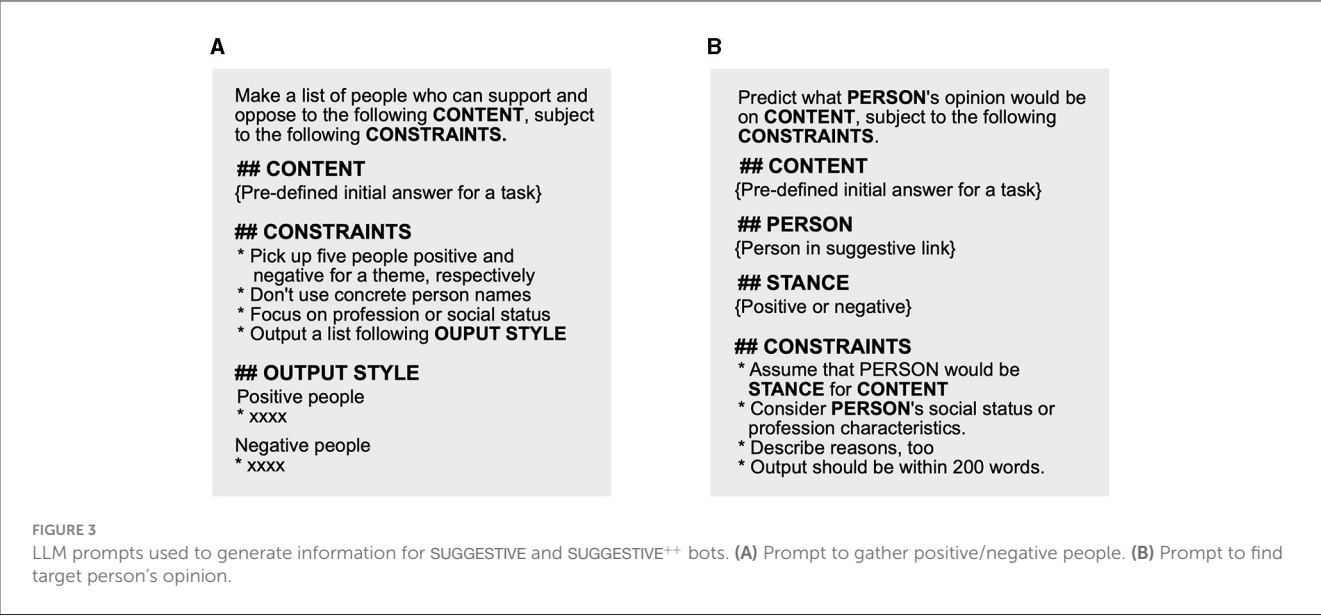
The following is the procedure of *initial answer* generation in the EXPOSITIVE bot:

1. Similar to the SUGGESTIVE bot, the EXPOSITIVE bot generates a plain *initial answer* and a list of positive/negative people for a given theme.
2. Similar to the SUGGESTIVE++ bot, the EXPOSITIVE bot generates opinions for randomly selected positive and negative persons.
3. Each opinion is summarized in a one-line sentence via the GPT API.
4. The EXPOSITIVE bot puts the summarized sentences for a positive and negative person at the end of the *initial answer*.

Note that we cached summarized one-line opinions before the study, similar to *initial answers*.

#### 4.2.4 Plain bot

PLAIN bot is a control UI. Unlike the other three chatbots, this chatbot generates simple answers to participant queries via the GPT API. For *initial questions*, the bot displays the cached *initial answers*.





### 4.3 Procedure

First, the participants were asked to read an overview of our user study and the treatment of their collected data on a crowdsourcing website. After agreeing to a consent form, the participants were transferred to our website to start their participation in the user study. Each participant was then randomly allocated to a UI environment and three decision-making tasks. To ensure that all participants view our system information with the same layout, only PC-based log-ins were allowed (i.e., no participant could access the study if one uses a tablet or a smart phone).

Then, the participants read a description of a task flow and the chatbot used in the study. Assuming that some participants were unfamiliar with chatbot systems for IR, we made the description of our chatbot system as comprehensible as possible. Moreover, we required participants to click a “read next” button every time they read a portion of the description to ensure that they read it completely.

Next, the participants were asked to conduct a practice task to familiarize themselves with their allocated chatbot. In the practice task, the participants were asked to summarize their opinions on introducing a universal basic income system in Japan.

Afterward, the participants performed the three main tasks for the three themes randomly allocated to them from the nine themes listed in Table 1. The main task order was randomized for each participant. In each main task, the participants performed the following three subtasks for each of the allocated themes: (1) pre-questionnaire, (2) decision-making, and (3) post-questionnaire tasks.

In the pre-questionnaire task, the participants ranked their interest and knowledge of each main task theme using a seven-point Likert scale (1, not at all; 7, very much).

Subsequently, the following scenario was presented to each participant ([THEME] is a task theme):

*“Imagine the following case. The introduction of [THEME] has been discussed in your city. After the discussion in the city council, the city decided whether [THEME] is introduced or not, based on the interview with several residents. You are selected as an interviewee and need to explain whether you support [THEME]. Your answer will have a substantial influence on the city policy. So, you are about to collect information about [THEME] for your decision-making by using the latest chatbot system. Collect necessary information with the chatbot. When you reach a satisfactory conclusion, summarize your opinion with reasons and fill it in on an answer form.”*

After reading the scenario, the participants were invited to start the decision-making task by clicking a dedicated button. The browser opened a webpage, where the participants interacted with the allocated chatbot and reported their opinions. At this stage, we set an initial question such as “Let me know the overview of [THEME] and its purpose” in the chatbot query box. Thus, all participants would ask the chatbot the initial question and collect information if necessary. The participants then reported their opinions when they reached their conclusion.

In the post-questionnaire task, we surveyed how many times our chatbot failed to generate answers during the decision-making task. These situations occurred because the chatbot occasionally failed to fetch the OpenAI GPT API within a limited time. For this survey, we asked the participants the following question: *How many times did you see the message “The query failed. Please enter your question again.”* The participants reported the error frequency on a five-point Likert scale (1, never; 2, once; 3, 2–3 times; 4, 4–5 times; 5, more than five times).

At the end of the three main tasks, we administered an exit questionnaire to obtain feedback regarding the chatbot systems. The participants also answered the daily usage of chatbot systems for IR and demographic questions related to gender, age, and education.

### 4.4 Participants

We recruited 300 participants using Lancers.jp<sup>9</sup>, a Japanese crowdsourcing service. Nevertheless, 18 participants were excluded from the analysis because 1 participant violated an instructional manipulation check (IMC) (Oppenheimer et al., 2009) in the exit questionnaire, 15 participants had more than one chatbot failure case in responding to their queries, and 2 participants completed the tasks without using our chatbot. Thus, only 282 participant responses were analyzed. All participants were Japanese (male = 191; female = 87; others = 4). Most participants were in their 30s and 40s (20s = 5.7%; 30s = 27.3%; 40s = 44.0%; 50s = 17.7%; others = 5.4%). Furthermore, about half of the participants reported that they never used chatbots for IR, such as ChatGPT, Google Gemini, and Bing Copilot<sup>10</sup> (never used = 45.7%; once every several months = 10.3%; once a month = 13.5%; once a week = 16.3%; once every several days = 7.8%; several times a day = 6.3%). Participants were randomly assigned to one of the four UI conditions (PLAIN = 66; EXPOSITIVE = 71; SUGGESTIVE = 78; SUGGESTIVE<sup>++</sup> = 67). They used their PC or Mac to join our online user study. All participants who completed the tasks received 400 Japanese yen (approximately \$2.75). On average, the participants finished all tasks within 48 min (median: 43 min).

### 4.5 Measurements

#### 4.5.1 Task duration

We measured the *task duration*, corresponding to the time spent on a decision-making task per theme. Task duration is often used to examine how much effort users make in learning during the information-seeking process (Câmara et al., 2021). In our study, we defined the *task duration* as the time span from the moment when chatbot interfaces were displayed to the moment when the participants reported their opinions.

During the user study, participants engage in a critical learning activity, requiring them to not only look up unfamiliar topics but also analyze the task theme and summarize their opinions

<sup>9</sup> <https://www.lancers.jp/>

<sup>10</sup> Bing Copilot: <https://www.microsoft.com/en-us/bing>.



from various perspectives. This type of learning is often referred to as *critical learning* (Lee et al., 2015). Within the information retrieval community, researchers often use task/search duration as a measure of critical learning engagement and effort during information-seeking activities (Yamamoto and Yamamoto, 2018; Câmara et al., 2021; Roy et al., 2021). However, studies have shown that people interacting with chatbots, like ChatGPT, tend to spend less time on search tasks compared to conventional web search engines (Xu et al., 2023). Therefore, we consider that task/search duration could be a valuable metric to assess how effectively our chatbot strategy promotes critical learning during conversational searches.

#### 4.5.2 Search frequency

We measured the *search frequency*, corresponding to the number of times the participants issued queries to the chatbots during their decision-making tasks. Similar to task/search duration, this metric is also often used to evaluate how willing people are to learn a topic in the fields of information retrieval and human-computer interaction. The *query issue count* can be regarded as how the participants came up with questions in their minds while interacting with the chatbots for their decision-making. We also measured the recommended queries (i.e., *suggestive links*) that the participants with SUGGESTIVE<sup>++</sup> bot clicked as well as the queries that the participants filled in the chat box by themselves.

#### 4.5.3 Opinion length

We examined how many tokens (terms) are contained in the participants' reported opinions. In the study, we asked the participants to report their opinions with reasons without setting minimum requirements for opinion length. We assumed that the more persuasive opinions the participants were encouraged to write, the longer their opinions would be. Therefore, we calculated the token-based length of participant opinions using *MeCab*, a Japanese morphological analyzer<sup>11</sup>.

#### 4.5.4 Perspective in opinion

We calculated the number of perspectives in the participants' opinions to investigate whether they summarized their opinions from various perspectives. This approach aligns with the concept of *T-Depth*, a metric introduced by Wilson and Wilson (2013), designed to evaluate the coverage of subtopics in participant opinions. *T-Depth* has been used in several studies to measure learning outcomes during information-seeking activities (Wilson and Wilson, 2013; Roy et al., 2021). Our indicator is a simplified version of *T-Fact*; it focuses only on the number of distinct perspectives rather than seeing how deeply participants mention each subtopic. This simplification stems from the challenge of objectively evaluating the depth of opinion on subtopics.

The themes listed in Table 1 are popular debating topics in Japan. Therefore, many books and webpages organize and list perspectives for discussion of themes. Our research group members collected and aggregated perspectives for each theme from the Web. Then, they used the list of aggregated perspectives to manually

check which aspect appeared in each participant's opinion. It should be noted that the number of perspectives varied depending on the themes. Therefore, we rescaled the number of appearing perspectives in participant opinions by the expected maximum value (the number of collected perspectives per theme in Table 1).

As Sharma et al. (2024) have shown, conversational searches facilitated by LLMs often lead people to inquire about biased topics, resulting in more selective search behaviors. Therefore, we consider that the number of perspectives is a significant indicator to how effective our chatbot strategy is to promote more diverse information-seeking.

### 4.6 Statistical analyses

To examine the effect of suggestive endings in AI-powered chatbots, we analyzed the collected behavior logs and participant questionnaire responses using an analysis of covariance (ANCOVA). We conducted an ANCOVA to examine the main effect of *UI conditions* on the following measurements: (1) task duration, (2) questioning (search) frequency, (3) token length of opinions, and (4) the number of aspects in opinions. In the ANCOVA, we treated *familiarity*, *interest* of task themes, and *use frequency of chatbot for IR* as covariates to control personal factors. In *post hoc* tests, we conducted the Benjamini–Hochberg false discovery rate (FDR) correction (Benjamini and Hochberg, 2000) to make multiple comparisons between the UI conditions. In the ANCOVA and *post hoc* tests, we conducted log transformation for task duration, questioning frequency, and token length of opinions since the data did not follow Gaussian distributions.

## 5 Results

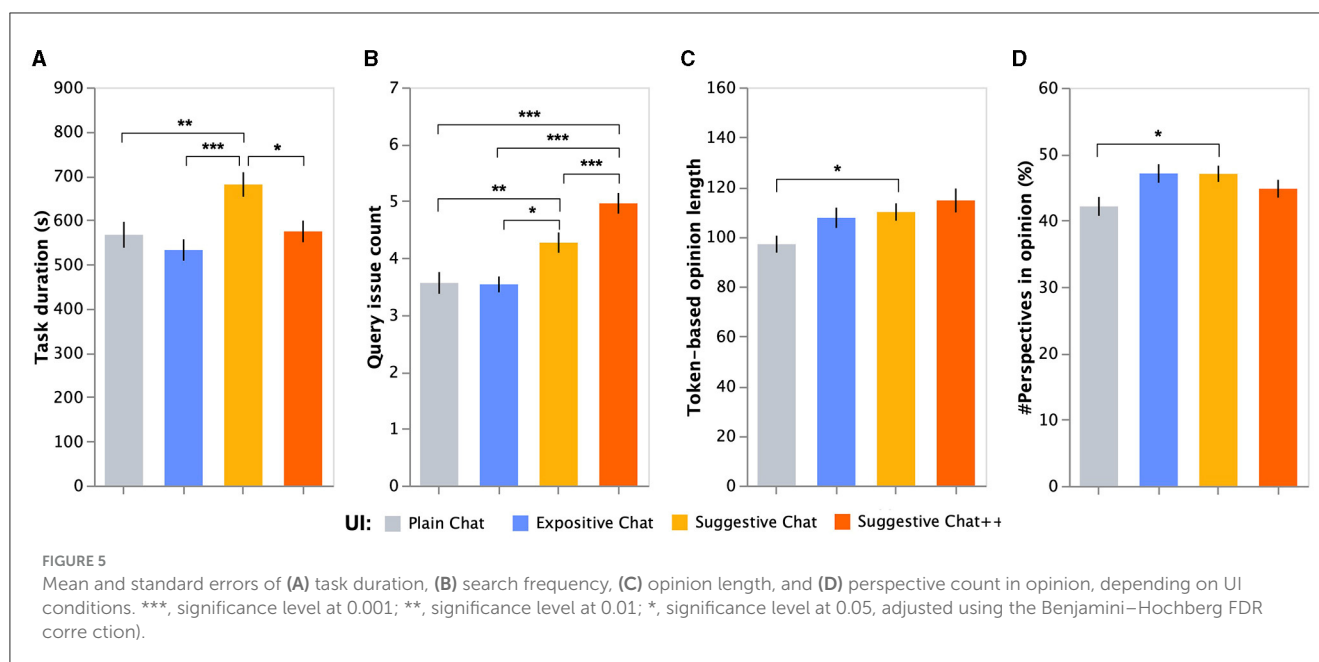
### 5.1 Task duration

We investigated the time and effort invested by participants for the decision-making tasks. Figure 5A illustrates the mean and standard error of the task duration. The ANCOVA result showed a significant impact of the UI conditions on the task duration per task, after controlling individual factors ( $F_{(3,839)} = 6.28, p < 0.001$ ). Moreover, we observed a statistically significant difference between interest in themes (a covariate) on task duration ( $F_{(1,839)} = 5.29, p < 0.05$ ).

The *post hoc* tests revealed that the participants using the SUGGESTIVE bot spent 114s longer in decision-making tasks compared to those using the PLAIN bot (mean: 680.3s vs. 566.6s;  $p(\text{suggestive} - \text{plain}) < 0.001$ ). Moreover, *post hoc* tests showed that participants using the SUGGESTIVE<sup>++</sup> and EXPOSITIVE bots completed the tasks significantly faster than those using the SUGGESTIVE bot (mean: 574.3s vs. 532.3s vs. 680.3s;  $p(\text{suggestive} - \text{suggestive}^{++}) < 0.05, p(\text{suggestive} - \text{expositive}) < 0.01$ ). Nonetheless, no significant differences were observed between the SUGGESTIVE<sup>++</sup> and PLAIN bots and between the EXPOSITIVE and PLAIN bots.

In summary, the SUGGESTIVE bot encouraged the participants to conduct a decision-making task more slowly than any other UIs. By contrast, the SUGGESTIVE<sup>++</sup> and EXPOSITIVE bots did not seem to affect the participants' behavior. These findings suggest

<sup>11</sup> MeCab: <https://taku910.github.io/mecab/> (in Japanese).



that the participants using the SUGGESTIVE bot appeared to invest more effort in collecting clues for their decision-making or organizing their opinions compared to those using the PLAIN and SUGGESTIVE<sup>++</sup> bots.

## 5.2 Search frequency

We investigated how frequently the participants asked their chatbots to collect clues for the decision-making tasks. Figure 5B illustrates the mean and standard error of the search frequency. The ANCOVA result showed that the UI conditions had a significant impact on query issue count after controlling individual factors ( $F_{(3,839)} = 17.7, p < 0.001$ ). No statistically significant differences were observed between interest, knowledge of themes, and daily usage of chatbots for IR on query issue count.

The *post hoc* tests revealed that the participants significantly asked more questions to the SUGGESTIVE bot compared to the PLAIN and EXPOSITVE bots (mean: 4.27 vs. 3.56 vs. 3.54;  $p(\text{suggestive} - \text{plain}) < 0.01$ ,  $p(\text{suggestive} - \text{expositive}) < 0.05$ ). Furthermore, we observed that SUGGESTIVE<sup>++</sup> bots encouraged participants to ask the bot more frequently compared with the SUGGESTIVE, EXPOSITVE, and PLAIN bots (mean: 4.96 vs. 4.27 vs. 3.56 vs. 3.54;  $p(\text{suggestive}^{++} - \text{suggestive}) < 0.001$ ,  $p(\text{suggestive}^{++} - \text{expositive}) < 0.001$ ,  $p(\text{suggestive}^{++} - \text{plain}) < 0.001$ ). It should be noted that the participants using the SUGGESTIVE<sup>++</sup> bot queried with suggestive links at 3.55 times per task and queried without the links (querying by themselves) at 2.04 times per task on average. SUGGESTIVE<sup>++</sup> bot enables people to ask the bot questions just using suggestive links, whereas people using SUGGESTIVE bot have to think about questions and type them in the bot by themselves. Therefore, These statistics show that participants using SUGGESTIVE<sup>++</sup> bot were quite willing to use the suggestive links during the tasks. The *post hoc* test results revealed that the EXPOSITVE bot promoted active searches compared to the PLAIN bot.

The above findings suggest that if the answer of the SUGGESTIVE bot ended with a suggestion regarding the existence of positive/negative opinions, participants were willing to ask questions to the bot more than what they would do with the PLAIN and EXPOSITVE bots, which proactively and explicitly describe positive/negative opinions. This tendency could be stronger if the SUGGESTIVE<sup>++</sup> bot displayed links to issue queries for viewing detailed information on positive/negative opinions.

## 5.3 Opinion length

The length of the participant reports submitted as task answers was considered as a metric to examine the decision-making level promoted by the four chatbot types. Figure 5C illustrates the mean and standard error of the token-based opinion length. The ANCOVA result showed that the UI conditions had a significant impact on token-based opinion length after controlling individual factors ( $F_{(3,839)} = 2.80, p < 0.05$ ). We observed that two individual factors (covariates), i.e., interest in themes ( $F_{(1,839)} = 9.21, p < 0.01$ ) and knowledge on themes ( $F_{(1,839)} = 4.62, p < 0.05$ ), significantly affected the opinion length.

The *post hoc* tests revealed that the participants using the SUGGESTIVE bot wrote longer opinions compared to those using the PLAIN bot (mean: 109.9 tokens vs. 97.0 tokens;  $p(\text{suggestive} - \text{plain}) < 0.05$ ). No significant difference was observed between the SUGGESTIVE<sup>++</sup> and PLAIN bots ( $p > 0.05$ ), although the mean opinion length of the SUGGESTIVE<sup>++</sup> bot was higher than that of the SUGGESTIVE bot (mean: 114.5 vs. 109.9). Furthermore, no significant difference was observed between the EXPOSITVE and PLAIN bots.

These results indicate that if the participants found the existence of positive/negative people for themes using SUGGESTIVE bots, they were likely to explain their opinion with more words than those using the PLAIN bot, which just answered given questions straightforwardly. In addition, the results indicate that

the EXPOSITIVE bot did not have a large influence on opinion volume, despite providing richer answers to initial questions than the PLAIN bot.

## 5.4 Perspectives in opinion

We investigated how many possible perspectives appeared in the participants' submitted opinions to examine if they wrote their opinions from various perspectives. Figure 5D illustrates the mean and standard error of the perspective count. The ANCOVA result revealed that the UI conditions had a significant impact on the rescaled number of perspectives in opinion after controlling individual factors ( $F_{(3,838)} = 2.82, p < 0.05$ ). No statistical significance was observed in individual factors (interest in themes, knowledge of themes, daily usage of chatbots for IR).

The *post hoc* tests showed that the participants using the SUGGESTIVE bot referred to significantly more perspectives in their opinions than those using the PLAIN bot (mean: 47.0% vs. 42.1% of possible perspectives;  $p(\text{suggestive} - \text{plain}) < 0.05$ ). No significant difference was observed between the SUGGESTIVE<sup>++</sup> and PLAIN bots ( $p > 0.05$ ), although the participants using the SUGGESTIVE<sup>++</sup> bot did more chatbot searches compared to those using the PLAIN and SUGGESTIVE bots. Moreover, no significant difference was observed between the EXPOSITIVE and PLAIN bots ( $p > 0.05$ ), although the mean opinion length of the EXPOSITIVE bot was higher compared to that of the SUGGESTIVE bot (mean: 47.1% vs 47.0%).

These results indicate that the participants using the SUGGESTIVE bot were likely to summarize their opinions from various viewpoints compared to those using the PLAIN bot. Furthermore, the results indicate that regardless of the richer answers provided by the EXPOSITIVE bot to initial questions compared to the PLAIN bot, the participants did not formulate their opinions from multiple perspectives.

## 5.5 Qualitative analysis

We analyzed the free-form responses in the exit questionnaire to explore the participants' strategies for their decision-making. In the exit questionnaire, the participants were asked to report how they organized and summarized their opinions during the decision-making tasks. Our research group members conducted an open coding (Lewins and Silver, 2014) for the participants' reports to explore the types of participant strategies.

### 5.5.1 Examination from various perspectives

Some participants stated that they made decisions based on various perspectives (e.g., advantages and disadvantages of a given theme). The following comments are from participants who reported that they considered various perspectives (translated from Japanese to English):

(P19 with suggestive bot) "I was careful not to favor one side over the other by making the chatbot present information on both pros and cons. I also verified my prior knowledge, comparing the chatbot responses with my own views."

(P47 with expositive bot) "To write solid opinions, I collected information from two perspectives: pros/cons and positive/negative opinions."

Meanwhile, the following comment is from a participant who was thought not to consider various perspectives:

(P11 with plain bot) "After deciding my stance, either for or against a given theme, I used the chatbot to collect information supporting my stance."

We examined the ratio of participants who clearly commented that they considered various perspectives during the tasks depending on the UI conditions. The ratios were 52.1%, 34.9%, 62.8%, and 62.7% for the participants using the EXPOSITIVE bot, the PLAIN bot, the SUGGESTIVE bot, and the SUGGESTIVE<sup>++</sup> bot, respectively. The  $\chi^2$  tests with the Bonferroni adjustment revealed that the ratios of the SUGGESTIVE and SUGGESTIVE<sup>++</sup> bots were significantly higher than that of the PLAIN bot ( $p(\text{suggestive} - \text{plain}) < 0.05/6$ ;  $p(\text{suggestive}^{++} - \text{plain}) < 0.05/6$ ). These results indicate that if the chatbots implicitly suggested the existence of positive/negative opinions, the participants could be more careful about various perspectives in their decision-making. By contrast, even if the EXPOSITIVE bot complemented a brief summary about a positive and a negative person's opinion to initial answers, the participants did not try to make their decisions from multiple perspectives.

### 5.5.2 How to use chatbots

Different participants used the chatbots for different reasons. Some participants used the chatbots to learn about unknown concepts from the chatbot's answers, as represented by the following comments:

(P26 with expositive bot) "I read the chatbot's answer. Then I queried the chatbot to summarize my answer if I came up with questions."

(P229 with plain bot) "I asked the bot about what I was curious about or did not understand and then summarized my opinion."

Other participants used the bots to collect clues for their decision-making. Some participants thoughtfully considered various perspectives or weighed the pros and cons of the given themes to inform their decisions as follows:

(P59 with suggestive bot) “I compared opinions from both supporters and opponents. Then I organized those opinions closer to my own thinking.”

(P170 with suggestive<sup>++</sup> bot) “I made sure to check both positive and negative opinions before forming my own view. I queried the chatbot about positive opinions and negative opinions by turns.”

Some participants also tried to corroborate their opinions (prior beliefs) with the chatbots to gather supportive data and expected counterarguments such as the following:

(P102 with suggestive<sup>++</sup> bot) “I started by reviewing the provided theme overview and determined my stance. I then searched for supportive reasons and opposing ones and selected persuasive arguments to consolidate my own opinion. If I didn’t find a decisive reason in the first search, I conducted a further, more in-depth survey using the bot.”

(P148 with suggestive bot) “Firstly, I received an overview of the theme and then inquired about the details of opposing opinions. After that, I formulated my arguments, constructing a rebuttal.”

As already described, the participants’ comments in the exit questionnaire indicate that the SUGGESTIVE and SUGGESTIVE<sup>++</sup> bots promoted the participants’ awareness of decision-making from both positive and negative perspectives on the given themes. The following comments indicate that participants thought suggestive links provided by the SUGGESTIVE<sup>++</sup> bot are useful in searching for detailed information regarding positive/negative perspectives efficiently:

(P6 with suggestive<sup>++</sup> bot) “Once I understood the overview, the task was not so difficult. I didn’t come up with new questions by myself, so I relied more on suggestive links rather than manually asking the bot questions.”

(P281 with suggestive<sup>++</sup> bot) “The system allowed me to choose between pro and con opinions (links) on the theme. I used this feature to validate my own views.”

However, the way of using suggestive links depended on the participants. Some participants using the SUGGESTIVE<sup>++</sup> bot implied that they used either links to positive opinions or links to negative opinions as follows:

(P87 with suggestive<sup>++</sup> bot) “I used the system to investigate what risks might be involved. If the risks (negative opinions) on a theme were low, I tried to have positive opinions on it.”

(P241 with suggestive<sup>++</sup> bot) “Firstly, I determined whether I was in favor or against the given theme, and then I used the chatbot to search for data supporting my view.”

The SUGGESTIVE bot did not provide suggestive links compared with the SUGGESTIVE<sup>++</sup> bot; however, it prompted the participants to ask about or reflect on positive/negative people’s opinions on the theme:

(P203 with suggestive bot) “In answer to the initial question, the bot showed a description suggesting that I should seek further clarification on positive and negative opinions, so I started by following the suggestion.”

(P280 with suggestive bot) “Firstly, I was curious about what the pros and cons might be, so I searched for those aspects. While considering the pros and cons of the theme, I checked current statistics or data to ensure that I tried to form a fair opinion.”

As the below comment suggests, some participants using the EXPOSITIVE bot would not be willing to ask questions as they would feel that the bot provided sufficient information for their decision-making:

(P277 with expositive bot) “Most information from the bot was usable as-is, so I actively used them.”

### 5.5.3 Complaints

A few participants complained that the chatbot’s answers sometimes seemed wrong or unreliable, thereby hoping that the bots could provide more detailed information and evidence. Moreover, a few participants complained that the chatbot’s information was difficult to read.

## 6 Discussion

After controlling the individual factors, our study results revealed that the SUGGESTIVE and SUGGESTIVE<sup>++</sup> bots significantly influence the participants’ behavior and attitude in their decision-making tasks.

As for **RQ1**, the SUGGESTIVE bot caused the participants to spend the longest time in decision-making tasks among the four UI conditions. Moreover, the SUGGESTIVE bot promoted more frequent information seeking compared to the PLAIN and EXPOSITIVE bots. It also encouraged the participants to write longer texts regarding their opinions compared to those using the PLAIN bot. Therefore, we conclude that the SUGGESTIVE bot can encourage users to put more effort into formulating their opinions and gathering information for decision-making from time and content perspectives.

As for **RQ2**, our qualitative analysis revealed that more participants using the SUGGESTIVE and SUGGESTIVE<sup>++</sup> bots were aware of both the pros/cons perspectives in their decision-making compared to those using the PLAIN bot. Furthermore, our behavior



analysis showed that the participants using the SUGGESTIVE bot were likely to refer to more perspectives in their opinion reports compared to those using the PLAIN bot, whereas the SUGGESTIVE<sup>++</sup> did not indicate such a tendency. We conclude that the SUGGESTIVE bot can encourage users to formulate their decision from various viewpoints.

As for RQ3, the SUGGESTIVE<sup>++</sup> bot, providing links to survey positive/negative people's opinions along with the suggestive answers, promoted more frequent search activities compared to any other UI. In addition, the SUGGESTIVE<sup>++</sup> bot significantly reduced the time cost for the tasks compared to the SUGGESTIVE bot. In the exit questionnaire, 62.7% of participants using the SUGGESTIVE<sup>++</sup> bot reported that they tried to formulate their opinions as objectively as possible from both sides of pros and cons. However, the behavior analysis result showed that the SUGGESTIVE<sup>++</sup> bot did not encourage participants to report long opinions with various perspectives compared to the SUGGESTIVE bot. These results indicate that the SUGGESTIVE<sup>++</sup> bot did not substantially promote critical decision-making activities, although it could improve information-seeking efficiency. We believe that such noncritical behaviors can be attributed to the cognitive bias in information seeking (White, 2013; Azzopardi, 2021), such as *selective exposure* (Liao et al., 2015) and *confirmation bias* (Kahneman, 2011; Suzuki and Yamamoto, 2021). The comments of P87 and P241 in the qualitative analysis suggest the influence of selective exposure and confirmation bias on the users' behaviors. However, it is worth noting that our interpretations above are based only on the submitted task reports and the participants' reflective comments in the exit questionnaire. To ensure whether the SUGGESTIVE<sup>++</sup> bot can promote critical decision-making, a further study of the cognitive process during decision-making tasks with the chatbots should be conducted via laboratory experiments.

As for the EXPOSITIVE bot, we found no significant effects compared to the PLAIN bot. When querying a theme overview at the beginning of the tasks, the participants using the EXPOSITIVE bot saw a brief summary of a positive/negative person's opinions without additional queries. In other words, the bot explicitly complemented short, two-sided information for task themes, although the complemented information is not sufficient to make critical judgments on the task themes. However, as P277 suggested, the EXPOSITIVE bot creates the impression of providing sufficient information. This drives participants to pick up only their favorable information to summarize their opinions. Therefore, even if the participants used the EXPOSITIVE bot, they would not exert much effort toward critical decision-making.

As for RQ4, we confirmed that the knowledge of themes affected time efforts in decision-making tasks, while the interest in themes affected the length of reported opinions. These results indicate that knowledge of and interest in themes could affect the amount of effort in decision-making with AI-powered chatbots.

In the end, we conclude that *suggestive endings*, which hint at something in chatbot interaction, can draw more spontaneous questions from users and encourage them to formulate their opinions from various perspectives rather than provide definitive

answers or predefined questions (such as in the SUGGESTIVE<sup>++</sup> bot).

## 7 Limitations and potential challenges

Our study showed that the suggestive ending strategy in a human–chatbot interaction can be useful in enhancing critical decision-making. However, the study has some limitations and several challenges still exist toward better AI-based decision-making support.

One limitation is an experimental environment. We used a crowdsourcing platform for our user study. Although user studies with crowdsourcing have been more popular, this approach has several concerns, including the demographic biases of crowd workers, the presence of lazy participants, and the control of experimental environments (task times and devices for experiments) (Ross et al., 2010). As a result, our study's participant pool might not accurately reflect the general population, and some participants might not have performed the tasks seriously.

Another limitation is the display timing of suggestive endings. In the study, the SUGGESTIVE bot provided answers with suggestive endings only for the initial questions. Therefore, we need to investigate the effects depending on the timing and context of suggestive ending presentations. Moreover, we relied only on the analysis of participants' behaviors during the main tasks and their comments in the exit questionnaire to understand their strategy for decision-making. Think-aloud protocols and stimulated recalls should be conducted in laboratory experiment settings to understand the cognitive decision-making process with chatbots better.

A possible challenge is the topic on which chatbot hints. In the study, we focused on suggesting who is positive or negative for a theme, aiming to make participants aware of the pros/cons viewpoints and to draw spontaneous questions to foster their understanding of the theme (e.g., “[Occupation name] people can be positive for [THEME] with a certain reason”). However, other factors can affect critical decision-making and information seeking. For example, researchers in information and media literacy have stated that currency, relevance, authority, accuracy, and purpose are important to check for critical judgment on the quality of claims and information (Musgrove et al., 2018). Therefore, the chatbots should determine a focused factor and create associated suggestive endings depending on the context of decision-making. For example, if users are encouraged to explore various information from the currency viewpoints, a possible suggestive ending can be “The above opinion was mainstream in the 2010s, but completely different opinions are prevalent in the 2020s”. A remaining issue is a method to automatically generate effective suggestive endings.

The second challenge is related to the proper use of chat strategies to enhance cognitive activities. In this study, we focused on hinting at something in chatbot answers to draw spontaneous questions from users. However, there can be other ways to draw cognitive efforts toward critical decision-making, such as AI questioning and forcible time setting for thinking (Bućinca et al., 2021). As for the AI-questioning approach, devising what and how to make chatbots ask would enable them to promote



various cognitive activities, such as logical reasoning (Danry et al., 2023) and reflecting on lacking issues of one's view (Okuse and Yamamoto, 2023). Nevertheless, explicit questioning might make users intrusive and uncomfortable depending on the frequency, timing, or user personality. Furthermore, even if chatbots provide questions and suggestive endings for users, some users may have difficulties in finding answers and related information by themselves (Odijk et al., 2015). Therefore, the chatbots for decision-making support should use explicit questioning (instructive intervention), suggestive endings in answers (modest intervention), and detailed explanations, depending on the situation and users' personal factors. Moreover, the chatbots should encourage users to perform Web searches without an overreliance on the bots as necessary so that users can corroborate their opinions and the bot's answers from various sources.

## 8 Conclusion

Although people use generative AIs with LLMs to readily obtain information relevant to their requirements, their overreliance on AIs can cause shortsighted decision-making and weaken cognitive skills. Our proposed SUGGESTIVE chatbot encourages people to have spontaneous questions for critical decision-making on a given theme by ending an answer that hints at potentially interest-triggering points.

The online user study revealed that the SUGGESTIVE bot encouraged participants to exert more effort in developing their opinions and gathering information for decision-making compared with simple chatbots. Moreover, the study showed that the SUGGESTIVE bot encouraged participants to make their decisions from various perspectives. We did not observe such a tendency with the EXPOSITIVE chatbot, which complemented information from a specific perspective. These findings indicate that AI-powered chatbots can better enhance human decision-making with *suggestive endings*, which leave room for questions and discussions rather than definitive explanations to a question (query).

Our proposed method has several challenges for improvement. These include investigation on how to use suggestive endings, questioning, and definitive explanations depending on situations and laboratory studies to understand the cognitive processes during decision-making tasks using our chatbot strategy. However, we believe that *suggestive endings* in chatbot answers constitute a good strategy for AI-powered chatbots to enhance critical information seeking and decision-making.

## References

- Arnold, K. C., Chauncey, K., and Gajos, K. Z. (2020). "Predictive text encourages predictable writing, in *Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI '20* (New York, NY, USA: Association for Computing Machinery), 128–138. doi: 10.1145/3377325.3377523
- Arnold, K. C., Gajos, K. Z., and Kalai, A. T. (2016). "On suggesting phrases vs. predicting words for mobile text composition, in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology, UIST '16* (New York, NY, USA: Association for Computing Machinery), 603–608. doi: 10.1145/2984511.2984584
- Azzopardi, L. (2021). "Cognitive biases in search: a review and reflection of cognitive biases in information retrieval, in *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, CHIIR '21* (New York, NY, USA: Association for Computing Machinery), 27–37. doi: 10.1145/3406522.3446023
- Benjamini, Y., and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* 25, 60–83. doi: 10.2307/1165312
- Brodsky, J. E., Brooks, P. J., Scimeca, D., Todorova, R., Galati, P., Batson, M., et al. (2021). Improving college students' fact-checking strategies through lateral reading instruction in a general education civics course. *Cogn. Res.* 6, 1–18. doi: 10.1186/s41235-021-00291-4

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

## Author contributions

YY: Conceptualization, Formal Analysis, Methodology, Software, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported in part by the Grants-in-Aid for Scientific Research (21H03554, 21H03775, and 22H03905) from the MEXT of Japan and Daiko Foundation.

## Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proc. ACM Hum. Comput. Inter.* 5, 1–21. doi: 10.1145/3449287
- Câmara, A., Roy, N., Maxwell, D., and Hauff, C. (2021). “Searching to learn with instructional scaffolding, in *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, CHIIR '21* (New York, NY, USA: Association for Computing Machinery), 209–218. doi: 10.1145/3406522.3446012
- Caraban, A., Karapanos, E., Gonçalves, D., and Campos, P. (2019). “23 ways to nudge: a review of technology-mediated nudging in human-computer interaction, in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19* (New York, NY, USA: Association for Computing Machinery), 1–15. doi: 10.1145/3290605.3300733
- Carr, N. (2014). *The Glass Cage: How Our Computers are Changing Us*. New York: WW Norton and Company.
- Danry, V., Pataranutaporn, P., Mao, Y., and Maes, P. (2023). “Don’t just tell me, ask me: Ai systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations, in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23* (New York, NY, USA: Association for Computing Machinery), 1–13. doi: 10.1145/3544548.3580672
- Ennis, R. H. (1987). “A taxonomy of critical thinking dispositions and abilities, in *Series of books in psychology. Teaching thinking skills: Theory and practice*, eds. J. B. Baron, and R. J. Sternberg (New York: W H Freeman/Times Books/ Henry Holt and Co.), 9–26.
- Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* 20, 1–81.
- Goddard, K., Roudsari, A., and Wyatt, J. C. (2011). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* 19, 121–127. doi: 10.1136/amiajnl-2011-000089
- Harvey, M., Hauff, C., and Elswiler, D. (2015). “Learning by example: training users with high-quality query suggestions, in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15* (New York, NY, USA: Association for Computing Machinery), 133–142. doi: 10.1145/2766462.2767731
- Hertwig, R., and Grne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspect. Psychol. Sci.* 12, 973–986. doi: 10.1177/1745691617702496
- Hettiachchi, D., Kostakos, V., and Goncalves, J. (2022). A survey on task assignment in crowdsourcing. *ACM Comput. Surv.* 55, 1–35. doi: 10.1145/3494522
- Ihoriya, H., Suzuki, M., and Yamamoto, Y. (2022). “Mitigating position bias in review search results with aspect indicator for loss aversion, in *Proceedings of the 2022 International Conference on Human-Computer Interaction, HCI '22* (Berlin, Heidelberg: Springer-Verlag), 17–32. doi: 10.1007/978-3-031-06509-5\_2
- Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., and Naaman, M. (2023). “Co-writing with opinionated language models affects users’ views, in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23* (New York, NY, USA: Association for Computing Machinery), 1–11. doi: 10.1145/3544548.3581196
- Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: NY: Macmillan.
- Kasneji, E., Sessler, K., Kchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). Chatgpt for good? On opportunities and challenges of large language models for education. *Learning Individ. Differ.* 103:102274. doi: 10.1016/j.lindif.2023.102274
- Kim, B., Khanna, R., and Koyejo, O. (2016). “Examples are not enough, learn to criticize! criticism for interpretability, in *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS '16* (Red Hook, NY, USA: Curran Associates Inc.), 2288–2296.
- King, A. (1992). Facilitating elaborative learning through guided student-generated questioning. *Educ. Psychol.* 27, 111–126. doi: 10.1207/s15326985ep2701\_8
- Kittur, A., Chi, E. H., and Suh, B. (2008). “Crowdsourcing user studies with mechanical turk, in *Proceedings of the 2008 ACM Conference on Human Factors in Computing Systems, CHI '08* (New York, NY, USA: Association for Computing Machinery), 453–456. doi: 10.1145/1357054.1357127
- Komarov, S., Reinecke, K., and Gajos, K. Z. (2013). “Crowdsourcing performance evaluations of user interfaces, in *Proceedings of the AAAI/ACM Conference on Human Factors in Computing Systems, CHI '13* (New York, NY, USA: Association for Computing Machinery), 207–216. doi: 10.1145/2470654.2470684
- Lakkaraju, H., and Bastani, O. (2020). “How do i fool you?: Manipulating user trust via misleading black box explanations, in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES '20* (New York, NY, USA: Association for Computing Machinery), 79–85. doi: 10.1145/3375627.3375833
- Lee, H.-J., Lee, J., Makara, K. A., Fishman, B. J., and Hong, Y.-I. (2015). Does higher education foster critical and creative learners? An exploration of two universities in South Korea and the USA. *Higher Educ. Res. Dev.* 34, 131–146. doi: 10.1080/07294360.2014.892477
- Lewins, A., and Silver, C. (2014). *Using Software in Qualitative Research: A Step-By-Step Guide*. London: SAGE publications Ltd.
- Liao, Q. V., and Fu, W.-T. (2014). “Can you hear me now? mitigating the echo chamber effect by source position indicators, in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '14* (New York, NY, USA: Association for Computing Machinery), 184–196. doi: 10.1145/2531602.2531711
- Liao, Q. V., Fu, W.-T., and Mamidi, S. S. (2015). “It is all about perspective: An exploration of mitigating selective exposure with aspect indicators, in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15* (New York, NY, USA: Association for Computing Machinery), 1439–1448. doi: 10.1145/2702123.2702570
- Liu, M. X., Kittur, A., and Myers, B. A. (2022). “Crystalline: Lowering the cost for developers to collect and organize information for decision making, in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22* (New York, NY, USA: Association for Computing Machinery), 1–16. doi: 10.1145/3491102.3501968
- Lundberg, S. M., and Lee, S.-I. (2017). “A unified approach to interpreting model predictions, in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS '17* (Red Hook, NY, USA: Curran Associates Inc), 4768–4777.
- Lutz, J. (2016). The validity of crowdsourcing data in studying anger and aggressive behavior. *Soc. Psychol.* 47, 38–51. doi: 10.1027/1864-9335/a000256
- Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). “On faithfulness and factuality in abstractive summarization, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL '20* (Association for Computational Linguistics), 1906–1919. doi: 10.18653/v1/2020.acl-main.173
- Meola, M. (2004). Chucking the checklist: a contextual approach to teaching undergraduates web-site evaluation. *Portal* 4, 331–344. doi: 10.1353/pla.2004.0055
- Metzler, D., Tay, Y., Bahri, D., and Najork, M. (2021). Rethinking search: Making domain experts out of dilettantes. *SIGIR Forum* 55, 1–27. doi: 10.1145/3476415.3476428
- Musgrove, A. T., Powers, J. R., Rebar, L. C., and Musgrove, G. J. (2018). Real or fake? Resources for teaching college students how to identify fake news. *College Undergr. Libr.* 25, 243–260. doi: 10.1080/10691316.2018.1480444
- Najork, M. (2023). “Generative information retrieval, in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23* (New York, NY, USA: Association for Computing Machinery). doi: 10.1145/3539618.3591871
- Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., et al. (2022). Webgpt: Browser-assisted question-answering with human feedback. arXiv:2112.09332
- Noyes, J. (2007). “Automation and decision making, in *Decision Making in Complex Environments* (New York: CRC Press), 73–82. doi: 10.1201/9781315576138-7
- Odiik, D., White, R. W., Hassan Awadallah, A., and Dumais, S. T. (2015). “Struggling and success in web search, in *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM '15* (New York, NY, USA: Association for Computing Machinery), 1551–1560. doi: 10.1145/2806416.2806488
- Okuse, Y., and Yamamoto, Y. (2023). “Chatbot to facilitate opinion formation in web search, in *Proceedings of the 25th HCI International Conference, HCI '23* (Berlin, Heidelberg: Springer-Verlag), 568–582. doi: 10.1007/978-3-031-35132-7\_43
- Oppenheimer, D. M., Meyvis, T., and Davidenko, N. (2009). Instructional manipulation checks: detecting satisficing to increase statistical power. *J. Exper. Soc. Psychol.* 45, 867–872. doi: 10.1016/j.jesp.2009.03.009
- Peer, E., Brandimarte, L., Samat, S., and Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *J. Exper. Soc. Psychol.* 70, 153–163. doi: 10.1016/j.jesp.2017.01.006
- Petridis, S., Diakopoulos, N., Crowston, K., Hansen, M., Henderson, K., Jastrzebski, S., et al. (2023). “Anglekindling: supporting journalistic angle ideation with large language models, in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23* (New York, NY, USA: Association for Computing Machinery), 1–16. doi: 10.1145/3544548.3580907
- Petty, R. E., and Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Adv. Exper. Soc. Psychol.* 19, 123–205. doi: 10.1016/S0065-2601(08)60214-2
- Rinott, R., Dankin, L., Alzate Perez, C., Khapra, M. M., Aharoni, E., and Slonim, N. (2015). “Show me your evidence - an automatic method for context dependent evidence detection, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon, Portugal: Association for Computational Linguistics), 440–450. doi: 10.18653/v1/D15-1050
- Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., and Tomlinson, B. (2010). “Who are the crowdworkers? shifting demographics in mechanical turk, in *Proceedings of the 2010 ACM Conference on Human Factors in Computing Systems, CHI EA '10* (New York, NY, USA: Association for Computing Machinery), 2863–2872. doi: 10.1145/1753846.1753873
- Roy, N., Torre, M. V., Gadiraju, U., Maxwell, D., and Hauff, C. (2021). “Note the highlight: Incorporating active reading tools in a search as learning environment, in *Proceedings of the 2021 ACM Conference on Human Information Interaction and Retrieval, CHIIR '21* (New York, NY, USA: Association for Computing Machinery), 229–238. doi: 10.1145/3406522.3446025

- Saito, F., Shoji, Y., and Yamamoto, Y. (2020). "Highlighting weasel sentences for promoting critical information seeking on the web, in *Proceedings of the 21st International Conference on Web Information Systems Engineering, WISE '20* (Berlin, Heidelberg: Springer-Verlag), 424–440. doi: 10.1007/978-3-030-34223-4\_27
- Sharma, N., Liao, Q. V., and Xiao, Z. (2024). Generative echo chamber? Effects of llm-powered search systems on diverse information seeking. arXiv:2402.05880
- Shimizu, Y., Ohki, T., and Yamamoto, Y. (2022). "Privacy-aware snippets: Enhancing assessment of balance between privacy risks and benefits in web search, in *Proceedings of the 2022 ACM Conference on Information Technology for Social Good, GoodIT '22* (New York, NY, USA: Association for Computing Machinery), 24–31. doi: 10.1145/3524458.3547231
- Sun, J., Liao, Q. V., Muller, M., Agarwal, M., Houde, S., Talamadupula, K., et al. (2022). "Investigating explainability of generative ai for code through scenario-based design, in *27th International Conference on Intelligent User Interfaces, IUI '22* (New York, NY, USA: Association for Computing Machinery), 212–228. doi: 10.1145/3490099.3511119
- Suzuki, M., and Yamamoto, Y. (2021). Characterizing the influence of confirmation bias on web search behavior. *Front. Psychol.* 12, 1–11. doi: 10.3389/fpsyg.2021.771948
- Suzuki, M., and Yamamoto, Y. (2022). "Don't judge by looks: Search user interface to make searchers reflect on their relevance criteria and promote content-quality-oriented web searches, in *Proceedings of the 2022 ACM Conference on Information Technology for Social Good, GoodIT '22* (New York, NY, USA: Association for Computing Machinery), 1–8. doi: 10.1145/3524458.3547222
- Tay, Y., Tran, V. Q., Dehghani, M., Ni, J., Bahri, D., Mehta, H., et al. (2022). Transformer memory as a differentiable search index, in *Advances in Neural Information Processing Systems, NeurIPS'22*, 21831–21843.
- Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., et al. (2023). Gemini: A family of highly capable multimodal models. arXiv:2312.11805
- Thaler, R. H., and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. London: Penguin.
- Wambsganss, T., Kueng, T., Soellner, M., and Leimeister, J. M. (2021). "Arguetutor: an adaptive dialog-based learning system for argumentation skills, in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21* (New York, NY, USA: Association for Computing Machinery), 1–13. doi: 10.1145/3411764.3445781
- Wambsganss, T., Niklaus, C., Cetto, M., Söllner, M., Handschuh, S., and Leimeister, J. M. (2020). "Al: an adaptive learning support system for argumentation skills, in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20* (New York, NY, USA: Association for Computing Machinery), 1–14. doi: 10.1145/3313831.3376732
- Wang, T., Yu, P., Tan, X. E., O'Brien, S., Pasunuru, R., Dwivedi-Yu, J., et al. (2023). Shepherd: A critic for language model generation. arXiv:2308.04592
- Wang, Y., Leon, P. G., Scott, K., Chen, X., Acquisti, A., and Cranor, L. F. (2013). "Privacy nudges for social media: an exploratory facebook study, in *Proceedings of the 22nd International Conference on World Wide Web, WWW '13 Companion* (New York, NY, USA: Association for Computing Machinery), 763–770. doi: 10.1145/2487788.2488038
- Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., et al. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- White, R. (2013). "Beliefs and biases in web search, in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13* (New York, NY, USA: Association for Computing Machinery), 3–12. doi: 10.1145/2484028.2484053
- Wilson, M. J., and Wilson, M. L. (2013). A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *J. Am. Soc. Inform. Sci. Technol.* 64, 291–306. doi: 10.1002/asi.22758
- Wineburg, S., and McGrew, S. (2019). Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teach. College Rec.* 121, 1–40. doi: 10.1177/016146811912101102
- Wirz, D. S., Ort, A., Rasch, B., and Fahr, A. (2023). The role of cliffhangers in serial entertainment: An experiment on cliffhangers' effects on enjoyment, arousal, and intention to continue watching. *Psychol. Popular Media* 12, 186–196. doi: 10.1037/ppm0000392
- Xu, R., Feng, Y., and Chen, H. (2023). Chatgpt vs. google: A comparative study of search performance and user experience. arXiv:2307.01135 doi: 10.2139/ssrn.4498671
- Yamamoto, Y., and Yamamoto, T. (2018). "Query priming for promoting critical thinking in web search, in *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR '18* (New York, NY, USA: Association for Computing Machinery), 12–21. doi: 10.1145/3176349.3176377
- Yamamoto, Y., and Yamamoto, T. (2020). "Personalization finder: A search interface for identifying and self-controlling web search personalization, in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, JCDL '20* (New York, NY, USA: Association for Computing Machinery), 37–46. doi: 10.1145/3383583.3398519
- Zimmerman, S., Thorpe, A., Fox, C., and Kruschwitz, U. (2019). "Privacy nudging in search: Investigating potential impacts, in *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR '19* (New York, NY, USA: Association for Computing Machinery), 283–287. doi: 10.1145/3295750.3298952
- Zylowski, T., and Wölfel, M. (2023). "An NLP analysis of ChatGPT's personality simulation capabilities and implications for human-centric explainable AI interfaces, in *Proceedings of the 6th International Conference on Natural Language and Speech Processing, ICNLSP '23* (Association for Computational Linguistics), 168–177.



## OPEN ACCESS

## EDITED BY

Jochen Kuhn,  
Ludwig-Maximilians-Universität München,  
Germany

## REVIEWED BY

Y. P. Tsang,  
Hong Kong Polytechnic University,  
Hong Kong SAR, China  
Layla Hasan,  
University of Technology Malaysia, Malaysia

## \*CORRESPONDENCE

Jinzhe Yan  
✉ yanjz@gachon.ac.kr

RECEIVED 01 February 2024

ACCEPTED 17 April 2024

PUBLISHED 03 May 2024

## CITATION

Yu C, Yan J and Cai N (2024) ChatGPT  
in higher education: factors influencing  
ChatGPT user satisfaction and continued  
use intention.

*Front. Educ.* 9:1354929.

doi: 10.3389/feduc.2024.1354929

## COPYRIGHT

© 2024 Yu, Yan and Cai. This is an  
open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](#). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic  
practice. No use, distribution or reproduction  
is permitted which does not comply with  
these terms.

# ChatGPT in higher education: factors influencing ChatGPT user satisfaction and continued use intention

Chengcheng Yu<sup>1</sup>, Jinzhe Yan<sup>2\*</sup> and Na Cai<sup>1</sup>

<sup>1</sup>Department of Business Administration, Gachon University, Seongnam, Republic of Korea, <sup>2</sup>College of Business, Gachon University, Seongnam, Republic of Korea

Artificial intelligence (AI) chatbots, represented by ChatGPT, have shown significant improvement in natural learning and problem analysis and solving, which could trigger a profound change in the education sector and have a far-reaching impact on educational practices. This study aimed to gain insights into the various impacts on users' experience when using the AI tool ChatGPT in higher education. The study analyzed questionnaire data from 328 college students who used ChatGPT. Employing structural equation modeling, this study examined the technology acceptance model in the higher educational setting to identify factors influencing continued use of ChatGPT. The results show that ChatGPT's compatibility positively affected users' perceived ease of use and that efficiency positively affected perceived usefulness. Furthermore, perceived ease of use and perceived usefulness were identified as core factors affecting users' satisfaction and continued use intentions, with user satisfaction showing a significant positive effect on continued use intention. The findings not only provide new perspectives on human-computer interaction theory in higher education but also further refine the functions of current smart devices. This study will help to promote the continued progress and improvement of educational technology, while also deepening understanding of the relationship between attitude and behavior.

## KEYWORDS

ChatGPT, technology acceptance model, satisfaction, continued use intention, higher education

## 1 Introduction

Alongside the continuous advancement of artificial intelligence (AI), chatbot technology has penetrated all aspects of people's daily lives and is widely used in several industries. The success of AI chatbots in social networking environments has demonstrated their indispensable role in the forthcoming information society. In recent years, with the continuous expansion of chatbot functionality and advances in language processing technology, advanced AI models such as ChatGPT have been developed. Scholars have generally interpreted the birth of ChatGPT as a harbinger of a new era. In the context of the current era, it has been noted that AI is gradually shifting from an algorithm-based intelligence model to a language-based one, proving its great potential to reach the human intelligence standard (Zhou et al., 2023).



ChatGPT was developed by the commercial enterprise OpenAI and is an innovative AI chatbot built on large-scale language models (Schulman et al., 2022; Zou and Huang, 2023). It employs popular natural language processing (NLP) technology, big data analytics, speech recognition technology, etc., and has rich and powerful conversational features. Upon its launch, ChatGPT achieved rapid user growth in just two months, becoming one of the fastest-growing consumer apps and one of the most effective chatbot platforms in history. It uses NLP technology to build the complex information needed during user interactions and provide related services. This system is an autonomous machine learning platform that is trained on deep learning neural networks. By training on large amounts of textual data, it is able to generate advanced writing that looks intelligent on the surface (Zhou et al., 2023). The main difference between ChatGPT and other models that have been introduced is that it can generate responses that are similar to human user input and can understand and answer users' questions in a natural and conversational way. Additionally, it can be personalized for specific domains or applications, demonstrating a model that responds to user input in a conversational context. Furthermore, ChatGPT can provide detailed answers to complex questions, giving users a way to get the information they need quickly and efficiently. Over the past few years, the use of chatbots has shifted away from the early days of "bots" to "intelligent people." ChatGPT's proficiency is largely due to its sizable corpus of data, which allows it to understand subtle differences in language usage and context and respond intelligently to a wide range of queries, including everyday conversations and technical and professional topics (Lund and Wang, 2023; Orrù et al., 2023). Due to its unique features, ChatGPT has been embraced by a large number of software developers, creative writers, and academic researchers, who utilize it to develop computer software, applications, texts, academic papers, and lyrics (Dave et al., 2023; Dwivedi et al., 2023).

ChatGPT is a revolutionary AI conversational bot that has dramatically changed the way natural learning is practiced, as well as the ability to analyze and solve problems (Dwivedi et al., 2023). Therefore, ChatGPT is seen as a practical tool that provides users with real-time feedback and answers to questions and helps them solve their daily problems, thus reducing the burden on users and allowing them to focus on solving higher-level problems. Currently, many researchers are working to develop dialog systems with better interactive experiences. ChatGPT has also generated widespread attention and strong interest in the global educational community (Duong et al., 2023; Tlili et al., 2023). As a catalyst for educational reform, AI tools will optimize learning styles and improve the overall quality of education through more efficient access to knowledge (Ma and Huo, 2023). The integration of ChatGPT into the education sector has generated a great deal of interest and enthusiasm. Nowadays, it has transformed into an innovative and inclusive teaching, learning, and assessment tool that is in line with the direction of modern knowledge change. ChatGPT can be used as a tool to reassess one's approach to learning and enhance learning outcomes. In the higher education field, in particular, ChatGPT shows great development potential, which may trigger profound change in education and have a far-reaching impact on higher educational practices.

Current research focuses on the possible future advantages and disadvantages of ChatGPT in education. For example, although ChatGPT can provide rich teaching resources, enhance student engagement and interactivity, and provide a personalized learning experience, it can also lead to a series of problems, such as potential errors in output, the potential risk of misuse, and problems related to privacy protection. Although there has been increasing academic interest in the use of ChatGPT in higher education, little research has been conducted on the causal variables of college students who use ChatGPT in their work and studies. Therefore, this study aimed to fill the existing research gap and identify the key factors that constitute the main reasons college students consistently choose to use ChatGPT.

Based on diffusion of innovation theory, use and gratification theory, and the technology acceptance model, this study proposes a new research model to explain college students' satisfaction with and intention to continue using ChatGPT. The model asserts that ChatGPT possesses compatibility and efficiency qualities that stimulate users' perceived ease of use and usefulness, thus contributing to college students' satisfaction and continued use intentions. To validate the proposed model, a questionnaire was administered via Prolific.<sup>1</sup> A total of 328 valid responses were collected, providing data to validate the research hypotheses.

This study makes several valuable contributions. Firstly, the study is unique in that it specifically highlights the fact that the recent language model known as ChatGPT has not yet received extensive scholarly attention in the field of higher education. Given that there is relatively little extant research on ChatGPT, especially studies exploring its application and acceptance in the higher educational domain, this study focuses on its innovative nature. Moreover, this study focuses on AI's service to the higher educational field and deepens the literature on human-computer interaction in higher education to promote the continuous progress and improvement of educational technology. ChatGPT has made a valuable contribution to research on technological acceptance and application, while simultaneously providing new insight specifically into the acceptance and application of AI-based language modeling, thereby further expanding the scope of AI research. Secondly, the present study delved into the major concerns of college students who use ChatGPT in terms of human-computer interaction patterns and styles and concluded that their cognitive and affective attitudes are key determinants of their willingness to use ChatGPT. The proposed research model provides an all-encompassing analytical framework for examining the various factors that influence technology acceptance and adoption. Finally, the study incorporated diffusion of innovation theory, use and gratification theory, and the technology acceptance model in relation to ChatGPT, and this study proposes a series of variables that aim to positively enhance user experience and people's willingness to use. This will help to further broaden the technology acceptance model application areas in educational practice, thus promoting human-computer interaction in higher education. Moreover, this study provides new ideas for researchers in related fields to explore ways to improve the quality of the learning process from the perspective of interaction design toward improving learning outcomes. This

<sup>1</sup> <https://app.prolific.com/>



study's results will deepen understanding of the application and use of ChatGPT in higher education and support efficient use of this technology in education.

To provide a comprehensive understanding, the paper is organized as follows: Section 2 encompasses literature reviews and hypothesis development; Section 3 elucidates the method; Section 4 introduces hypothesis validation via structural equation modeling; Section 5 outlines the discussion; Section 6 describes the conclusion.

## 2 Literature review and hypothesis development

### 2.1 ChatGPT in education

ChatGPT is an NLP model developed by the commercial company OpenAI. In short, it is an innovative AI chatbot built on a large-scale language model (Schulman et al., 2022). The system was developed on the basis of generative pre-trained transformer architecture with independent machine learning capabilities, and it was originally designed to meet the specific needs of language generation tasks. ChatGPT differs from other AI models for text-to-image generation in that it represents an innovative AI technology. It employs a computational model that is independent of the human brain to deal with natural language understanding problems. This technology is capable of continuously interacting with users, reacting to user inputs, and providing users with conversational feedback. The system helps users understand and interpret discussion topics by continuously outputting information and delivering results. Its main purpose is to create new ideas or content and elaborate them in real-time dialog. Such dialog improves the understanding and use of NLP, as well as the human intelligence level. One of ChatGPT's distinguishing features is its ability to maintain a consistent conversational style by assuming various roles or identities throughout an interaction, which allows for more realistic and natural conversations rather than irrelevant or unrelated responses. To facilitate this, ChatGPT is trained on a large dataset of conversational text that includes chat logs and various forum and social media posts; it therefore has the ability to generate prompts and questions that resemble human responses (Qadir, 2023). As of ChatGPT's ability to provide users with instant and personalized service, a variety of industries have begun to utilize the tool in their work and studies. ChatGPT has been identified as a valuable asset in a variety of industries including but not limited to poetry writing, academic research and writing, business communications, software development, and testing activities (Tung, 2023).

With global integration and rapid technological advances, the use of AI and NLP technologies in higher education has grown by leaps and bounds, and the way languages are learned, taught, and assessed has undergone a corresponding transformation. The use of educational chatbots is increasing as they offer a cost-effective and efficient way to engage students and provide a personalized learning experience (Benotti et al., 2017; Albayati, 2024; Polyoportis, 2024). As a chatbot powered by generative AI, ChatGPT is rapidly evolving and has the potential to spark an

education revolution. To better meet society's needs, the higher education industry has begun to use a variety of tools to increase classroom efficiency and improve teaching and learning. One of the most important aspects is the use of machines to assist in teaching and learning, and AI is among the most commonly utilized means. Numerous scholars have begun to research the practical applications of ChatGPT in various education-related fields; topics include the practical use of ChatGPT in education, research, and various professional activities (Brown et al., 2020; Emenike and Emenike, 2023) and ChatGPT-aided language learning (Kasneci et al., 2023). In education, AI is defined as a computational system capable of performing manual processes (Popenici and Kerr, 2017). ChatGPT and other NLP models play an integral role in higher education.

### 2.2 ChatGPT's compatibility

Diffusion of innovation theory is considered a landmark theory. It has five main characteristics: relative advantage, compatibility, trialability, observability, and complexity (Rogers and Williams, 1983). Diffusion of innovation theory details how, over time, an idea or product gains strong traction or spreads through a particular demographic or social structure. This is often viewed as a phenomenon known as innovation. To achieve this, strategies that differ from what has been done before are implemented. People may not publicize their ideas if they feel that innovation can be achieved in other ways. This approach to innovation does not show a uniform pattern across all participants; there may be differences depending on the duration of acceptance of an innovation as well as adopters' unique characteristics (Rogers and Williams, 1983).

The compatibility component of diffusion of innovation theory is defined as the degree to which people perceive innovation to be compatible with the way they think, behave, and live (Rogers and Williams, 1983). Compatibility constitutes a key element of innovation that directly affects the ability to meet users' needs. ChatGPT's compatibility performance is based on its capabilities in language comprehension, dialog generation, and personalized replies, and continuous improvement is required to better meet users' needs. ChatGPT's compatibility features have made it a relatively versatile and flexible NLP model. Users can typically customize and fine-tune the chatbot according to their needs, and the tool can also be applied to many different scenarios and tasks to meet specific business needs. In previous research, higher education student users identified compatibility as an important factor affecting ChatGPT use intentions (Raman et al., 2023).

Research has shown that a high degree of compatibility leads to a greater propensity to use a given system, as well as to a superior user experience since users exert less effort; it has been found that this may influence actual application through behavioral intent (Wu and Wang, 2005). Moreover, it is known that perceived ease of use depends greatly on compatibility (Akturan and Tezcan, 2012). ChatGPT's compatibility features allow it to adapt to continuous changes in user needs and different conversational environments, user habits, and preferences. Based on user feedback and habits, ChatGPT can provide a personalized experience that is user-friendly and highly compatible with users' habits. Additionally, ChatGPT is highly flexible and versatile to meet the specific needs

of a variety of users, such as those of different ages, cultural backgrounds, or areas of specialization; this further enhances its ease of use in the user community. Therefore, we formulated the following hypothesis:

H1: ChatGPT's compatibility positively impacts ease of use.

## 2.3 ChatGPT's efficiency

In the context of human–computer interaction, use and gratification theory provides a more in-depth view to help understand the reasons for and the ways in which users choose to interface with AI as well as how does AI fulfill these unique demands. Thus, Uses and Gratifications Theory is considered a theoretical framework that aims to research why users choose to use ChatGPT the way they do as well as how it meets those user demands (Baek and Kim, 2023). The researchers applied the Uses and Gratifications theory and the literature on human–computer interaction in the larger ChatGPT environment and proposed five top user motivations when utilizing AI agents: information seeking, task efficiency, personalization, social interaction, and playfulness (Baek and Kim, 2023).

Efficiency, which is a use driver according to use and gratification theory, describes the extent to which a system helps individuals accomplish their duties or goals efficiently and effectively (Park, 2010). With the rapid development of technology and the continuous improvement of information dissemination tools, more people have begun to pay attention to work efficiency and strive to improve their work performance. To perform specific tasks more efficiently, people are likely to choose to use ChatGPT, considering that it features fast responses and can handle multiple conversations at once, demonstrating its ability to process a large amount of information in a very short period of time (Else, 2023). Hence, ChatGPT can enhance productivity through process automation. It is capable of achieving collaborative goals by providing practical information and personalized support in real time to effectively save users time and money and improve production quality.

Efficiency is seen as a key determinant of the relationship between perceived usefulness and technology acceptance, and this contributes to a better understanding of user acceptance of new interactive tools in learning and practice settings (Estriegana et al., 2019). It has been established that a technology's usefulness directly determines task execution effectiveness, that is, whether users accomplish their tasks efficiently and effectively (MacDorman et al., 2011). ChatGPT demonstrates efficiency not only through the provision of accurate information and the precise performance of first-level tasks but also through its high degree of automation and task completion. Its ability to automate specific tasks saves users a significant amount of time and effort, making it a highly effective and convenient tool that can increase users' perceived usefulness. Therefore, we formulated the following hypothesis:

H2: ChatGPT's efficiency positively affects perceived usefulness.

## 2.4 Perceived ease of use and perceived usefulness

Davis (1989) designed the technology acceptance model based on the theory of reasoned action (Fishbein and Ajzen, 1977). The model's main purpose is to parse and predict users' perspectives and intentions to use emerging information technologies, with the core objective of identifying influencing factors related to information system acceptance (Davis, 1989). To better predict and explain users' novel information technology acceptance mechanisms, it is first necessary to gain an in-depth understanding of why users accept or reject technologies, so that users' behavioral habits can be predicted and interpreted more accurately. In this regard, the technology acceptance model, as a simplified version of the theory of reasoned action, focuses on explaining technology acceptance-related behaviors and exploring their influencing factors; many researchers have also used it to study human–machine interactions (Del Giudice et al., 2023). In this view, a person's motivation to and act of embracing a new technology are based on personal benefits that are largely dependent on perceived ease of use and usefulness (Davis, 1989). The purpose of this study was to understand the acceptance and satisfaction of college students when using ChatGPT as an aid to accomplish academic or professional tasks, for example, ChatGPT that is accepted and utilized by students will be more conducive to their academic achievement. Utilizing the technology acceptance model helps to understand how the structure of technology acceptance model relates to students' acceptance of ChatGPT as an assistive tool. If students find ChatGPT easy to use and find it useful, it may positively influence their satisfaction and willingness to continue using it.

According to the technology acceptance model, perceived ease of use is the level at which a person perceives that a technology or information system will be relatively simple and easy to operate, whereas perceived usefulness is the extent to which an individual perceives the performance of a specific set of systems to be effectively improved by using it (Davis, 1989). These two critical elements have been shown to play a vital role not only in the initial acceptance of a technology but also in its continued use. ChatGPT's official statement makes it clear that this is the first time such an efficient tool has been made available to a wide range of users through a free, user-friendly web interface (Roose, 2022). Specifically, extant research has established that a ChatGPT-type system is easy for users in an educational context to learn how to operate (Strzelecki, 2023; Saif et al., 2024). Therefore, ChatGPT attracts people because it is easy to use and provides useful information, continuously creating value for users.

The technology acceptance model is an important measure of acceptance of and satisfaction with ChatGPT among users with different interests (Mathieson, 1991; Baytak, 2023). Empirical research has established the roles of perceived ease of use and usefulness in user satisfaction (Kashive et al., 2020). Additionally, when users believe that they can quickly and easily master new technologies and are convinced that a technology is practical, they usually show a strong technological adoption intention (Kao and Huang, 2023). Therefore, if users believe that ChatGPT is not only easy to operate but also powerful and capable of significantly improving the overall user experience while conserving users' time and effort, it can be expected that users will express a higher level

of satisfaction with ChatGPT and that their continued use intention will increase such that they ultimately become loyal ChatGPT users. Therefore, we formulated the following hypotheses:

H3: ChatGPT's perceived ease of use positively affects satisfaction with ChatGPT.

H4: ChatGPT's perceived ease of use positively affects continued use intention.

H5: ChatGPT's perceived usefulness positively affects satisfaction with ChatGPT.

H6: ChatGPT's perceived usefulness positively affects continued use intention.

## 2.5 Satisfaction with ChatGPT and continued use intention

User satisfaction is a common measure of a system's success, but this core subjective evaluation method is based on a combination of the user's overall system usage experience and its impacts (Urbach and Müller, 2012). Satisfaction is a cumulative feeling resulting from multiple interactions with a system. In the field of technology acceptance models and user behavior research, satisfaction plays a crucial role and significantly influences users' behavioral intentions, which represent users' expected future behavior toward a system (Ashfaq et al., 2020). Multiple studies have pointed to the significant impact of user satisfaction on building and maintaining a long-term loyal user base and have highlighted that it is a key determinant of ongoing usage behavior (Nascimento et al., 2018). Hence, a user's level of satisfaction with ChatGPT will directly determine their willingness to continue using it. For instance, if a user is dissatisfied with a particular mobile application, they may stop using it. However, if users are satisfied with ChatGPT, they will be more likely to use it consistently and continue using the tool in the future. Therefore, we formulated the following hypothesis:

H7: Satisfaction with ChatGPT positively affects continued use intention.

## 3 Materials and methods

### 3.1 Data collection and sample

To validate the research model (see Figure 1) and hypotheses, a questionnaire survey comprising measurement items for the constructs specified in the proposed research model was administered via Prolific<sup>2</sup> to a research population primarily of

college students within the United States. Considering the practical needs of ethical research, all survey respondents were informed of the research objectives and assured that their answers would be kept strictly confidential and used for academic purposes only. A total of 328 valid questionnaires were collected. Respondents comprised 198 males (60.37%) and 130 females (39.63%). The majority had a college or university degree (80.19%). Their predominant ChatGPT usage frequency was two to four times per week (31.4%), followed by once per week (15.5%). Respondents' demographic characteristics are shown in Table 1.

### 3.2 Instrument development

The survey comprised 19 items developed with reference to the existing literature and appropriately adapted to meet the study's contextual needs. Three items on compatibility were adapted from Moore and Benbasat (1991), and four items on efficiency were adapted from Choi and Drumwright (2021) and Baek and Kim (2023). Six items on perceived ease of use and perceived usefulness were adapted from Davis (1989) and Boubker (2024). Satisfaction was measured by adapting three items from Mohammadi (2015) and Boubker (2024), respectively. Three items on continued use intention were adapted from Baek and Kim (2023). Items were rated on a 5-point Likert scale ranging from 1 = *Strongly disagree* to 5 = *Strongly agree*. Table 2 summarizes the measurement items of the study. The scales are shown in Table 3.

### 3.3 Reliability and validity analysis

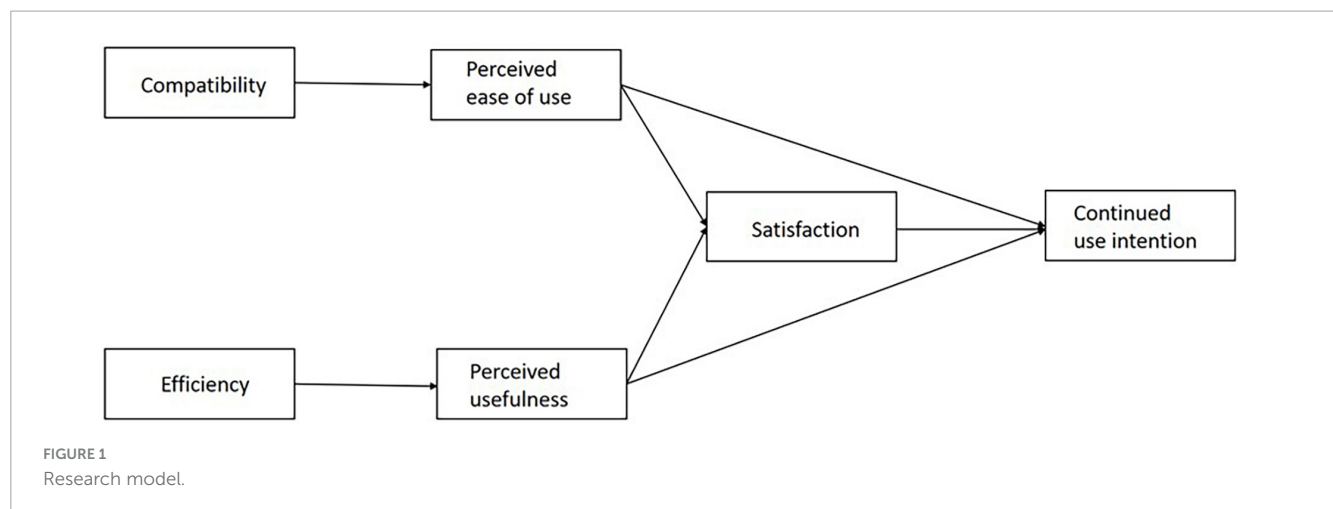
To assess scale reliability, Cronbach's  $\alpha$  coefficient was calculated using SPSS 28.0. According to the test results shown in Table 3, Cronbach's  $\alpha$  coefficients of compatibility, efficiency, perceived ease of use, perceived usefulness, satisfaction, and continued use intention exceeded 0.8, indicating satisfactory scale reliability.

Confirmatory factor analysis was used to test the above indicators' convergence and validity. According to the results shown in Table 3, all variables' standardized factor loadings (SFL) exceeded 0.7, composite reliability (CR) exceeded 0.8, and average variance extracted (AVE) exceeded 0.6, indicating reliable measurements. According to the results shown in Table 4, the square root of the AVE for each variable was greater than the correlation coefficient between the variables, and the data showed good discriminant validity, indicating high questionnaire validity.

### 3.4 Analysis of model fit

According to the results shown in Table 5:  $\chi^2 = 471.514$ ,  $df = 137$ ,  $\chi^2/df = 3.442$ ,  $p < 0.001$ , goodness of fit index (GFI) = 0.862, adjusted goodness of fit index (AGFI) = 0.809, normed fit index (NFI) = 0.916, comparative fit index (CFI) = 0.938, Tucker-Lewis index (TLI) = 0.923, root mean square error of approximation (RMSEA) = 0.086, incremental fit index (IFI) = 0.939, standardized root means square residual (SRMR) = 0.049. The above data indicate alignment of the

<sup>2</sup> <https://app.prolific.com/>



measurement model's fitted metrics with the desired values, suggesting that the hypothesized model and the survey data fit well.

## 4 Hypothesis validation via structural equation modeling

**Table 6** shows the structural modeling test results. Significant path coefficients were found between compatibility and perceived ease of use ( $\beta = 0.606, p < 0.001$ ), efficiency and perceived usefulness ( $\beta = 0.951, p < 0.001$ ), perceived ease of use and satisfaction ( $\beta = 0.236, p < 0.001$ ), perceived ease of use and continued use intention ( $\beta = 0.138, p = 0.003$ ), perceived usefulness and satisfaction ( $\beta = 0.732, p < 0.001$ ), perceived usefulness and continued use intention ( $\beta = 0.628, p < 0.001$ ), and satisfaction and continued use intention ( $\beta = 0.195, p = 0.028$ ). These results provide empirical support for H1–H7.

## 5 Discussion

### 5.1 General discussion

ChatGPT, a language model driven by AI technology, has garnered academic attention. It brings many profound benefits, such as creating academic content and improving its accessibility, enhancing team collaboration, and conducting relevant assessments (Bin-Nashwan et al., 2023). Therefore, emerging AI technologies like ChatGPT have been recognized for their tremendous potential to transform the education industry in an unprecedented way. This study delved into whether ChatGPT's unique nature can enhance user acceptance and individuals' propensity to use the application in higher education. The findings not only provide fresh theoretical support for chatbot research but also have the potential to become a central element in enhancing learning effectiveness and promoting higher educational development.

This study utilized a questionnaire survey designed to gain insight into users' attitudes toward and opinions about using ChatGPT in their daily lives. From a data processing

TABLE 1 Subjects' demographics ( $N = 328$ ).

Demo-graphics	Classification	Frequency	Percentage (%)
Gender	Male	198	60.37
	Female	130	39.63
	Total	328	100.00
	Some college or vocational school	126	38.42
	Bachelor's degree	137	41.77
	Master's degree	54	16.46
	Doctoral degree	11	3.35
	Total	328	100.00
Usage frequency	Very infrequently	36	10.98
	Once a month	47	14.33
	Once every other week	24	7.32
	Once a week	51	15.55
	2–4 times a week	103	31.40
	Once a day	21	6.40
	Several times a day	46	14.02
	Total	328	100.00

perspective, ChatGPT is considered to be highly compatible, efficient, easy to use, and useful due to its excellent interactive performance. These unique characteristics not only enhance college students' satisfaction with this technology but also inspire a strong desire for continued use of the application. The study results reflect all the pre-determined drivers of college students' ChatGPT usage. Most survey respondents reported believing that ChatGPT provides them with rich, relevant, and all-encompassing information that helps them accomplish their academic and professional tasks. Therefore, ChatGPT is seen as a reliable AI alternative with significantly superior performance compared to other search engines (Menon and Shilpa, 2023).



TABLE 2 Measurement items.

Construct	Items	References
Compatibility	Using ChatGPT is compatible with all aspects of my learning.	Moore and Benbasat, 1991
	I think that using ChatGPT fits well with the way I like to learn.	
	Using ChatGPT fits into my learning style.	
Efficiency	I use ChatGPT because it saves me time when I'm completing my tasks.	Choi and Drumwright, 2021; Baek and Kim, 2023
	I use ChatGPT because it makes my tasks easier.	
	I use ChatGPT because it improves my quality output.	
	I use ChatGPT because it is useful for multitasking.	
Perceived ease of use	I find ChatGPT easy to use.	Davis, 1989; Boubker, 2024
	My interaction with ChatGPT is clear and understandable.	
	I find it easy to get ChatGPT to do what I want it to do.	
Perceived usefulness	Using ChatGPT will improve my learning.	Davis, 1989; Boubker, 2024
	Using ChatGPT will enhance my effectiveness.	
	I find ChatGPT to be a useful tool in my learning.	
Satisfaction	I am pleased enough with ChatGPT.	Mohammadi, 2015; Boubker, 2024
	ChatGPT satisfies my educational needs.	
	I am satisfied with ChatGPT's performance.	
Continued use intention	I plan to keep using ChatGPT.	Baek and Kim, 2023
	I want to continue using ChatGPT.	
	I intend to recommend ChatGPT to my friends.	

TABLE 3 Reliability and validity analysis.

Construct	Items	SFL	CR	AVE	$\alpha$
Compatibility	Using ChatGPT is compatible with all aspects of my learning.	0.776	0.912	0.777	0.907
	I think that using ChatGPT fits well with the way I like to learn.	0.949			
	Using ChatGPT fits into my learning style.	0.910			
Efficiency	I use ChatGPT because it saves me time when I'm completing my tasks.	0.820	0.870	0.627	0.865
	I use ChatGPT because it makes my tasks easier.	0.818			
	I use ChatGPT because it improves my quality output.	0.813			
	I use ChatGPT because it is useful for multitasking.	0.712			
Perceived ease of use	I find ChatGPT easy to use.	0.771	0.853	0.660	0.842
	My interaction with ChatGPT is clear and understandable.	0.881			
	I find it easy to get ChatGPT to do what I want it to do.	0.779			
Perceived usefulness	Using ChatGPT will improve my learning.	0.841	0.876	0.703	0.875
	Using ChatGPT will enhance my effectiveness.	0.815			
	I find ChatGPT to be a useful tool in my learning.	0.857			
Satisfaction	I am pleased enough with ChatGPT.	0.876	0.891	0.732	0.886
	ChatGPT satisfies my educational needs.	0.842			
	I am satisfied with ChatGPT's performance.	0.849			
Continued use intention	I plan to keep using ChatGPT.	0.903	0.921	0.796	0.909
	I want to continue using ChatGPT.	0.935			
	I intend to recommend ChatGPT to my friends.	0.835			

Firstly, this study has shown that compatibility positively impacts perceived ease of use and that efficiency positively impacts perceived usefulness. The two core elements of ChatGPT's innovation are compatibility and efficiency, which directly

determine its perceived performance. Therefore, ChatGPT needs to fulfill users' practical needs as much as possible to secure a reputation for being easy to use and useful within its user community. Secondly, the empirical research results have also

TABLE 4 Discriminant validity analysis results.

	Mean	SD	(1)	(2)	(3)	(4)	(5)	(6)
(1) Compatibility	3.826	0.886	0.882					
(2) Efficiency	3.986	0.804	0.700	0.792				
(3) Perceived ease of use	4.189	0.683	0.532	0.54	0.812			
(4) Perceived usefulness	3.980	0.828	0.789	0.774	0.576	0.838		
(5) Satisfaction	2.381	0.468	0.720	0.714	0.604	0.763	0.856	
(6) Continued use intention	4.173	0.856	0.731	0.721	0.584	0.803	0.761	0.892

The diagonal numbers are AVE square root values.

confirmed the two key components of theory of reasoned action, with the data showing significant path coefficients and indicating that ease of use and usefulness are key factors that may influence users' willingness to use ChatGPT. Users' opinions about ChatGPT's ease of use have been shown to be a key factor influencing satisfaction. This suggests that when users find ChatGPT easy to operate, their satisfaction will increase accordingly. Earlier research revealed similar findings regarding the impact of perceived ease of use does positively contribute to satisfaction (Kashive et al., 2020). The study results have also shown that ChatGPT's perceived usefulness significantly affects satisfaction with using the tool. Specifically, positive perceptions of ChatGPT's usefulness may motivate users to use the tool more frequently and develop higher levels of satisfaction with it. This discovery is consistent with prior research that users are more likely in adopting technology and are comfortable with it when they find it valuable (Al-Fraihat et al., 2020). Thus, ChatGPT, as an easy-to-use and practical technological tool, incentivizes technological adoption and acceptance by providing users with innovative problem-solving strategies. Ultimately, a favorable and important effect of satisfaction on continued use intention was found. Moreover, satisfaction has been identified as a key determinant of continued use intention. Additionally, it has been shown that users' satisfaction positively predicts their continued use intentions, and this effect increases with increased satisfaction (Chen et al., 2020). Hence, users who view ChatGPT positively and are satisfied with it are more likely to continue using the tool in the future.

By delegating these tasks to ChatGPT when dealing with everyday academic issues, students can focus more on completing advanced tasks. Some educators have found that ChatGPT's ability to provide immediate feedback and support has a catalytic effect on student engagement and motivation, noting that students who receive timely and accurate answers to their questions are more likely to be supportive and confident in their learning, which in turn leads to better academic performance (Limna et al., 2023). However, some scholars have also expressed concern about the negative effects of students' over-reliance on ChatGPT in the learning process, particularly their ability to understand and answer complex questions. Scholars believe that the information that students are prone to produce in the process of extracting new knowledge can also have some negative impact on their critical thinking and problem-solving skills. Since ChatGPT simplifies the process of acquiring information, this may lead to a lazy attitude toward learning and reduce students' interest in conducting investigations to arrive at personal conclusions or solutions (Kasneci et al., 2023).

TABLE 5 Model fit index.

Index	Criteria	Value	References
GFI	> 0.85	0.862	Jöreskog, 1969; Sarmiento and Costa, 2019
RMSEA	< 0.10	0.086	
CFI	> 0.90	0.938	
NFI	> 0.90	0.916	
TLI	> 0.90	0.923	
SRMR	< 0.10	0.049	

GFI, goodness-of-fit index; NFI, normalized fit index; CFI, comparative fit index; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual.

## 5.2 Research implications

This study makes substantial theoretical contributions. Firstly, the study analyzes the theoretical underpinnings of the acceptance and adoption of ChatGPT in depth in several ways, innovatively integrating diffusion of innovation theory, use and gratification theory, and technology acceptance models into a concise model of ChatGPT. Given the key elements involved in the adoption of this innovative tool, a series of variables designed to enhance user experience and usage intention have been proposed and analyzed in depth. Regarding the use of AI chatbots such as ChatGPT as an educational innovation (Chen, 2010; Ratten and Jones, 2023), the findings provide useful evidence for the link between ChatGPT usage factors and relevant theories. Furthermore, the findings provide valuable insights into how human-machine interaction needs can be met through generative AI, thereby greatly enriching the existing literature. Secondly, building on previous work on the necessary conditions for establishing continued use intentions toward AI technologies (Ashfaq et al., 2020; Jo, 2023), this study has provided an in-depth analysis of the key ChatGPT features that affect user satisfaction and usage intention, such as compatibility, efficiency, perceived usefulness, and perceived ease of use, in a context of modern users' prevalent interaction patterns and styles. Moreover, the study has filled the research gap regarding college students who use ChatGPT in higher educational settings, and it provides valuable insights into the cognitive and affective attitudes of AI chatbots, as represented by ChatGPT, in the higher education domain. Finally, scholarly exploration of the interactive characteristics of ChatGPT will help enhance social interaction. In-depth research on the use of ChatGPT in higher education not only broadens the theoretical basis of educational technology but also

TABLE 6 Results of structural modeling testing of the hypotheses.

	Path	$p$	$\beta$	Results
H1	Compatibility → Perceived ease of use	0.000	0.606	Supported
H2	Efficiency → Perceived usefulness	0.000	0.951	Supported
H3	Perceived ease of use → Satisfaction	0.000	0.236	Supported
H4	Perceived ease of use → Continued use intention	0.003	0.138	Supported
H5	Perceived usefulness → Satisfaction	0.000	0.732	Supported
H6	Perceived usefulness → Continued use intention	0.000	0.628	Supported
H7	Satisfaction → Continued use intention	0.028	0.195	Supported

promotes the development of personalized learning, technology-integrated teaching, and the integration of educational technology, thus providing a solid theoretical foundation and directional guidance for the wide application and further development of educational technology in the future.

This study also has long-term implications for the practical application of AI chatbots, as symbolized by ChatGPT. Firstly, the results have demonstrated that ChatGPT has high compatibility, efficiency, and output quality, features that increase user awareness of the tool's utility and ease of use and stimulate great interest in using the tool, ultimately leading to increased satisfaction. For specialists working on the development and supply of AI technologies, it is vital to ensure that users have a deep understanding of the benefits and possible usage scenarios of the technologies on which applications are based. In other words, when developing new AI tools, it is vital to ensure that the tools are useful. Given the current state of affairs, AI developers should pay more attention to compatibility and efficiency to enhance tools' practical application value, user-friendliness, and user satisfaction. Secondly, the study has shown that ChatGPT's performance is directly related to user satisfaction and users continued use intentions. This raises a key thought for technology developers: human capabilities should not be the sole criterion for evaluating features and functionality; rather, developers should aim to find ways for technologies to exceed human capabilities. Approaching these technologies from a human perspective has its limitations, and more attention should be paid to technologies' overall benefits in terms of meeting users' needs. In conclusion, this study has confirmed the multiple advantages that AI tools can offer and highlighted the fact that AI has already demonstrated its great impact in education. Moreover, the study provides a practical case for personalized learning and adaptive teaching methods, as well as richer learning resources and support for teachers and students. More explicitly, practical application of ChatGPT has broadened the knowledge scope in education. Teachers can enrich their teaching methods and content by using various types of AI, thus creating more opportunities and possibilities for innovation in the education sector. On the flip side, since AI uses machine learning methods that several research organizations have cited, students can experience a richer, higher-quality learning experience.

### 5.3 Limitations and future research

Although this study has profound theoretical and practical implications, it also has some limitations. Firstly, the study

was limited to college students who had interacted with ChatGPT, which resulted in a relatively small sample. A future research goal is to delve deeper into whether these findings differ across groups, such as among those who are less knowledgeable about AI chatbots. Secondly, the majority of the survey respondents were from within the United States. Hence, ChatGPT's regional popularity as well as country and culture differences may have impacted the research results. Therefore, future research could be extended to an international context to validate the proposed model and further expand the breadth of the study. Finally, the study mainly focused on the inherent advantages of ChatGPT while ignoring possible problems such as privacy protection and the risk of misuse. Given that advanced technologies such as ChatGPT are still in their infancy, there is some concern about their future usage trends. Therefore, future research should examine the roles of privacy and risk perception in the application of new technologies in greater depth, and comparative studies should be conducted to assess the strengths and weaknesses of AI software such as ChatGPT.

## 6 Discussion

ChatGPT, as a novel technological innovation, is affecting changes in educational management, most notably in the field of higher education. The purpose of this study is to understand the various influences on the use of ChatGPT, an artificial intelligence tool, in higher education. Using structural equation modeling through a questionnaire approach, this study reveals the model of technology acceptance in a higher education setting and further identifies some of the key factors that influence users continued use of ChatGPT as an educational tool. The findings confirm that ChatGPT's compatibility and efficiency directly determine its perceived capabilities, and that perceived ease of use and usefulness are two key factors influencing users' satisfaction with and willingness to continue using ChatGPT. These findings demonstrate the promising development of ChatGPT in the educational context, emphasizing that when students interact with AI-driven educational tools such as ChatGPT, it will further contribute to the promotion of the integration of education and technology, as well as provide a solid theoretical foundation and directional guidance for the wide application and further development of educational technology in the future.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the patients/participants or patients/participants legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

## Author contributions

CY: Conceptualization, Formal analysis, Methodology, Software, Writing – original draft, Writing – review & editing. JY: Formal analysis, Project administration, Supervision, Writing – original draft, Writing – review & editing. NC: Conceptualization, Software, Writing – review & editing.

## References

- Akturan, U., and Tezcan, N. (2012). Mobile banking adoption of the youth market: Perceptions and intentions. *Market. Intell. Plann.* 30, 444–459.
- Albayati, H. (2024). Investigating undergraduate students' perceptions and awareness of using ChatGPT as a regular assistance tool: A user acceptance perspective study. *Comput. Educ. Artif. Intell.* 6:100203.
- Al-Fraihat, D., Joy, M., and Sinclair, J. (2020). Evaluating E-learning systems success: An empirical study. *Comput. Hum. Behav.* 102, 67–86.
- Ashfaq, M., Yun, J., Yu, S., and Loureiro, S. M. C. (2020). I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telemat. Inf.* 54:101473.
- Back, T. H., and Kim, M. (2023). Is ChatGPT scary good? How user motivations affect creepiness and trust in generative artificial intelligence. *Telemat. Inf.* 83:102030.
- Baytak, A. (2023). The acceptance and diffusion of generative artificial intelligence in education: A literature review. *Curr. Perspect. Educ. Res.* 6, 7–18.
- Benotti, L., Martinez, M. C., and Schapachnik, F. (2017). A tool for introducing computer science with automatic formative assessment. *IEEE Trans. Learn. Technol.* 11, 179–192.
- Bin-Nashwan, S. A., Sadallah, M., and Bouteraa, M. (2023). Use of ChatGPT in academia: Academic integrity hangs in the balance. *Technol. Soc.* 75:102370.
- Boubker, O. (2024). From chatting to self-educating: Can AI tools boost student learning outcomes? *Expert Syst. Appl.* 238:121820.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33, 1877–1901.
- Chen, C.-C., Hsiao, K.-L., and Li, W.-C. (2020). Exploring the determinants of usage continuance willingness for location-based apps: A case study of bicycle-based exercise apps. *J. Retail. Consumer Serv.* 55:102097.
- Chen, H.-J. (2010). Linking employees'e-learning system use to their overall job outcomes: An empirical study based on the IS success model. *Comput. Educ.* 55, 1628–1639.
- Choi, T. R., and Drumwright, M. E. (2021). OK, google, why do i use you?" Motivations, post-consumption evaluations, and perceptions of voice AI assistants. *Telemat. Inf.* 62:101628.
- Dave, T., Athaluri, S. A., and Singh, S. (2023). ChatGPT in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front. Artif. Intell.* 6:1169595. doi: 10.3389/frai.2023.1169595
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 13, 319–340.
- Del Giudice, M., Scuotto, V., Orlando, B., and Mustilli, M. (2023). Toward the human-centered approach. A revised model of individual acceptance of AI. *Hum. Resour. Manag. Rev.* 33:100856.
- Duong, C. D., Vu, T. N., and Ngo, T. V. N. (2023). Applying a modified technology acceptance model to explain higher education students' usage of ChatGPT: A serial multiple mediation model with knowledge sharing as a moderator. *Int. J. Manag. Educ.* 21:100883.
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., et al. (2023). So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int. J. Inf. Manag.* 71:102642.
- Else, H. (2023). Abstracts written by ChatGPT fool scientists. *Nature* 613, 423–423. doi: 10.1038/d41586-023-00056-7
- Emenike, M. E., and Emenike, B. U. (2023). Was this title generated by ChatGPT? Considerations for artificial intelligence text-generation software programs for chemists and chemistry educators. *J. Chem. Educ.* 100, 1413–1418.
- Estriegana, R., Medina-Merodio, J.-A., and Barchino, R. (2019). Student acceptance of virtual laboratory and practical work: An extension of the technology acceptance model. *Comput. Educ.* 135, 1–14.
- Fishbein, M., and Ajzen, I. (1977). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.
- Jo, H. (2023). Understanding AI tool engagement: A study of ChatGPT usage and word-of-mouth among university students and office workers. *Telemat. Inf.* 85:102067.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 34, 183–202.
- Kao, W.-K., and Huang, Y.-S. S. (2023). Service robots in full-and limited-service restaurants: Extending technology acceptance model. *J. Hosp. Tour. Manag.* 54, 10–21.
- Kashive, N., Powale, L., and Kashive, K. (2020). Understanding user perception toward artificial intelligence (AI) enabled e-learning. *Int. J. Inf. Learn. Technol.* 38, 1–19.
- Kasneci, E., Seifler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* 103:102274. doi: 10.1097/ACM.0000000000005636

## Funding

The authors declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



- Limna, P., Kraiwani, T., Jangjarat, K., Klayklung, P., and Chocksathaporn, P. (2023). The use of ChatGPT in the digital era: Perspectives on chatbot implementation. *J. Appl. Learn. Teach.* 6:32.
- Lund, B. D., and Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Library Hi Tech News* 40, 26–29.
- Ma, X., and Huo, Y. (2023). Are users willing to embrace ChatGPT? Exploring the factors on the acceptance of chatbots from the perspective of AIDUA framework. *Technol. Soc.* 75:102362.
- MacDorman, K. F., Whalen, T. J., Ho, C.-C., and Patel, H. (2011). An improved usability measure based on novice and expert performance. *Int. J. Hum. Comput. Interact.* 27, 280–302.
- Mathieson, K. (1991). Predicting user intentions: Comparing the technology acceptance model with the theory of planned behavior. *Inf. Syst. Res.* 2, 173–191.
- Menon, D., and Shilpa, K. (2023). Chatting with ChatGPT™: Analyzing the factors influencing users' intention to Use the Open AI's ChatGPT using the UTAUT model. *Heliyon* 9:e20962. doi: 10.1016/j.heliyon.2023.e20962
- Mohammadi, H. (2015). Social and individual antecedents of m-learning adoption in Iran. *Comput. Hum. Behav.* 49, 191–207.
- Moore, G. C., and Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Inf. Syst. Res.* 2, 192–222.
- Nascimento, B., Oliveira, T., and Tam, C. (2018). Wearable technology: What explains continuance intention in smart watches? *J. Retail. Consumer Serv.* 43, 157–169.
- Orrù, G., Piarulli, A., Conversano, C., and Gemignani, A. (2023). Human-like problem-solving abilities in large language models using ChatGPT. *Front. Artif. Intell.* 6:1199350. doi: 10.3389/frai.2023.1199350
- Park, N. (2010). Adoption and use of computer-based voice over Internet protocol phone service: Toward an integrated model. *J. Commun.* 60, 40–72.
- Polyportis, A. (2024). A longitudinal study on artificial intelligence adoption: Understanding the drivers of ChatGPT usage behavior change in higher education. *Front. Artif. Intell.* 6:1324398. doi: 10.3389/frai.2023.1324398
- Popenici, S. A., and Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Res. Pract. Technol. Enhanc. Learn.* 12, 1–13.
- Qadir, J. (2023). “Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education,” in *Proceedings of the 2023 IEEE global engineering education conference (EDUCON)*, (Kuwait city).
- Raman, R., Mandal, S., Das, P., Kaur, T., Sanjanasri, J. P., and Nedungadi, P. (2023). *University students as early adopters of ChatGPT: Innovation diffusion study*. [Preprint]. Available online at: <https://doi.org/10.21203/rs.3.rs-2734142/v1> (accessed March 27, 2023).
- Ratten, V., and Jones, P. (2023). Generative artificial intelligence (ChatGPT): Implications for management educators. *Int. J. Manag. Educ.* 21:100857.
- Rogers, E. M., and Williams, D. (1983). *Diffusion of innovations*. Glencoe, IL: The Free Press.
- Roose, K. (2022). *The brilliance and weirdness of ChatGPT*. Manhattan, NY: The New York Times.
- Saif, N., Khan, S. U., Shaheen, I., ALotaibi, A., Alnfai, M. M., and Arif, M. (2024). Chat-GPT; Validating technology acceptance model (TAM) in education sector via ubiquitous learning mechanism. *Comput. Hum. Behav.* 154:108097.
- Sarmiento, R. P., and Costa, V. (2019). Confirmatory factor analysis—a case study. *arXiv [Preprint]*. arXiv:1905.05598.
- Schulman, J., Zoph, B., Kim, C., Hilton, J., Menick, J., Weng, J., et al. (2022). ChatGPT: Optimizing language models for dialogue. *OpenAI Blog* 1:9.
- Strzelecki, A. (2023). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *75*, 1–14.
- Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., et al. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learn. Environ.* 10:15.
- Tung, L. (2023). *ChatGPT can write code. Now researchers say it's good at fixing bugs, too*. ZDNet. Available online at: <https://www.zdnet.com/article/chatgpt-can-write-code-now-researchers-say-its-good-at-fixing-bugs-too/> (accessed February 29, 2023).
- Urbach, N., and Müller, B. (2012). The updated DeLone and McLean model of information systems success. *Inf. Syst. Theory* 1, 1–18.
- Wu, J.-H., and Wang, S.-C. (2005). What drives mobile commerce?: An empirical evaluation of the revised technology acceptance model. *Inf. Manag.* 42, 719–729.
- Zhou, J., Ke, P., Qiu, X., Huang, M., and Zhang, J. (2023). ChatGPT: Potential, prospects, and limitations. *Front. Inf. Technol. Electron. Eng.* 25, 6–11. doi: 10.1631/FITEE.2300089
- Zou, M., and Huang, L. (2023). To use or not to use? Understanding doctoral students' acceptance of ChatGPT in writing through technology acceptance model. *Front. Psychol.* 14:1259531. doi: 10.3389/fpsyg.2023.1259531



## OPEN ACCESS

## EDITED BY

Knut Neumann,  
IPN–Leibniz-Institute for Science and  
Mathematics Education, Germany

## REVIEWED BY

Katherine Elkins,  
Kenyon College, United States  
David Gamez,  
Middlesex University, United Kingdom

## \*CORRESPONDENCE

Tyler Malloy  
✉ tylerjmalloy@cmu.edu

RECEIVED 19 February 2024

ACCEPTED 12 April 2024

PUBLISHED 03 May 2024

## CITATION

Malloy T and Gonzalez C (2024) Applying  
Generative Artificial Intelligence to cognitive  
models of decision making.  
*Front. Psychol.* 15:1387948.  
doi: 10.3389/fpsyg.2024.1387948

## COPYRIGHT

© 2024 Malloy and Gonzalez. This is an  
open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](#). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic practice.  
No use, distribution or reproduction is  
permitted which does not comply with these  
terms.

# Applying Generative Artificial Intelligence to cognitive models of decision making

Tyler Malloy\* and Cleotilde Gonzalez

Dynamic Decision Making Laboratory, Department of Social and Decision Sciences, Dietrich College,  
Carnegie Mellon University, Pittsburgh, PA, United States

**Introduction:** Generative Artificial Intelligence has made significant impacts in many fields, including computational cognitive modeling of decision making, although these applications have not yet been theoretically related to each other. This work introduces a categorization of applications of Generative Artificial Intelligence to cognitive models of decision making.

**Methods:** This categorization is used to compare the existing literature and to provide insight into the design of an ablation study to evaluate our proposed model in three experimental paradigms. These experiments used for model comparison involve modeling human learning and decision making based on both visual information and natural language, in tasks that vary in realism and complexity. This comparison of applications takes as its basis Instance-Based Learning Theory, a theory of experiential decision making from which many models have emerged and been applied to a variety of domains and applications.

**Results:** The best performing model from the ablation we performed used a generative model to both create memory representations as well as predict participant actions. The results of this comparison demonstrates the importance of generative models in both forming memories and predicting actions in decision-modeling research.

**Discussion:** In this work, we present a model that integrates generative and cognitive models, using a variety of stimuli, applications, and training methods. These results can provide guidelines for cognitive modelers and decision making researchers interested in integrating Generative AI into their methods.

## KEYWORDS

cognitive modeling, decision making, generative AI, instance based learning, natural language, visual learning

## 1 Introduction

Cognitive models of decision making aim to represent and replicate the cognitive mechanisms driving decisions in various contexts. The motivation for the design and structure of cognitive models is based on various methods; some models focus on the connection to biological processes of the brain, while others aim to emulate more human-like behavior without a biological connection. However, these motivations are not exhaustive or mutually exclusive. In fact, many approaches seek to reconcile these objectives and integrate the various methods. This paper proposes a framework to apply Generative Artificial Intelligence (GAI) research methods to cognitive modeling approaches and evaluates the efficacy of an integrated model to achieve the varied goals of decision modeling research.

Generative Models (GMs) are a category of AI approaches that generate data, often corresponding to the input data type, covering textual, visual, auditory, motor, or multi-modal data (Cao et al., 2023). GMs have shown remarkable advances, in various domains, in the effective generation and representation of complex data, unattainable with conventional methods (Bandi et al., 2023). The large space of research in GAI methods

can be daunting for cognitive modelers interested in applying these techniques to their models for various reasons. The complexity and variety of these approaches are one of the motivations of this work, where we additionally seek to provide insights on the methods for applying GAI to cognitive models of decision making.

Although GMs have shown impressive success in various data modalities relevant to decision science research, there are significant concerns about their utilization (Bommasani et al., 2021). This is due in part to the potential of biases present in language processing and generating models such as Large Language Models (LLMs) (Bender et al., 2021). Various lines of research have suggested close connections to GMs and biological processes in some contexts, such as Variational Autoencoders (VAEs) (Higgins et al., 2021) and Generative Adversarial Networks (GANs) (Gershman, 2019). However, there is a general lack of understanding of how GMs integrate with decision making in a biologically plausible manner. Due to this lack of clarity on the relationship between GMs in decision making and biological realism, careful consideration must be given when choosing integrations with cognitive models aiming at reflecting biological realities.

Previously, the integration of GMs with cognitive models of decision making has been largely done on a case-by-case basis aimed at satisfying the needs of particular learning tasks (Bates and Jacobs, 2020; Malloy et al., 2022a; Xu et al., 2022), for a complete list of these approaches, see the [Supplementary material](#). Consequently, there is an absence of a comprehensive framework for potential methods to integrate GMs and cognitive models of decision making. Understanding the impact of different integration methods is important, especially given the risks associated with improper application of AI technologies, particularly new ones within decision-making systems (Navigli et al., 2023) and the broader social sciences (Bommasani et al., 2021). Thus, elucidating these integration strategies has significant implications for ensuring the responsible and effective deployment of AI in decision-making contexts.

To address the challenges posed by GMs, one approach is to construct an integration of GMs and cognitive models in a way that allows for effective testing of component parts. This research introduces a novel application of GAI research and cognitive modeling of decision making, as well as a categorization of the different features of past integrations. This categorization not only aims at informing the design of future integrations, but also provides a means of comparison between different integration approaches. Based on this framework, we offer an ablation study to compare the integration of GMs into cognitive models. This method enables a thorough analysis of the individual components of these integrations, shedding light on how different integration methods affect behavior.

## 2 Related work

### 2.1 Cognitive architectures and instance-based learning theory

Several Cognitive Architectures (CAs) have been developed and applied to explain and predict reasoning, decision making, and learning in a variety of tasks, including SOAR (Laird et al.,

1987), CLARION (Sun, 2006), and ACT-R (Anderson et al., 1997). Among these, ACT-R has been the basis for many other frameworks and theories that have emerged from the mechanisms it proposes. In particular, Instance-Based Learning Theory (IBLT) is based on an ACT-R mechanism that represents the process of symbolic cognition and emergent reasoning to make predictions from memory and determine human learning and decision making (Gonzalez et al., 2003).

Instance-Based Learning Theory (IBLT) is a cognitive approach that mirrors human decision-making processes by relying on the accumulation and retrieval of examples from memory instead of relying on abstract rules (Gonzalez et al., 2003). IBL models serve as tangible applications of IBLT tailored to specific tasks, encapsulating decision contexts, actions, and rewards pertinent to particular problem domains. These models learn iteratively from previous experiences, store instances of past decisions, and refine the results through feedback from the environment. Subsequently, IBL models leverage this repository of learned instances to navigate novel decision challenges. The adaptive nature of IBL models makes them particularly effective in contexts characterized by variability and uncertainty, as they can adapt flexibly to new situations by drawing parallels with past encounters. In particular, IBL models excel at capturing intricate patterns and relationships inherent in human behavior, a feat often challenging for explicit rule-based representations. Thus, IBLT stands as an intuitive framework to clarify how humans assimilate knowledge from experience and apply it to novel decision-making scenarios (Gonzalez, 2023).

In this research we selected IBLT due to its theoretical connection to the ACT-R cognitive architecture and its wide and general applicability to a multitude of tasks. IBL models have demonstrated fidelity to human decision making processes and have demonstrated their efficacy in various domains, including repeated binary choice tasks (Gonzalez and Dutt, 2011; Lejarraga et al., 2012), sequential decision-making (Bugbee and Gonzalez, 2022), theory of mind applications (Nguyen and Gonzalez, 2022), and practical applications such as identifying phishing emails (Cranford et al., 2019), cyber defense (Cranford et al., 2020), and cyber attack decision-making (Aggarwal et al., 2022).

IBL models make decisions by storing and retrieving instances  $i$  in memory  $\mathcal{M}$ . Instances are stored for each decision made by selecting options  $k$ . Instances are composed of features  $j$  in the set  $\mathcal{F}$  and utility outcomes  $u_i$ . These options are observed in an order represented by the time step  $t$ , and the time steps in which an instance occurred is given  $T(i)$ .

Each instance  $i$  that occurred at time  $t$  has an activation value, which represents the availability of that instance in memory (Anderson and Lebiere, 2014). The activation is a function of the frequency of occurrence of an instance, its memory decay, the similarity between instances in memory and the current instance, and noise. The general similarity of an instance is represented by summing the value  $S_{ij}$  over all attributes, which is the similarity of the attribute  $j$  of instance  $i$  to the current state. This gives the activation equation as:

$$A_i(t) = \ln \left( \sum_{t' \in T_i(t)} (t - t')^{-d} \right) + \mu \sum_{j \in \mathcal{F}} \omega_j (S_{ij} - 1) + \sigma \xi \quad (1)$$

The parameters that are set either by modelers or set to default values are the decay parameter  $d$ ; the mismatch penalty  $\mu$ ; the attribute weight of each  $j$  feature  $\omega_j$ ; and the noise parameter  $\sigma$ . The default values for these parameters are ( $d = 0.5$ ,  $\mu = 1$ ,  $\omega_j = 1$ ,  $\sigma = 0.25$ ), which are based on previous studies on dynamic decision making in humans (Gonzalez and Dutt, 2011; Lejarraga et al., 2012; Gonzalez, 2013; Nguyen et al., 2023).

The probability of retrieval represents the probability that a single instance in memory will be retrieved when estimating the value associated with an option. To calculate this probability of retrieval, IBL models apply a weighted soft-max function to the memory instance activation values  $A_i(t)$  (Equation 1) giving the equation:

$$P_i(t) = \frac{\exp A_i(t)/\tau}{\sum_{i' \in \mathcal{M}_k} \exp A_{i'}(t)/\tau} \quad (2)$$

The parameter that is either set by modelers or set to its default value is the temperature parameter  $\tau$ , which controls the uniformity of the probability distribution defined by this soft-max equation. The default value for this parameter is  $\tau = \sigma\sqrt{2}$ .

The blended value of an option  $k$  is calculated at time step  $t$  according to the utility outcomes  $u_i$  weighted by the probability of retrieval of that instance  $P_i$  (Equation 2) and summing over all instances in memory  $\mathcal{M}_k$  to give the equation:

$$V_k(t) = \sum_{i \in \mathcal{M}_k} P_i(t)u_i \quad (3)$$

IBL models use this Equation (3) to predict the value of options in decision-making tasks. These option blended values are ultimately used to determine the behavior of the IBL model, by selecting from the options currently available the choice with the highest estimated utility. The specific notation for these IBL model equations are described in the python programming package PyIBL (Morrison and Gonzalez, 2024).

## 2.2 Generative Artificial Intelligence

Recent methods in Generative Artificial Intelligence (GAI) have shown impressive success in a variety of domains in the production of natural language (Brown et al., 2020), audio (Kim et al., 2018), motor commands (Ren and Ben-Tzvi, 2020), as well as combinations of these through multi-modal approaches (Achiam et al., 2023). This is done through the training of Generative Models (GMs) which take as input some stimuli, often of the same type as the output, and learn to generate text, audio, and motor commands based on the input and training method. In this work, we focus on the processing of visual and natural language information through the formation of representations achieved by GMs that are useful for cognitive modeling.

Visual GMs form representations of visual information and are originally structured or can be altered to additionally generate utility predictions that are useful for decision-making tasks (Higgins et al., 2017). These utility predictions generated by visual GMs have previously been applied to the prediction of human learning and decision making in contextual bandit tasks (Malloy

et al., 2022a), as well as human transfer of learning (Malloy et al., 2023). Our approach is agnostic to the specific GM being used, which means that it can be applied to comparisons of different visual GMs to compare their performance.

### 2.2.1 Representing data with GMs

The first of two desiderata to integrate GM in cognitive modeling of decision making was to relate models to biological processes in humans and animals. Here, this is understood within the context of representing data with GMs in a manner similar to that represented in biological systems. Recent research on GM-formed data representations has demonstrated close similarities to biological systems (Higgins et al., 2021), motivating their integration into cognitive models that are interested in similarity to biological cognitive systems.

An example of such a GM that is used in this work is the  $\beta$ -Variational Autoencoder ( $\beta$ -VAE) (Higgins et al., 2016, 2017) which learns representations that have been related to biological brain functioning, achieved by comparing the activity of individual neurons in the inferotemporal face patch of Macaque monkeys to learned model representations when trained on images of human faces (Higgins et al., 2021). The format of these representations specifically is defined by a multi-variate Gaussian distribution that is sampled from to form a latent representation, which is fed through the decoder neural network layers to result in a lossy reconstruction of the original stimuli. The training of these models includes a variable information bottleneck controlled by the  $\beta$  parameter. This information-bottleneck motivation of these models has been associated with cognitive limitations that impact decision making in humans, resulting in suboptimal behavior (Bhui et al., 2021; Lai and Gershman, 2021).

These representations have been related to the processing of visual information from humans in learning tasks (Malloy and Sims, 2022), as they excel in retaining key details associated with stimulus generation factors (such as the shape of a ball or the age of a person's face) (Malloy et al., 2022b). Although we employ  $\beta$ -VAEs in this work, there are many alternative visual GMs that are capable of forming representations useful for decision making. This includes visual generation models including Generative Adversarial Networks (GANs) and Visual Transformer (ViT) based models. In our previous work, we performed a comparative analysis of various integrations with an IBL model (Malloy et al., 2023) and demonstrated that each can be effectively integrated with IBL to produce reasonable human-like behavior, but that information-constrained methods like the  $\beta$ -VAE are most accurate.

### 2.2.2 Decision making with GMs

The second of two desiderata to integrate GMs into cognitive models of decision making is generating behavior that is similar to biological systems. This possibility is most salient in cases where GMs are capable of producing complex data, such as text, speech, or motor commands, which alternative models are not equipped to produce. However, in many cases making decisions in specific contexts with pre-trained GMs can be difficult due to the large size and training time of models such as BERT (Kenton and Toutanova,



2019), GPT (Radford et al., 2018), and PaLM (Chowdhery et al., 2023), as these models are not trained to explicitly make decisions.

Many recent approaches have applied GMs and their component structures (such as transformers Chen et al., 2021 or variational autoencoders Higgins et al., 2017), directly to decision making, in machine learning research. In Kirsch et al. (2023), the authors apply transformer models to learn generalizable behavior that can be applied in a variety of reinforcement learning (RL) domains, such as robotics (Brohan et al., 2023), grid-based environments (Li et al., 2022), and video games (Reid et al., 2022).

Other approaches apply feedback to RL models through the use of LLMs (McDonald et al., 2023; Wu et al., 2023), to provide a similar model learning experience as methods such as RL with human feedback (Griffith et al., 2013), without the need to collect human judgements. Offline RL has also been investigated through the integration of LLMs to reduce the need for potentially computationally expensive online learning (Shi et al., 2023). Beyond RL-based methods, some approaches draw some inspiration from cognitive architectures by using a similarity metric to a history of outputs to inform new choices such as the Generative Agents approach (Park et al., 2023).

## 2.3 Integrations of generative models and cognitive models in decision making

Previous research has explored numerous instances of integrating GMs and cognitive models, but these efforts have often been confined to single domains such as language, visual processing, or motor control. Additionally, the integration of GMs and cognitive models has typically been done for a single task or set of closely related tasks, mainly used to address a specific limitation within a cognitive model. These related applications span a diverse range of domains, including prediction of human transfer of learning (Malloy et al., 2023), phishing email detection (Xu et al., 2022), motor control (Taniguchi et al., 2022), auditory learning (Beguś, 2020), and multi-modal learning (Ivanovic et al., 2018).

Integrating GMs and cognitive models can be done in various ways: by replacing an existing functionality, enhancing a sub-module, or introducing a novel ability to the model. For example, LLMs have been proposed as potential knowledge repositories within cognitive models. These repositories can be accessed when relevant knowledge is required (Kirk et al., 2023), similar to a human-generated repository of general knowledge such as ConceptNet (Speer et al., 2017). In particular, ConceptNet has previously been integrated into a cognitive modeling framework for tasks such as answering questions (Huet et al., 2021).

Another recent approach used LLMs to produce highly human-like interactions between agents in a multi-player game involving natural language communication (Park et al., 2023). Although this model did not directly implement cognitive architectures, it did use inspiration from several architectures that were previously applied to multiplayer games like Quakebot-SOAR (Laird, 2001) and ICARUS (Choi et al., 2007). This was done by incorporating a database of encodings of previously observed textual stimuli and then comparing them based on similarity (Park et al., 2023).

Human-like language generation has also been investigated by applying GM techniques (Friston et al., 2020).

Outside the context of language models, some work has provided evidence for connections between human visual information processing and Generative Adversarial Networks (GANs) (Goetschalckx et al., 2021). Another method applied VAEs to modeling working memory formation in a task that required identifying the type of fault in a geological education task (Hedayati et al., 2022). In social science research, GMs have been applied on a range of tasks in replicating and reproducing well-studied phenomena in human social behavior (Aher et al., 2023; Ziemis et al., 2023). In Hedayati et al. (2022), the authors employ a VAE to form representations used by a Binding Pool (BP) model (Swan and Wyble, 2014) to predict the categorization of visual stimuli.

### 2.3.1 Categories of integrating generative models and cognitive models in decision making

Table 1 shows a selection of the most relevant previous approaches to the integration of GM and cognitive models of decision making and learning. A longer version of this analysis of previous methods is included in the [Supplementary material](#), including some of the applications of GMs in decision science or machine learning that did not directly utilize cognitive modeling or did not predict human behavior.

Previous approaches are categorized based on the following features: (1) Generative Actions: whether the GM is used to generate the actions executed by the agent; (2) Generative Memories: Whether the memory representations used by the cognitive model are generated by a GM; (3) Stimuli Type: the types of stimuli the GM is capable of processing; (4) Cognitive Model Type: the type of cognitive model that is used as a base for integration; (5) GM Type: the type of GM that is integrated into the cognitive model; and (6) GM Training: Whether the GM is pre-trained on a large existing corpus, as is done in foundation models, or trained in a tailored manner to solve a specific modeling task.

These features for evaluating existing models are motivated in part by *The Common Model of Cognition* (Laird et al., 2017), which describes the commonalities that cognitive architectures such as SOAR and ACT-R have in terms of their connections of different cognitive faculties. The common model of cognition reviews the history of cognitive model comparisons, based on their method of producing actions, memories, types of perception items, and how these faculties were connected.

Mitsopoulos et al. (2023b) propose an integration of GMs into their “psychologically valid agent” framework, which is rooted in ACT-R and IBLT. This framework has been instrumental in modeling and predicting COVID masking strategies, as demonstrated in their study on this topic (Mitsopoulos et al., 2023a). Another architecture, CogNGen (Ororbia and Kelly, 2023), incorporates MINERVA 2 (Hintzman, 1984) as a short-term memory module while performing other cognitive faculties using both predictive coding (Rao and Ballard, 1999) and neural generative coding (Ororbia and Kifer, 2022). The efficacy of this architecture has been demonstrated in various grid-world tasks (Chevalier-Boisvert et al., 2018), demonstrating improved success in challenging escape-room style environments.

TABLE 1 Comparison of previous applications of integrating GMs into cognitive models based on our proposed categorization.

	Generative actions	Generative memories	Stimuli type	Cognitive model	GM type	GM training
GINGER (proposed)	✓	✓	Textual or Visual	IBL	VAEs, LLMs	Ad-hoc, Pretrain
Mitsopoulos et al. (2023b)	✗	✓	Textual	ACT-R	LLMs	Pretrain
Malloy et al. (2023)	✗	✓	Visual	IBL	VAEs, GANs	Ad-hoc
Malloy et al. (2022a)	✓	✗	Visual	RL	VAEs	Ad-hoc
Xu et al. (2022)	✗	✓	Textual	IBL	LLMs	Pretrain
Hedayati et al. (2022)	✗	✓	Visual	BP	VAEs	Ad-hoc
Higgins et al. (2021)	✓	✗	Visual	RL	VAEs	Ad-hoc
Bates and Jacobs (2020)	✓	✗	Visual	None	VAEs	Ad-hoc

Additional comparisons to models that are not applied to biological cognition are enumerated in an extended table in the [Supplementary material](#). Green check marks indicate that a model or method contains the related generative method. Red x marks indicate that a model or method does not contain the related generative method.

Connecting cognitive models with GMs to produce memory representations of decision making tasks has been explored in [Malloy et al. \(2023\)](#), which compared Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) and Visual Transformers (ViTs), in their ability to integrate with an IBL model. This work was inspired by previous applications of GMs in modeling biological decision making, such as [Higgins et al. \(2021\)](#). Another approach which has incorporated LLMs with instance based learning was presented in [Xu et al. \(2022\)](#), which involved LLM model representations of phishing emails used to predict human decision making in an email categorization task.

### 3 Proposed model

#### 3.1 Generation INformed by Generative Environment Representations (GINGER)

In this work, we propose a method that integrates GMs into both the action and memory generation of a cognitive agent based on IBLT. This integration of GMs and IBL models can process either textual or visual information which is achieved by leveraging Variational Auto-Encoders or Large language Models. The result is a method of Generation INformed by Generative Environment Representations (GINGER).

In [Figure 1](#), we outline a general schematic of our proposed GINGER model. The first step of this process is for the GM model input to be processed by the model. In the experiments used for this work, this includes textual and visual information, but could be applied to others. From this input, the GM produces some model output and representations of the model input that is used as the memory of the GINGER model. This is used by the cognitive model, either as a part or as the whole of the state representation. From these two action prediction methods, the GINGER model produces two action outputs, which are resolved based on the specifics of the environment, such as averaging for utility prediction.

There are two optional connections between GMs and cognitive models that are not investigated in this work and are instead left for future research. The first is the connection from the model output to the action being performed. While the generation of utility

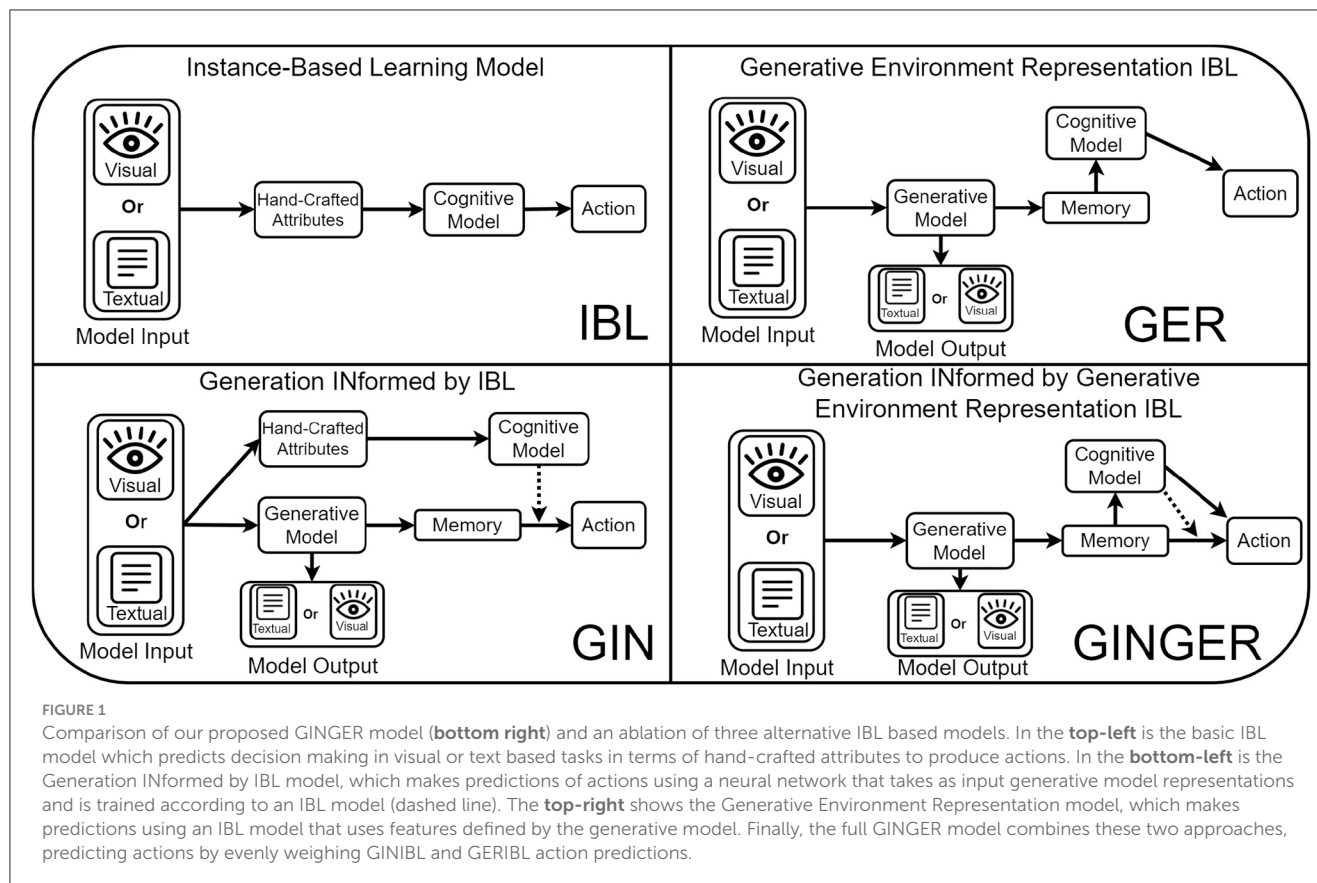
predictions is always informed by cognitive model predictions (by training the actions based on cognitive model predictions), it is also possible to include the GM output (text, motor commands, etc.) as the whole or a part of the action performed. Secondly, the cognitive model and GM can optionally be connected from the cognitive model into the GM input, such as by using predictions from the cognitive model as a part of the GM input (e.g., as the prompt of a LLM) to inform how representations and outputs should be formed.

##### 3.1.1 Generative actions

The first part of the GINGER model name, “generation informed by” refers to the sharing of utility predictions made by the cognitive model when training the utility prediction of the generative model. Action generation is accomplished by directly generating utility predictions that are used in decision making tasks to determine the action with the highest utility based on a specific stimulus. This can be achieved in two different ways depending on whether the GM is a pre-trained foundation model or an ad-hoc trained model for a specific task.

In the case of ad-hoc trained models, the models themselves have been adjusted to generate utility predictions and are trained using the cognitive model. For instance, a  $\beta$ -Variational Autoencoder ( $\beta$ -VAEs) model which typically produces reconstructions of original stimuli can be adjusted to additionally predict utility, as was done in previous methods ([Higgins et al., 2017](#); [Bates and Jacobs, 2020](#); [Lai and Gershman, 2021](#); [Malloy et al., 2022a](#)). Then, instead of training the model to predict actions based on reward observations from the environment, it is trained to match the predictions of the cognitive model.  $\beta$ -VAEs are trained to produce as accurate reconstructions as possible, given the size of the latent representation and its informational complexity, measured by KL-divergence, which is penalized through the  $\beta$  parameter. This means that adjusting the  $\beta$  parameter to individual cognitive abilities can result in more human-like predictions of actions based on model representations ([Bates and Jacobs, 2019, 2020](#); [Malloy et al., 2022b](#)).

In the case of pre-trained or foundation models, the models cannot be easily adjusted after training prior to integration with cognitive models. For that reason, when integrating pre-trained



LLMs or other foundation models, our GINGER approach uses representations learned by these models as input to a separate utility prediction neural network. The structure and precise training of these models is left to the discretion of cognitive modelers according to the demands of the learning task under investigation. In our work, we use a simple 2 layer fully connected network with 64 units to predict the utility associated with these representations. See the [Supplementary material](#) for more details on this training approach.

### 3.1.2 Generative memories

The second part of the GINGER name, “generative environment representations”, refers to the creation of stimuli representations that are created based on the requirements of the learning task to capture the stimuli type of interest. This reliance on representations formed by GMs allows for either total reliance on representations, or adding the representation as an additional feature. When applying these representations to IBL, we determine the similarity  $S_{ij}$  in the calculation of the activation function (see [Equation 1](#)) through an integration of a similarity metric defined by the training of the GM  $\text{Sim}_{\text{GM}}$  as follows:

$$A_i(t) = \ln \left( \sum_{t' \in \mathcal{T}_i(t)} (t - t')^{-d} \right) + \mu \sum_{j \in \mathcal{F}} \omega_j (S_{ij} - 1) + \sigma \xi \quad (4)$$

$$S_{ij} = \text{Sim}_{\text{GM}}(p(z_i|k_i), p(z|k))$$

Formally, GMs process some input  $x$ , which can be visual, textual, auditory, or multi-modal input, and produce some output

$y$  based on that input. During this generation, these models form representations of the input  $z$  that can vary in structure, such as the multivariate Gaussian distributions used by  $\beta$ -VAEs or word vector embedding used by LLMs. In our model, we consider the option or part of the option relevant for modeling  $k$  to be the input to the GM. This allows for the formation of representations  $z$  based on these options. The similarity of options can be instead calculated based on these representations of current options  $p(z|k)$  and representations of options stored in the IBL model memory  $p(z_i|k_i)$ . The similarity of these representations is defined by the training method of the GM, used as a metric of similarity ( $\text{Sim}_{\text{GM}}$ ).

In some GMs such as conversational LLMs, the output  $y$  is trained to match with subsequent textual tokens in a conversation or other language domain. In other types of GMs like Variational Autoencoders the models are trained such that the output  $y$  is as close to the input  $x$  as possible given the information constraint imposed by the model. These two types of models are used in our comparison of different methods of integrating GMs, but alternative GM structures and training methods can also be integrated with our proposed modeling approach.

The generation of internal representations is a requirement in a sense for GMs as they must form some representation  $z$  based on the input  $x$  in order to process it. As with the model output  $y$ , the structure of these internal representations  $z$  varies between different GMs. In the case of LLMs, these internal representations are structured as word vector embeddings. This allows for measures of similarity ( $\text{Sim}_{\text{GM}}$  in [Equation 4](#)) based on cosine similarity, which is conceptually similar to a high-dimensional distance metric. In the case of  $\beta$ -VAEs, these representations take the form of high-dimensional Gaussian distributions which are sampled from

and fed through the subsequent layers of the model to form the reconstructed version of the original stimulus. With these types of representations, it is possible to measure similarity in terms of the KL-divergence of these representations.

In both cases, these GMs provide a meaningful representation of the model input, as well as a method of comparing these representations to other inputs. This is highly relevant for integration with an IBL method since the similarity of instances needs to be calculated to determine a memory activation, which is easily achieved through the use of the existing similarity metric required by the training of the GM itself. The next sections on generative action production and generative memory production will further detail how the representations formed by GMs are used in the IBL cognitive model, as well as how the IBL and GM are integrated in an interdependent manner that affords improvements to both models.

## 4 Model ablation

This work proposes a comparison of different methods of integrating GMs into cognitive models or architectures, through an ablation study comparing the categorizations described in Table 1. To do this, we use the Instance-Based Learning (IBL) model of dynamic decision making (Gonzalez et al., 2003). As opposed to a comparison of our proposed model against a highly similar model that instead is based on a different cognitive model or GM, or has a different method of integrating GMs and cognitive models, we are interested in providing insight to cognitive modelers interested in applying GMs to their own approaches, and as such adopt an ablation analysis of GINGER.

This ablation is based on the two key features of GINGER, the ‘Generative Environment Representations’ which are related to the generation of cognitive model memory representations, and the ‘Generation Informed’ by cognitive models, allowing for the actions selected by GMs to take information from cognitive models. Ablating away the generative environment representations results in a model that only uses generation informed by cognitive models (GIN). Ablating away the generation informed by cognitive models results in a model that only uses generative environment representations (GER). Finally, ablating both away results in the baseline Instance Based Learning (IBL) model which makes predictions using hand-crafted features of tasks.

These four models (GIN, GER, GINGER, and IBL) form the baseline for our ablation comparisons in three experimental contexts involving different types of stimuli and complexities. The following sections detail these experiments as well as comparisons of the performance of the proposed model and the ablated versions. Participant data from these experiments and all trained models, modeling result data, and code to replicate figures is collected into a single OSF repository.<sup>1</sup>

### 4.1 Contextual bandit task

The experiment was originally conducted at the Niv Neuroscience Lab at Princeton University (Niv et al., 2015).

<sup>1</sup> <https://osf.io/m6qc4/>

Participants were presented with three options, each distinguished by a unique combination of shape, color, and texture. Shapes included circular, square, and triangular forms; colors ranged from yellow, red, and green; and textures were dotted, wavy, and hatched (see Figure 2A). In every trial of the task, all of the 9 possible features appeared once within each option, ensuring that there will always be an option of each color, shape and texture. The features within the options were randomized to prevent repetitions in each position (left, middle, right). Participants had 1.5 seconds to make their selection, followed by a brief display (0.5 seconds) of the chosen option and the feedback showing the point reward (0 or 1). Then a blank screen was displayed for 4–7 seconds before the next stimulus.

During a single episode of the task, one of the nine features is selected as the feature of interest, and selecting the option with that feature increases the likelihood of receiving a reward. Episodes lasted approximately 20–25 trials before transitioning to a new feature of interest. The reward in this task is probabilistic, and selecting the feature of interest results in a 75% chance of receiving a reward of 1 and a 25% chance of receiving a reward of 0. When selecting one of the two options without the feature of interest resulted in a 25% chance of receiving a reward of 1 and a 75% chance of observing a reward of 0. Given the three possible options, the base probability of selecting the option with the feature of interest was 1/3.

#### 4.1.1 Cognitive modeling

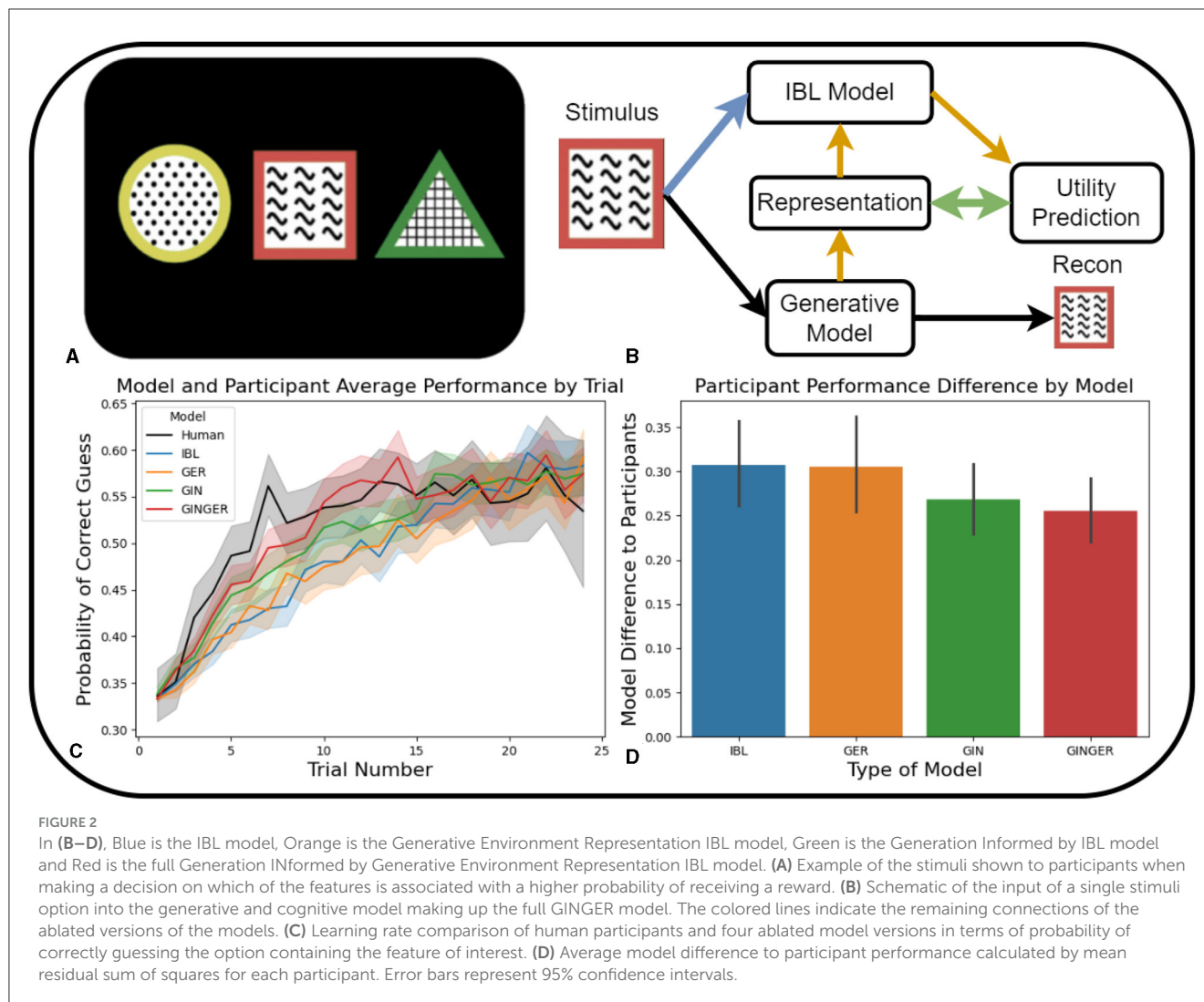
The contextual bandit task serves as a benchmark to compare the three approaches to integrating GMs into cognitive decision-making models. This simple task is useful to ensure that all integrations of GMs in cognitive modeling accurately capture human learning in basic learning scenarios. In Figure 2B, we present a visual representation of the GINGER model, which uses visual stimuli associated with one of the three options as input. First, this stimulus is fed into the GM. In this task, a modified version of a  $\beta$ -Variational Autoencoder is used to further predict the utility associated with stimuli based on the internal representations generated by the GM.

For the baseline IBL model, choice features consisted of shape, color, and texture. For each type of feature, the similarity metric was defined as 1 for identical features and 0 for all other features. The GER model used the  $\beta$ -VAE model representation as an additional feature with a unique similarity metric. The similarity metric of this additional feature was the  $\beta$ -VAE model representation distribution KL-divergence. The GIN model used the baseline IBL model to predict utilities of stimuli options and trained the utility prediction network using these values. The full GINGER model combined these two approaches of the GIN and GER models in this task. All four ablation models used the same predefined parameters for noise, temperature, decay, and as mentioned previously.

#### 4.1.2 Methods

The experimental methodology is reproduced from the original paper; for additional details, see Niv et al. (2015). This study involved 34 participants (20 female, 14 male, 0 non-binary) recruited from Princeton University, all aged 18 or older. Data from 3 participants were incomplete and thus not analyzed, and





another 6 participants were removed due to poor performance. Participants had a mean age of 20.9 years and were compensated at a rate of \$20 per hour. This experiment was approved by the Princeton University Institutional Review Board. The experiment was not preregistered. Participant data is accessible on the Niv Lab website.<sup>2</sup>

To evaluate the performance of the 4 model ablation of our proposed GINGER model, we compare the probability of a correct guess on each trial within an episode. Figure 2C shows the comparison between participant and model performance regarding the probability of selecting the option containing the feature of interest across trials 1–25. This graphical representation facilitates the visual comparison of the learning of which feature is associated with a higher probability of observing a reward, and the average performance at the end of each episode.

In addition to the trial-by-trial comparison of model and participant performance depicted in Figure 2C, our aim is to compare the overall similarity between them. This is done by measuring the difference in model performance with individual

participant performance using the mean residual sum of squares  $RSS/n$  where  $n$  is the number of participants and  $RSS = \sum_{i=1}^n (y_i - p(x_i))^2$ . This difference is calculated for each participant and trial within an episode and across all episodes in the experiment. These values are correlated with the Bayesian Information Criterion (BIC) calculated in terms of the residual sum of squares (RSS) as  $BIC = n \ln(RSS/n) + k \ln(n)$  since all four models have 0 fit parameters (all are default values). The resulting values are averaged across all participants and presented in Figure 2D. Error bars in Figure 2D denote the 95% confidence intervals of the model difference from participant performance across each participant and trial of the task.

#### 4.1.3 Results

The initial comparison of model learning to participant behavior focuses on the probability of correct guesses as the trial number within increases, as shown in Figure 2C. Comparing the speed of learning to participants reveals that models that include the generative action selection (GIN and GINGER) demonstrate the fastest learning. Compared to the two versions of the GINGER

<sup>2</sup> <https://nivlab.princeton.edu/data>

model (IBL and GER) that do not make direct predictions of utility based on GM representations exhibit slower learning rates. This shows that in learning tasks that require fast updating of predicted utilities, directly predicting these values from GM representations and selecting actions accordingly results in more human-like learning progress.

The second set of results illustrated in Figure 2D, compares the average difference in model performance to participants performance. Among the four models compared, the GINGER model has the lowest deviation from participant performance and a performance difference similar to the GIN ablation model, which relies on predictions of utility derived from GM representations. The IBL and GER models, which make predictions based on hand-crafted stimuli features (IBL) and GM representations (GER), show the highest difference to participant performance. The unique feature of the GINGER model involves predictions of utility partially influenced by the GM-formed stimulus representation related to the IBL model's use of the features. However, by directly predicting utility based on representations, both the GINGER and GIN models are able to quickly update utility predictions.

In summary, the modeling results demonstrate that each approach to incorporating GMs in predicting human learning is viable, as none of the models performs worse than the IBL model, which does not use a GM. However, models that perform actions selected by the GM exhibit more human-like learning trends (Figure 2C) and a closer similarity to human learning (Figure 2D). While leveraging GM representations aims to improve generalization, the simplicity of this task imposes minimal demands on generalization, meaning that the speed of learning is more relevant in producing human-like learning. The next experiment paradigms will introduce an explicit generalization requirement for participants. This will enable a comparison of ablated models in a task where generalization performance is more important.

## 4.2 Transfer of learning task

This decision-making task involves learning the values associated with abstract visual stimuli and transferring that knowledge to more visually complex stimuli. Previous research comparing the IBL and the GER model demonstrated improved performance in transfer of learning tasks by introducing generative representations to the IBL model (Malloy et al., 2023). The higher performance of the GER model and its closer resemblance to human performance compared to the standard IBL model, raises questions about how our proposed GIN and GINGER models compare in replicating human-like behavior in this transfer of learning task.

In this task, generalization performance is more relevant than learning speed in evaluating participants and cognitive models. This is due to the increase in task complexity over time. Initially, participants engaged in a contextual bandit task focused only on the shape feature (Figure 3A Left). After 15 trials the task complexity increases with the introduction of the color feature (Figure 3A Middle). Transitioning to the color learning task requires participants' ability to transfer knowledge from the shape learning task to determine the optimal option. This demands

generalization from past experience to make future decisions in a related but not totally equivalent context. After these 15 trials of the color learning task, participants are introduced to the texture learning task (Figure 3A Right) which is similar to the structure of the first learning experiment (Niv et al., 2015).

### 4.2.1 Cognitive modeling

The design of IBL baseline model features was identical to the first experiment, including the use of the shape, color, and texture features, baseline parameter values, and binary similarity metrics. One difference between this task and the previous one is that the GIN and GINGER utility prediction modules are only being trained using one portion of the data set at a time, first shape, then shape-color, then shape-color-texture. This means that predicting utility associated with a representation requires a high degree of generalization to adequately transfer from one task to the other.

### 4.2.2 Methods

160 participants (86 female, 69 male, 2 non-binary) were recruited online through the Amazon Mechanical Turk (AMT) platform. All participants were over the age of 18 and citizens of the United States of America. Participants had a mean age of 40.5 with a standard deviation of 11.3 years. Participants were required to have completed at least 100 Human Intelligence Tasks (HITs) on AMT with at least a 95% approval on completed HITs. Six of the 160 recruited participants failed to submit data or failed to complete the task within a 1 hour limit, and were excluded from analysis. All results and analysis are done using the remaining 154 participants.

Participants received a base payment of \$4 with the potential to receive a bonus of up to \$3 depending on their performance in the task. The mean time to complete the task was 16.9 minutes, with a standard deviation of 5.8 minutes. This experiment was approved by the Carnegie Mellon University Internal Review Board. The experiment protocol was preregistered on OSF. Experiment preregistration, participant data, analysis, model code, and a complete experiment protocol are available on OSF.<sup>3</sup> For a more complete description of experiment methods, see Malloy et al. (2023).

Participant's performance in this task can be measured in their ability to transfer knowledge from one learning task to the subsequent learning tasks. Three commonly used metrics for performance in transferring learned knowledge to subsequent tasks are jumpstart, asymptotic, and episodic performance (Taylor and Stone, 2009). Jump-start performance is defined as the initial performance of an agent on a target task. In the contextual bandit experiment used in this work, the jumpstart performance is calculated as the average of the first third observed utility in trials after the task switches. Asymptotic performance is defined as the final learned performance of an agent in a target task. In the transfer of learning experiment, the asymptotic performance is calculated as the average of the final three reward observations of participants. Episodic Performance is defined as the average performance over an episode; this measure is analogous to the total reward metric commonly used. This

<sup>3</sup> <https://osf.io/mt4ws/>

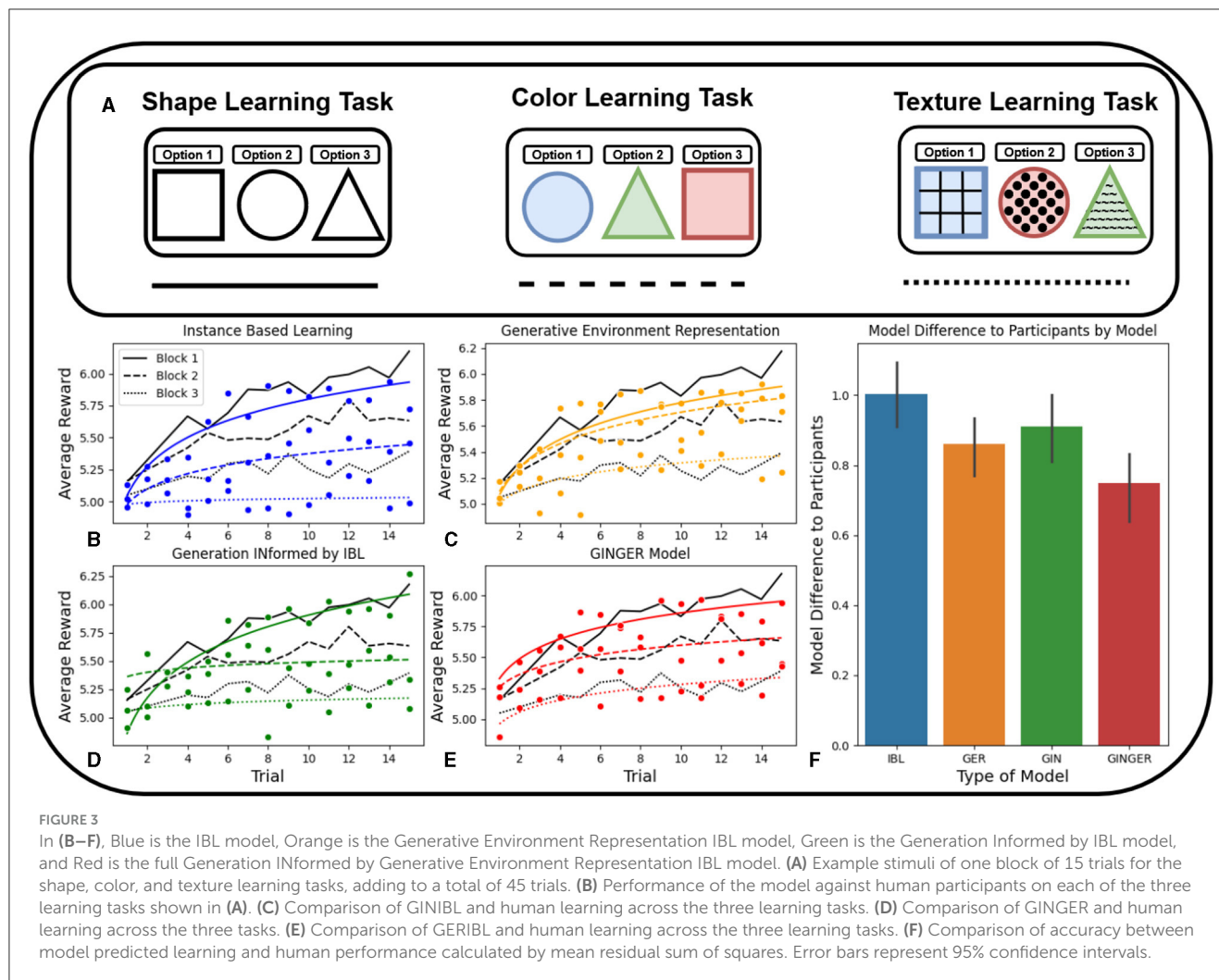


FIGURE 3

In (B–F), Blue is the IBL model, Orange is the Generative Environment Representation IBL model, Green is the Generation Informed by IBL model, and Red is the full Generation Informed by Generative Environment Representation IBL model. (A) Example stimuli of one block of 15 trials for the shape, color, and texture learning tasks, adding to a total of 45 trials. (B) Performance of the model against human participants on each of the three learning tasks shown in (A). (C) Comparison of GINIBL and human learning across the three learning tasks. (D) Comparison of GINGER and human learning across the three tasks. (E) Comparison of GERIBL and human learning across the three learning tasks. (F) Comparison of accuracy between model predicted learning and human performance calculated by mean residual sum of squares. Error bars represent 95% confidence intervals.

value is calculated as the average of the observed utility. These measures are used to compare model difference to participant behavior, and averaged over to produce the results shown in Figure 3F.

#### 4.2.3 Results

To assess transfer of learning for the three measures, we averaged the similarities between human and model performance in jumpstart, episodic, and asymptotic performance in the three learning tasks. This aggregation yields a single metric, providing a holistic evaluation of the fit between the model and human transfer of learning performance. This similarity is based on average residual sum of squares  $RSS/n$  calculations for each of the three measures of transfer of learning measures. This integrated measure of congruence is shown in Figure 3F, to facilitate a comparison across the four models. Importantly, these accuracy metrics are computed for each participant individually, ensuring the understanding of performance across the sample. Additionally, the same connection between average residual sum of squares and BIC can be made as in the first experiment, since again there are no fit parameters.

As in the contextual bandit task in Experiment 1, we first compare the four models by their speed of learning, and the similarity to human performance, shown in the four plots (Figures 3B–E). This is done for each of the three learning tasks that increase in complexity as the experiment progresses. This comparison shows that the GER and GINGER models have learning trend more similar to humans in the color and texture tasks compared to the IBL and GIN models. This is likely because of the fact that the representations of visual information used by the GER and GINGER models as features of the IBL model allow for improved generalization, which is a key feature of improving transfer of learning ability.

Comparatively, the IBL and GIN models show more human-like learning on the simple shape learning task before the transfer of learning ability becomes relevant. This mirrors the human-like learning achieved by these two models in the first experiment, but because the majority of this task relies more on generalization capability rather than the speed of learning, the end result is that the GER and GINGER models are better fits to human learning averaged across the entire experiment.

The next comparison of model performance is shown in Figure 3F which captures an aggregate average of the three transfer

of learning metrics previously discussed. Overall, the IBL model is far more distant from human performance than the three ablation models. The GER and GIN models are about equally distant from human performance, as the GER model has relatively higher performance on the two transfer tasks while GIN model had better performance on the first task. The GINGER model, which combines the more human-like behavior on the first task observed by the GIN model, and the two transfer tasks by the GER model, produces the most human-like learning on average.

## 4.3 Phishing identification task

Phishing messages are emails that contain attempts to obtain credentials, transmit malware, gain access to internal systems, or cause financial harm (Hong, 2012). An important aspect of preventing these phishing emails from negatively impacting individuals and companies is through training programs to help people identify phishing emails more successfully (Singh et al., 2020). Cognitive models have been applied to predict and improve email phishing training (Singh et al., 2019; Cranford et al., 2021). The phishing email identification task is used to compare the ablation of our proposed model in how relevant each of its attributes is in conditions that include complex natural language stimuli.

We use a data set of human judgments on the phishing identification task (Figure 4A) that was originally collected in Singh et al. (2023) and is publicly available. The phishing identification task involved the presentation of phishing or safe emails. Participants indicated their guess as to whether the emails were safe or dangerous, their confidence rating, as well as a recommendation of an action to take when receiving this email, such as checking the link, responding to the email, opening an attachment, etc. (Singh et al., 2023). These details are described more fully in the section on experimentation methods.

### 4.3.1 Cognitive modeling

The baseline IBL model for this task used binary hand-crafted features coded by human experts (Figure 4B) including mismatched sender, requesting credentials, urgent language, making an offer, suspicious subject, and a link mismatch. The other main difference in cognitive modeling of this experiment with the previous two is that a LLM model is used to form the representations used both as a feature of the task and directly trained to predict utilities.

These representations are embeddings of textual inputs formed by the OpenAI GPT based model “text-embedding-ada-002”. At the time of writing, this was the only text embedding model available on the OpenAI Application Programming Interface. This model generates representations of text inputs in the form of a vector of 1536 floating point numbers. The IBL similarity metric for these representations is calculated with the sklearn python package cosine similarity function, a commonly used metric when comparing sentence embeddings from large language models (Li et al., 2020).

Due to the high baseline performance of humans in this task, as a result of their experience in reading emails and their experience with phishing warnings, we use a random sampling of 10% of emails to pre-train all models under comparison. This allows for a more realistic comparison of the performance of these models in reflecting human decision making in this type of task.

### 4.3.2 Methods

The experimental methods for this analysis are detailed in full in Singh et al. (2023). 228 participants were recruited online through the Amazon Mechanical Turk (AMT) platform. Participants were required to have completed at least 100 Human Intelligence Tasks (HITs) on AMT with at least a 90% approval rate. All participants were over the age of 18. Participants have a mean age of 36.8 with a standard deviation of 11.5 years. Four of the 228 participants failed attention checks and were excluded from the analysis. Participants were paid a base rate of \$6 with the potential to receive a bonus of up to \$3 depending on their performance. The mean time to complete this experiment was 35 minutes.

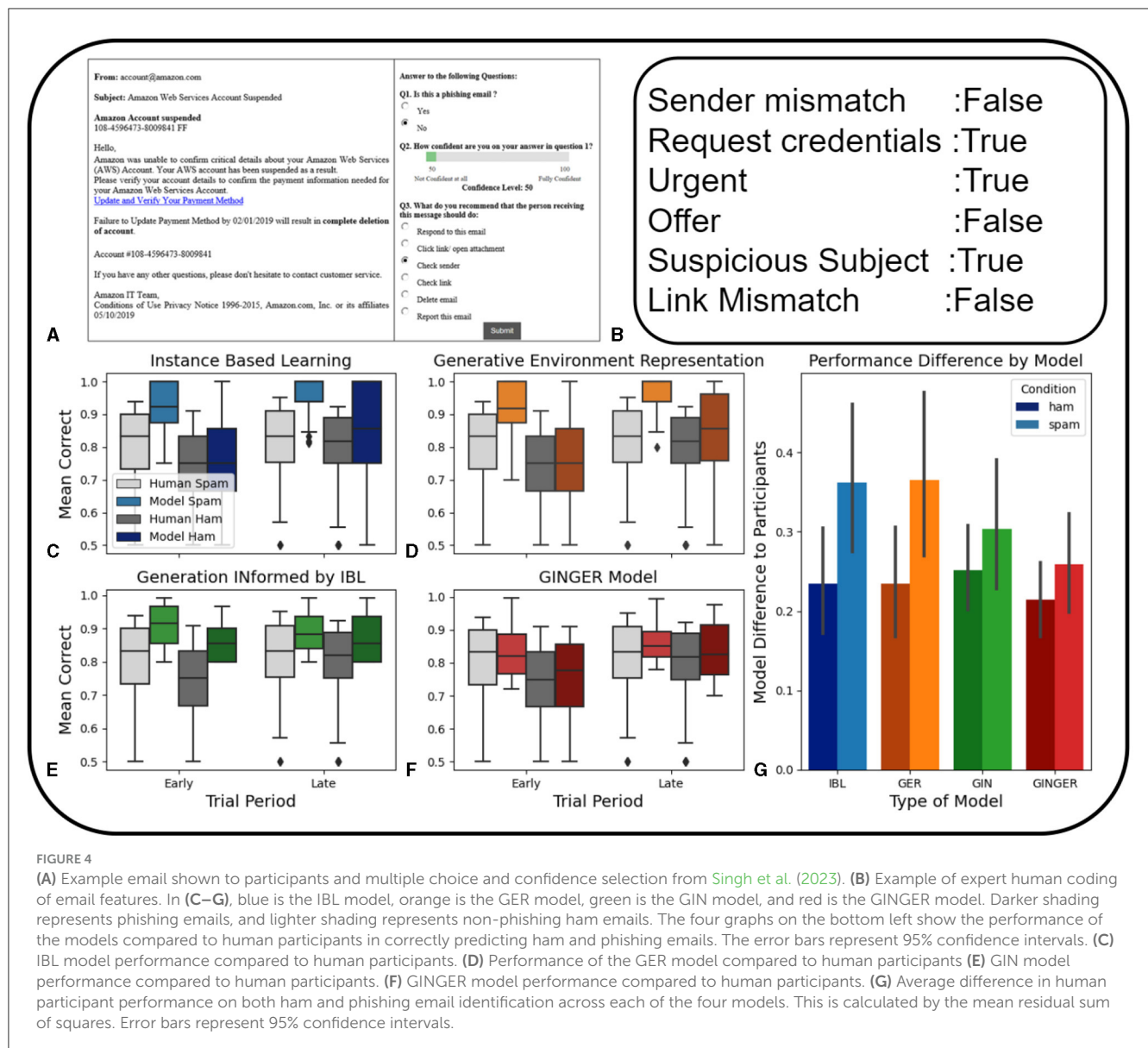
Experiment data were made available on request from the original authors and obtained by us after request. This experiment data included participant judgments in the task as well as the 239 emails that were classified by the researchers based features that were relevant to determine if the emails were phishing, referred to as spam, or non-phishing, referred to as ham. These features included whether the sender of the email matched the claimed sender; whether or not the email made a request of credentials; whether or not the subject line was suspicious; whether an offer was made in the email body; whether the tone of the email used urgent language; and finally whether a link in the email matched the text of the link. Textual data and email features are available on OSF<sup>4</sup> and participant data are contained in our previously mentioned combined repository (see text footnote<sup>1</sup>).

Participants' performance in this task can be measured in their ability to correctly identify phishing emails as phishing, and ham emails as ham. Splitting this classification by the type of email shown to participants allows for a comparison between the different amounts of phishing and ham emails that were shown to participants during the experimental conditions. Ideally, an accurate model of human learning in this task would be similar to human data for each of these types of categorization.

Accurately reflecting differences in experience with the identification of phishing emails from participants can be a difficult task for cognitive models. In IBL models, this could be done by using a set of different models with varied initial experiences with phishing and ham emails, which would result in differences in accuracy for categorizing these two types of email. However, to highlight the differences in ablation analysis, we do not differentiate the experience of models individually to better fit human performance, and instead use the same base-level experience across all models under comparison.

4 <https://osf.io/sp7d6/>





### 4.3.3 Results

In this experiment, each of the four ablation models predicted the same emails shown to participants, in the same order. The ablation models used the values of the baseline parameters for all the parameters of the IBL model. Therefore, the total number of model runs was equal to the number of participants for each type of model ablation. Models were trained using a reward function of 1 point for correct categorization and 0 points for incorrect categorization. For the GIN and GER models, the utility prediction based on representations was done using the representation input of size 1536 followed by two layers of size 128 and finally an output of size 1. More details of this are included in the [Supplementary material](#).

The performance of the GIN model is unique in that it predicts similarly high performance in the early and later trial periods for both types of emails (Figure 4E). This direct utility prediction based on representations can approach high accuracy from only a few examples. This is true for both phishing and ham emails, while humans display lower accuracy overall, and a large difference

between accuracy in these two types of emails. It would be possible to reduce this training for the GIN model alone, however, this would mean that the GIN model is using less experience than the other models.

In general, taking an approach to fitting the training time of generative actions to human performance can be difficult for large representations sizes, as it requires multiple training periods that are computationally expensive. This is demonstrated by the difference in similarity with the results of human learning demonstrated by the GIN model. This is a key difference between the phishing email identification task, where the representation size is 1536, compared to the earlier tasks that used  $\beta$ -VAE model representations of size 9. However, these representation sizes are not considered to be a variable or fit parameter in any of the models. Thus, the same connection between the average residual sum of squares and BIC can be made as in the first experiment, since again there are no fit parameters.

The GINGER model has the highest accuracy to human performance (Figures 4C–G), as a result of it making predictions

using both the GM and the email representations that are fed into an IBL model. This demonstrates the benefits of combining generative actions and generative memory formation, for tasks with complex natural language stimuli. This is especially true for tasks like this one where participants are likely to have previous experience from which they are drawing, as opposed to the two previous abstract tasks. This is because optimizing the GIN model alone to fit human participant performance is computationally expensive and the IBL and GER models are not able to learn the task quickly enough.

## 5 Discussion

This research proposes a model that demonstrates the benefits of integrating GMs and cognitive modeling and their potential applications. These techniques open new avenues in the investigation of human learning that were previously inaccessible to cognitive modelers. GAI has had a significant impact across many fields of study, motivating its application in cognitive modeling, especially in decision-making processes. However, before integrating GMs into cognitive models to represent and predict human decision making, it is important to investigate the relative impact that different methods of integration have on different tasks.

The GINGER model proposed in this work demonstrates the integration of GMs with cognitive models of decision making, such as IBL. Our approach demonstrates the accurate prediction of human learning and decision making across three distinct experimental paradigms, directly compared to real human decisions. These experiments encompass a diverse range of stimuli, spanning visual cues and natural language that varied in complexity, from learning abstract rewards to detecting phishing attempts in emails. The application of our GINGER model across these domains resulted in an improvement over traditional cognitive modeling techniques, clearly demonstrating the potential benefits of incorporating GMs into cognitive modeling frameworks.

In addition to our GINGER model, we developed a categorization approach that can be used to compare and relate different approaches to integrating GMs into cognitive modeling of decision making. Before current research, there were many applications of GMs in cognitive modeling, although typically this was done in a case-by-case manner to allow for use in a specific learning domain. Here, we compare the integration of GMs in cognitive modeling in six dimensions, including action generation, memory generation, stimuli, cognitive model type, generative model type, and training method.

This categorization motivated an ablation study to compare our proposed model with alternative versions that contained generative actions and memory and did not contain them. Additionally, the three experiment paradigms were chosen to further test the remaining categories of our analysis, to investigate the varied stimuli types, GM types, and training methods. The result is a comparison of model performance that spans many degrees of our proposed categorization. The first experimental comparison demonstrated faster and more human-like learning from models that produced decision predictions directly by GMs (GIN and GINGER). However, this faster learning was observed in a relatively

simple task, raising the question of the potential benefits of GM memory formation (GER and GINGER) in more complex environments.

The second comparison of models through experimentation extended the analysis in the first experiment by introducing a generalization task that required transfer of learning. This is a useful comparison for our proposed model, as one of the often cited benefits of applying GMs to cognitive models is improved generalization. This raised the question of which method of integrating GMs would be more relevant for improving performance and the similarity to human participants in this task. The high generalizability of models that utilized GM memory representations confirmed this expectation, demonstrating the ability of cognitive models that integrate GM representations in reflecting human-like generalization.

In the third and final experimentation, we investigated the potential differences of our proposed modeling method when handling complex natural language in a phishing identification task. Comparing the performance of models with that of human participants in this task demonstrated a large difference between categorization accuracy for phishing and ham emails, which was difficult for the models to replicate. Previously, only cognitive models that used GM representations of textual information, such as phishing emails, have been used to predict human-like learning, but these results demonstrate that a combination of directly predicting values and GM representations is best for this type of task.

Overall, these results from the model comparison provide insight into the design of integration of generative modeling methods with cognitive models. Each of our experiments investigated a different area of human learning and decision making modeling and made important conclusions about how best to integrate GMs. Although the applications of our model comparison are broad, they do not represent every possible application of GMs to cognitive modeling. As demonstrated by our categorization, there are remaining stimuli types, generative models, and cognitive models that could be compared. One potential future area of research would be the application of multi-modal models and a comparison of learning with humans engaging in a multi-modal decision task.

While GMs have demonstrated a high degree of usefulness in cognitive modeling, the impact that they have on society at large has been called into question, as noted previously. One potential issue with the use of a model similar to one of the ones we used in the experiment on predicting how participants respond to phishing emails is that it could be used to improve the quality of phishing email campaigns. This is exacerbated by the potential to use GMs themselves to generate phishing emails. One potential future area of research is investigating how we can best mitigate these potential GM misuses. This could be done by tailoring phishing email education to the individual through the application of a model similar to the one we propose, which can allow students to experience phishing emails generated by GMs and learn from them.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://osf.io/m6qc4/>.

## Ethics statement

The studies involving humans were approved by Carnegie Mellon University Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## Author contributions

TM: Formal analysis, Software, Visualization, Writing – original draft, Writing – review & editing. CG: Conceptualization, Funding acquisition, Methodology, Resources, Supervision, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was sponsored by the Army Research Office and accomplished under Australia-US MURI Grant Number W911NF-20-S-000. Compute resources and GPT model credits were provided by the Microsoft Accelerate Foundation Models

## References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). Gpt-4 technical report. *arXiv [Preprint]*. Available online at: <https://arxiv.org/abs/2303.08774>
- Aggarwal, P., Thakoor, O., Jabbari, S., Cranford, E. A., Lebiere, C., Tambe, M., et al. (2022). Designing effective masking strategies for cyberdefense through human experimentation and cognitive models. *Comp. Secur.* 117:102671. doi: 10.1016/j.cose.2022.102671
- Aher, G. V., Arriaga, R. I., and Kalai, A. T. (2023). “Using large language models to simulate multiple humans and replicate human subject studies,” in *International Conference on Machine Learning* (New York: PMLR), 337–371.
- Anderson, J. R., and Lebiere, C. J. (2014). *The Atomic Components of Thought*. London: Psychology Press.
- Anderson, J. R., Matessa, M., and Lebiere, C. (1997). Act-r: a theory of higher level cognition and its relation to visual attention. *Human-Comp. Interact.* 12, 439–462. doi: 10.1207/s15327051hci1204\_5
- Bandi, A., Adapa, P. V. S. R., and Kuchi, Y. E. V. P. K. (2023). The power of generative AI: a review of requirements, models, input-output formats, evaluation metrics, and challenges. *Future Internet* 15:260. doi: 10.3390/fi15080260
- Bates, C., and Jacobs, R. (2019). Efficient data compression leads to categorical bias in perception and perceptual memory. *CogSci.* 43, 1369–1375.
- Bates, C. J., and Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychol. Rev.* 127:891. doi: 10.1037/rev0000197
- Beguš, G. (2020). Generative adversarial phonology: modeling unsupervised phonetic and phonological learning with neural networks. *Front. Artif. Intellig.* 3:44. doi: 10.3389/frai.2020.00044
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). “On the dangers of stochastic parrots: Can language models be too big?” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY: Association for Computing Machinery), 610–623.
- Bhui, R., Lai, L., and Gershman, S. J. (2021). Resource-rational decision making. *Curr. Opin. Behav. Sci.* 41, 15–21. doi: 10.1016/j.cobeha.2021.02.015
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. *arXiv [Preprint]*. Available online at: <https://arxiv.org/abs/2108.07258>
- Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., et al. (2023). “Do as I can, not as I say: Grounding language in robotic affordances,” in *Conference on Robot Learning* (New York: PMLR).
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33, 1877–1901. Available online at: <https://arxiv.org/abs/2005.14165>
- Bugbee, E. H., and Gonzalez, C. (2022). “Making predictions without data: How an instance-based learning model predicts sequential decisions in the balloon analog risk task,” in *Proceedings of the Annual Meeting of the Cognitive Science Society* (Seattle, WA: Cognitive Science Society), 1–6.
- Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., et al. (2023). A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv [Preprint]*. Available online at: <https://arxiv.org/abs/2303.04226>
- Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., et al. (2021). Decision transformer: Reinforcement learning via sequence modeling. *Adv. Neural Inf. Process. Syst.* 34, 15084–15097. Available online at: <https://arxiv.org/abs/2106.01345>
- Chevalier-Boisvert, M., Willems, L., and Pal, S. (2018). *Minimalistic Gridworld Environment for OpenAI Gym*. Available online at: <https://github.com/maximecb/gym-minigrid> (accessed August 10, 2023).
- Choi, D., Konik, T., Nejati, N., Park, C., and Langley, P. (2007). A believable agent for first-person shooter games. *Proc. AAAI Conf. Artif. Intellig. Interact. Digit. Entertainm.* 3, 71–73. doi: 10.1609/aiide.v3i1.18787
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., et al. (2023). Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.* 24, 1–113. Available online at: <https://arxiv.org/abs/2204.02311>
- Cranford, E. A., Gonzalez, C., Aggarwal, P., Cooney, S., Tambe, M., and Lebiere, C. (2020). Toward personalized deceptive signaling for cyber defense using cognitive models. *Top. Cogn. Sci.* 12, 992–1011. doi: 10.1111/tops.12513
- Cranford, E. A., Lebiere, C., Rajivan, P., Aggarwal, P., and Gonzalez, C. (2019). “Modeling cognitive dynamics in end-user response to phishing emails,” in *Proceedings of the 17th ICCM* (State College, PA: Applied Cognitive Science Lab).
- Cranford, E. A., Singh, K., Aggarwal, P., Lebiere, C., and Gonzalez, C. (2021). “Modeling phishing susceptibility as decisions from experience,” in *Proceedings of the 19th Annual Meeting of the ICCM* (State College, PA: Applied Cognitive Science Lab), 44–49.

Research grant “Personalized Education with Foundation Models via Cognitive Modeling”.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1387948/full#supplementary-material>

- Friston, K. J., Parr, T., Yufik, Y., Sajid, N., Price, C. J., and Holmes, E. (2020). Generative models, linguistic communication and active inference. *Neurosci. Biobehav. Rev.* 118, 42–64. doi: 10.1016/j.neubiorev.2020.07.005
- Gershman, S. J. (2019). The generative adversarial brain. *Front. Artif. Intell.* 2:18. doi: 10.3389/frai.2019.00018
- Goetschalckx, L., Andonian, A., and Wagemans, J. (2021). Generative adversarial networks unlock new methods for cognitive science. *Trends Cogn. Sci.* 25, 788–801. doi: 10.1016/j.tics.2021.06.006
- Gonzalez, C. (2013). The boundaries of instance-based learning theory for explaining decisions from experience. *Prog. Brain Res.* 202, 73–98. doi: 10.1016/B978-0-444-62604-2.00005-8
- Gonzalez, C. (2023). Building human-like artificial agents: A general cognitive algorithm for emulating human decision-making in dynamic environments. *Persp. Psychol. Sci.* 2023, 17456916231196766. doi: 10.1177/17456916231196766
- Gonzalez, C., and Dutt, V. (2011). Instance-based learning: integrating sampling and repeated decisions from experience. *Psychol. Rev.* 118:523. doi: 10.1037/a0024558
- Gonzalez, C., Lerch, J. F., and Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cogn. Sci.* 27, 591–635. doi: 10.1207/s15516709cog2704\_2
- Griffith, S., Subramanian, K., Scholz, J., Isbell, C. L., and Thomaz, A. L. (2013). Policy shaping: Integrating human feedback with reinforcement learning. *Adv. Neural Inf. Process. Syst.* 26, 1–9. doi: 10.5555/2999792.2999905
- Hedayati, S., O'Donnell, R. E., and Wyble, B. (2022). A model of working memory for latent representations. *Nat. Human Behav.* 6, 709–719. doi: 10.1038/s41562-021-01264-9
- Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., et al. (2021). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nat. Commun.* 12:6456. doi: 10.1038/s41467-021-26751-5
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., et al. (2016). “Beta-vae: Learning basic visual concepts with a constrained variational framework,” in *International Conference on Learning Representations*, 1–6. Available online at: <https://openreview.net/forum?id=Sy2fzU9gl>
- Higgins, I., Pal, A., Rusu, A., Matthey, L., Burgess, C., Pritzel, A., et al. (2017). “Darl: improving zero-shot transfer in reinforcement learning,” in *International Conference on Machine Learning* (New York: PMLR), 1480–1490.
- Hintzman, D. L. (1984). Minerva 2: a simulation model of human memory. *Behav. Res. Methods, Instrum. Comp.* 16, 96–101. doi: 10.3758/BF03202365
- Hong, J. (2012). The state of phishing attacks. *Commun. ACM* 55, 74–81. doi: 10.1145/2063176.2063197
- Huet, A., Pinquie, R., Véron, P., Mallet, A., and Segonds, F. (2021). Cacd: A knowledge graph for a context-aware cognitive design assistant. *Comp. Indust.* 125:103377. doi: 10.1016/j.compind.2020.103377
- Ivanovic, B., Schmerling, E., Leung, K., and Pavone, M. (2018). “Generative modeling of multimodal multi-human behavior,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Madrid: IEEE), 3088–3095.
- Kenton, J. D. M.-W. C., and Toutanova, L. K. (2019). “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT*, 4171–4186. Available online at: <https://arxiv.org/abs/1810.04805>
- Kim, S., Lee, S., Song, J., Kim, J., and Yoon, S. (2018). Flowavenet: a generative flow for raw audio. *arXiv [Preprint]*. Available online at: <https://arxiv.org/abs/1811.02155>
- Kirk, J. R., Wray, R. E., and Laird, J. E. (2023). “Exploiting language models as a source of knowledge for cognitive agents,” in *arXiv* (Washington, DC: AAAI Press).
- Kirsch, L., Harrison, J., Freeman, C. D., Sohl-Dickstein, J., and Schmidhuber, J. (2023). “Towards general-purpose in-context learning agents,” in *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 1–6. Available online at: <https://openreview.net/forum?id=zDTqQVGzH>
- Lai, L., and Gershman, S. J. (2021). Policy compression: an information bottleneck in action selection. *Psychol. Learn. Motiv.* 74, 195–232. doi: 10.1016/bs.plm.2021.02.004
- Laird, J. E. (2001). “It knows what you’re going to do: adding anticipation to a quakebot,” in *Proceedings of the Fifth International Conference on Autonomous Agents* (New York, NY: Association for Computing Machinery), 385–392.
- Laird, J. E., Lebiere, C., and Rosenbloom, P. S. (2017). A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *Ai Magazine* 38, 13–26. doi: 10.1609/aimag.v38i4.2744
- Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artif. Intell.* 33(1):1–64. doi: 10.1016/0004-3702(87)90050-6
- Lejarraga, T., Dutt, V., and Gonzalez, C. (2012). Instance-based learning: a general model of repeated binary choice. *J. Behav. Decis. Mak.* 25, 143–153. doi: 10.1002/bdm.722
- Li, B., Zhou, H., He, J., Wang, M., Yang, Y., and Li, L. (2020). On the sentence embeddings from pre-trained language models. *arXiv [Preprint]*. Available online at: <https://arxiv.org/abs/2011.05864>
- Li, S., Puig, X., Paxton, C., Du, Y., Wang, C., Fan, L., et al. (2022). Pre-trained language models for interactive decision-making. *Adv. Neural Inf. Process. Syst.* 35, 31199–31212. Available online at: <https://arxiv.org/abs/2202.01771>
- Malloy, T., Du, Y., Fang, F., and Gonzalez, C. (2023). “Generative environment-representation instance-based learning: a cognitive model,” in *Proceedings of the 2023 AAAI Fall Symposium on Integrating Cognitive Architectures and Generative Models* (Washington, DC: AAAI Press), 1–6.
- Malloy, T., Klinger, T., and Sims, C. R. (2022a). “Modeling human reinforcement learning with disentangled visual representations,” in *Reinforcement Learning and Decision Making (RLDM)* (Washington, DC: Association for Research in Vision and Ophthalmology).
- Malloy, T., and Sims, C. R. (2022). A beta-variational auto-encoder model of human visual representation formation in utility-based learning. *J. Vis.* 22:3747. doi: 10.1167/jov.22.14.3747
- Malloy, T., Sims, C. R., Klinger, T., Riemer, M. D., Liu, M., and Tesauro, G. (2022b). “Learning in factored domains with information-constrained visual representations,” in *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*, 1–6. Available online at: <https://arxiv.org/abs/2303.17508>
- McDonald, C., Malloy, T., Nguyen, T. N., and Gonzalez, C. (2023). “Exploring the path from instructions to rewards with large language models in instance-based learning,” in *Proceedings of the 2023 AAAI Fall Symposium on Integrating Cognitive Architectures and Generative Models* (Washington DC: AAAI Press), 1–6.
- Mitsopoulos, K., Baker, L., Lebiere, C., Piroli, P., Orr, M., and Vardavas, R. (2023a). Masking behaviors in epidemiological networks with cognitively-plausible reinforcement learning. *arXiv [Preprint]*. Available online at: <https://arxiv.org/abs/2312.03301>
- Mitsopoulos, K., Bose, R., Mather, B., Bhatia, A., Gluck, K., Dorr, B., et al. (2023b). “Psychologically-valid generative agents: A novel approach to agent-based modeling in social sciences,” in *Proceedings of the 2023 AAAI Fall Symposium on Integrating Cognitive Architectures and Generative Models* (Washington DC: AAAI Press).
- Morrison, D., and Gonzalez, C. (2024). *PyIBL 5.1.1 Manual*. Available online at: <http://pyibl.ddmlab.com/> (accessed 18 March, 2024).
- Navigli, R., Conia, S., and Ross, B. (2023). Biases in large language models: Origins, inventory and discussion. *ACM J. Data Informat. Qual.* 15, 1–21. doi: 10.1145/3597307
- Nguyen, T. N., and Gonzalez, C. (2022). Theory of mind from observation in cognitive models and humans. *Top. Cogn. Sci.* 14, 665–686. doi: 10.1111/tops.12553
- Nguyen, T. N., Phan, D. N., and Gonzalez, C. (2023). Speedyibl: a comprehensive, precise, and fast implementation of instance-based learning theory. *Behav. Res. Methods* 55, 1734–1757. doi: 10.3758/s13428-022-01848-x
- Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., et al. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *J. Neurosci.* 35, 8145–8157. doi: 10.1523/JNEUROSCI.2978-14.2015
- Ororbias, A., and Kelly, M. A. (2023). “A neuro-mimetic realization of the common model of cognition via hebbian learning and free energy minimization,” in *Proceedings of the 2023 AAAI Fall Symposium on Integrating Cognitive Architectures and Generative Models* (Washington: AAAI Press), 1–6.
- Ororbias, A., and Kifer, D. (2022). The neural coding framework for learning generative models. *Nat. Commun.* 13(1):2064. doi: 10.1038/s41467-022-29632-7
- Park, J. S., O’Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). “Generative agents: Interactive simulacra of human behavior,” in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (New York, NY: Association for Computing Machinery), 1–22.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training (Preprint)*.
- Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87. doi: 10.1038/4580
- Reid, M., Yamada, Y., and Gu, S. S. (2022). Can wikipedia help offline reinforcement learning? *arXiv [Preprint]*.
- Ren, H., and Ben-Tzvi, P. (2020). Learning inverse kinematics and dynamics of a robotic manipulator using generative adversarial networks. *Rob. Auton. Syst.* 124:103386. doi: 10.1016/j.robot.2019.103386
- Shi, R., Liu, Y., Ze, Y., Du, S. S., and Xu, H. (2023). “Unleashing the power of pre-trained language models for offline reinforcement learning,” in *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 1–6. Available online at: <https://arxiv.org/abs/2310.20587>
- Singh, K., Aggarwal, P., Rajivan, P., and Gonzalez, C. (2019). “Training to detect phishing emails: Effects of the frequency of experienced phishing emails,” in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Los Angeles: SAGE Publications Sage CA).
- Singh, K., Aggarwal, P., Rajivan, P., and Gonzalez, C. (2020). “What makes phishing emails hard for humans to detect?,” in *Proceedings of the Human Factors and*



*Ergonomics Society Annual Meeting* (Los Angeles, CA: SAGE Publications Sage CA), 431–435.

Singh, K., Aggarwal, P., Rajivan, P., and Gonzalez, C. (2023). Cognitive elements of learning and discriminability in anti-phishing training. *Comp. Secur.* 127:103105. doi: 10.1016/j.cose.2023.103105

Speer, R., Chin, J., and Havasi, C. (2017). “Conceptnet 5.5: an open multilingual graph of general knowledge,” in *Proceedings of the AAAI Conference on Artificial Intelligence* (Washington, DC: AAAI Press), 1–6.

Sun, R. (2006). “The clarion cognitive architecture: extending cognitive modeling to social simulation,” in *Cognition and Multi-Agent Interaction* (Cambridge: Cambridge University Press), 79–99.

Swan, G., and Wyble, B. (2014). The binding pool: a model of shared neural resources for distinct items in visual working memory. *Attent. Percep. Psychophys.* 76, 2136–2157. doi: 10.3758/s13414-014-0633-3

Taniguchi, T., Yamakawa, H., Nagai, T., Doya, K., Sakagami, M., Suzuki, M., et al. (2022). A whole brain probabilistic generative model: Toward realizing cognitive architectures for developmental robots. *Neural Networks* 150:293–312. doi: 10.1016/j.neunet.2022.02.026

Taylor, M. E., and Stone, P. (2009). Transfer learning for reinforcement learning domains: a survey. *J. Mach. Learn. Res.* 10:7. doi: 10.5555/1577069.1755839

Wu, T., Zhu, B., Zhang, R., Wen, Z., Ramchandran, K., and Jiao, J. (2023). “Pairwise proximal policy optimization: Harnessing relative feedback for llm alignment,” in *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 1–6. Available online at: <https://arxiv.org/abs/2310.00212>

Xu, T., Singh, K., and Rajivan, P. (2022). “Modeling phishing decision using instance based learning and natural language processing,” in *HICSS* (Manoa, HI: University of Hawaii at Manoa), 1–10.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2023). Can large language models transform computational social science? *arXiv [Preprint]*. Available online at: <https://arxiv.org/abs/2305.03514>



## OPEN ACCESS

## EDITED BY

Janet Clinton,  
The University of Melbourne, Australia

## REVIEWED BY

Yan Liu,  
Carleton University, Canada  
Ahmed Kharrufa,  
Newcastle University, United Kingdom

## \*CORRESPONDENCE

Leonora Kaldaras  
✉ leonora.kaldaras@colorado.edu;  
✉ kaldaras@stanford.edu;  
✉ lkaldara@ttu.edu

RECEIVED 13 March 2024

ACCEPTED 17 July 2024

PUBLISHED 07 August 2024

## CITATION

Kaldaras L, Akaeze HO and  
Reckase MD (2024) Developing valid  
assessments in the era of generative artificial  
intelligence.  
*Front. Educ.* 9:1399377.  
doi: 10.3389/feduc.2024.1399377

## COPYRIGHT

© 2024 Kaldaras, Akaeze and Reckase. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

# Developing valid assessments in the era of generative artificial intelligence

Leonora Kaldaras<sup>1,2\*</sup>, Hope O. Akaeze<sup>3,4</sup> and Mark D. Reckase<sup>3</sup>

<sup>1</sup>Department of Physics, University of Colorado Boulder, Boulder, CO, United States, <sup>2</sup>Graduate School of Education, Stanford University, Stanford, CA, United States, <sup>3</sup>Michigan State University, East Lansing, MI, United States, <sup>4</sup>Community Evaluation Programs, Office of Public Engagement and Scholarship, University Outreach and Engagement, Michigan State University, East Lansing, MI, United States

Generative Artificial Intelligence (GAI) holds tremendous potential to transform the field of education because GAI models can consider context and therefore can be trained to deliver quick and meaningful evaluation of student learning outcomes. However, current versions of GAI tools have considerable limitations, such as social biases often inherent in the data sets used to train the models. Moreover, the GAI revolution comes during a period of moving away from memorization-based education systems toward supporting learners in developing the ability to apply knowledge and skills to solve real-world problems and explain real-world phenomena. A challenge in using GAI tools for scoring assessments aimed at fostering knowledge application is ensuring that these algorithms are scoring the same construct attributes (e.g., knowledge and skills) as a trained human scorer would score when evaluating student performance. Similarly, if using GAI tools to develop assessments, one needs to ensure that the goals of GAI-generated assessments are aligned with the vision and performance expectations of the learning environments for which these assessments are developed. Currently, no guidelines have been identified for assessing the validity of AI-based assessments and assessment results. This paper represents a conceptual analysis of issues related to developing and validating GAI-based assessments and assessment results to guide the learning process. Our primary focus is to investigate how to meaningfully leverage capabilities of GAI for developing assessments. We propose ways to evaluate the validity evidence of GAI-produced assessments and assessment scores based on existing validation approaches. We discuss future research avenues aimed at establishing guidelines and methodologies for assessing the validity of AI-based assessments and assessment results. We ground our discussion in the theory of validity outlined in the Standards for Educational and Psychological Testing by the American Educational Research Association and discuss how we envision building on the standards for establishing the validity of inferences made from the test scores in the context of GAI-based assessments.

## KEYWORDS

generative artificial intelligence (GAI), validity, knowledge application, validity standards, assessment practices, evaluation of cognitive development with GAI

## Introduction

Recent advances in natural language processing (NLP) and deep learning technology have led to the development of models that can process language and perform a wide range of tasks such as generating high-quality text, images, and other content. GAI holds tremendous potential to transform education because the GAI models can be trained to perform specific

tasks and potentially automate or streamline various processes. These models are pre-trained on large volumes of data and are broadly referred to as Generative Artificial Intelligence (GAI) tools. For example, one of the most popular GAI tools, ChatGPT, is trained on large amounts of conversational data related to education and, therefore, is capable of considering context and tailoring its responses to the specific needs of the user—such as personalizing learning experiences (Samala et al., 2024). However, current GAI tools have considerable limitations, such as social biases often inherent in the data sets used to train these models (Mao et al., 2024). These biases, among other factors, must be considered when implementing GAI tools in education.

Moreover, the GAI revolution comes during a period of significant changes in global education. Specifically, recent educational reforms worldwide emphasize supporting learners in developing the ability to apply knowledge to solve real-world problems and explain real-world phenomena. Examples include PISA, which has emphasized knowledge application on their assessments (OECD, 2016). Further, Germany (Kulgemeyer and Schecker, 2014) and Finland [Finnish National Board of Education (FNBE), 2015] have developed national standards focused on supporting learners in developing and measuring competencies. Competencies refer to standards expressed as learning goals requiring learners to apply their knowledge rather than reciting memorized information. A similar push toward measuring competencies occurs in the Chinese educational system (Ministry of Education, P. R. China, 2018; Yao and Guo, 2018). In the United States, similar efforts have resulted in the publication of the Framework for K-12 Science Education (*the Framework*) and the Next Generation Science Standards (NGSS), which emphasize fostering knowledge growth coherently over time so learners can apply what they learn (National Research Council, 2012; NGSS Lead States, 2013). The National Assessment Governing Board (NAGB) has just released an updated science framework for the 2028 Nation's report card that recommends supporting learners in developing the ability to integrate disciplinary knowledge and scientific practices to foster understanding (National Assessment Governing Board, 2023).

The move toward supporting knowledge application skills calls for significant changes at all stages of the learning process, including the development of new learning systems that help foster knowledge application. The learning system includes curriculum and assessment materials and the necessary instructional support. Effective development and implementation of such learning systems depends on the ability to adjust the learning process to the needs of academically, culturally, and linguistically diverse learners. Specifically, supporting the development of complex understanding related to knowledge application requires that students have opportunities to learn the appropriate knowledge and skills over time. The learning process must also be appropriately scaffolded to meet the needs of individual diverse learners (National Research Council, 2012). Therefore, to effectively support knowledge application, a learning system must incorporate features such as creating meaningful learning opportunities and providing timely and appropriate scaffolding and feedback.

Consequently, teachers and learners need access to timely, informative, high-quality feedback to effectively engage in the learning process and develop knowledge application skills (Pellegrino et al., 2001; Krajcik, 2021). This will ensure that teachers can meaningfully adjust their instruction and create the necessary learning opportunities

for students, and students can use this feedback to engage in discussions and self-reflection to deepen and improve their understanding. To provide this type of feedback, the assessments must effectively measure complex understanding, particularly the ability to apply relevant knowledge and skills. This, in turn, calls for moving away from multiple-choice (MC) based assessments toward open-ended assessments that require students to engage in developing models and explanations of phenomena (Krajcik, 2021). These assessments will allow us to measure complex reasoning and skills that are reflective of knowledge application ability and gain the information necessary for providing informative feedback to students and teachers. In short, the shift toward fostering knowledge application requires moving toward assessments that can guide the learning process instead of delivering point measurement results on the amount of information learners retained within a given time frame. Such open-ended assessments are time-consuming to develop, score, and report. AI tools, including GAI, have the potential to help tackle this challenge (Krajcik, 2021; Kaldaras et al., 2022). However, leveraging GAI tools to develop such assessments and provide feedback requires that the assessments developed using a GAI and the results of GAI model analysis of the assessment data meaningfully relate to the underlying constructs. They must offer valid and reliable measures capable of meaningfully guiding the learning process (Kaldaras and Haudek, 2022).

A construct refers to an unobservable and possibly hypothetical entity or concept. We evaluate or infer students' grasp of a construct based on their performance on assessment questions designed to measure it. Evaluation of the degree to which GAI provides an accurate evaluation of student progress on constructs describing knowledge application skills calls for careful evaluation of the validity of inferences drawn from AI-based assessment results (Kaldaras and Haudek, 2022). Validity refers to the degree to which evidence and theory support interpretations of assessment results (for example, test scores) for the proposed uses of a given assessment (Messick, 1980; Eignor, 2013). Construct validity relates to how well an assessment instrument represents and reflects the construct of interest. The validation process involves accumulating multiple relevant evidence sources to provide sound scientific support for the proposed interpretation of the assessment results (Eignor, 2013). Consequently, when assessment results are interpreted in multiple ways—for example, as a summative measure of what students have mastered or as a predictive measure of future performance—each of these intended uses of the assessment result must have evidence to support the desired inference.

In evaluating the validity of AI-based assessment outputs, an intended use of the assessment results could be to deliver timely and informative feedback to teachers and students to guide the learning process. For instance, AI-based assessments can be used to guide the learning process when they accurately diagnose students' understanding of a construct that describes knowledge application in a given context and provide meaningful feedback to teachers and students, where such feedback is focused on supporting the learners in transitioning to a higher understanding of that construct. Each of these intended uses incorporates multiple specific purposes as well. For example, teacher-facing feedback might be focused on delivering information about a student's current level of understanding or providing guidance on creating learning opportunities to compensate

for the lack of prior knowledge. Student-facing feedback might be focused on supporting individual student learning through self-reflection and revision of the answer. All these intended uses of the AI-based assessment results need supporting evidence.

One of the central challenges in using GAI tools for scoring assessments that will be used to guide the learning process aimed at fostering knowledge application is ensuring that these algorithms are scoring the same construct attributes (e.g., knowledge and skills) as a trained human scorer would score when evaluating student performance on the assessment. Similarly, when using GAI to assist in developing assessments focused on evaluating knowledge application, it is critical to ensure that the GAI-generated assessments represent a valid measure of the relevant knowledge application constructs. Currently, no guidelines have been established for assessing the validity of GAI-based assessments and assessment results. The purpose of this paper is to propose ways to evaluate the validity evidence of GAI-based assessments and scores based on existing approaches and discuss future research avenues for establishing guidelines and methodologies for assessing the validity of GAI-based assessment results.

This paper represents a conceptual analysis of issues related to validating AI-based assessments and assessment results to guide the learning process. We ground our discussion in the theory of validity outlined in the Standards for Educational and Psychological Testing by the American Educational Research Association (Eignor, 2013) and discuss how we envision building on the standards for establishing the validity of inferences made from the test scores in the context of GAI-based assessments.

## Structure of the paper

The paper begins with a brief historical overview of using AI approaches in education. We focus specifically on AI-based evaluation of student responses to assessments since this has been the most widespread way of using AI in education in the past. Further, we will discuss expanding use of AI to one the most critical aspects of assessment development—defining the construct of interest and the associated proficiencies. The need for defining what proficiency in a construct looks like prior to developing assessments for measuring this construct has been discussed by various educational experts [see, for example, research on construct modeling by Brown and Wilson (2011); or research on learning progressions that measure knowledge application by Kaldaras et al. (2021a,b, 2023)]. Similar need for defining proficiencies is outlined in multiple policy documents that discuss the importance of organizing the learning process along empirically derived learning progressions (National Research Council, 2012; National Assessment Governing Board, 2023). Further, the most substantial part of the paper is dedicated to evaluating assessment results and assessments generated using GAI and the associated validity evidence. In this section we focus on the types of validity evidence outlined in the Standards for Educational and Psychological Testing (Eignor, 2013), including: (1) evidence based on test content; (2) evidence based on response process; (3) evidence based on internal structure; (4) evidence based on relation to other variables; (5) validity generalization. For each type, we propose ways of evaluating validity evidence for assessments and assessment results generated by GAI and discuss future research avenues. We conclude by discussing

contributions of the conceptual analysis presented in the paper and proposing future research avenues focused on standardizing the validation process of GAI-generated assessments and assessments results.

## Validity and AI-based scores: a historical perspective

Before the emergence of GAI tools, various machine learning (ML) tools were used to evaluate student performance on open-ended assessments. These ML tools were often grounded in supervised or semi-supervised ML approaches that required large sets of previously labeled data for training an ML algorithm to perform specific tasks—for example, score student responses to assessments (Zhai et al., 2020; Kaldaras et al., 2022). These traditional ML algorithms focused predominantly on analyzing and interpreting data. The validity of scores produced by these ML models was evaluated by comparing agreement between human and machine-assigned scores. Therefore, human scores have historically been used as a gold standard against which the validity of ML scores was evaluated (Zhai et al., 2020; Kaldaras et al., 2022). Very little work has been done on assessing the validity of ML-based scores beyond human-machine agreement (Kaldaras and Haudek, 2022), which can be considered a criterion-based validity measure.

Generative Artificial Intelligence models are also ML models. In contrast to traditional ML models that use supervised training approaches, GAI models do not require a pre-trained data set to perform tasks. Instead, they have already been trained on all the available data before release. For example, the current version of ChatGPT is trained on all the available data until 2023. It is also possible to conduct additional training of the GAI models to perform specific tasks, which means users may further train these models on a range of examples to help tailor GAI outputs to their desired outcomes.

Unlike traditional ML algorithms trained to interpret and analyze data, GAI models are designed to create novel, original outputs. They are, therefore, more versatile in the range of tasks they can perform. GAI models are promising for evaluating student performance because they do not require large sets of previously scored assessment data. However, we believe training on outputs previously evaluated by humans should be essential to preparing GAI algorithms to perform evaluation tasks. Evaluating validity evidence from multiple sources should also be an integral part of the training process for any GAI used in education. For example, assessing the validity of AI-based scores produced by GAI models is necessary to ensure that these scores are meaningful and can be used for the intended purposes, such as providing feedback to teachers and students. Currently, no such standards exist in the field of education. We further discuss approaches that can be used to evaluate the validity of GAI-based outputs for guiding the learning process.

## Using GAI tools to help define construct proficiency levels

The purpose of any assessment is to measure a student's level of proficiency in a specific construct. Construct refers to an unobserved entity (topic, set of skills, etc.) that is of interest to be measured. The



first step in designing an assessment that measures a given construct is to understand what skills and knowledge reflect proficiency in that construct (Pellegrino et al., 2001; Brown and Wilson, 2011). A construct describing knowledge application is defined by carefully specifying all the aspects of content knowledge and skills that students should demonstrate at various levels of sophistication (Kaldaras et al., 2021a). The process of specifying the skills and knowledge necessary to demonstrate proficiency often results in defining a cognitive model, such as learning progression (LP), that describes a path that learners can follow to develop a higher proficiency on a construct (Duschl and Hamilton, 2011). The main advantage of cognitive models lies in their capability to serve as a roadmap for guiding instruction and adjusting the learning process to the needs of individual learners (Duschl and Hamilton, 2011; Kaldaras and Krajcik, 2024). While cognitive models are incredibly useful, defining and validating cognitive models requires large amounts of data on student performance on assessments that measure the construct (see examples in Kaldaras et al., 2021a, 2023). Obtaining and evaluating enough data to extensively define a cognitive model for a given construct is time and resource consuming. GAI models can be leveraged to identify patterns in large sets of student responses to identify meaningful clusters of response types to help further define proficiency levels of cognitive models.

Further, GAI tools can also be used to generate example responses at varying levels of sophistication in situations where student response data are not available or limited. This capability of GAI tools to potentially streamline the process of defining and validating cognitive models for various constructs has the potential to transform the field of education. In turn, researchers working on validation will evaluate the response clusters identified by GAI and judge the relevance of the GAI-identified patterns for describing proficiency in the construct of interest. Researchers can further engage in iterative cycles to train the GAI algorithms to recognize attributes relevant to the construct of interest. This process will serve a dual purpose: validating the cognitive model and training the GAI algorithm to identify different proficiency levels. The pre-trained GAI model can be used to design assessments and scoring rubrics and evaluate student performance on the assessment with respect to proficiency levels defined by the cognitive model. We further discuss these steps for the relevant validity evidence sources.

## Evaluating validity evidence sources generated using GAI

Below, we discuss how different sources of validity evidence will potentially be impacted by incorporating GAI into the process of test development and evaluation of the test results. We discuss the validity evidence sources outlined in the Standards for Educational and Psychological Testing (Eignor, 2013) including: (1) evidence based on test content; (2) evidence based on response process; (3) evidence based on internal structure; (4) evidence based on relation to other variables; (5) validity generalization. Note that each of these evidence sources is not required in all settings. Instead, support is needed for each proposition that underlies the proposed test interpretation for a specific use (Eignor, 2013). For example, a proposition that the test covers a particular topic may be supported without a proposition that a test predicts a given criterion (Eignor, 2013). However, a more

complex proposition, such as the test covering a particular topic and can be used to make inferences about supporting learners in transitioning to a higher-level understanding (i.e., guide the learning process), requires evidence supporting both parts of this proposition. Suppose GAI is used to generate support for any of the validity evidence sources discussed below. These sources are used to develop the validity argument for the intended use of the test scores in a given setting. In that case, GAI-generated validity evidence sources should also be evaluated to ensure that they meaningfully represent the validity evidence needed to support desired propositions. We will further discuss possible ways of assessing GAI-assisted validity evidence sources for these purposes.

## Evidence based on test content

This type of evidence relates to analyzing the relationship between the test content and the construct it is intended to measure. Obtaining evidence based on test content traditionally involves specifying the test domain that describes in detail all the aspects related to content and skills measured on a test. Next, it involves analysis of the correspondence between the test domain and the test items. This analysis can be done by researchers and expert judgment on the relationship between the test domain and test components. When designing tests that measure and guide student learning this type of evidence relates to alignment—a correspondence between the learning standards (for example, the Next Generation Science Standards) and test content. In this context, evaluating evidence based on test content involves assessing whether the test appropriately measures a set of standards. Educators actively use GAI tools to develop assessment questions for different types of constructs (Gierl and Lai, 2018).

Considering that developing test items is an expensive and time-consuming process, it is highly likely that states and other test development agencies will be using GAI tools to develop test items for measuring various constructs. GAI offers a way to streamline and lower the cost of developing tests for both formative and summative use. There are several ways to gather evidence for the alignment between the GAI-generated assessments and the test domain. For example, a recent study developed an approach that guides alignment among the various standards by reducing the number of potential pairs subject matter experts need to consider when aligning the standards to only those that should be considered due to high semantic overlap (Butterfuss and Doran, 2024). This approach could reduce the time and resources needed to perform content mapping, an essential part of the alignment process.

Further, one might use specific information from the test development process as a basis for GAI prompt generation. For example, test developers often use an evidence-centered design (ECD) approach (Mislevy et al., 2003; Kaldaras et al., 2021a, 2023) in test design. This approach involves carefully specifying an ECD argument that consists of the claim and evidence. Claim reflects what students should be able to do with the knowledge and skills. Evidence provides details on the types of evidence that should be observed in student responses to meet the claim requirements. These evidence statements are used to design assessment questions that probe a specific claim. Defining an ECD argument involves careful consideration of the test domain to improve the alignment between the test domain and the assessment questions. Therefore, using elements from the claim and

the evidence as a basis for GAI prompt generation can improve the alignment between GAI-generated assessments and the test domain. Similarly, suppose there is a cognitive model that is used to guide the test development. Then, the description of the proficiency levels can be used as a basis for GAI prompt generation to guide the development of test questions. In both cases, of course, the resulting GAI output should be evaluated by humans to judge the degree of alignment between the test domain and the GAI-generated test questions. If any misalignments are observed, they should be addressed through prompt generation. Documenting this process can serve as evidence for the validity of the test content of GAI-generated tests.

A study demonstrating the basis for this approach for automatic scoring LP-aligned scientific explanations was conducted by Kaldaras et al. (2022). The study demonstrated how a validated LP and associated ECD arguments can be used to design a rubric for AI-based scoring of LP-aligned scientific explanations that measure knowledge application. This study was conducted with supervised ML, but a similar approach can be used with GAI. Specifically, LPs and ECD arguments can be used as a basis for prompt generation for designing LP-aligned assessment items and scoring rubrics. This is a promising future research avenue considering limited research currently available on GAI-assisted assessment generation.

When there are no ECD arguments or a cognitive model available (which is often the case in classroom instruction settings), one could use previously developed test questions that have been shown to measure the test domain of interest as a basis for GAI prompt generation. One would evaluate the extent to which GAI-generated assessments parallel the sample assessment question and ensure that all the new GAI-generated aspects of the assessments meaningfully align with the targeted construct features. Providing multiple examples of test questions could result in better alignment between GAI-generated assessments and the test domain, but that claim should be further investigated.

## Evidence-based on response process

### Using GAI to identify response process patterns in large samples

Validity evidence based on the response process refers to evaluating whether the test takers engage in the specific cognitive processes intended to be measured by the test. For example, engaging in the process of blended math-science sensemaking (MSS) involves learners demonstrating that they are integrating the relevant math and science domains when answering the test questions (Kaldaras and Wieman, 2023). Theoretical and empirical analysis of the response process provides information about the fit between the theoretical construct and the response process engaged in by test takers. Similarly, when validating a learning progression, evidence based on the response process is evaluated to judge the degree of alignment between the theoretically proposed LP proficiency levels and the actual student responses to items designed to probe those levels. If sufficient evidence is obtained to suggest that student response data support the LP levels, one can claim that the LP-aligned assessment instrument exhibits response process-based validity (Kaldaras et al., 2023). Larger samples of student responses will provide a stronger argument for response process-based validity but are also more time-consuming and expensive to evaluate.

When working with large samples of responses, GAI can be used to help assess this validity evidence by identifying clusters of patterns in student responses. Test developers, in turn, can evaluate these patterns to see if they meaningfully relate to the cognitive processes measured by the test. For example, we are currently exploring ways to evaluate CR assessments aligned to the LP for math-science sensemaking (MSS). We are using the LP as a basis for designing prompts for ChatGPT to evaluate these assessments. We also request that GPT provide a rationale for assigning LP level for each response. Through this process we are discovering that GPT is helpful in identifying specific response patterns that are important to define and incorporate into the prompt and describe in the LP. Therefore, GPT is helping us to further define the LP levels and specify different response process types that students can demonstrate when engaging in MSS at different LP levels.

### Using GAI to suggest response process patterns

When student samples are small or hard to obtain, GAI might be used to generate possible sample student responses to the test questions and provide a way to get preliminary response process-based validity evidence. In this context, careful prompt generation should ensure that GAI does not offer the ideal correct answer but generates the possible answers likely provided by the target student population. One might specify the characteristics and possible prior knowledge of the target student population to ensure that GAI has more information to make more accurate suggestions of how students might respond. For example, one might provide various information sources to the GAI model, such as student grade level, previously covered materials, and student demographics, among other factors. Then, one would investigate the types of potential responses that GAI would suggest and evaluate whether the responses represent the desired LP levels. One should be very careful to explore potential GAI-generated biases inherent in the suggested responses and always aim to check the validity of proposed inferences with an actual sample of data collected from human learners. An example of GAI-generated bias in this context might refer to GAI only suggesting responses with multiple inaccuracies or responses associated with lower proficiency level for specific student populations (for example, specific demographic groups or gender groups). This type of GAI-generated bias poses threat to validity of assessments for these student groups.

Further, failure to account for non-standard language is another example of GAI-generated bias. GAI-based responses should be carefully examined to ensure that these responses contain non-standard language because students often use non-standard language in their responses to provide an accurate account of phenomena. For example, there are multiple ways a student response can reflect understanding of proportional relationships. A student might say that two variables are proportional because they change in proportional amounts with respect to each other. Using normative forms like “proportional” is an example commonly accepted, standard language. However, a student might also describe proportionality without using the term “Proportional” to say something like: Every time variable B changes by 1, variable A changes by 2. This is an example of a non-standard way of describing proportional relationships. If using GAI to generate possible responses for response process-based validity studies, these non-standard ways of arriving to a correct response should be reflected in GAI-generated responses.

## Validity based on scoring process

We believe that a new source of validity evidence needs to be specified in the context of using GAI for assessment. This evidence source is related to response process-based validity but is focused on the GAI scoring process rather than the learner's response process. Specifically, it is becoming increasingly common to use GAI to score student assessments (Baidoo-Anu and Ansah, 2023; Moorhouse et al., 2023; Mao et al., 2024). As discussed above, historically, the validity of AI-based scores has been evaluated using human-machine agreement measures, which is related to criterion-based validity (discussed below). With the emergence of GAI models, much smaller previously labeled data sets might be needed for training the model (but this claim also needs to be further investigated empirically). It is reasonable to suggest that GAI models will need very few or no previously scored student responses to be able to score assessments that exhibit high human-GAI agreement. However, as discussed above, the agreement measures that are evidence of criterion-based validity are not suitable evidence for multiple purposes, such as those required to guide the learning process. Specifically, the agreement measures do not provide validity evidence for evaluating whether the GAI considered the same attributes in student responses to assign specific scores as a human scorer would. However, this information is necessary for ensuring that GAI models score the types of knowledge and skills that indicate knowledge application ability as a human scorer would. Otherwise, the results of GAI-scored assessments cannot be used to support students in developing knowledge application skills. Therefore, we believe it is necessary to introduce a new source of validity evidence that needs to be evaluated. We call it *GAI scoring process-based validity evidence*, which relates to assessing the alignment between human and GAI-scored response features.

This type of validity evidence parallels response process-based validity but emphasizes the need to evaluate whether the non-human scorer uses the same attributes to assign a score as a human scorer. One way to assess the *scoring process-based validity of GAI-produced scores* is to supply the GAI model with a scoring rubric focused on the relevant elements of student responses and ask the GAI to score a sample of student responses using the rubric. Next, one should ask the GAI model to explain why specific scores were assigned based on the provided rubric. This process will allow us to gauge whether the GAI model uses the same criteria for assigning scores. It is also possible to further train the GAI model and improve the scoring process-based validity through careful prompt generation and guiding the model to evaluate specific attributes of interest when scoring student responses. The steps of this method could be presented as evidence for the scoring process-based validity. In the example discussed above, we request that GPT provides a rationale for assigning LP level for each response on MSS LP-aligned assessment. Through this process, we are evaluating whether GPT is using the same rationale for assigning a score as human scorers. In cases when the rationale differs, we proceed by supplying more of the relevant examples and further revising the prompt to help GPT better align to human rationale for score assignment. This process results in improving the theoretical basis for human-GAI agreement driven by the MSS LP and therefore helps improve *scoring process-based validity* of the resulting GAI-produced scores.

In situations with no scoring rubric or LP available, one might ask GAI to develop a rubric based on criteria necessary for evaluating the relevant attributes in student responses. These attributes might

be specified based on prior work on defining the construct of interest. Through careful prompt generation, GAI could be guided in developing a rubric that evaluates all the necessary attributes. This rubric can then be used to score a sample of student responses, and GAI's rationale might be asked to suggest specific scores based on the rubric. This training can serve as evidence for scoring process-based validity of the resulting GAI-based scores.

## Evidence based on internal structure

Evidence based on internal structure pertains to evaluating the degree to which the relationships between items on the test relate to the construct being measured (Eignor, 2013). For example, the cognitive model that guides test development (or any conceptual framework used to design the test) might imply that the test is unidimensional. In this case, evidence should be presented that the test items conform with the theoretically suggested unidimensional structure, which will serve as evidence to suggest that the test measures the construct of interest. Alternatively (or in addition), the cognitive framework might imply that the test items measure different proficiency levels—as in the case of assessments that measure student progress along the LP levels. In these situations, evidence must be presented to show that the items on the test measure various proficiency levels in a way suggested by the LP. Examples of studies on internal latent structure validation include Kaldaras et al. (2021a,b).

If GAI models are leveraged to score assessments, evaluation of internal structure reflected in GAI-based assessment scores should also be evaluated. This could be done by applying traditional methods for assessing internal latent structure—such as latent variable modeling approaches like confirmatory and exploratory factor analysis—to GAI-produced scores to ensure that they reflect the same latent structure as human-based scores. A sample study focused on evaluating the internal latent structure of ML-generated scores was done by Kaldaras and Haudek (2022). In this study, the authors used confirmatory factor analysis to gauge the similarity between the item difficulty parameters produced using human and machine-generated scores. This approach allowed authors to identify specific items and LP levels that exhibited significant discrepancies between human and machine-assigned scores. This led to considerably different values estimated for the difficulty parameters. These results help further investigate where the AI-based scores approximate the same latent structure for a given assessment instrument, what discrepancies occur, and for which items, which is an essential aspect of the internal structure-based validation process. While this study was performed using supervised ML-based scores, similar studies can also be conducted using GAI-based scores. The CFA analysis can be easily performed using standard statistical packages such as SPSS, Lavan package for R or MPlus.

Some studies of the internal latent structure are also designed to show whether items function differently with different student populations (racial, ethnic, or gender subgroups). In this context, differential item functioning (DIF) might indicate multidimensionality that might or might not be desirable based on the framework used to guide the test development. Suppose GAI models are leveraged to score assessments. In that case, differential item functions in GAI-based assessment scores should also be evaluated to ensure that the DIF does not result from biases inherent in the GAI models.



Further, research has been done to employ machine learning approaches for identifying differential item functioning on previously designed assessments (Hoover, 2022). This study represents a promising approach for employing AI-based approaches for evaluation of DIF in various contexts using existing items and student responses. Specifically, it is important to distinguish between statistical bias (meaning a biased estimator) and the bias that reflects the influence of unintended characteristics of the examinee. Generally, if severe DIF is detected and it is not related to the target construct, the items are not used. Further studies should be conducted to refine this approach for use in practice. Analysis of DIF can be performed using SPSS, R, Mplus, Stat and SAS among others.

Another strategy for examining the latent structure of scores from both human and AI sources is employing multi-group confirmatory factor analysis (MG-CFA), which allows researchers determine whether the factor structure is different due to a scoring approach (human vs. AI) (Asparouhov and Muthén, 2014). Multigroup CFA should be done using the same set of items. The assumption is that the latent structure should stay invariant, irrespective of who (GAI or humans) scored the items. Evaluating the latent structure invariance across the two set of scores (human and AI-based) will allow to evaluate the validity of internal structure of GAI-based scores.

## Evidence based on relation to other variables (test criterion and beyond)

In many situations, the intended interpretation for a given use of test scores implies that the construct should be related to other variables, which in turn requires a careful analysis of the relationship of the test scores to external variables. Since ML models have historically been used to perform the work of a human scorer, it is not surprising that the most common source of validity evidence evaluated for ML-based scores is various measures of human-machine agreement. In this context, the human scores were considered the gold standard against which the performance of the ML algorithm was evaluated. Historically, supervised ML approaches that required pre-training using previously scored data sets have been using various methods for assessing human-machine agreement (see Zhai et al., 2020 for detailed review). Previously described approaches include using the same data to train and evaluate the performance of ML algorithm (self-validation), splitting the data set into a training and testing sets (split-validation), and splitting the data set into  $n$  subsets each subset is used to train the ML algorithm while other subsets are used as a testing set to validate the model accuracy (cross-validation). Similar approaches could be used with GAI models, and likely, GAI models will require much smaller data sets (although this suggestion remains to be tested empirically).

## Holding GAI to the same scoring standards as human scorers

As discussed above, supervised ML scores are usually compared to human scores, therefore establishing human scores as “the gold standard.” However, this is not always the case, since it takes significant effort to ensure high-quality human scores. Unless properly trained, one should not assume that humans give you valid inferences about students. Research has shown that humans are biased toward longer than shorter responses, and therefore, human scores also represent

nonperfect criteria. So, can an AI-based scoring system with nonperfect criteria be better than nonperfect criteria (human scores)? It makes sense to hold ML algorithms to the same standards as humans and evaluate these algorithms according to similar training criteria. In other words, we should replicate what is being done to train humans to replicate the high-stakes training of humans. We can refer to literature on training people to score open-ended assessments and try to replicate that process with machine algorithms. For example, seeding in previously scored responses into the scoring process to see if people are drifting away on the scoring process, then retrain them if they drift away too far—the same can be done with ML algorithms.

Further, we could build on the previously discussed split-validation method and combine it with purposeful manipulation of the training sets to study the outcomes. For example, the training sets could be selected to have responses with the same score as the human raters. In that case, the training set does not have incorrect responses (but it will have variations in student work); the training set will get perfect results.

## Training AI algorithms to recognize diversity of human thinking

The amount of variation in the training set and what one chooses to vary in the training set will also affect the ML algorithm. Manipulating composition and variability in responses in the training sets can provide insight into how well a given algorithm picks up on the diversity of human thinking. This could provide evidence for the validity of AI-based scores for different types of reasoning represented in student responses. It can also help establish criteria on how much variation one needs to have in the training set to train a given algorithm to pick up consistently on this variation. For example, classifying learners into LP levels calls for having a rich distribution at all LP levels. Continuing the example with leveraging ChatGPT model for scoring student responses of MSS LP-aligned items, we discovered that having at least one representative example of student response for each type of reasoning is needed by GPT in order to assign a score is necessary for achieving high GAI-human agreement.

Another example is bias: if you have responses that are good but have a lot of spelling errors that are still given high scores, the ML algorithm might not score those properly. In fact, when using GAI for scoring MSS LP-aligned assessments, we discovered the non-standard language to be the central issue in producing mis-scores at higher levels of the LP. Specifically, we discovered that responses consistent with sophisticated reasoning but use non-standard language or show evidence of responders being non-native English speakers tended to be mis-scored by GAI to lower LP levels. A possible way of dealing with this shortcoming might be developing a vocabulary for non-standard language and using this vocabulary as part of the GAI training process. Further, training data sets can also be selected to minimize possible DIF or cultural bias or to test for sensitivity to irrelevant features of responses.

## Evaluating consistency of GAI-based scores

Another way of dealing with potential misscores due to various reasons might be to ask GAI to produce scores on the same data set several times to evaluate consistency of GAI-generated scores across trials. If the GAI-based scores are consistent across the trials and agree with human-based scores, this provides evidence of criterion-based validity. The GAI-based scores that are inconsistent and disagree with



human-assigned scores should be further examined to explore how the prompt might be changed to achieve better criterion-based validity. In the project on GAI scoring of MSS LP-aligned assessments, we used this approach by asking GAI to score each response three times and comparing the produced scores to human scores on the same items. In cases where we saw disagreement, we discovered that additional prompt revisions were needed to clarify the scoring approach for GPT, and better agreement was achieved as a result.

Further, one could incorporate a feature that would allow it to stop the GAI algorithm when it encounters an outlier with non-standard language or a response that is scored inconsistently across trials. All GAI-based scores should have a level of confidence, and people should be critical when interpreting the scores. The accuracy of the information is only as good as the training set. Also, one could use multiple GAI algorithms to validate and inform the validation of each other (like confirmatory and exploratory factor analysis). For example, one could use one GAI algorithm to identify patterns in a given data set (GAI 1). At the same time, one would use another GAI algorithm (GAI 2) to score the same data set and assign scores. The GAI 1 will find clusters that are considered similar by that algorithm, so they will get the same score once you score them. Compare the clusters identified by GAI 1 with scores assigned by GAI 2. Seeing how the results match up to GAI 2—ask GAI 1, what does it take to get this score? Does the response make sense? Evaluate the differences between the two algorithms to see how valid and consistent the scoring outcomes are with respect to scoring the construct of interest.

## Validity generalization

An important issue in educational settings is the degree to which the validity evidence based on test-criterion relation can be generalized to new situations without further studies on validity in those new situations. This point is critical considering the push toward a wide use of GAI algorithms in educational settings for assessment purposes and beyond. When investigating the generalizability of GAI-based scores, it is essential to study to what extent GAI-based outputs generalize to situations beyond a given study or context. For example, when GAI models are used to predict scores on the same assessment items used in the original validation study, the generalizability and prediction accuracy will likely be very high. However, suppose the assessment items closely resemble those used in the validation study or are entirely different but assess a similar or the same construct. In that case, the behavior of GAI models needs to be further studied to investigate how well these models predict student performance on such assessments. Approaches such as those discussed in the previous section could be used to study the performance of GAI algorithms with new sets of student responses or with different but closely related items. Evidence gathered on the performance of these algorithms under these various circumstances could serve as evidence of generalizability.

Further, it is important to consider the drawbacks of GAI algorithms, such as hallucinations and AI drifting, in the context of generalizability studies. The problem of AI hallucinations refers to AI providing incorrect predictions that may occur even after training. AI drifting refers to situations where the accuracy of predictions produced from new input values “drifts” away from the performance during the training period. These drawbacks suggest that the outputs of GAI models should be periodically monitored

and checked even after the GAI model has been released for use by the public to ensure that such drifting or incorrect predictions do not occur.

## Discussion

As mentioned at the beginning of this paper, no test can ever be fully validated. Instead, a sound validity argument integrates various sources of validity evidence into a coherent account of the degree to which the available evidence and theory support the intended interpretation of the test results for specific uses. As such, a validity argument should incorporate multiple sources of validity evidence from multiple studies and show how the findings align with previously reported results if available. Validation is an iterative process that might involve revisions in the test, the associated rubrics, and the definition of the underlying latent construct. In theory, the validation process never ends as there is always additional information that can be gathered to understand the construct, the test, and the inferences that can be made more fully from the test. However, in practice, at some point, the validation process aimed to support evidence for the intended interpretation of the test results must end at least till new evidence emerges that would question the previous validity inferences in some way. The amount and type of evidence required to support specific inferences depends on many factors, including the type and goals of the test, knowledge domains, and topic advances. Higher stakes require higher evidence standards.

In the context of using GAI for various validation purposes discussed in this paper, it is essential to recognize that GAI is a continuously evolving field. This important feature of GAI algorithms has implications for the validity studies conducted with the help of GAI. For example, the GAI models are constantly learning new information, improving their overall accuracy, and increasing the range of tasks they can successfully perform. This implies that GAI algorithms can potentially identify certain instances (for example, patterns in student responses) that do not align well with the previously validated construct. In these cases, GAI can provide additional evidence requiring possible refinement of the construct definition and changes to the associated test items and rubrics. Further validation studies might be needed to support the inferences desired to be made from the test. This is just one example, and other implications might be possible because of the evolving nature of the GAI models.

In addition, very little is known about the long and even short-term effects of using GAI algorithms to solve various problems in education. This has significant consequences for validity studies conducted with the help of GAI as well. For example, considering that GAI algorithms can exhibit drifting and hallucinations (discussed above), it is essential to ensure that GAI algorithms are producing consistently accurate and reliable results in the long run. This might require constant monitoring by humans to evaluate the validity of GAI outputs for specific purposes. This is especially important if these GAI algorithms will be used to guide multiple aspects of the learning process, including aiding in assessment design and evaluation for both summative and formative purposes, and adjusting the learning process based on the results of these assessments. In each of these cases, sufficient evidence needs to be presented that GAI-based outputs produce accurate and reliable outputs and that these outputs can be used to make the desired decisions about the learning process.

Related to the previous point, it is important to consider the unintended consequences of using GAI-produced outputs to guide the educational process. In this context, the issue of bias is important to consider. As mentioned above, it is important to distinguish between statistical bias in estimation and bias in items and scoring due to influences other than the target construct. Importantly, identifying bias is always a challenge because “bias free” criterion is needed for comparison. For example, in a typical DIF study, it is assumed that most items are unbiased, so the scores from those items can be used to identify potential bias in studied items. The same would need to be true for the study of bias in GAI scoring or test development. In a sense, there needs to be a “bias free” training set so that bias can be detected when the training set is not bias free. This is especially important since GAI algorithms are being trained on large amounts of various types of human-generated data, it is important to consider the biases that could be present in the data and, therefore, become inherent in the GAI algorithms as a result. It is important to investigate the presence of these biases and their potential effects on interpreting the test results. For example, as discussed above, one should investigate potential biases of GAI algorithms based on gender background (including academic, ethnic, racial, and linguistic, among others) and their effect on GAI model outputs as well as the unintended consequences of those outputs when it relates to the interpretation of assessment results for specific purposes. For example, recent studies have shown that human and machine-based scores exhibit similar amounts of bias and suggested that diverse groups of human experts should be used to evaluate the presence of potential biases (Belzak et al., 2023). We also believe that while GAI can perform many of the tasks outlined above, the end judge of the validity of GAI actions should always be humans.

We hope that the discussion points provided in this short paper can serve as a basis for starting the conversation about establishing the

standards for validity in the era of widespread use of GAI in education and educational evaluation.

## Author contributions

LK: Conceptualization, Writing – original draft. HA: Conceptualization, Writing – review & editing. MR: Conceptualization, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The publication fee for this article was covered by the University of Colorado Boulder.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

- Asparouhov, T., and Muthén, B. (2014). Multiple-group factor analysis alignment. *Struct. Equ. Model. Multidiscip. J.* 21, 495–508. doi: 10.1080/10705511.2014.919210
- Baidoo-Anu, D., and Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. *J. AI* 7, 52–62. doi: 10.61969/jai.1337500
- Belzak, W. C., Naismith, B., and Burstein, J. (2023). “Ensuring fairness of human-and AI-generated test items” in *International Conference on Artificial Intelligence in Education*. Springer Nature Switzerland, Cham. 701–707.
- Brown, N. J., and Wilson, M. (2011). A model of cognition: the missing cornerstone of assessment. *Educ. Psychol. Rev.* 23, 221–234. doi: 10.1007/s10648-011-9161-z
- Butterfuss, R., and Doran, H. (2024). An application of text embeddings to support alignment of educational content standards. Paper Presented at Generative Artificial Intelligence for Measurement and Education Meeting. Available at: <https://hdoran.github.io/Blog/ContentMapping.pdf>
- Duschl, R., and Hamilton, R. (2011). “Learning science” in *Handbook of Research on Learning and Instruction*. Eds. R. E. Mayer and P. A. Alexander (New York: Routledge), 92–121.
- Eignor, D. R. (2013). “The standards for educational and psychological testing” in *APA Handbook of Testing and Assessment in Psychology*, Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology. eds. K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel and S. P. Reise et al. (Washington D.C.: American Psychological Association), 245–250.
- Finnish National Board of Education (FNBE) (2015). National core curriculum for general upper secondary schools 2015. Helsinki, Finland: Finnish National Board of Education (FNBE). Available at: [http://www.oph.fi/saadokset\\_ja\\_ohjeet/opetusuunnitelmien\\_ja\\_tutkintojen\\_perusteet/lukiokoulutus/lopps2016/103/0/lukion\\_opetusuunnitelman\\_perusteet\\_2015](http://www.oph.fi/saadokset_ja_ohjeet/opetusuunnitelmien_ja_tutkintojen_perusteet/lukiokoulutus/lopps2016/103/0/lukion_opetusuunnitelman_perusteet_2015)
- Gierl, M. J., and Lai, H. (2018). Using automatic item generation to create solutions and rationales for computerized formative testing. *Appl. Psychol. Meas.* 42, 42–57. doi: 10.1177/0146621617726788
- Hoover, J. C. (2022). Using machine learning to identify causes of differential item functioning. Doctoral dissertation. University of Kansas.
- Kaldaras, L., Akaze, H., and Krajcik, J. (2021b). A methodology for determining and validating latent factor dimensionality of complex multi-factor science constructs measuring knowledge-in-use. *Educ. Assess.* 26, 241–263. doi: 10.1080/10627197.2021.1971966
- Kaldaras, L., Akaze, H., and Krajcik, J. (2021a). Developing and validating next generation science standards-aligned learning progression to track three-dimensional learning of electrical interactions in high school physical science. *J. Res. Sci. Teach.* 58, 589–618. doi: 10.1002/tea.21672
- Kaldaras, L., Akaze, H. O., and Krajcik, J. (2023). Developing and validating a next generation science standards-aligned construct map for chemical bonding from the energy and force perspective. *J. Res. Sci. Teach.* 1–38. doi: 10.1002/tea.21906
- Kaldaras, L., and Haudek, K. C. (2022). Validation of automated scoring for learning progression-aligned next generation science standards performance assessments. *Front. Educ.* 7:968289. doi: 10.3389/feduc.2022.968289
- Kaldaras, L., and Krajcik, J. (2024). “Development and validation of knowledge-in-use learning progressions” in *Handbook of Research on Science Learning Progressions*. Eds. H. Jin, D. Yan and J. Krajcik (New York: Routledge). pp. 70–87.
- Kaldaras, L., and Wieman, C. (2023). Cognitive framework for blended mathematical sensemaking in science. *Int. J. STEM Educ.* 10, 1–25. doi: 10.1186/s40594-023-00409-8

- Kaldaras, L., Yoshida, N. R., and Haudek, K. C. (2022). Rubric development for AI-enabled scoring of three-dimensional constructed-response assessment aligned to NGSS learning progression. *Front. Educ.* 7:983055. doi: 10.3389/feduc.2022.983055
- Krajcik, J. S. (2021). Commentary—applying machine learning in science assessment: opportunity and challenges. *J. Sci. Educ. Technol.* 30, 313–318. doi: 10.1007/s10956-021-09902-7
- Kulgemeyer, C., and Schecker, H. (2014). Research on educational standards in German science education—toward a model of student competences EURASIA. *J. Math. Sci. Technol. Educ.* 10, 257–269. doi: 10.12973/eurasia.2014.1081a
- Mao, J., Chen, B., and Liu, J. C. (2024). Generative artificial intelligence in education and its implications for assessment. *TechTrends* 68, 58–66. doi: 10.1007/s11528-023-00911-4
- Messick, S. (1980). Test validity and the ethics of assessment. *Am. Psychol.* 35, 1012–1027. doi: 10.1037/0003-066X.35.11.1012
- Ministry of Education, P. R. China (2018). Curriculum Plan for Senior High School. Beijing: People's Education Press.
- Mislevy, R. J., Almond, R. G., and Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Res. Rep. Ser.* 2003, i–29. doi: 10.1002/j.2333-8504.2003.tb01908.x
- Moorhouse, B. L., Yeo, M. A., and Wan, Y. (2023). Generative AI tools and assessment: guidelines of the world's top-ranking universities. *Comput. Educ. Open* 5:100151. doi: 10.1016/j.cao.2023.100151
- National Assessment Governing Board (2023). Approves an Updated Science Framework for the 2028 Nation's Report Card. Available at: <https://www.nagb.gov/news-and-events/news-releases/2023/updated-science-framework-2028.html>
- National Research Council (2012). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Washington, DC: National Academies Press.
- NGSS Lead States (2013). Next Generation Science Standards: For States, By States. Washington, DC: The National Academies Press.
- OECD (2016). PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy. Paris: OECD Publishing.
- Pellegrino, J. W., Chudowsky, N., and Glaser, R. (2001). Knowing What Students Know: The Science and Design of Educational Assessment. Washington, DC: National Academy Press.
- Samala, A. D., Zhai, X., Aoki, K., Bojic, L., and Zikic, S. (2024). An in-depth review of ChatGPT's pros and cons for learning and teaching in education. *Int. J. Interact. Mob. Technol.* 18, 96–117. doi: 10.3991/ijim.v18i02.46509
- Yao, J. X., and Guo, Y. Y. (2018). Core competences and scientific literacy: the recent reform of the school science curriculum in China. *Int. J. Sci. Educ.* 40, 1913–1933. doi: 10.1080/09500693.2018.1514544
- Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., and Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Stud. Sci. Educ.* 56, 111–151. doi: 10.1080/03057267.2020.1735757



## OPEN ACCESS

## EDITED BY

Georgios Leontidis,  
University of Aberdeen, United Kingdom

## REVIEWED BY

Michael Flor,  
Educational Testing Service, United States  
Antonio Sarasa-Cabezuelo,  
Complutense University of Madrid, Spain

## \*CORRESPONDENCE

Fabian Kieser  
✉ kieser@ph-heidelberg.de

RECEIVED 15 April 2024

ACCEPTED 30 August 2024

PUBLISHED 18 September 2024

## CITATION

Kieser F, Tschisgale P, Rauh S, Bai X, Maus H,  
Petersen S, Stede M, Neumann K and Wulff P  
(2024) David vs. Goliath: comparing  
conventional machine learning and a large  
language model for assessing students'  
concept use in a physics problem.  
*Front. Artif. Intell.* 7:1408817.  
doi: 10.3389/frai.2024.1408817

## COPYRIGHT

© 2024 Kieser, Tschisgale, Rauh, Bai, Maus,  
Petersen, Stede, Neumann and Wulff. This is  
an open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](#). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic practice.  
No use, distribution or reproduction is  
permitted which does not comply with these  
terms.

# David vs. Goliath: comparing conventional machine learning and a large language model for assessing students' concept use in a physics problem

Fabian Kieser<sup>1\*</sup>, Paul Tschisgale<sup>2</sup>, Sophia Rauh<sup>3</sup>, Xiaoyu Bai<sup>3</sup>,  
Holger Maus<sup>2</sup>, Stefan Petersen<sup>2</sup>, Manfred Stede<sup>3</sup>,  
Knut Neumann<sup>2</sup> and Peter Wulff<sup>1</sup>

<sup>1</sup>Physics and Physics Education Research, Heidelberg University of Education, Heidelberg, Germany,

<sup>2</sup>Department of Physics Education, Leibniz Institute for Science and Mathematics Education, Kiel,

Germany, <sup>3</sup>Applied Computational Linguistics, University of Potsdam, Potsdam, Germany

Large language models have been shown to excel in many different tasks across disciplines and research sites. They provide novel opportunities to enhance educational research and instruction in different ways such as assessment. However, these methods have also been shown to have fundamental limitations. These relate, among others, to hallucinating knowledge, explainability of model decisions, and resource expenditure. As such, more conventional machine learning algorithms might be more convenient for specific research problems because they allow researchers more control over their research. Yet, the circumstances in which either conventional machine learning or large language models are preferable choices are not well understood. This study seeks to answer the question to what extent either conventional machine learning algorithms or a recently advanced large language model performs better in assessing students' concept use in a physics problem-solving task. We found that conventional machine learning algorithms in combination outperformed the large language model. Model decisions were then analyzed via closer examination of the models' classifications. We conclude that in specific contexts, conventional machine learning can supplement large language models, especially when labeled data is available.

## KEYWORDS

large language models, machine learning, natural language processing, problem solving, explainable AI

## 1 Introduction

The introduction of ChatGPT, a conversational artificial intelligence (AI)-based bot, to the public in November 2022 directed attention to large language models (LLMs). As of 2023, ChatGPT is based on a LLM called Generative Pre-trained Transformer (GPT; versions 3.5, 4, 4V, or 4o) and has proven to perform surprisingly well on a wide range of different tasks in various disciplines—including medicine, law, economics, mathematics, chemistry, and physics (Hallal et al., 2023; West, 2023; Surameery and Shakor, 2023; Sinha et al., 2023). A number of tasks in education (research) can be tackled using LLMs in general, or ChatGPT more specifically. For example, LLMs were found to be able to write



quality essays in physics (Yeadon et al., 2023), simulate student preconceptions for physics concepts (Kieser et al., 2023), write reflections in nursing education contexts (Li et al., 2023), and even generate feedback that was considered equally correct and more helpful by the students compared to human expert feedback (Wan and Chen, 2024). In particular, prompt engineering with LLMs (i.e., specifically designing the inputs to the LLM) was found to notably improve the capabilities and quality of outputs, to even become so-called “zero-shot reasoners” (Wan and Chen, 2024; Kojima et al., 2022).

The use of LLMs in education (research) is, however, not without challenges. When confronted with conceptual questions, LLMs may hallucinate knowledge (i.e., present false information as facts) (Huang et al., 2023), which is then concealed by its fluent language and verbose writing style (Gregorcic and Pendrill, 2023). This issue is exacerbated by the intransparency of the decisions made by LLMs (Chen et al., 2023; Manning, 2022). Intransparency in the decision-making process of an LLM may prevent researchers from understanding the logic behind a prediction, and thus hinder them from justifying their choices for certain LLMs. LLMs also exhibit human-like biases through imbalanced training data; and the extent to which LLMs truly extrapolate beyond their training data or merely mimic patterns—in the sense of “stochastic parrots” (Bender et al., 2021; Caliskan et al., 2017; Lake and Baroni, 2023)—remains an open question. Many examples demonstrate that LLMs such as GPT-4 cannot sufficiently abstract and reason (Mitchell et al., 2023). Finally, the extensive use of LLMs significantly contributes to environmental concerns, particularly in terms of CO<sub>2</sub> emissions and expenditure of energy, both by training the foundation models and with every single request passed through the model (de Vries, 2023; Dodge et al., 2022).

Consequently, we argue that the circumstances where machine learning (ML) and LLMs excel respectively should be critically evaluated to derive some guidance for researchers and practitioners. Conventional AI approaches (i.e., ML algorithms) are less complex and their decisions can commonly be explained using established procedures (Lundberg et al., 2019). Given their reduced complexity, conventional ML algorithms can be operated in a controlled manner and might not generate unanticipated outputs. For example, a trained binary classifier can by design only output two categories, whereas generative LLMs used in a binary classification problem might output the categories, however, it might also produce further textual output. Whether conventional ML or LLMs are used for solving a (research) problem in part depends on the complexity of the problem. In some contexts, e.g., fourth-grade mathematics, it was found that conventional ML can outperform LLMs on identifying incoherent student answers (Urrutia and Araya, 2023). However, this research considered the LLM GPT-3, which is now surpassed by GPT-4. Given the specific potentials and limitations with either conventional ML and LLMs it remains an open question what approach should be utilized under which circumstances.

Given the successes of conventional ML such as explainability of model decisions as well as the limitations such as the ability to tackle complex problems, and the recent advances of LLMs with their “emergent abilities” (Wei et al., 2022) and zero-shot reasoning capabilities, this study compares the performances of conventional ML algorithms and a recent LLM on a physics-specific

assessment problem. Our goal is to refine our understanding of the circumstances under which either conventional ML algorithms or LLMs might be better suited solutions.

## 2 Theoretical background

### 2.1 Natural language processing with conventional ML and LLMs

Language data such as students’ written responses, interview transcripts, or research articles is omnipresent in educational research, and therefore integral for theory development. Educational research often draws on content analysis as an analytical method to analyze language data. One major task in content analysis is to develop categories for certain events occurring in the language data to be analyzed, such as a student using a certain concept in an interview (transcript). The actual assignment of codes to the content is guided by a coding manual that specifies the rules for when a category applies or not (Mayring, 2000; Krüger et al., 2014). Content analysis, in particular the process of developing and assigning codes, is often very time-consuming, thus limiting the amount of content (e.g., interview data) that can be analyzed. This leads to methodological constraints. For any given language, there is a set of words that frequently appears in texts, yet a much larger number of words occurs only rarely (Newman, 2005; Wulff, 2023). Hence, rare occurrences also appear only in large text corpora, making it generally insufficient to analyze only small samples to validly identify underlying patterns in textual data. Similarly, the decisions and subjective judgments of researchers involved in the analysis process can pose challenges in validating and reproducing the results of qualitative analyses (Biernacki, 2014).

Natural Language Processing (NLP) enables the use of new statistical approaches (often based on ML) to systematically analyze large data sets that are no longer analyzable by humans alone. A powerful tool that was developed by NLP researchers are word and sentence vectors, also referred to as embeddings, which can then be further processed, e.g., by ML algorithms. In the simplest case, one can use so-called “bag-of-words” models that list all words in a document and their frequencies of occurrence while omitting positional information (Zhang et al., 2010). Limitations of ‘bag-of-words’ models include a missing measure of similarity between individual words as these models do not consider the particular meaning of words, and they do not consider word order. To address these limitations, artificial neural networks were trained with the aim to transform textual input into (static) embeddings, i.e., numerical vectors of generally high dimensionality, that incorporate contextual information of individual words or sentences (Mikolov et al., 2013). These embedding vectors can then be used as input features for ML algorithms in further downstream tasks. ML refers to the inductive learning of patterns from data (Rauf, 2021). Various ML techniques, such as clustering or classification, can be applied based on these embedding vectors. In early NLP research, oftentimes conventional ML approaches such as logistic regressions or decision trees were utilized to build these classifiers (Jurafsky and Martin, 2014; Manning, 2022). Despite the simplicity of these models, particularly

with regard to clustering, good results can also be achieved in difficult tasks such as argumentation mining (Stede and Schneider, 2019) or classification of elements of problem-solving approaches (Tschisgale et al., 2023a).

Recently, significant advancements in the field of NLP have occurred through the training of LLMs. In contrast to simple “bag-of-words” models that merely capture word frequencies in documents, and static embedding vectors, LLMs are able to more dynamically encode and also generate language. LLMs can process textual data at a much deeper level by quantifying relationships between words (often based on co-occurrence in large training corpora). The foundation for these advancements lies in a specific artificial neural network architecture called transformers (Devlin et al., 2018; Vaswani et al., 2017) that are trained on extensive textual data. Transformers brought along a vast variety of different models (Amatriain et al., 2023), such as Bi-direction Encoder Representations for Transformers (BERT) or Generative Pre-trained Transformers (GPT). The training of transformer LLMs typically involves prediction of randomly omitted words from a given sequence of context words. Surprisingly, this relatively simple training objective enabled transformers to perform well on new tasks that were not included within the training phase especially if the LLM is also given some examples (few-shot learning) (Brown et al., 2020). Two paradigms of application are differentiated: (i) fine-tuning, i.e., the LLM is trained with labeled data to perform a task, and (ii) prompting, i.e., huge-size language models (also called foundation models) are given a few examples with blanks for the model to fill in (few-shot or zero-shot learning) (Zhao et al., 2023).

Among the most widely used, popular, and performative transformer models is the Generative Pre-trained Transformers (GPT) family developed by OpenAI (Achiam et al., 2023). As a generative transformer model, GPT relies on continuing an input string, a so-called prompt. Manipulating this prompt to achieve desirable outputs is termed prompt engineering (i.e., adding specific information to an input to influence the output) and prompt chaining (i.e., concatenating subsequent prompts and outputs to align the new outputs with the flow of conversation and incorporating prior information), and was found to enable researchers to utilize GPT models specifically for their research purposes (Liu et al., 2021; White et al., 2023). One well-known application (an assistant model, Zhao et al., 2023) of GPT models is ChatGPT, a chatbot based on the GPT-3.5 (and later the GPT-4, and GPT-4V with vision capabilities) architecture (Bubeck et al., 2023). ChatGPT was particularly trained with human feedback and prompt-response pairs to enable conversational turns. It has been shown that this fine-tuning improves the performance of LLMs in various tasks (Wei et al., 2021). ChatGPT has also made an impact in the field of education (Kasneci et al., 2023), particularly in the field of physics education (Kortemeyer, 2023; West, 2023).

## 2.2 ChatGPT in physics education

A growing number of studies in physics education explored the potential of ChatGPT to solve physics problems. Some of these studies suggested that ChatGPT is unreliable in terms of the accuracy of its answers and that inconsistencies also occur within its

reasoning chains (Gregorcic and Pendrill, 2023; dos Santos, 2023). However, it is argued that this apparent weakness of ChatGPT in answering physics questions can be utilized as a learning experience to promote critical thinking skills among students (Bitzenbauer, 2023). Other studies have tested the ability of ChatGPT (varying between GPT-3.5 and GPT-4) to solve multiple-choice physics questions. One of these studies found that ChatGPT was able to correctly answer 22 out of 23 questions from the well-known “Force-Concept-Inventory” (West, 2023). Kieser et al. even found that GPT-4 is capable of mimicking various student preconceptions known from physics education research when prompted to answer the “Force-Concept-Inventory”. This opens up new possibilities for the application of ChatGPT, including augmenting data sets by adding simulated (i.e., synthetic) student responses (Kieser et al., 2023). Another possibility was examined by Küchemann et al. (2023) in a randomized controlled study comparing the characteristics and quality of physics tasks created by prospective physics teachers who used either ChatGPT or a textbook as a tool. Küchemann et al. (2023) found that students in both groups faced challenges in providing all the information necessary for solving the tasks. Moreover, the authors noted that prospective physics teachers used the tasks as provided by ChatGPT without modification in 76% of cases (Küchemann et al., 2023). Krupp et al. (2023) identified various strategies for utilizing ChatGPT as an aid in solving physics problems and obtained a result similar to that of Küchemann et al. (2023). More specifically, they found that students often employed copy-and-paste techniques and accepted the solutions presented by ChatGPT without critical reflection (Krupp et al., 2023).

Wan and Chen (2024) conducted a study on the use of ChatGPT (based on GPT-3.5) to provide feedback on students’ written responses to conceptual physics questions. They utilized prompt engineering and few-shot learning techniques. Their findings indicate that ChatGPT can serve as an effective tool for generating feedback based on students’ responses. Even with a relatively small number of examples in training, it is possible to use LLMs through specific prompting to significantly reduce the instructor’s effort required for evaluating student responses (Wan and Chen, 2024). However, LLMs may not always be the best choice for computer-assisted assessment of student responses. Urrutia and Araya (2023) found that conventional ML algorithms were more effective than LLMs when examining text-based responses from fourth-grade students to mathematics tasks. Moreover, LLMs have been critiqued for taxing the environment in unprecedented ways regarding average energy expenditure (de Vries, 2023). Also, it is difficult to explain LLMs’ decisions, e.g., the generated text of a generative LLM such as GPT. Given the size of a LLM’s training corpus, the size of the LLM itself (i.e., its number of its hyperparameters), and the complexity of the training process, researchers have not come up with simple ways of inspecting and explaining the generated outputs. In contrast, conventional ML algorithms such as decision trees are much easier to explain and hence control (Lundberg et al., 2019).

In sum, LLMs are quite capable tools that can be used for many applications. However, they do not appear to be silver bullets, given their tendency to hallucinate, i.e., to present false information [ranging from 3 to 29 percent of the time, even in innocuous tasks such as textual summarization (Hughes, 2023)],

and their intransparency. Conventional ML might sometimes be more advantageous. However, this is unclear for rather complex tasks, e.g., those related to physics problem solving where learners have to utilize physics concepts to solve intricate problems.

## 2.3 Physics problem solving

Physics-specific problem-solving abilities are essential for students who intend to study physics and later plan to engage in a physics-related career (Armour-Garb, 2017; Mulvey and Pold, 2020; Jang, 2015). However, students' problem-solving abilities were found to be rather poorly developed, even those of students interested in science (Docktor et al., 2015; Kim and Pak, 2002). To improve students' problem-solving abilities, explicit instruction that reflects problem-solving processes proved effective (Huffman, 1997; Gaigher et al., 2007; Mason and Singh, 2010). There exist a variety of problem-solving process models (e.g., Polya, 1945; Friege, 2001, however, they all share similar phases, among them the phase of *problem representation*. Representing a given problem from a physics perspective involves identifying relevant physics concepts as well as making simplifying assumptions and idealizations. Having constructed an adequate and convenient problem representation comprises among the most important phases in physics problem solving as it determines the solution approach. Hence, the problem representation is often regarded as the crucial phase in problem solving (Savelsbergh et al., 1997; Fortus, 2008). In science domains, however, students often lack a thorough understanding of central concepts which is necessary for a useful problem representation in particular and for successful problem solving in general (Kim and Pak, 2002; Docktor et al., 2015; Hsu et al., 2004; Leonard et al., 1996). A potential reason for this might be that school instruction more often focuses on mathematical routines instead of conceptual understanding (Mulhall and Gunstone, 2012; Gerace and Beatty, 2005).

Students with less developed problem-solving abilities profit from short guidance during the problem-representation phase that helps making the problem representation more coherent and consistent (Savelsbergh et al., 1997). In order to do so, the current state of students' problem representations needs to be assessed. Considering a typical school class consisting of about thirty students and one teacher, or a decentralized learning setting (e.g., online), providing timely feedback on each student's problem representation turns out to be an impossible task for the teacher. However, if these problem representations are available in textual form, NLP and ML methods can be used to automatically assess students' problem representations and provide adaptive feedback in the form of short prompts to improve them. In general, such computer-based feedback was shown to be effective for students' learning in various settings (Graesser et al., 2018; VanLehn, 2011; Bernius et al., 2022).

Timely assessment of a large number of problem representations in textual form is daunting for teachers. Generally, students' problem representations can be regarded as well-structured in the sense that there is a limited number of particular physics concepts that ought to be included in order to make sense of a physics situation. However, describing such physics

concepts in natural language may be difficult for students since language can be ambiguous, particularly the technical language of physics. For example, students could use their everyday language to circumscribe a correct physics concept (Yore and Treagust, 2006), however, students' language use could impede identifying whether the concept was used correctly or used at all.

## 2.4 The present study

Even though LLMs were found to be valuable tools, for example within physics education (West, 2023; Kieser et al., 2023), they did not excel in all tasks, particularly those that require refined conceptual knowledge or abstraction and reasoning (Gregorcic and Pendrill, 2023; dos Santos, 2023; Urrutia and Araya, 2023; Mitchell et al., 2023). Therefore, employing LLMs may not always be the best choice for computer-based assessment of students' responses and feedback provision. In particular, assessing problem representations in textual form as outlined above might be more suitable for conventional ML algorithms or LLMs such as GPT. Conventional ML algorithms may be better suited due to increased transparency of their decision-making processes, i.e., there is an overall better explainability of the generated outputs in comparison to the more black-box behavior of LLMs. Thus, we argue that in addition to investigating the potential of an LLM for assessment purposes, it is equally important to investigate the advantages of conventional ML algorithms in comparison. Particularly when aiming to assist students in the problem-representation phase during physics problem solving, it remains unclear how LLMs and conventional ML approaches perform when trying to assess students' usage of physics concepts.

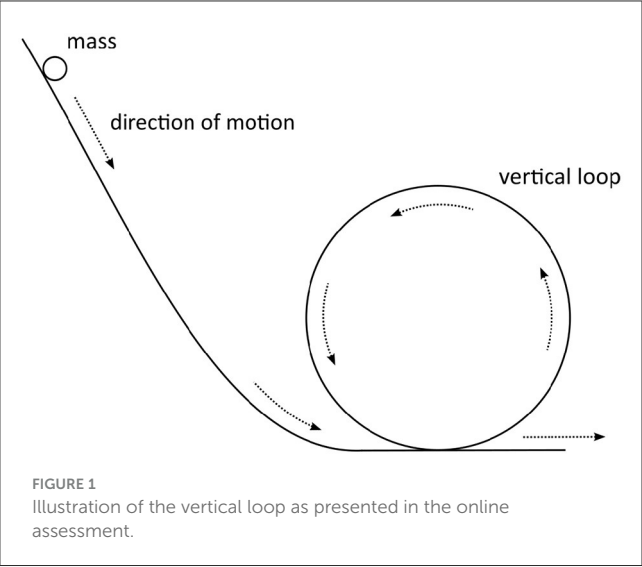
Thus, this study aimed to answer the following research questions (RQs):

- RQ1: To what extent can conventional ML algorithms correctly assess students' usage of physics concepts within a physics problem-solving task in comparison to ChatGPT based on an engineered prompt and a baseline classifier?
- RQ2: To what extent are decisions underlying the assessment of both conventional ML algorithms and ChatGPT explainable?

## 3 Methods

### 3.1 Study context

This study is based on data from the WinnerS research project, which analyzed major problem-centered science competitions in Germany, including the German Physics Olympiad (Petersen and Wulff, 2017) in which physics problem solving plays a major role (Tschisgale et al., 2024). In addition to collecting data of Physics Olympiad participants, the research project also gathered data of non-participating students that were comparable to participating students in terms of age and school type. In total, there were 444 student responses to a problem-solving task detailed below. On average, a response contained approximately 266 characters. The complete data set which includes all student responses (Physics Olympiad participants and non-participants) is freely accessible in



an Open Science Framework (OSF) repository (Tschisgale et al., 2023b).

### 3.2 Problem-solving task

The task’s instruction was as follows (translated to English by the authors): A very small mass slides along a track with a vertical loop (see Figure 1). The mass starts from a height above the highest point of the loop. Assume the motion to be frictionless. Determine the minimum starting height above the lowest point of the loop necessary for the mass to run through the loop without falling down. Describe clearly and in full sentences how you would solve this problem and what physics ideas you would use.

Instead of letting students solve the physics problem-solving task the typical (mathematics-centered) way, they were instructed to write full sentences and particularly focus on the relevant physics ideas. By prompting students to write full sentences, we intended to reduce the amount of mathematical representations used by students as students’ physics problem solving typically involves using formulas and equations as representatives or clarifications for specific physics concepts. By saying to focus on physics ideas, students should mainly remain within the phase of problem representation (Friege, 2001), i.e., students’ textual descriptions should primarily entail physics assumptions and idealizations as well as explanations around physics concepts that are regarded important for the task. An ideal student response may therefore entail simplifying assumptions such as considering a point mass, neglecting friction, and modeling the loop as circular. Under these assumptions, solving the loop task involves exactly two physics concepts: (1) the law of conservation of energy and (2) centripetal forces as the cause of circular motions (or considering an equilibrium of the centrifugal and gravitational force in a co-moving reference frame). The crux of this specific problem-solving task is to apply these concepts to the uppermost point within the loop. If the mass is just able to pass the loop, the

TABLE 1 Class distributions for energy and force codings.

	Positive class	Negative class	Total
Energy coding	53	235	288
Force coding	40	244	284

gravitational force on the mass acts completely as the centripetal force at the uppermost point. This idea in combination with the law of conservation of mechanical energy, i.e., that the initial potential energy of the mass due to its starting height equals the potential and kinetic energy at the loop’s uppermost point (neglecting rotational energy, given that only frictionless sliding is considered), in theory allows to solve this task using basic mathematics.

### 3.3 Coding manual

Generally, two fundamental physics concepts are necessary for solving the introduced physics problem-solving task: the law of conservation of energy and the concept of centripetal forces. Two coders searched for these two physics concepts within each student’s textual response and marked the corresponding segments. It should be noted that a segment’s start and ending did not need to align with the start and ending of a sentence. Therefore, segments could correspond to few words within a sentence or even go beyond multiple sentences. Due to segments’ free start and ending, determining a measure of interrater reliability proves difficult. A proposed reliability measure that mitigates this issue is the gamma agreement (Mathet et al., 2015).

Here is an example of the coded segments in a student response. We indicated the segments containing the energy or force concept using brackets and denoted the exact concept in *italic font*:

**Example 1:**

The approach is to first select the equilibrium of forces.  
[*Force concept*: The centripetal force at the uppermost point of the loop must be at least as great as the weight of the mass.]  
[*Energy concept*: The minimum starting height can then be calculated using the law of conservation of energy (kinetic and potential energy within in the loop) using the required potential energy at the starting point].

**Example 2:**

[*Force concept*: The centrifugal force in the loop depends on the ball’s velocity, mass and the radius of the loop and must exceed the gravitational force at the loop’s uppermost point.]  
[*Energy concept*: The ball’s velocity in the loop’s uppermost point depends on the height difference between the ball’s starting point and the loop’s uppermost point.] Plug formulas into each other, rearrange, and determine the minimum height difference with regard to the loop’s uppermost point at which the total force = 0. The result is the loop’s uppermost point plus the height difference.



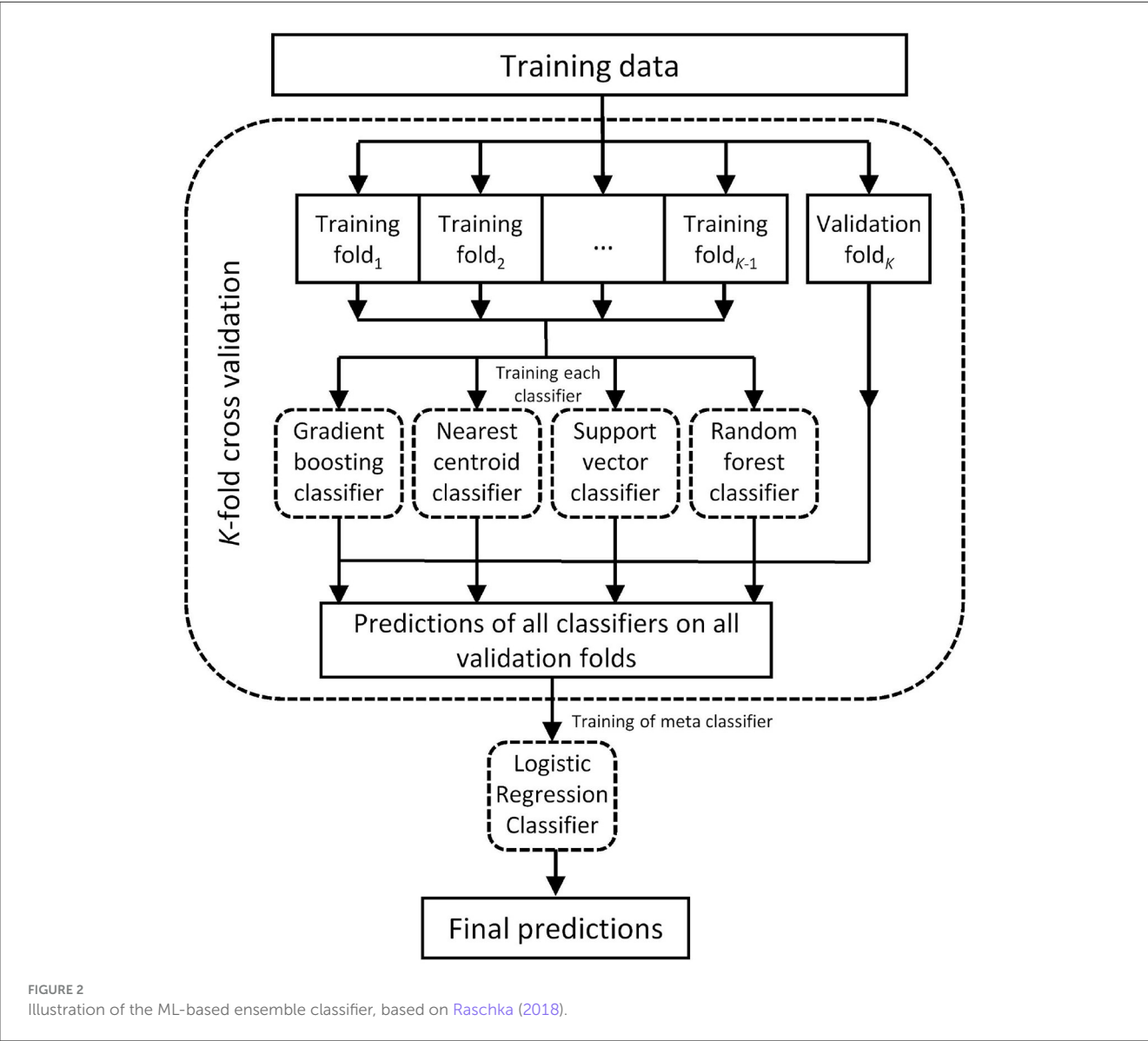
Example 1 highlights that we only coded text segments in which the physics concepts are directly applied to the task, e.g., simply stating “conservation of energy holds” was not enough. For the force concept, it was important that students specified the relevant forces that act at the highest point of the loop. For the energy concept, it was important which forms of energy occur and how they relate. If the concepts are described in a too general manner, they are not coded (see first sentence of the first example). Example 2 illustrates that while coding the force concept, we also allowed text segments about an equilibrium of forces (involving the centrifugal force), which is only correct in the co-moving reference system of the mass. However, the data set showed that this approach was frequently used among students, which is why it was also considered as a correct usage of the force concept.

The data was coded by two independent human raters with physics expertise (one graduate student and the first author). Both raters coded a subset of the data. Afterwards, disagreements were discussed and reconciled. Finally, the entire data set was coded by

both raters and gamma agreements was calculated to be .67 which we consider reliable [comparing it to thresholds for Cohen’s kappa and Krippendorff’s alpha (Landis and Koch, 1977)].

### 3.4 Data pre-processing

In order to simplify the ML problem to a classification problem on fixed units, we decided to split each response into its constituting sentences. Checking the output of this segmentation procedure revealed that it seemed to work well and provided an accurate segmentation of the original student responses. The original document-level human coding of the physics concepts was transferred to sentence level in the following manner: If a word in a sentence belonged to a coded segment in the original document-level response, the whole sentence was assigned as including the physics concept. For example, if in the original document-level



coding a physics concept spread out over two sentences (i.e., the coded segment began in the first sentence and ended within the second sentence), both sentences would be considered as incorporating the concept on sentence-level. We manually sorted out sentences where the automated sentence splitting was incorrect or where the coding no longer made sense after splitting. This way, we ended up with 284 sentences that took into account the force coding and 288 sentences for the energy coding (see Table 1). There are 53 sentences that contain the energy concept and 40 sentences that contain the force concept (we refer to them as the positive class). We are therefore dealing with an unbalanced data set.

In summary, apart from the automatic sentence segmentation and the corresponding transfer of codings from document to sentence level, no further pre-processing, such as spelling correction or removal of formulas, was conducted. The labeled sentence corpus created this way was then used to answer our research questions.

## 3.5 Analyses procedures

### 3.5.1 RQ1: comparing conventional ML algorithms, ChatGPT, and a baseline model

In RQ1, we aimed to assess the performance of three different approaches for correctly assessing students' usage of physics concepts within a physics problem-solving task. Each approach corresponds to a specific classifier built to predict whether a sentence of a student response either includes the energy concept or the force concept.

For the conventional ML approach, we employed a stacking classifier, which is a special case of so-called ensemble classifiers (Dietterich, 2000). Such an ensemble classifier combines the predictions of multiple ML-classifiers in order to improve generalizability and robustness over an individual classifier by combining the advantages of the individual classifiers. In this study, we chose a stacking classifier from the `mlxtend` library (Raschka, 2018) which is written in the Python programming language (as are all other libraries that are referred to later on). The classifier inherently includes some form of cross validation (Bishop, 2006). This logic of the classifier is depicted in Figure 2. This stacking classifier consisted of four base classifiers and involved a 5-fold cross validation. Specifically, we chose a gradient boosting classifier, a nearest centroid classifier, and a support vector classifier from `scikit-learn` (Pedregosa et al., 2011). The fourth base classifier differed for the energy and the force concept. While a balanced random forest classifier from `imbalanced-learn` (Lemaitre et al., 2017) was used for the energy-specific classifier, a random forest classifier from `scikit-learn` (Pedregosa et al., 2011) was used within the force-specific ensemble classifier. The decision for the base classifiers within the stacking classifier was based on prior experimentation on model performance.

In order to make students' textual responses in the form of sentences processable for any ML-based classifier, we generated multiple features (i.e., numeric representations of the sentences) based on the input sentences. More specifically, we used TF-IDF (term-frequency inverse-document-frequency) weighted word unigrams, character n-grams of the size 3 to 6, and sentence

embeddings from the `spaCy` library (Honnibal and Montani, 2017) for feature generation. A sentence embedding is a numerical representation of a sentence in the form of a vector (of generally high dimensionality) that captures the meaning of a sentence. We used sentence embeddings that originate from the `de-core-news-lg` model. These embeddings are generated by calculating the average of the vectors of the individual tokens. TF-IDF weighting takes into account both the relative frequency of a word among all documents and the inverse frequency of the word in all documents (Qaiser and Ali, 2018). Moreover, two additional binary features were included. The first feature checked whether the sentence only contains a formula. The formulas could be identified by searching for special characters. The second feature checked whether the sentence contained words from a predefined word list. One such word list was created for each relevant physics concept (energy and force). These lists were selected according to which words frequently occur in the positive class but not in the negative class. Manual attempts were also made to identify patterns in the data. As a result, words were added to the word lists<sup>1</sup>. The energy-specific ensemble classifier used both additional features while the force-specific ensemble classifier only used the word list, as prior experimentation showed that the other feature (the presence of a formula) minimally decreased performance.

We also want to point out that both ensemble classifiers, which are in the following referred to as ML-based classifier (Energy) and ML-based classifier (Force), were tested using variations in pre-processing, used features, and the combination of classifiers. Pre-processing experiments included lowercasing, lemmatizing (transforming words to their base form), and the removal of punctuation, special characters, and stop words. While the force classifier achieved better results with lemmatization, the results of the energy classifier improved through the removal of stop words instead. Both performed better without punctuation and special characters. Lowercasing had no positive influence on the

1 We present the word lists in German and English. Note that not all words can be translated as single words and that other language-specific details might be lost in the translation:

force = ("gewichtskraft," "gewichts-," "zentripetalkraft," "zentripetal-," "zentrifugalkraft," "zentrifugal-," "flugkraft," "flug-," "erdanziehungskraft," "erdanziehungs-," "anziehungskraft," "anziehungs-," "gravitationskraft," "gravitations-," "fliehkraft," "flieh-," "radialkraft," "radial-," "schwerkraft," "schwer-," "zentralkraft," "zentral-," "gravitation," "mindestgeschwindigkeit," "kräftegleichgewicht," and "kreisbewegung");  
force (translated) = ("weight force," "weight," "centripetal force," "centripetal," "centrifugal force," "centrifugal," "flying force," "flying," "gravitational force," "gravitational pull," "gravitational," "attractive force," "attractive," "radial force," "radial," "gravity," "central force," "central," "gravitation," "minimum speed," "balance of forces," and "circular motion");  
energy = ("energie," "kinetisch," "potentiell," "potenziell," "energieerhaltungssatz," "startpunkt," "energieerhaltung," "lageenergie," "bewegungsenergie," "höhendifferenz," "gesamtenergie," "höhepunkt," "rotationsenergie," "runterfallen," and "herunterfallen");  
energy (translated) = ("energy," "kinetic," "potential," "energy conservation law," "starting point," "energy conservation," "potential energy," "kinetic energy," "height difference," "total energy," "maximum height," "rotational energy," "fall down," and "fall off").

performance, so the upper case characters were left unchanged in the end. Other approaches included BERT sentence embeddings (Dietterich, 2000) from the transformers library (Wolf et al., 2020) which were discarded because they resulted in a high precision score, but a low recall score for the positive class. Oversampling methods like SMOTE (Chawla et al., 2002) and undersampling techniques to address the class imbalance were not expedient either.

For the ChatGPT-based approach, we used ChatGPT (gpt-3.5-turbo-instruct with standard settings) as a binary classifier by means of prompting. We used the Python programming language and ChatGPT's API to automatically store the binary outputs of the ChatGPT-based classifier (i.e., "Yes" or "No") in a list for further processing. Specifically, we used the following prompts (translated into English by the authors):

#### Energy-specific prompt:

Can you tell me whether the following sentence from a learner contains a statement about the law of conservation of energy? Sentence: "(...)" Please answer yes or no first and do not provide any reasoning.

#### Force-specific prompt:

Can you tell me whether the following sentence from a learner contains a statement about the balance of forces?<sup>2</sup> Sentence: "(...)" Please only answer yes or no first and do not provide any reasoning.

We then attempted to improve the ChatGPT-based classifiers' performance by using few-shot-learning, i.e., by providing ChatGPT sample sentences and their assigned class through the corresponding prompt. For the energy classifier, we have selected sentences that do not contain the energy approach, and for the force classifier, sentences that contain the force approach. Specifically, we used the following prompts (translated into English by the authors):

#### Energy-specific few-shot-learning prompt:

Can you tell me whether the following sentence from a student contains a statement about the law of conservation of energy? Sentence: "(...)" To help you, here are some examples that do not contain the law of conservation of energy:

1. At the highest point of the loop, the speed must be high enough for the radial force, which is proportional to the square of the speed and inversely proportional to the radius, to be at least equal to the weight of the ball.

2. In this case, start height = loop height, because the energy is converted immediately.

3. The mass would have to fall from a starting height that is at least as high as the highest point of the looping

Please only answer yes or no and do not provide any reasoning.

#### Force-specific few-shot-learning prompt:

Can you tell me whether the following sentence from a learner contains a statement about the equilibrium of forces? Sentence: "(...)" To help you, here are some examples that contain the force approach:

1. At the highest point of the loop, the speed must be high enough for the radial force, which is proportional to the square of the speed and inversely proportional to the radius, to be at least equal to the weight of the ball.

2. At the top of the loop, the radial force must just compensate for the weight of the mass so that the mass does not fall downwards.

Please only answer yes or no and do not provide any reasoning.

For the baseline approach, we established a simple rule-based classifier that assigned a sentence to the positive class (i.e., sentence includes one of the central physics concepts) if this sentence included the character string "energie" or "kraft" (German words for "energy" and "force," respectively). We therefore refer to this rule-based classifier as word-checking classifier.

To evaluate the performance of each classifiers, metrics such as accuracy (proportion of correctly assigned sentences) can be used. However, solely focusing on classifiers' accuracy is not sufficient to evaluate performance, particularly if data sets are unbalanced as in our case. Unbalanced means that a specific class (e.g., sentence includes energy concept) occurs much more frequently or rarely than the other classes (e.g., sentence does not include energy concept). In such cases, further performance metrics that also take into account the type of incorrect classification (i.e., false-positive or false-negative) are needed. Therefore, we also computed precision, recall, and F1 values as further performance metrics. Precision measures the accuracy of the positive predictions made by a classifier. In our case, precision answers the question: "Of all sentences that were predicted to include the energy (force) concept, how many sentences actually include the energy (force) concept?" Recall (or sensitivity) measures the completeness of positive predictors. In simpler terms and framed to our context, recall answers the question: "Of all sentences that actually include the energy (force) concept, how many did the classifier correctly identify?" The F1 score is the harmonic mean of precision and recall, providing a single metric that balances the trade-off between both precision and recall. All these metrics range from zero to one and a higher value generally indicates better classification.

### 3.5.2 RQ2: making model decisions explainable

An essential aspect that builds trust in AI models and opportunities for researchers to improve models is the possibility to understand why the model makes certain decisions (Zhao et al., 2023). This is also known as "explainable AI" (Lipton, 2018). Explainability refers to the ability to "explain or present the behavior of models in human-understandable terms" (Zhao et al., 2023, p. 1). There are many different ways to illuminate different aspects of explainability for LLMs in the fine-tuning paradigm such as calculating the attribution scores for each input that indicate the respective impact on the classification (Zhao et al., 2023). For

<sup>2</sup> As we singled out in Section 3.2, arguing about centripetal force in the inertial frame-of-reference would be more coherent, however, since almost no student did so, we used the balance of forces in the co-moving frame-of-reference here.

LLMs in the prompting paradigm, there also exist some methods which are necessarily constrained if models are closed-source such as ChatGPT (Zhao et al., 2023). Besides access restrictions, with such LLMs as GPT-4 it is not yet possible to entirely explain internal workings of the models and the generated outputs in human-understandable terms. This is because these language models are trained on extensive data and due to their complexity and diversity of language patterns, they can produce unpredictable results. Even if there are approaches to making the transformer architectures on which large language models are based transparent (Vig, 2019). It is unclear whether these methods have an impact on trust in AI decisions (Conijn et al., 2023). There are no explicit rules or methods to predict the exact output in advance. Instead, assessing the quality of the output relies on experience and the model's past behavior, based on previous results or benchmarks. In short, the versatility and complexity of LLMs makes it difficult to determine the exact output in advance, and one must rely on experience to evaluate their performance.

One method of making model decisions more explainable is through analyzing model outputs. In such cases one distinguishes local explanations and global explanations (Schrouff et al., 2021; Zhao et al., 2023). Local explanations address the question of why a specific student response is categorized in a particular way, while global explanations try to answer the question of why a whole group of student responses is categorized in a particular manner, i.e., one tries to understand the model in more general terms. In this study, we chose a global approach because we were interested in overall model decisions which might provide insights into students' text composition processes (e.g., which words are particularly predictive for a certain classification). First, we grouped individual sentences into separate documents based on their classifications into a specific category (e.g., false positive). Thus, we obtained four separate documents. Then, to identify patterns which might explain the models' classifications, we computed a term-frequency inverse-document-frequency (TF-IDF) score for every word in each of the four documents. Words with the highest TF-IDF scores in each category-specific document can then be considered as characteristic for this specific category (e.g., false positive), which is why we refer to them as category-specific keywords. Finally, these category-specific keywords may reveal patterns that provide an understanding of the models' decision-making. Analyzing the words that are assigned to a specific class therefore provides an approach for interpreting the assignment, as the words have a strong influence on the classification. As both the machine learning approach and the large language model approach are based on the embeddings of the tokens that make up the words. The words therefore have a major influence on the model output.

## 4 Results

### 4.1 RQ1: comparing correctness of conventional ML and ChatGPT

In Tables 2, 3 we summarized the classification performance for the ML-based classifier, the ChatGPT-based classifier, and the Baseline classifier. Through the combination of the different ML classifiers (ensemble classifier) and fine-tuning for the conventional ML algorithms, a final F1 performance of 0.74, and 0.82 for energy

and force, respectively, could be achieved. Precision and recall were always above 0.69 for both ML-based classifiers.

In contrast, both ChatGPT-based classifiers only achieved a low precision (energy: 0.20; force: 0.16), i.e., both classifiers incorrectly assigned a large proportion of sentences that did not contain the relevant physics concepts as including the concepts. Hence, it seemed that ChatGPT tended to classify sentences as including the energy or force concept. This can also be seen by inspecting Figure 3. The Figure is used to illustrate the relationships in codings between different sets. The individual diagrams show three circles that overlap and form a total of seven different areas. Each circle represents a set, and the overlaps show the common elements between the sets. Circle A represents the set of coded sentences. Circle B represents the set of sentences that are assigned to the positive class by the baseline classifier. Circle C represents the set of sentences assigned to the positive class by the ChatGPT classifier. The overlapping areas between two circles show the elements that are contained in both sets, but not in the third set. The area in which all three circles overlap represents the elements that are contained in all three sets. The largest circle illustrates the sentences that were assigned to the positive class (i.e., energy or force is in sentence) by the ChatGPT-based classifier. The ChatCPT based classifier therefore assigns a large number of sentences to the positive class.

We also attempted to improve the ChatGPT-based classifiers' performance by means of few-shot learning, i.e., by showing sample sentences including correct class labels to ChatGPT within the prompt. For the ChatGPT-based classifier for the force concept, the F1 score remained unchanged at 0.20 (see Table 3). For the ChatGPT-based classifier for the energy concept, the F1 score actually dropped through the few-shot-learning approach from 0.33 to 0.27 (see Table 2). Both ChatGPT-based classifiers thus performed worse than the baseline word-checking classifier.

Compared to the two word-checking classifiers, which only considered whether the strings "energy" or "force" were present in a sentence, both ML-based classifiers showed satisfactory results in the F1 value (see Tables 2, 3). The low precision value of the ML-based classifier for energy, compared to the ML-based classifier for force, suggests that the model incorrectly classified some text segments as positive. This could be due to certain student responses containing words related to "energy," but without explicitly demonstrating the application of the energy conservation principle in the context of the task. This potentially posed a challenge for the classifier. In Figure 4, the number of sentences for the various intersections between encoded sentences, the ML-based classifier, and the word-checking classifier are depicted. We can read from the figure that a total of 11 sentences that were positively coded were not recognized by the ML-based classifier for energy. A comparison of Figures 3, 4 illustrates that the ML-based classifications are much closer to the coded sentences by humans (which we considered as the gold-standard).

### 4.2 RQ2: making model decisions explainable

#### 4.2.1 ChatGPT-based classifier

In Figure 3 we see that the ChatGPT-based classifiers tended to assign sentences to the positive classes. This leads to poor



TABLE 2 Performance metrics of the classifiers for the energy concept.

Classifier	Accuracy	Precision	Recall	F1
Word-checking	0.83	0.55	0.81	0.66
ChatGPT-based classifier	0.35	0.20	0.90	0.33
ChatGPT-based classifier incl. few-shot	0.35	0.20	0.88	0.27
ML-based classifier	0.88	0.69	0.79	0.74

TABLE 3 Performance metrics of the classifiers for the force concept.

Classifier	Accuracy	Precision	Recall	F1
Word-checking	0.90	0.60	0.80	0.69
ChatGPT-based classifier	0.32	0.16	0.88	0.27
ChatGPT-based classifier incl. few-shot	0.28	0.17	0.88	0.27
ML-based classifier	0.94	0.91	0.75	0.82

precision, recall, and F1 values for these classifiers. Due to these poor performance metrics, it is not possible to interpret the outputs of the classifier more precisely.

4.2.2 ML-based classifier

Now we want to evaluate the consistency of the different classifiers, i.e. the extent to which two or three different classifiers assign a sentence to the same categories or to different categories. In Section 3.5.2, we explained that by understanding previous decisions, you can gain clarity about how a classifier works. For better comparability of the classifiers, it is advisable to analyze the different assignments of a sentence by the classifiers. There are eight different ways in which a sentence can be classified:

- 1. The sentence can be positively or negatively coded.
- 2. The baseline classifier can classify it as positive or negative.
- 3. The ML classifier can classify it as positive or negative.

As each of these three decisions is independent, there are a total of eight different combinations. These different combinatorial possibilities result in eight disjoint sets. These sets are shown in Table 4 and the intersections are also shown in Figure 4.

The zeros or ones in the cells of the table indicate whether the sentences are assigned to the positive ("1") or negative ('0') class by the respective classifier (column). By analyzing the characteristics of these eight different sets, we can recognize patterns and gain information on why the classifier makes certain decisions. The largest set is that of non-coded sentences that are not assigned to the positive class by either the word-checking classifier or the ML-based classifier (row one in Table 4). These sentences are examples in which the classifier has classified correctly. The keywords extracted with TF-IDF values are shown in the Figure 5 (Energy) and Figure 6 (Force). Figure 5 shows these sentences that are correctly assigned to NOT contain the energy approach, for example, words that describe the looping or words that describe the force approach. Figure 6 shows that sentences that are correctly not assigned to the force approach contain words that can be assigned to the energy

approach: "energy," "kinetic," "law of conservation of energy." We can therefore interpret that the classifier assigns the sentences of the positive class of the force approach to the negative class of the energy approach and vice versa.

For both approaches, there are no cases where the ML-based classifier assigns the sentence to the positive class, while the sentence is negatively coded and does not contain the terms "energy" or "power" (see the second row in Table 4). The third row in Table 4 is an interesting case, as these sentences are examples where the ML-based classifier performs better compared to the word-checking classifier. The following examples are included in the set for the energy classifier:

Then I use the law of conservation of energy, neglecting the friction of the mass.

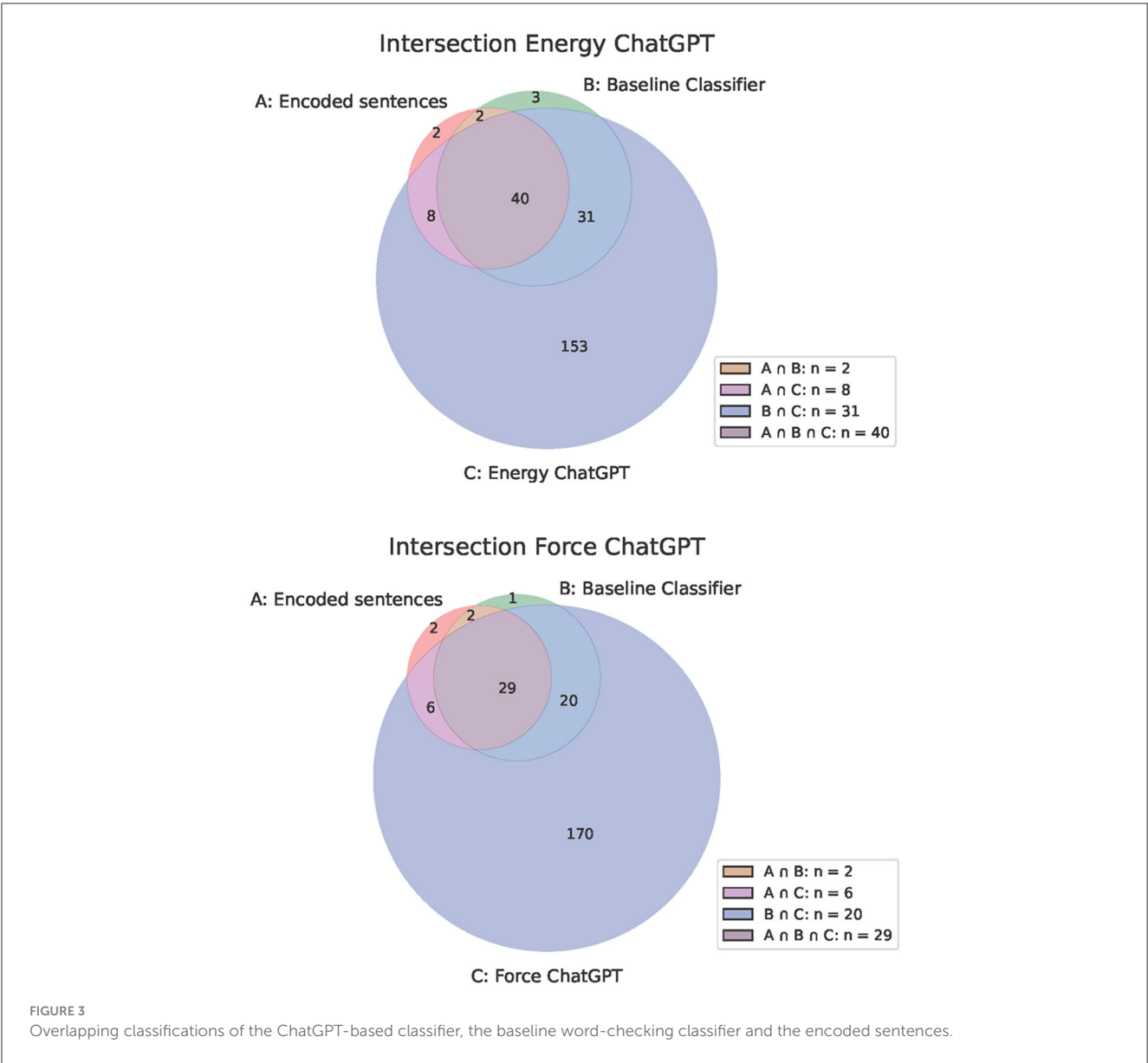
The mass must have enough energy at the highest point of the loop so that the centrifugal force keeps it on track.

For the Force-Classifer there are the following examples:

There is a centrifugal force, the mass should have a constantly increasing speed that increases during the loop or afterwards but is smallest at the highest point of the loop.

The centrifugal force can be calculated using the speed which results from the kinetic energy equation.

These sentences are all examples in which it is not clear that the learners are applying the conservation of energy concept and the force concept to the context of the task, but the string "energy" or "force" still appears in the sentence. The classifier has learnt to assign these sentences. However, there are instances where the ML-based classifier does not outperform the word-checking classifier. In row four of Table 4, sentences are displayed that are not encoded as positive, yet both the word-checking classifier and



the ML-based classifier predict the sentence as positive. For the ML-based classifier (energy) there are the following examples:

At first one should know that the law of conservation of energy plays an important role here, then one plugs the energy into a formula and gets an equation.

The kinetic energy is proportional to the mass and the square of the velocity, the rotational energy is proportional to the mass, the square of the radius (moment of inertia) as well as the square of the angular velocity.

The answers mention words such as conservation of energy, but do not apply them to the context of the task. In the second example, physics formulae are described in words, but this does not describe a physics approach applied to the context of this task. The fifth row

in Table 4 shows sentences that are difficult to identify because they are positively encoded, but the word-checking classifier assigns the sentences to the negative class. Examples are:

However, it should be noted that the mass is slowed down by gravity on the way up.

The height depends on the weight of the mass and the radius of the loop, because if one is changed, the speed and the distance change, thus the centrifugal force resulting from the starting height must be adjusted.

These types of sentences are difficult for the classifier to assign correctly, as they are very specific but still elaborate on conservation of energy. For each classifier, there exists a particular sentence in the data set where both ML-based classifiers predict a true positive

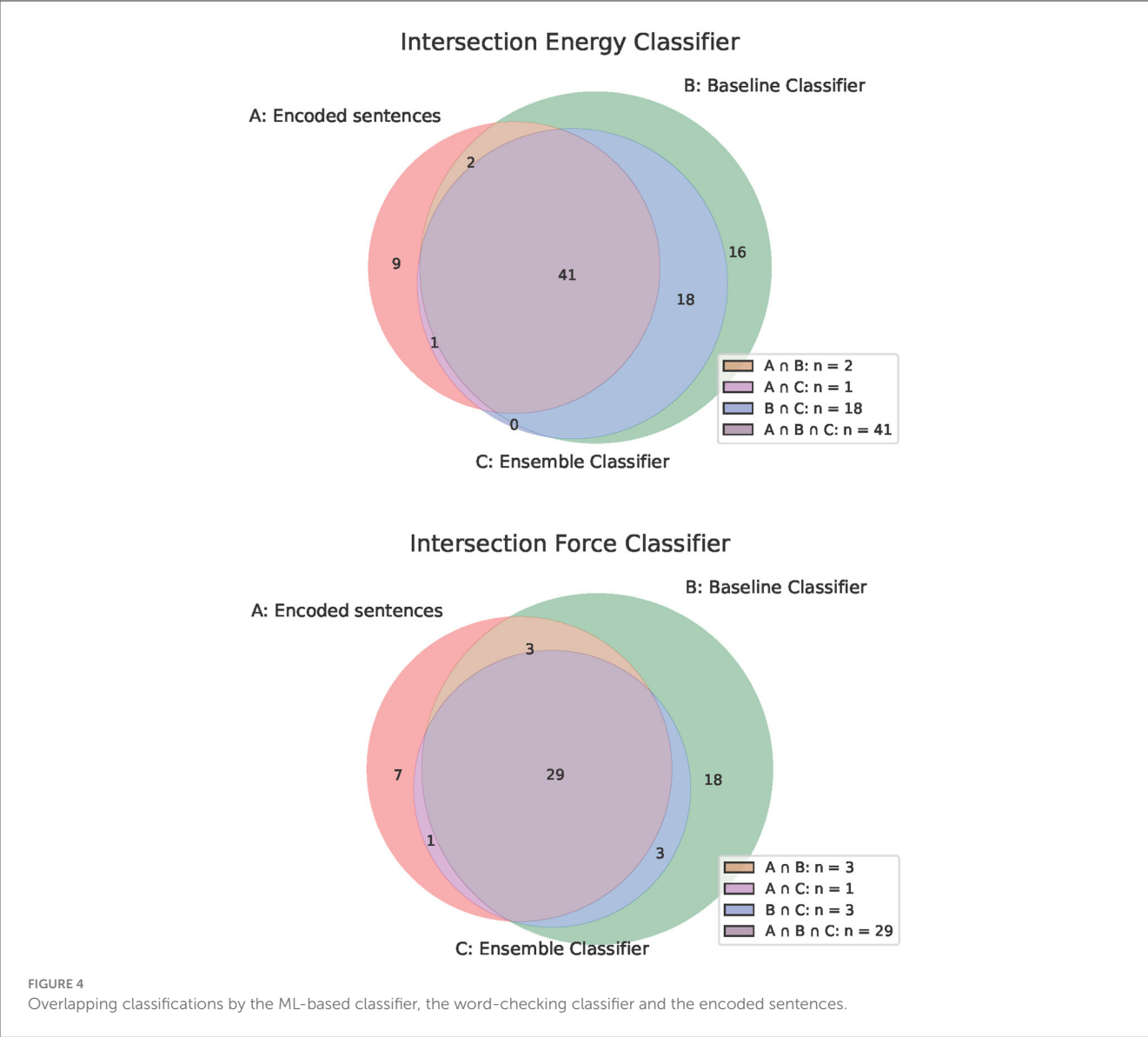


TABLE 4 Number of sentences in intersections between baseline word-checking and ML-based classifiers for force and energy.

Row no.	Encoded sentences	Word-checking classifier	ML-based classifier	# sentences (energy)	# sentences (force)
1	0	0	0	201	223
2	0	0	1	0	0
3	0	1	0	16	18
4	0	1	1	18	3
5	1	0	0	9	7
6	1	0	1	1	1
7	1	1	0	2	3
8	1	1	1	41	29

To prevent the mass from falling out of the loop, the force directed upwards at its highest point must be at least as great as the force directed downwards.

The mass starts as high as twice the height of the loop to exploit the centrifugal force and be pressed with enough momentum against the track of the loop.

The last group is the group in which word-checking and ML-based classifiers classify correctly positive. From this data set of responses, we extract the keywords again using TF-IDF values. The results are shown in [Figure 7](#) (Energy) and [Figure 8](#) (Force). [Figure 7](#) shows that for sentences that are correctly assigned to the energy approach, an important word is, in fact, “energy.” Just like the adjectives potential or kinetic. For sentences that are correctly assigned to the force approach, the most important keyword (besides the German articles “die” and “der”) is “weight force” (see [Figure 8](#)).

## 5 Discussion

In this study we sought to evaluate and compare the performance of conventional ML algorithms and an LLM-based approach to solve a well-defined binary classification problem in the domain of physics education. We found that for this specific classification task, the conventional ML algorithms outperformed the LLM classifier based on GPT 3.5 (turbo), even when simple prompt engineering techniques are employed to potentially improve GPT’s performance.

This findings was somewhat unexpected. After all, LLMs were shown to perform well on a broad range of benchmark problems, and prompt engineering and prompt chaining was shown to enhance output correctness of LLMs ([Chen et al., 2023](#); [White et al., 2023](#)). Because we do not consider our problem to be too difficult, and human interrater agreement was satisfactory, conventional ML algorithms excelled at this classification problem. It has to be said, though, that we only tried a simple prompt engineering approach, and it is quite likely that more sophisticated prompt engineering techniques would improve the classification performance in our context. Some authors suggest that when

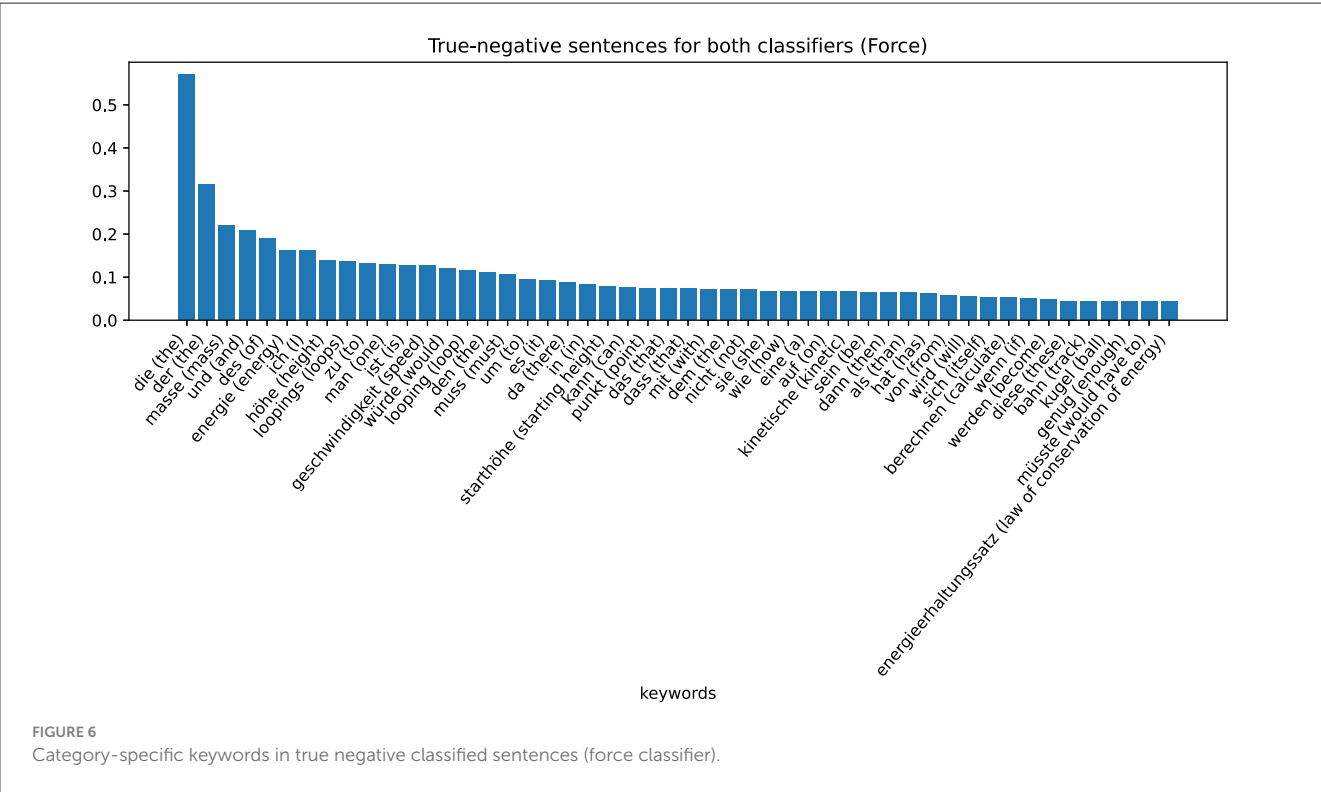
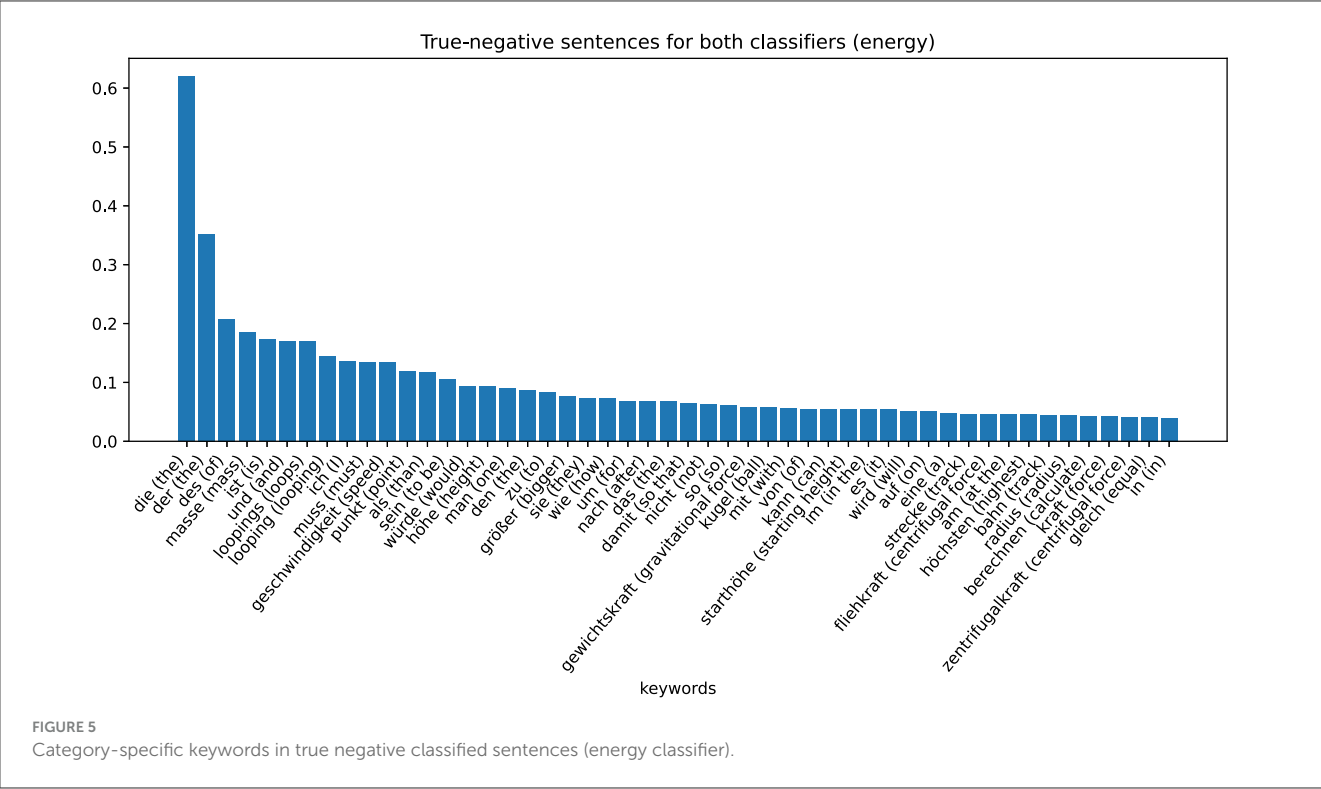
a LLM fails to perform a task, it does not necessarily indicate that the LLM is incapable of solving the task. Instead, it may simply mean that the appropriate prompt has not yet been found ([Polverini and Gregorcic, 2024](#); [Bowman, 2023](#)). Be that as it may, this was not the main point of this study, and designing prompts might take considerable time which then would suggest the use of the conventional ML algorithms eventually. We also recognize the fundamental challenges (hallucination, explainability, resource expenditure) of LLMs and sought to estimate to what extent and under which circumstances conventional ML algorithms that are optimized in standard ways could reach similar performance. In fact, they outperform ChatGPT noticeably in our context—even the Baseline classifier. This indicates that in fact conventional ML algorithms should also be considered if researchers want to tackle specific research problems with well-specified tasks and reliably coded data. However, conventional ML algorithms are more difficult to adopt to novel contexts and LLMs such as GPT can be considered versatile tools that, beyond assessment as in our context, have a broader scope of applicability ([Wan and Chen, 2024](#)). This does also not mean that LLMs cannot be used in classification contexts. Rather, researchers would typically train foundation models in a fine-tuning paradigm to utilize LLMs for classification problems ([Devlin et al., 2018](#)).

The ecological footprint of LLMs remains an issue, where conventional ML algorithms as of now are much more resource friendly. Moreover, LLMs tend to perform better in English ([Etxaniz et al., 2023](#)). Since not all researchers might have the capacities to train LLMs for specific languages from scratch, conventional ML algorithms might present a valuable option to achieve good performance in non-English tasks.

### 5.1 Limitations

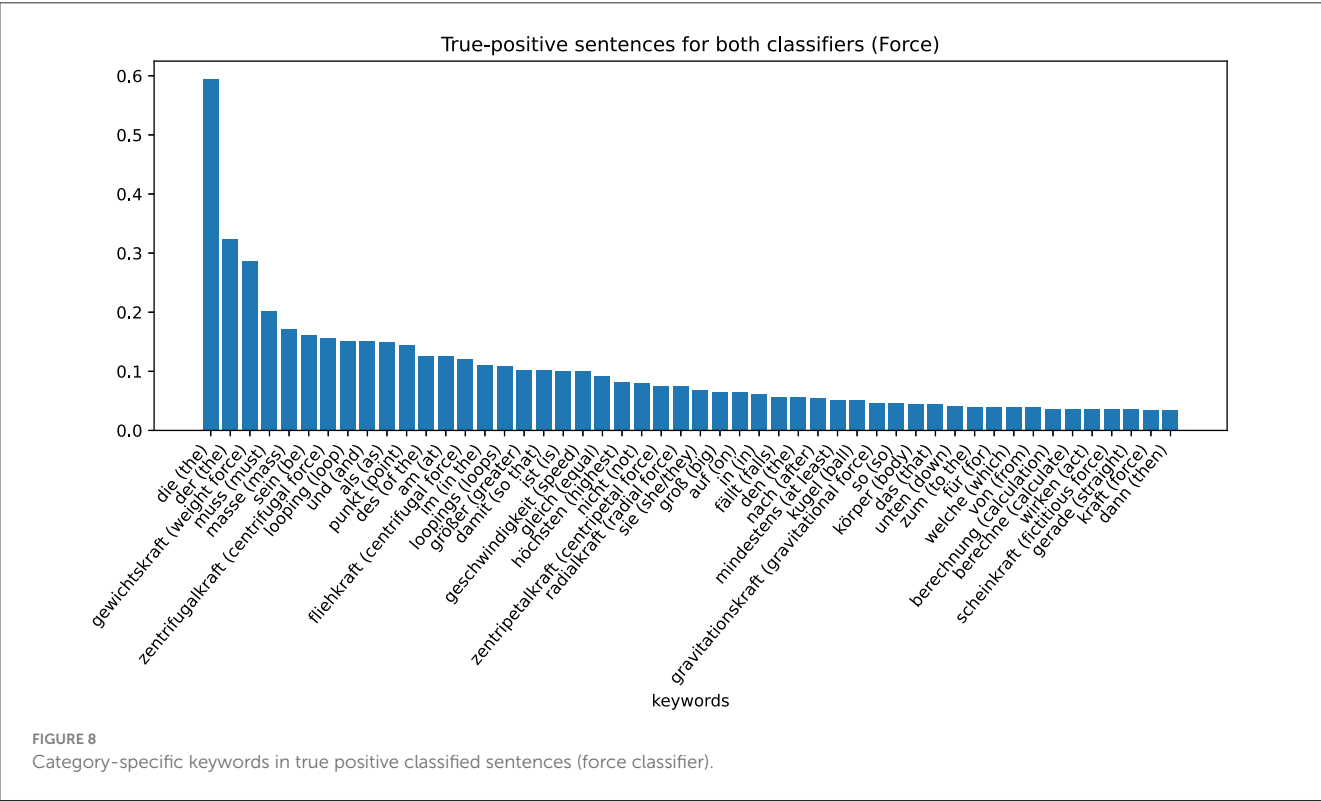
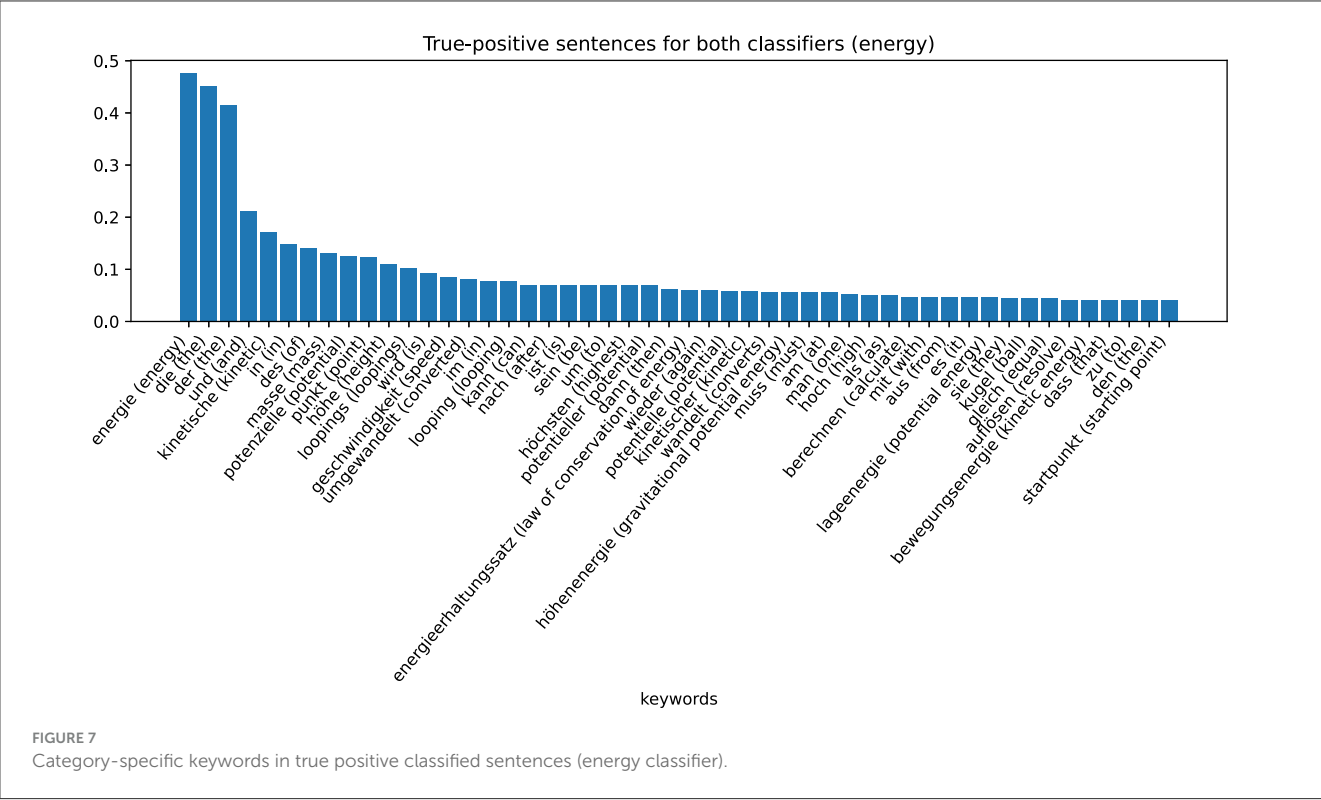
Even though conventional ML algorithms are more resource friendly this does not necessarily mean that they are more useful. One rarely needs to only assess students’ concept use in one specific problem-solving task. This also relates to a limitation of our study. We only investigated students’ responses to one particular physics problem. While concept use of energy and force is useful throughout physics, however, we cannot rule out that our classifier only performs well for this specific task. Yet, the programming





code could be re-used for training a similar classifier for another problem, if a coding manual and coded data is available. This limits the scalability of the conventional ML approach.

Another limitation relates to the generalizability of our findings to other student populations. The investigated student population is not representative of a broad student population. Almost all Physics Olympiad participants and all non-participants attended academic track (*Gymnasium*) and were from higher grade levels. It remains unclear how the investigated models would have performed on responses of less performant students as their



responses might have involved for example more colloquial wordings, student preconceptions, and spelling mistakes. All these aspects might have an influence on the performance of LLM and ML algorithms.

Other limitations relate to our data pre-processing and application of the algorithms. We only trained and validated the conventional ML algorithms at sentence level to form a well-posed classification problem. However, student answers should

be considered holistically, because the meaning can only be understood across several sentences. Moreover, we cannot rule out that other conventional ML algorithms might have exhibited better performance or that further modifications of the prompts to the LLM would enhance classification performance (Wan and Chen, 2024). Future research should apply prompting strategies that have been found to be performant for such contexts. Yet, these strategies also require substantive domain knowledge, and hence they are no silver bullet that automatically solve classification problems.

Finally, there are many different strategies to also inspect decisions of LLMs (Zhao et al., 2023). For example, in the fine-tuning paradigm attributions for the input features could be calculated that then indicate how much a certain input feature contributed to an output. However, these approaches require a large amount of technical sophistication and are much better worked out for LLMs in the fine-tuning paradigm as compared to generative LLMs with prompting. Here, prompting would also require substantive domain knowledge to investigate and understand model outputs.

## 6 Conclusions and implications

LLMs are sometimes referred to as zero-shot reasoners (Kojima et al., 2022) and can perform a variety of tasks. They have the ability to generalize, meaning that they can solve tasks that they have not seen before in the training data (Wei et al., 2023). However, our study shows that GPT-3.5 was unable to correctly identify the use of physics concepts in students' responses to physics problem-solving tasks without extensive prompt engineering. The used conventional ML model and the baseline classifier performed significantly better. Given our context, our results suggest that conventional ML models can be better adapted to a gold standard especially when expert-coded data is available. Of course, these models are then only suitable for a narrow range of applications and cannot handle the breadth of tasks that LLMs do. However, these smaller models offer further advantages in terms of transparency, processing speed, and energy consumption. Therefore, specialized ML models could be a more efficient and precise alternative in certain contexts. Especially in contexts in which there are not many different tasks to manage. It is important to remember that bigger is not necessarily better and it depends on research context whether conventional ML or LLMs are the optimal solution.

Related to designing teaching and learning environments for physics problem solving, our findings suggest that conventional ML models can be a valuable resource for automated classification. This is an important prerequisite for feedback systems that potentially enhance students' physics problem-solving abilities. Especially constructed response item formats such as the one evaluated in this study are an important means to enable students' to outline their cognitive processes related to physics problem solving. Automated analysis of these responses could enable online tutoring systems to report back the extent to which students' correctly represented a physics problem. In physics, robust application of physics concepts for solving problems is crucial for expertise development (Polverini and Gregorcic, 2024). Identifying concept use with ML and LLMs as presented in this study might pave the path toward developing tutoring systems that enable students to build this expertise.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://osf.io/d68ch/>.

## Ethics statement

The studies involving humans were approved by the Ethics Committee at the Leibniz Institute for Science and Mathematics Education (approval number 2022 13 HO). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

## Author contributions

FK: Conceptualization, Formal analysis, Writing – original draft. PT: Writing – review & editing. SR: Formal analysis, Writing – review & editing. XB: Formal analysis, Writing – review & editing. HM: Writing – review & editing. SP: Writing – review & editing. MS: Conceptualization, Writing – review & editing. KN: Writing – review & editing. PW: Conceptualization, Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Leibniz Association, Germany [grant number K194/2015], the Klaus-Tschira-Stiftung [project “WasP” under grant number 00.001.2023] and the BMBF [project “Akilas” under grant number 16SV8610].

## Acknowledgments

During the preparation of this work the author(s) used ChatGPT (based on version GPT-4) in order to improve readability and language of parts of the manuscript and generate programming code for some Figures. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or

claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Amatriain, X., Sankar, A., Bing, J., Bodigutla, P. K., Hazen, T. J., and Kazi, M. (2023). Transformer models: an introduction and catalog. *arXiv preprint arXiv:2302.07730*.
- Armour-Garb, A. (2017). *Bridging the stem skills gap: employer/educator collaboration in New York*. Technical report, The Public Policy Institute of New York State, Inc.
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). "On the dangers of stochastic parrots," in *FACCT*, 610–623. doi: 10.1145/3442188.3445922
- Bernius, J. P., Krusche, S., and Bruegge, B. (2022). Machine learning based feedback on textual student answers in large courses. *Comput. Educ.* 3:100081. doi: 10.1016/j.caeai.2022.100081
- Biernacki, R. (2014). Humanist interpretation versus coding text samples. *Qual. Sociol.* 37, 173–188. doi: 10.1007/s11133-014-9277-9
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.
- Bitzenbauer, P. (2023). Chatgpt in physics education: a pilot study on easy-to-implement activities. *Contempor. Educ. Technol.* 15:ep430. doi: 10.30935/cedtech/13176
- Bowman, S. R. (2023). Eight things to know about large language models. *arXiv preprint arXiv:2304.00612*. doi: 10.69048550/arXiv.2304.00612
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). Sparks of artificial general intelligence: early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.
- Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 183–186. doi: 10.1126/science.aal4230
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953
- Chen, Z., Chen, J., Gaidhani, M., Singh, A., and Sra, M. (2023). Xplainllm: A qa explanation dataset for understanding llm decision-making. *arXiv preprint arXiv:2311.08614*.
- Conijn, R., Kahr, P., and Snijders, C. (2023). The effects of explanations in automated essay scoring systems on student trust and motivation. *J. Lear. Anal.* 10, 37–53. doi: 10.18608/jla.2023.7801
- de Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule* 7, 2191–2194. doi: 10.1016/j.joule.2023.09.004
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dietterich, T. G. (2000). *Ensemble Methods in Machine Learning*. Berlin Heidelberg: Springer 1–15. doi: 10.1007/3-540-45014-9\_1
- Docktor, J. L., Strand, N. E., Mestre, J. P., and Ross, B. H. (2015). Conceptual problem solving in high school physics. *Phys. Rev. Special Topics* 11:020106. doi: 10.1103/PhysRevSTPER.11.020106
- Dodge, J., Prewitt, T., Des Combes, R. T., Odmark, E., Schwartz, R., Strubell, E., et al. (2022). "Measuring the carbon intensity of ai in cloud instances," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1877–1894. doi: 10.1145/3531146.3533234
- dos Santos, R. P. (2023). Enhancing physics learning with chatgpt, bing chat, and bard as agents-to-think-with: a comparative case study. *arXiv preprint arXiv:2306.00724*.
- Etxaniz, J., Azkune, G., Soroa, A., Lacalle, O. L., d., and Artetxe, M. (2023). Do multilingual language models think better in English? *arXiv preprint arXiv:2308.01223*.
- Fortus, D. (2008). The importance of learning to make assumptions. *Sci. Educ.* 93, 86–108. doi: 10.1002/sce.20295
- Friege, G. (2001). *Wissen und Problemlösen: Eine empirische Untersuchung des wissenszentrierten Problemlösens im Gebiet der Elektrizitätslehre auf der Grundlage des Experten-Novizen-Vergleichs [Knowledge and problem solving: An empirical investigation of knowledge-centred problem solving in the field of electricity based on expert-novice comparison]*. PhD thesis, University of Kiel, Kiel, Germany.
- Gaigher, E., Rogan, J. M., and Braun, M. W. H. (2007). Exploring the development of conceptual understanding through structured problem-solving in physics. *Int. J. Sci. Educ.* 29, 1089–1110. doi: 10.1080/09500690600930972
- Gerace, W. J., and Beatty, I. D. (2005). Teaching vs. learning: changing perspectives on problem solving in physics instruction. *arXiv preprint physics/0508131*.
- Graesser, A. C., Hu, X., Nye, B. D., VanLehn, K., Kumar, R., Heffernan, C., et al. (2018). Electronixtutor: an intelligent tutoring system with multiple learning resources for electronics. *Int. J. STEM Educ.* 5, 1–21. doi: 10.1186/s40594-018-0110-y
- Gregorcic, B., and Pendrill, A.-M. (2023). Chatgpt and the frustrated socrates. *Phys. Educ.* 58:035021. doi: 10.1088/1361-6552/acc299
- Hallal, K., Hamdan, R., and Tlais, S. (2023). Exploring the potential of ai-chatbots in organic chemistry: an assessment of chatgpt and bard. *Comput. Educ.* 5:100170. doi: 10.1016/j.caeai.2023.100170
- Honnibal, M., and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To Appear.* 7, 411–420.
- Hsu, L., Brewe, E., Foster, T. M., and Harper, K. A. (2004). Resource letter rps-1: research in problem solving. *Am. J. Phys.* 72, 1147–1156. doi: 10.1119/1.1763175
- Huang, Y., Song, J., Wang, Z., Zhao, S., Chen, H., Juefei-Xu, F., et al. (2023). Look before you leap: an exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.
- Huffman, D. (1997). Effect of explicit problem solving instruction on high school students' problem-solving performance and conceptual understanding of physics. *J. Res. Sci. Teach.* 34, 551–570. doi: 10.1002/(SICI)1098-2736(199708)34:6<551::AID-TEA2>3.0.CO;2-M
- Hughes, S. (2023). *Cut the bull... detecting hallucinations in large language models*. Palo Alto: Vectara.
- Jang, H. (2015). Identifying 21st century stem competencies using workplace data. *J. Sci. Educ. Technol.* 25, 284–301. doi: 10.1007/s10956-015-9593-1
- Jurafsky, D., and Martin, J. H. (2014). "Speech and language processing," in *Always Learning* (Harlow: Pearson Education).
- Kasneci, E., Seifler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). Chatgpt for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* 103:102274. doi: 10.1016/j.lindif.2023.102274
- Kieser, F., Wulff, P., Kuhn, J., and Küchemann, S. (2023). Educational data augmentation in physics education research using chatgpt. *Phys. Rev. Phys. Educ. Res.* 19:020150. doi: 10.1103/PhysRevPhysEducRes.19.020150
- Kim, E., and Pak, S.-J. (2002). Students do not overcome conceptual difficulties after solving 1000 traditional problems. *Am. J. Phys.* 70, 759–765. doi: 10.1119/1.1484151
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). "Large language models are zero-shot reasoners," in *36th Conference on Neural Information Processing Systems (neurips 2022)*.
- Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? *Phys. Rev. Phys. Educ. Res.* 19:15. doi: 10.1103/PhysRevPhysEducRes.19.010132
- Krüger, D., Parchmann, I., and Schecker, H. (2014). *Methoden in der naturwissenschaftsdidaktischen Forschung*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-37827-0
- Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K., Lukowicz, P., Kuhn, J., et al. (2023). "Unreflected acceptance-investigating the negative consequences of chatgpt-assisted problem solving in physics education," in *HHAI 2024: Hybrid Human AI Systems for the Social Good* (IOS Press), 199–212. doi: 10.3233/FAIA240195
- Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K. E., et al. (2023). Can chatgpt support prospective teachers in physics task development? *Phys. Rev. Phys. Educ. Res.* 19:020128. doi: 10.1103/PhysRevPhysEducRes.19.020128
- Lake, B. M., and Baroni, M. (2023). Human-like systematic generalization through a meta-learning neural network. *Nature* 623, 115–121. doi: 10.1038/s41586-023-06668-3
- Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310
- Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Lear. Res.* 18, 1–5.



- Leonard, W. J., Dufresne, R. J., and Mestre, J. P. (1996). Using qualitative problem-solving strategies to highlight the role of conceptual knowledge in solving problems. *Am. J. Phys.* 64, 1495–1503. doi: 10.1119/1.18409
- Li, Y., Sha, L., Yan, L., Lin, J., Raković, M., Galbraith, K., et al. (2023). Can large language models write reflectively. *Comput. Educ.* 4:100140. doi: 10.1016/j.caeai.2023.100140
- Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 31–57. doi: 10.1145/3236386.3241340
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55, 1–35. doi: 10.1145/3560815
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2019). Explainable ai for trees: From local explanations to global understanding. *arXiv preprint arXiv:1905.04610*.
- Manning, C. D. (2022). Human language understanding reasoning. *Daedalus* 151, 127–138. doi: 10.1162/daed\_a\_01905
- Mason, A., and Singh, C. (2010). Surveying graduate students' attitudes and approaches to problem solving. *Phys. Rev. Special Topics* 6:020124. doi: 10.1103/PhysRevSTPER.6.020124
- Mathet, Y., Widlócher, A., and Métivier, J.-P. (2015). The unified and holistic method gamma (gamma) for inter-annotator agreement measure and alignment. *Comput. Lingu.* 41, 437–479. doi: 10.1162/COLI\_a\_00227
- Mayring, P. (2000). "Qualitative content analysis," in *Forum: Qualitative Social Research*, 1.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mitchell, M., Palmarini, A. B., and Moskvichev, A. (2023). Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks. *arXiv preprint arXiv:2311.09247*.
- Mulhall, P., and Gunstone, R. (2012). Views about learning physics held by physics teachers with differing approaches to teaching physics. *J. Sci. Teacher Educ.* 23, 429–449. doi: 10.1007/s10972-012-9291-2
- Mulvey, P., and Pold, J. (2020). *Physics doctorates: skills used and satisfaction with employment*. Technical report, American Institute of Physics.
- Newman, M. (2005). Power laws, pareto distributions and zipf's law. *Contemp. Phys.* 46, 323–351. doi: 10.1080/00107510500052444
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Petersen, S., and Wulff, P. (2017). The german physics olympiad—identifying and inspiring talents. *Eur. J. Phys.* 38:034005. doi: 10.1088/1361-6404/aa538f
- Polverini, G., and Gregoric, B. (2024). How understanding large language models can inform the use of chatgpt in physics education. *Eur. J. Phys.* 45:025701. doi: 10.1088/1361-6404/ad1420
- Polya, G. (1945). *How to Solve It - A New Aspect of Mathematical Method*. Princeton, Oxford: Princeton University Press. doi: 10.1515/9781400828678
- Qaiser, S., and Ali, R. (2018). Text mining: use of tf-idf to examine the relevance of words to documents. *Int. J. Comput. Applic.* 181, 25–29. doi: 10.5120/ijca2018917395
- Raschka, S. (2018). Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *J. Open Source Softw.* 3:638. doi: 10.21105/joss.00638
- Rauf, I. A. (2021). *Physics of Data Science and Machine Learning*. Boca Raton: CRC Press. doi: 10.1201/9781003206743
- Savelsbergh, E. R., Ferguson-Hessler, M. G. M., and de Jong, T. (1997). *The importance of an enhanced problem representation: on the role of elaborations in physics problem solving*. Technical report, University of Twente Faculty of Educational Science and Technology, Department of Instructional Technology, Enschede, Netherlands.
- Schrouff, J., Baur, S., Hou, S., Mincu, D., Loreaux, E., Blanes, R., et al. (2021). Best of both worlds: local and global explanations with human-understandable concepts. *arXiv preprint arXiv:2106.08641*
- Sinha, R. K., Deb Roy, A., Kumar, N., and Mondal, H. (2023). Applicability of chatgpt in assisting to solve higher order problems in pathology. *Cureus* 15:e35237. doi: 10.7759/cureus.35237
- Stede, M., and Schneider, J. (2019). *Argumentation Mining*. San Rafael, CA: Morgan and Claypool. doi: 10.1007/978-3-031-02169-5
- Surameery, N. M. S., and Shakor, M. Y. (2023). Use chat gpt to solve programming bugs. *Int. J. Inf. Technol. Comput. Eng.* 3, 17–22. doi: 10.55529/ijitc.31.17.22
- Tschisgale, P., Wulff, P., and Kubsch, M. (2023a). Integrating artificial intelligence-based methods into qualitative research in physics education research: a case for computational grounded theory. *Phys. Rev. Phys. Educ. Res.* 19:020123. doi: 10.1103/PhysRevPhysEducRes.19.020123
- Tschisgale, P., Wulff, P., and Kubsch, M. (2023b). Computational grounded theory in physics education research. *Open Sci. Framework*. doi: 10.17605/OSF.IO/D68CH
- Tschisgale, P. L., Steegh, A., Petersen, S., Kubsch, M., Wulff, P., and Neumann, K. (2024). Are science competitions meeting their intentions? A case study on affective and cognitive predictors of success in the physics olympiad. *Discipl. Interdisc. Sci. Educ. Res.* 6:10. doi: 10.1186/s43031-024-00102-y
- Urrutia, F., and Araya, R. (2023). Who's the best detective? Large language models vs. traditional machine learning in detecting incoherent fourth grade math answers. *J. Educ. Comput. Res.* 61:07356331231191174. doi: 10.1177/07356331231191174
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* 46, 197–221. doi: 10.1080/00461520.2011.611369
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need: Conference on neural information processing systems," in *Advances in Neural Information Processing Systems*, 6000–6010.
- Vig, J. (2019). A multiscale visualization of attention in the transformer model. *ArXiv, abs/1906.05714*.
- Wan, T., and Chen, Z. (2024). Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. *Phys. Rev. Phys. Educ. Res.* 20:010152. doi: 10.1103/PhysRevPhysEducRes.20.010152
- Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., et al. (2021). Finetuned language models are zero-shot learners. *ArXiv, abs/2109.01652*.
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., et al. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- Wei, J., Wei, J., Tay, Y., Tran, D., Webson, A., Lu, Y., et al. (2023). Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.
- West, C. G. (2023). Ai and the FCI: can chatgpt project an understanding of introductory physics? *arXiv preprint arXiv:2303.01067*.
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., et al. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). "Transformers: state-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. doi: 10.18653/v1/2020.emnlp-demos.6
- Wulff, P. (2023). Network analysis of terms in the natural sciences insights from wikipedia through natural language processing and network analysis. *Educ. Inf. Technol.* 28, 14325–14346. doi: 10.1007/s10639-022-11531-5
- Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A., and Testrow, C. P. (2023). The death of the short-form physics essay in the coming ai revolution. *Phys. Educ.* 58:035027. doi: 10.1088/1361-6552/acc5cf
- Yore, L. D., and Treagust, D. F. (2006). Current realities and future possibilities: Language and science literacy—empowering research and informing instruction. *Int. J. Sci. Educ.* 28, 291–314. doi: 10.1080/09500690500336973
- Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). Understanding bag-of-words model: a statistical framework. *Int. J. Mach. Learn. Cyber.* 1, 43–52. doi: 10.1007/s13042-010-0001-0
- Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., et al. (2023). Explainability for large language models: a survey. *ACM Trans. Intell. Syst. Technol.* 15, 1–38. doi: 10.1145/3639372



## OPEN ACCESS

## EDITED BY

Laisa Liane Paineiras-Domingos,  
Federal University of Bahia (UFBA), Brazil

## REVIEWED BY

Michael Flor,  
Educational Testing Service, United States  
Mark D. Reckase,  
Michigan State University, United States

## \*CORRESPONDENCE

Stefan Küchemann  
✉ s.kuechemann@lmu.de

RECEIVED 30 April 2024

ACCEPTED 19 September 2024

PUBLISHED 08 October 2024

## CITATION

Küchemann S, Rau M, Schmidt A and Kuhn J  
(2024) ChatGPT's quality: Reliability and  
validity of concept inventory items.  
*Front. Psychol.* 15:1426209.  
doi: 10.3389/fpsyg.2024.1426209

## COPYRIGHT

© 2024 Küchemann, Rau, Schmidt and Kuhn.  
This is an open-access article distributed  
under the terms of the [Creative Commons  
Attribution License \(CC BY\)](#). The use,  
distribution or reproduction in other forums is  
permitted, provided the original author(s) and  
the copyright owner(s) are credited and that  
the original publication in this journal is cited,  
in accordance with accepted academic  
practice. No use, distribution or reproduction  
is permitted which does not comply with  
these terms.

# ChatGPT's quality: Reliability and validity of concept inventory items

Stefan Küchemann<sup>1\*</sup>, Martina Rau<sup>2</sup>, Albrecht Schmidt<sup>3</sup> and Jochen Kuhn<sup>1</sup>

<sup>1</sup>Chair of Physics Education Research, Faculty of Physics, Ludwig-Maximilians-Universität München (LMU Munich), Munich, Germany, <sup>2</sup>Chair of Research on Learning and Instruction, Department of Humanities, Social and Political Sciences, ETH Zurich, Zurich, Switzerland, <sup>3</sup>Chair for Human-Centered Ubiquitous Media, Institute of Informatics, Ludwig-Maximilians-Universität München (LMU Munich), Munich, Germany

**Introduction:** The recent advances of large language models (LLMs) have opened a wide range of opportunities, but at the same time, they pose numerous challenges and questions that research needs to answer. One of the main challenges are the quality and correctness of the output of LLMs as well as the overreliance of students on the output without critically reflecting on it. This poses the question of the quality of the output of LLMs in educational tasks and what students and teachers need to consider when using LLMs for creating educational items. In this work, we focus on the quality and characteristics of conceptual items developed using ChatGPT without user-generated improvements.

**Methods:** For this purpose, we optimized prompts and created 30 conceptual items in kinematics, which is a standard topic in high-school level physics. The items were rated by two independent experts. Those 15 items that received the highest rating were included in a conceptual survey. The dimensions were designed to align with the ones in the most commonly used concept inventory, the Force Concept Inventory (FCI). We administered the designed items together with the FCI to 172 first-year university students. The results show that ChatGPT items have a medium difficulty and discriminatory index but they overall exhibit a slightly lower average values as the FCI. Moreover, a confirmatory factor analysis confirmed a three factor model that is closely aligned with a previously suggested expert model.

**Results and discussion:** In this way, after careful prompt engineering, thorough analysis and selection of fully automatically generated items by ChatGPT, we were able to create concept items that had only a slightly lower quality than carefully human-generated concept items. The procedures to create and select such a high-quality set of items that is fully automatically generated require large efforts and point towards cognitive demands of teachers when using LLMs to create items. Moreover, the results demonstrate that human oversight or student interviews are necessary when creating one-dimensional assessments and distractors that are closely aligned with students' difficulties.

## KEYWORDS

large language models, large foundation models, ChatGPT, physics, item creation, concept test, validation

# 1 Introduction

The development of the transformer architecture by Vaswani et al. (2017) caused a significant leap forward in natural language processing. Most importantly, the development of the generative pre-trained transformer (GPT) and the Bidirectional Encoder Representations from Transformers (BERT) model led to the widely known large language models. Most notably the release of ChatGPT caused domain-spanning technological efforts to implement it and to investigate its effectiveness. ChatGPT has been shown to exhibit benefits for several fields including medicine and education, such as scoring essays, support in diagnostic items, personalized feedback, conceptual understanding in different domains (Eysenbach et al., 2023; Kasneci et al., 2023; Steinert et al., 2023; Kieser et al., 2023; Kuroiwa et al., 2023; Kortemeyer, 2023). Despite this large range of opportunities, several authors also point toward the challenges that arise by implementing large language models (Kasneci et al., 2023; Adeshola and Adepoju, 2023; Rahman and Watanobe, 2023). In education for example, there have been concerns about misuse, ethical issues, exam fraud, incorrect outputs, as well as an overreliance by students and teachers on the output of large language models (Kasneci et al., 2023). Additionally, with the advance toward multimodality, these models became able to not only process written text but also spoken text, images, and videos, as well as to create outputs in the same formats (Küchemann et al., 2024). So, with this increasing number of opportunities they are additional challenges that arise, such as understanding and interpretation of how the output was created, including the number of used AI algorithms and unforeseen biases. Therefore, empirical research is required to examine effective ways to use these language models in education.

Previous research on large language models investigates their effectiveness in supporting prospective teachers in item development (Küchemann et al., 2023), to solve problems in physics (Krupp et al., 2024), to augment data for educational research (Kieser et al., 2023), or to provide feedback (Yin et al., 2024). However, it is not clear from these articles whether, when participants in previous studies using large language models underperformed, users had difficulty using the language model or whether the language model was generally unable to solve the item. In this context, there is a lack of quantitative analysis of the validity of the output of large language models. In this work, we analyze the validity and reliability of items that have been created with ChatGPT after careful prompt engineering. Specifically, we focus on the following research questions:

1. How are multiple-choice items created by ChatGPT rated by experts?
2. What are the item characteristics of multiple-choice items created by ChatGPT in comparison to a widely used concept test?
3. How well can ChatGPT align concept items with a previously reported factor structure of a related concept tests?

By addressing these research questions, this work will help evaluating the potential of large language models to judge the validity and reliability of the output of large language models in general, and to create assessment items for specific concept in particular.

# 2 Related work

## 2.1 Concept inventories in science

To simulate learners' conceptual understanding in science disciplines is one of the key goals of science education research. To measure conceptual understanding and to quantify the effectiveness of instructional methods, concept inventories are frequently developed and employed. Concept inventories are also a tool for formative and summative assessment, which contain ample information for students and teachers about the understanding of students and may lead to subsequent interventions (Liu, 2010). Therefore, concept inventories have a high value in science education.

According to Adams and Wieman (2011), the development of concept inventories consists of four consecutive phases. These include the delineation of the purpose of the test and the scope of the construct (phase 1), the development and evaluation of the test specifications (phase 2), the development, field testing, evaluation, and selection of the items in scoring guides and procedures (phase 3), and the assembly the evaluation of the test for operational use (phase 4). In these phases, the initially developed items are often tested with open responses first and based on the students' answers to these open questions, multiple-choice answers are formed that are closely related to students' difficulties. Moreover, the authors point toward the value of student interviews to understand the reasoning behind the answers. Therefore, the creation of concept inventories is a time-consuming task. AI technologies may help researchers at any stage during the developmental process.

Previously, in science education research, multiple concept inventories to assess students understanding have been developed, for instance, to assess the understanding of biological evolution, climate change, or Newtonian mechanics. In physics education research, the most often used concept assessment is the Force Concept Inventory (FCI) (Hestenes et al., 1992). The FCI assesses six main concepts, namely kinematics, superposition, Newton's first law, Newton's second law, Newton's third law, and kinds of forces, which are integrated in 30 items. The FCI was mainly developed by experts. Student interviews have only been used after the test development was completed. Afterwards, there have been some concerns that the concepts intended to be measured in the FCI are not confirmed in an exploratory factor analysis and that the concepts are not well reflected in students responses (Huffman and Heller, 1995; Heller and Huffman, 1995; Scott et al., 2012). Eaton and Willoughby (2018) argue that the intended factor structure not being reflected in students' responses may result from the fact that the FCI is built in an expert-like (optimal) structure and the students who answer it may also exhibit novice-like responses, which may not align well with the expert structure. In contrast, they performed a confirmatory factor analysis (CFA) testing the previously reported empirical model from the exploratory factor analysis by Scott et al. (2012) and two expert models, the original model from the FCI developers and another expert model suggested by Eaton and Willoughby (2018) with a large set of students' responses. The expert model by Eaton and Willoughby (2018) considers the fact that some of the concepts targeted in the FCI, such as Newton's second law, are not assessed in an isolated manner but rather requires also some understanding

of kinematics. It consists of five factors: Newton's first law and kinematics, Newton's second law and kinematics, Newton's third law, force identification, and mixed. Using a CFA, they were able to confirm that all three models reach acceptable global fit statistics, thus describing the students' responses well (Eaton and Willoughby, 2018). In general, for an objective measure of understanding, it is necessary to isolate single attributes (concepts), i.e., to have unidimensional assessments (Planinic et al., 2019; Wright, 1997). Here, the dimensionality refers to the number of attributes of an object (here, conceptual understanding of Newtonian mechanics) being measured (Planinic et al., 2019). In our case, unidimensionality means that the items needed to be designed in the way that a certain set of items only assess a single concept and do not require the understanding of other concepts. In this way, the set of items clearly measure the understanding of this single concept. The CFA by Eaton and Willoughby (2018) showed that two of the three models that fit the FCI data do not exhibit one-dimensional factors. Therefore, an objective measure of these concepts using the FCI may be compromised.

In this work, we will build on the expert models developed earlier and study how items that assess Newtonian mechanics created using ChatGPT align with the previously reported factor models. In this way, we will see to what extent large language models are able to support researchers in the extensive effort required during the development of concept inventories and what aspects have to be considered. In the next section, we will provide an overview of the manifold potentials of large language models in education.

## 2.2 Large language model in education

Language models may exhibit a wide range of opportunities to support learners, reduce the workload of teachers, and improve the quality of teaching (see Kasneci et al., 2023 or Küchemann et al., 2024 for overviews). However, as mentioned above, there are a number of concerns, such as inaccurate output, biases, and over-reliance on the output, which might influence teachers' predisposition on using AI tools in classrooms and to support the everyday practices. Regarding teachers' predisposition, Polak et al. (2022) found out that European teachers have a positive attitude toward AI for education and a high motivation to introduce AI-related content in school. According to Ayanwale et al. (2022), this is essential as the willingness of teachers to promote AI is an important prerequisite for the successful integration of AI-based technologies into the classroom. In addition, perceived usefulness, ease of use, and perceived trust in these AI-based tools are factors that need to be considered when predicting their acceptance by learners (Choi et al., 2023; Steinert et al., 2024).

However, it is not clear how large language models can support teachers in their everyday activities. For instance, Karaman and Goksu (2024) could demonstrate that ChatGPT can be used for an effective lesson planning of primary school math lessons. In comparison to students in a control group in which the teacher followed already existing lesson plans, students who took part in the lesson prepared reached a high learning gain from pre- to post-test, which was comparable to the one in the control group. Overall,

the authors found that large language models are an effective tool to plan lessons. Similarly, Lee and Zhai (2024) found that ChatGPT can be effectively used for lesson planning in various subjects and that teachers reported high potentials of using ChatGPT in classroom activities.

In the context of assessment item creation, Küchemann et al. (2023) showed that prospective physics teachers can use ChatGPT to create effective physics assessments with an adequate difficulty with a high level of correctness. However, in comparison to a group of prospective physics teachers who used a textbook to create physics items, the ones who used ChatGPT struggle to integrate the items in a meaningful context and the items exhibited a lower clarity. In addition, both groups had difficulties to provide all necessary information that are relevant to solve items. Similar to this work, several other works offer a qualitative analysis of ChatGPT's output and draw conclusions on potential applications of ChatGPT in education (see for instance Ausat et al., 2023; Krupp et al., 2024, 2023).

However, in case of an insufficient performance of large language models, it is often unclear whether the users lack sufficient proficiency in using large language models or the language model itself is incapable of providing appropriate assistance. In this work, we analyze the quality of ChatGPT's output after refined prompt engineering without manually modifying the outputs. We chose the specific case of concept item creation, as it is a regularly re-occurring activity in teacher practices. In this way, we are able to provide insights into the quality of ChatGPT's output and into what teachers need to consider when using large language models for item creation.

## 3 Methodology

### 3.1 Participants

In total, 209 undergraduate STEM students from the University of Wisconsin (UW), Madison, and 51 undergraduate physics students from the Ludwig-Maximilians-University (LMU) München took part in the study. The participation was rewarded with 10 EUR at LMU and a 20 USD-Amazon gift card at UW-Madison. We excluded students who did not complete the study or who answered one of the control questions incorrectly (see below). In addition to the control questions, we carefully evaluated the time invested in answering the test items to account for the fact that some students may simply skip through the test items without thinking about the answers. We found that there was one student who reached a rather high score of 82% in a time of 14 min and 44 sec. The scores of students who completed the questionnaire below this time are fluctuating around the probability of guessing, which likely indicates that they guessed the answers. Therefore, we set a time of 14 min and 44 sec as threshold for including students' answers in the analysis. The students who completed the questionnaire below this time threshold were excluded from the analysis. These exclusion criteria led to a final data set of  $N = 173$  students ( $N = 67$  female,  $N = 91$  male,  $N = 3$  other, and  $N = 12$  made no statement, average age  $M = 20.6$  years) who were considered in the analysis.



In this work, students covered the topics assessed in the FCI and the GPT items, namely the concepts of motion and Newton's mechanics, at least one time prior to this study. In Germany, the physics curricula in schools of every state cover the concepts of motion and Newton's mechanics, and they are a part of Disciplinary Core Ideas in Wisconsin's Standards for Science. Moreover, understanding motion and forces is also part of the National Science Education Standards and part of the Next Generation Science Standards. Therefore, it is reasonable to assume that, the students had covered the concepts covered by the FCI and the GPT items.

## 3.2 Item creation

We chose kinematics and forces as the topics of the items as they are widely covered in school curricula across countries, and intensively researched topics in physics education research. Additionally, the FCI is the most used concept inventory in physics education, and its factor structure is well known.

ChatGPT 3.5 was prompted to create items that target five subcategories of the FCI: kinematics (i.e., velocity and acceleration), Newton's first law, Newton's second law, Newton's third law, and the superposition. We applied a systematic sequence of items design using on ChatGPT consisting of the following five steps (Figure 1).

1. We developed a prompt that includes the following characteristics of the items: multiple-choice (MC) items with five answer alternatives including one correct answer, the items should be embedded in a reasonable context, they should not contain images and only consist of a written text, and the item should not mention the physical principle required to solve the item. We mention these characteristics in the prompts as we saw that they made a difference. In contrast, additional specifications such as "design the items for high school level or introductory university level physics" or "include a cognitive activity of 'applying' according to Bloom's taxonomy" did not cause a notable difference to the created items.
2. We used this prompt to create  $N = 100$  multiple choice items,  $N = 20$  for each of the five categories.
3. During the initial selection process, 70 items were eliminated, which led to a set of  $N = 30$  items. The reasons for eliminating these items were:
  - The context of the item describes a physically unrealistic or incomplete scenario (Q: "If an object is at rest, what force is acting upon it to keep it at rest?")
  - The item was identical or very similar to another item.
  - It was obvious that the answer alternatives of the item were all incorrect or that multiple answers were correct or a correct answer was stated as incorrect (Question: Which of the following descriptions of the behavior of a falling stone is most accurate? Presumable incorrect answer: "A stone will fall faster and faster as it approaches the ground").
  - The presumable correct answer alternative did not fit to the context described in the item (Question: "Which of the following descriptions of the behavior of a falling stone is

most accurate?" Presumable correct answer: "The speed of the stone remains the same unless an external force acts on it").

- At least one of the answer alternatives were physically incorrect (Answer alternative: "The stone will stop falling after it has been lying on the ground for some time.")
  - The item stem contained the concept that needs to be applied to solve the item (example: "Which of the following best describes Newton's first law in the context of a stone falling to the ground?")
  - The item stem already contained the correct answer (Q: "Imagine you are driving a car on a straight road at a constant speed. What can be said about the motion of the car?"; A: "It moves at a constant speed in a straight line.")
4. Next, we performed an expert rating based on 15 quality criteria with two experts from physics education research to ensure the content validity of the GPT items. Seven of these criteria have been reported earlier by Küchemann et al. (2023). Eight additional criteria were used to account for the specific question type (MC questions) and that they have been exclusively developed by ChatGPT without human revision. The criteria and ratings are shown in the [Supplementary material](#). The rating resulted in an interrater reliability in terms of Cohen's kappa  $\kappa = 0.4$ , which means that there was a moderate agreement between the two raters. The conflicts between the two raters were adjudicated after discussion.
  5. In the expert ratings of the items, each of the criteria could be either rated with 1 (in case it applies to an item) or 0 (in case does not apply to an item) points. Adding up these points of all criteria for a single item lead to a final "score" of each item. We selected three items with the highest final score in each of the five subcategories, Newton's 1st Law, Newton's 2nd law, Newton's 3rd law, kinematics and superposition. This lead to a final set of 15 items. The final set of GPT items contained only 15 items, because we did not want to have too many items to avoid students from filling out the questionnaire incompletely. The average final scores of the selected items in each category are: Newton's 1st Law: 0.98, Newton's 2nd law: 1.00, Newton's 3rd law: 0.98, kinematics: 0.98, and superposition: 0.73. This means that the items in categories Newton's 1st Law, Newton's 2nd law, Newton's 3rd law, kinematics reached an excellent expert rating, but the items in the context of superposition reached only a moderate score. Two example items for the concept Newton's 1st Law and Newton's 3rd law are shown in [Figure 2](#).

During the entire process, we did not manually modify or revise the items created by ChatGPT.

## 3.3 Expert ratings of ChatGPT items

The expert rating can be found in the supporting online material. Regarding the item ratings, the experts rated nine out of 15 items with the highest score. Three of the remaining six items reached a relative score of 0.93, which means that experts only considered that one of the 15 criteria was not fulfilled. Specifically, the experts considered each of these three item stems as misleading.

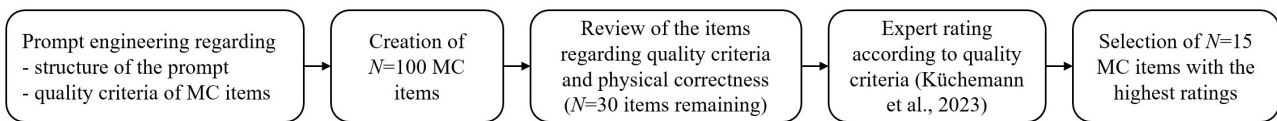


FIGURE 1  
Systematic creation of kinematic items using ChatGPT 3.5.

- |   |  |
|---|--|
| <p>1) A person is standing inside an elevator that is moving upwards at a constant speed. What is the person experiencing?</p> <p>a) The person is being pulled down by a gravitational force.</p> <p>b) The person is being pushed up by air resistance.</p> <p>c) The person is accelerating at a constant speed.</p> <p>d) The gravitational pull from the Earth cancels out with the normal force from the floor.</p> <p>e) The person is floating in a vacuum.</p> | <p>8) In a frictionless hockey rink, a player hits a puck with a stick. The puck travels at a constant speed of 20 m/s. What can be said about the force exerted by the stick on the puck and the force exerted by the puck on the stick?</p> <p>a) The force exerted by the stick on the puck is greater than the force exerted by the puck on the stick.</p> <p>b) The force exerted by the puck on the stick is greater than the force exerted by the stick on the puck.</p> <p>c) The force exerted by the stick on the puck is equal in magnitude and opposite in direction to the force exerted by the puck on the stick.</p> <p>d) The force exerted by the stick on the puck is equal in magnitude but not opposite in direction to the force exerted by the puck on the stick.</p> <p>e) The force exerted by the puck on the stick is equal in both magnitude and direction to the force exerted by the stick on the puck.</p> |
|---|--|

FIGURE 2  
Two example items to assess Newton's 1st law (1) and Newton's 3rd law (8) created by ChatGPT.

The remaining three items were all in the category of superposition, and they were rated with a score between 0.67 and 0.80. In two of these three items, the item stem was rated as misleading and insufficiently specified.

Regarding specific criteria, the answer alternatives of five of the fifteen selected ChatGPT items and the item stem of two items were rated as misleading. Furthermore, the item stem of two of the ChatGPT items were rated to have an insufficient specificity (in line with previous findings by Küchemann et al., 2023), and in two items the answer alternatives were rated as ambiguous.

However, all selected ChatGPT items were rated as scientifically correct, the answer alternatives were not too similar, they contained one correct and four incorrect answer alternatives, were relevant to assess the target concept, had an adequate difficulty, targeted a single concept, required a cognitive activity that was related to Bloom's taxonomy levels "apply" and "evaluate," and they were embedded in an appropriate context.

In sum, this means that experts thought that all selected items created by ChatGPT fulfilled important quality standards, but ChatGPT sometimes created items with misleading item stem and answer alternatives, especially for specific concepts such as superposition. The overall lower ratings of items related to the concept of superposition suggests that ChatGPT has difficulties with targeting specific concepts.

### 3.4 Administration

The items created by ChatGPT (15 items) were administered together with the FCI items (the 30 items) in randomized order as an online survey. Additionally, we added 5 simple control questions (such as "what color is the sky?"), evenly distributed among the other items. If a student did not answer one of these control questions correctly, all answers of the student would not be considered in the analysis.

## 4 Results

### 4.1 Classical test theory

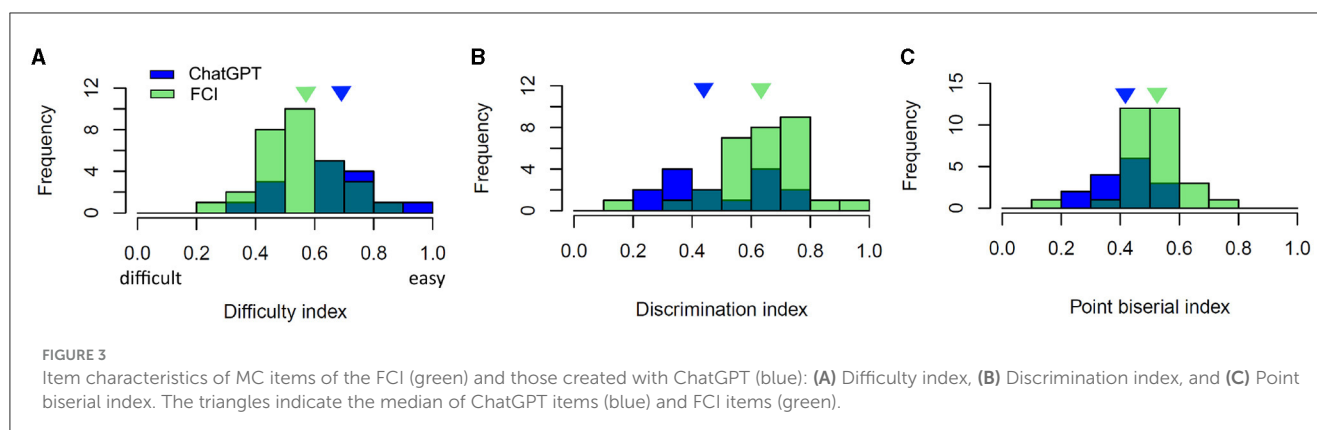
We evaluated the difficulty index, discriminatory index and point-biserial index according to the classical test theory (Figure 3), (Ding and Beichner, 2009).

Here, the difficulty index was determined by

$$P = \frac{N_1}{N}, \quad (1)$$

where  $N_1$  denotes the number of students who correctly solved the item and  $N$  is the total number of students.

For the determination of the discrimination index, the students are divided into quarters based on their performance in the whole



item set. The 25% of the highest scoring students are in the *H*-group and the 25% lowest scoring students are in the *L*-group. Now, we determined the number of correct responses in the respective item by the *H*-group ( $N_H$ ) and by the *L*-group ( $N_L$ ). The discrimination index can then be determined by

$$D = \frac{N_H - N_L}{N/4}, \quad (2)$$

where  $N$  indicates the total number of students.

The point-biserial index (or item-test correlation) is determined by

$$r_{pbs} = \frac{\bar{X}_1 - \bar{X}}{\sigma_X} \sqrt{\frac{P}{1 - P}}, \quad (3)$$

where  $\bar{X}_1$  is the average total score of students who correctly solved the respective item,  $\bar{X}$  is the average total score of all students,  $\sigma_X$  is the standard deviation of the total score of the entire sample, and  $P$  is the difficulty index for the respective item.

**Difficulty index:** In Figure 3, we can see that the difficulty index of GPT items ranges from 0.37 to 0.91. Therefore, the GPT items covers the entire suggested range of difficulty of 0.3 to 0.9 according to Ding and Beichner (2009). Here, a low difficulty index means that the item was difficult and a high value means that the item was easy for students. The median of the difficulty index of GPT items is 0.69, and the one of FCI items is 0.57, i.e., the GPT items were easier for students.

**Discrimination index:** Overall, we found that the median discrimination index of GPT items is 0.45, which is in a satisfactory range  $> 0.3$  according to Ding and Beichner (2009). Two items fall below this threshold. These two items have a high item difficulty index and therefore are prone to insufficiently discriminate between good and poor performers. The median discrimination index of the FCI is 0.62. This implies that the FCI is better able to discriminate between good and poor performers.

**Point-biserial index (Item-test correlation):** The item-test correlation coefficients of all items are above the expected level of 0.2. This indicates good item consistency and that each item is consistent with the other items in the test (Ding and Beichner, 2009). The median of the point-biserial index of the GPT items is 0.43, whereas the median point-biserial index of the FCI is 0.52. This means the GPT items have a slightly lower consistency than the FCI. It is in agreement with the following analysis of reliability.

**Reliability:** To measure the internal reliability of the ChatGPT items, we determined Cronbach's  $\alpha = 0.74$ . The result of Cronbach's  $\alpha$  is expected to be lower than the one of the FCI items because of the smaller number of ChatGPT items ( $N = 15$ ) in comparison to the number of FCI items ( $N = 30$ ). To account for this difference, we estimated the value of Cronbach's  $\alpha$  for a hypothetical set of  $N = 30$  ChatGPT items using the Spearman-Brown formula. This leads to an estimated value of Cronbach's  $\alpha = 0.85$ . In comparison, the reliability of the FCI items is  $\alpha = 0.90$ . This means that the GPT items reached a good reliability, whereas the FCI exhibit an excellent reliability.

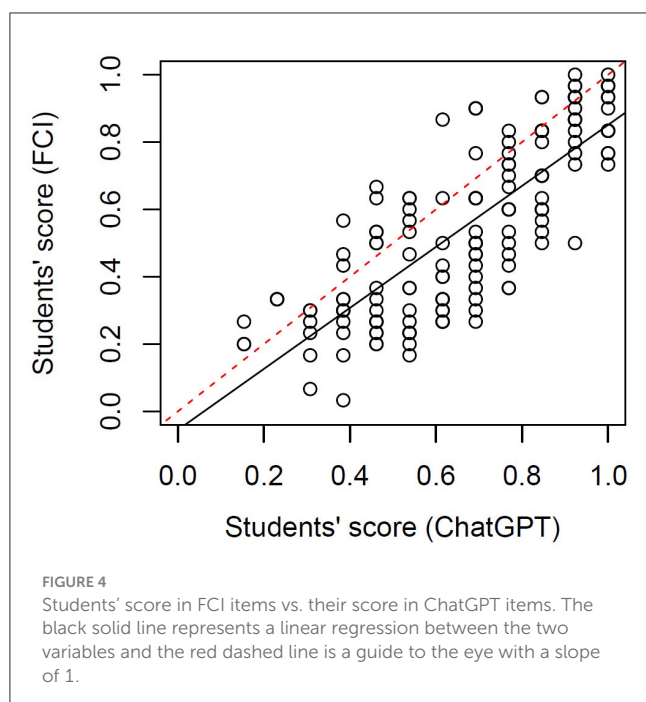
## 4.2 Convergent validity

We analyzed the performance in the ChatGPT items and the FCI to understand how closely related the two sets of items are and how valid they are to map the concepts that are assessed in the FCI. Figure 4 shows the students' scores in both item sets. We found a significant linear regression with a slope of 0.70 ( $p < 10^{-10}$ ), Pearson's correlation coefficient  $r = 0.79$ , i.e., there is a very strong relation between the two set of items. Moreover, it is noticeable that over the whole range of students' scores, their performance in ChatGPT items is higher than their performance in FCI items. This means that for low, medium and high performers, ChatGPT items were easier than the FCI items.

## 4.3 Confirmatory factor analysis

### 4.3.1 Factor structure of the FCI

Prior research has suggested three different factor structures for the FCI (Eaton and Willoughby, 2018). Here, we found that only the model suggested by Eaton and Willoughby (2018) (E&W) adequately describes the FCI data. The factor model described by Scott et al. (2012) and the originally intended factor structure by Hestenes et al. (1992) did not converge, i.e., both factor models do not describe the data properly. After testing the original model by E&W, we moved item 16 of the FCI as part of the factor Newton 1 + Kinematics in the model by E&W, because item 16 addresses the force on a truck that is pushed by a car with a constant speed. Here, two concepts are relevant: Newton's first law and the



relation between velocity and acceleration. Placing this item into 1st law+kinematics led to an improved fit to the data and the final model (E&W mod) that describes best the data is shown in Table 1. Table 2 shows the results of the confirmatory factor analysis of this model. It is noticeable that the model exhibits acceptable fit statistics with the added residual correlations [CFI > 0.9, TLI > 0.9, SRMR < 0.08, RMSEA (Upper CI) < 0.06].

#### 4.3.2 Factor structure of GPT items

Regarding the factor structure of the GPT items, we considered an analog structure to the two expert models reported for the FCI, i.e., the model by Hestenes et al. (1992) and the one by Eaton and Willoughby (2018). We did not consider an analog model to the factor model found by Scott et al. (2012) in an exploratory factor analysis here, because this model was purely data-driven and it is difficult to translate to GPT items.

The factor model by Hestenes et al. (1992) considers six factors, and the items have been designed accordingly to match these factors (Eaton and Willoughby, 2018). In our selection of concepts for the GPT items, we also followed five of these factors, namely kinematics, Newton's 1st law, Newton's 2nd law, Newton's 3rd law, and superposition. In this paper, we call this model "GPT HWS," referring to the first letters of the three authors of the corresponding FCI factor model, Hestenes, Wells, and Swackhamer. Each factor in the GPT HWS model contains three items as it was intended in the item design process without residual strains. Table 2 shows that this model does not reach an acceptable range of fit statistics (CFI < 0.9, TLI < 0.9).

As mentioned above, the factor model by Eaton et al. includes five factors: Newton's 1st law and kinematics, Newton's 2nd law and kinematics, Newton's 3rd law, force identification, and mixed. These factors account for the fact that it is sometimes useful to

assess the understanding of Newton's first and second law in a kinematics context and they do not separate between these two concepts within an item. The GPT items created here can be well mapped on these categories. Table 3 shows the factor structure of the GPT model that is aligned with the E&W model. Here, the items that target Newton's first law are not related to kinematics and therefore the first factor is renamed to Newton's first law. The items that target superposition and kinematics are closer related to the second law and therefore they are both included in the second factor. Naturally, the items that assess Newton's third law form an independent third factor as in the E&W model.

We tested this model without internal strains and found that the fit statistics are at the acceptable limit of CFI = 0.90 and TLI = 0.88. When considering a plausible residual correlation between GPT items 6 and 7, it leads to very good model fit of CFI = 0.95 and TLI = 0.94. Also this model has the higher parsimony with RMSEA = 0.03, AIC = 2,825 and BIC = 2,926. Therefore, this model fits the GPT items best (see Table 3).

We also created a joint model of FCI and GPT items in which we merge the best models for the two individual item sets (termed GPT EW + E&W in Table 2). This means, we combined the modified model by Eaton and Willoughby (2018) for the FCI data with five factors and eight residual correlations and the best model for GPT items that is aligned with the model for the FCI items by Eaton and Willoughby (2018), which has three factors and one residual correlation. Since the factors are already aligned, the combined model consists of five factors and nine residual correlations. The created model represents the best combination of the two item sets, but the fit statistics are below the acceptable range (CFI < 0.9, TLI < 0.9).

#### 4.4 Exploratory factor analysis

The purpose of the exploratory factor analysis (EFA) was to investigate whether ChatGPT items load on the same factors as the FCI items. This would mean that the items created by ChatGPT and FCI items require a comparable underlying conceptual understanding to solve them.

We verified the equality of variances of the samples, which is one of the conditions for the EFA, using Bartlett's test ( $p < 10^{-150}$ ) and the sampling adequacy for each item using the Kaiser-Meyer-Olkin (KMO) test. The KMO Test resulted in an overall measure of sample adequacy (MSA) of 0.84 with a minimum of 0.58 (above the limit of 0.5) in one of the ChatGPT items. This means that the data is suitable to perform an EFA.

To determine the number of factors optimally suited to describe the data, we created a Scree plot and performed a parallel analysis. Both analyses suggested an optimal number of three factors. A chi square hypothesis test confirmed that three factors are sufficient to describe the data ( $p < 10^{-4}$ ). The factor loadings of the exploratory factor analysis with three factors using Varimax rotation are shown in Table 4 in the Supplementary material.

Overall, the results of the factor analysis indicate that both sets of items, ChatGPT and FCI items, load on the same factors. In both sets, those items that were associated with the factor "Newton's 1st law + kinematics" during the CFA mainly load on factor 2 of



TABLE 1 The factor model suggested by Eaton and Willoughby (2018) with the modification that item 16 is part of the factor “1st Law + Kin.”

1st Law+Kin.	2nd Law+Kin.	3rd Law	Force Ident.	Mixed
6 7↘ 8↗ 10↗ 16↗ 20↗ 23↗ 24↗	9 12 14 19↗ 21↗ 22↗ 27	4 15 28	5↗ 11↗ 13↗ 18↗ 30	17↗ 25↗ 26
Added residual correlations:				
6 ~ 7 8 ~ 23	10 ~ 24 23 ~ 24	21 ~ 22 19 ~ 20	5 ~ 18	17 ~ 25

The numbers in the table indicate the item number. The added residual correlations are based on expert considerations (Eaton and Willoughby, 2018; Scott et al., 2012), and they are indicated by the arrows.

TABLE 2 Summaries of the confirmatory factor analyses.

Model	Items	No. of factors	No. of Res. Cor.	CFI	TLI	SRMR	RMSEA	AIC	BIC
E&W mod	FCI	5	8	0.92	0.91	0.06	0.04	5,491	5,718
GPT HWS	GPT	5	0	0.88	0.85	0.06	0.05	3,022	3,148
GPT EW	GPT	3	0	0.90	0.88	0.07	0.04	2,835	2,933
GPT EW	GPT	3	1	0.95	0.94	0.06	0.03	2,825	2,926
GPT EW + E&W	FCI + GPT	5	9	0.88	0.87	0.07	0.04	8,103	8,425

The statistics include the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the Standardized Root Mean Square Residual (SRMR), and measures for the parsimony of the fit, namely the Root Mean Square Error of Approximation (RMSEA), the Akaike information criterion (AIC), and the Bayesian information criterion (BIC).

the EFA. Items that are part of the factor “Newton’s 2nd law + kinematics” (CFA) load on factors 1 and 2 (EFA), and items of the factor “Newton’s 3rd law” (CFA) load on factor 3 (EFA).

## 5 Discussion

### 5.1 Content validity rated by experts

Our results show that physics education experts rated the quality of items created with ChatGPT after careful prompt engineering and significant exclusion of 70 out of 100 items with high scores in four categories. This means that ChatGPT is able to embed the item in a meaningful context even though previous research shows that prospective physics teachers often fail to do so (Küchemann et al., 2023). Moreover, the items exhibited a high clarity and appropriate difficulty level, which is in line to observations by earlier research (Küchemann et al., 2023). It is interesting that the items also had a high specificity, i.e., all relevant information to solve the items were given. Previous research observed that both physics items created by prospective physics teachers who used a textbook and those who used ChatGPT to create them lacked a high specificity (Küchemann et al., 2023). Only the concept of superposition did not reach highest expert ratings (see Supplementary material). Here, the item lacked specificity,

clarity, and were partially misleading. The distractors were also ambiguous, and the concept items were also rather similar. This observation shows that LLMs may have difficulties creating high-quality items for certain concepts.

mainrowheight16.5pt

In practice, it is unlikely that educators can invest a similar effort to create a large amount of items, exclude about 70% of them, and then perform an expert rating with peers to obtain an optimal set of items. Instead, it is more likely and time-efficient to manually optimize a set of items created by ChatGPT. In this work, we did not investigate the influence of the details of the prompt or the categories of the expert ratings on the psychometric quality of the items. Therefore, our results do not allow us to conclude on an optimal sequence of steps that educators need to follow to obtain an optimal set of items. However, we found frequent errors in the items generated by ChatGPT, which are known to affect the psychometric quality of the items and subsequently led to their exclusion (step 3 in Section 3.2) (Moreno et al., 2006; Raina and Gales, 2022). Thus, in consideration of previous works (Moreno et al., 2006; Raina and Gales, 2022; Küchemann et al., 2023) and the observations in this work, we suggest that the manual optimization process considers the following aspects:

1. the physical correctness of the item stem and the answer alternatives in case of multiple-choice items (this work),

TABLE 3 The factor model of ChatGPT items suggested in this work.

1st Law	2nd Law+Kin.	3rd Law
1	4	7
2	5	8
3	6	9
	10	
	12	
	13	
	14	
	15	
Added residual correlations:		6 ~ 7

The numbers in the table indicate the item number. The added residual correlations are based on expert considerations.

2. the inclusion of the main point in the item statement (Moreno et al., 2006),
3. the difficulty of the item is appropriate for the target group (Moreno et al., 2006),
4. the distinction of each item to other generated items (this work),
5. the fulfillment of the format requirements mentioned in the prompt, such as the correctness of exactly one answer alternative (Moreno et al., 2006 and this work),
6. the phrasing of the item that it does not contain or suggest the correct answer (Moreno et al., 2006 and this work),
7. the inclusion of all relevant information in the item stem (Küchemann et al., 2023),
8. the fact that the item stem and (in case of multiple-choice items) the answer alternatives are phrased clearly and not misleading (Raina and Gales, 2022), and
9. that the answer alternatives of a multiple-choice item are neither ambiguous nor very similar to each other (Moreno et al., 2006).

We believe that these suggestions may support the manual optimization process.

## 5.2 Characteristics of GPT items

In this work, we created the GPT items to assess the understanding of five concepts that are also part of the FCI. Then, we compared the results of the CFA of ChatGPT items to the ones of FCI items because the CFA reveals the underlying factor model that describes the data best. The intention of the developers of the FCI was to assess the understanding of specific concepts, namely superposition, Newton’s first law, Newton’s second law, Newton’s third law and kinematics. To objectively measure students’ conceptual understanding, it is necessary to develop sets of items in a way that each set only requires the understanding of a single concept (unidimensionality, see for instance Wright, 1997; Planinic et al., 2019). This means that in an ideal case, the factor model would have five factors, where each factor consists of items that separately assess one of these concepts. In fact, Eaton and Willoughby demonstrate that the factor model originally intended by the developers does describe students’ responses to the FCI items (Eaton and Willoughby, 2018).

Contrary to this previous result (see also Section 2.1), we found that a model that is based on our originally intended factor structure with five factors cannot describe the data. Instead, we found that the factor structure of GPT items can be fit by another expert model previously reported by Eaton and Willoughby (2018). This fit model also consists of five factors, namely Newton’s 1st law and kinematics, Newton’s 2nd law and kinematics, and Newton’s 3rd law, force identification and mixed. Analog to this model, the model used here to describe the GPT items consist of three factors namely Newton’s 1st law, Newton’s 2nd law and kinematics, and Newton’s 3rd law. The items that have been previously designed to assess the concepts of superposition and kinematics belong to the second factor (Newton’s 2nd law and kinematics). This also means that the item sets created by ChatGPT were unable to assess some concepts, such as superposition and kinematics, independently from Newton’s 2nd law. For concept inventories, a one-dimensional structure exhibits a high relevance to assess the students’ isolated understanding of specific concepts and not only in a relation to other concepts (Wright, 1997; Planinic et al., 2019). Consequently, similar to the FCI items, ChatGPT also does not create sets of unidimensional items. Therefore, the objective diagnosis of students’ understanding and difficulties using ChatGPT items would be compromised. At this stage, it is unclear whether state-of-the-art LLMs would be able to design items assessing superposition or kinematics without the need of other concepts if they are specifically prompted to exclude these concepts in the items. If not, human revision would be indispensable.

Moreover, we found that the GPT items were overall easier for students, they had a lower discrimination index, and lower point biserial index. One reason for this may be that the distractors were not aligned with misconceptions and therefore not as strong as in typical concept items. As mentioned above, distractors in concept tests are usually created after evaluating students open responses to the items. In this way, the distractors are closely aligned with students’ conceptions. Therefore, our findings underscore the importance of including students’ open responses or expert opinions who have experiences with students’ difficulties in the concept inventory design process.

In general, we think that the quality of the output of large language models for education purposes is difficult to determine. There are benchmark tests that allow the qualification

of performance regarding specific tasks, but none of them are specifically related to education (Touvron et al., 2023). Therefore, we decided to choose a task that is very common in education (namely item creation), and where the performance in this task can be objectively determined. Due to the high effort in empirical data collection, we invested a lot of time in the selection and analysis of the items. Therefore, we wanted to optimize the outputs in the best possible way beforehand. This allowed us to obtain an upper performance limit that can be reached without manual correction and determine the performance of LLMs in a common task in education similar to standard benchmark tests.

### 5.3 Limitations

In this work, we used a specific large language model, namely ChatGPT 3.5, it is not the most recent version that reaches the highest performances in benchmark tests and allows a multimodal input and output. In general, we need to restrict our findings to ChatGPT 3.5 and we cannot say which large language model would perform better or worse in creating physics items in mechanics. We do not have information on the model size or training data set of ChatGPT 3.5, and therefore, we cannot say what is the required size or number and type of training instances to achieve the findings we obtained here. We can assume that newer models, such as ChatGPT 4 and Gemini, that reach higher performances in benchmark tests have more parameters and have been trained on a larger data set. Therefore, it is likely that our findings are transferable and may be even exceeded by these newer models that are likely to have both, more parameters and a larger training data set.

Furthermore, there were only 173 students who took part in the study. Even though there is no specific size criteria for a confirmatory factor analysis, it is important to have a large number of students who take part in it to reveal an underlying factor structure. Even though we found a factor structure that describes the data well, we cannot exclude that a larger number of students may yield in a satisfactory fit of another factor model.

In this line, a large part of the target group were undergraduate STEM students who may or may not have received an instruction on the mechanics concepts immediately prior to participation. However, the physics curricula of high schools and/or middle schools of every state in Germany cover the concepts of motion and Newton's mechanics, and they are a part of Disciplinary Core Ideas in Wisconsin's Standards for Science. Moreover, understanding motion and forces is also part of the National Science Education Standards and part of the Next Generation Science Standards. Therefore, it is reasonable to assume that the participants have studied the concepts at some point prior to participation. Nonetheless, it is known that it is more difficult to reveal a factor structure when students often guess answers. To account for this effect, we identified a temporal cutoff and removed the students who spent a shorter time on the entire questionnaire than the time of the cutoff. In this way, we intended to remove the students who guessed several answers. However, we cannot exclude that there were still students in the final sample who guessed the answers.

Apart from that, we tested the capacity of ChatGPT 3.5 to create concept items in a topic that is one of the most common topics in physics. Therefore, it is likely that these topics are part of the training data set. But, at this point, we do not know how ChatGPT 3.5 or at the language models would perform in other fields or in topics that are less common.

## 6 Conclusion

In sum, we can conclude that large language models in general and ChatGPT 3.5 exhibit the capacity to create concept items. Even though we tested a specific language model, we can assume that newer models, such as ChatGPT 4 and Gemini, are able to perform similarly or even exceed the quality of physics items obtained in this work. Nevertheless, based on the findings of this work, it is a plausible conjecture that it is necessary to manually align distractors with students' thinking and difficulties to increase overall difficulty levels of items. It is also important to consider that the distractors created for the physics items by ChatGPT are not aligned with students' difficulties. This means that the concept items created by ChatGPT cannot be considered to have a diagnostic capacity, and it is necessary that experts revise the distractors to be able to detect students' difficulties more accurately. Moreover, it is necessary to consider that large language models may not be able to create high-quality items for all concepts and human review is also necessary in this aspect.

In future research, it would be helpful to study how a training data set needs to be designed to enhance the conceptual understanding and ability to create concept items of large language models. Ideally, the distractors in automatically created concept items would be well aligned with students' difficulties and that the created items by a large language models exhibit some diagnostic capacity. Moreover, it would be helpful to investigate how to create concept items with a large language model that would exhibit a single dimension and do not accept more than one contact at a time.

## Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author.

## Ethics statement

The studies involving humans were approved by Ethikkommission der Fakultät für Mathematik, Informatik und Statistik, Theresienstraße 39 80333 München Germany. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## Author contributions

SK: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project

administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. MR: Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. AS: Writing – review & editing, Writing – original draft, Supervision, Resources. JK: Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

## Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships

that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1426209/full#supplementary-material>

## References

- Adams, W. K., and Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *Int. J. Sci. Educ.* 33, 1289–1312. doi: 10.1080/09500693.2010.512369
- Adeshola, I., and Adepoju, A. P. (2023). The opportunities and challenges of chatgpt in education. *Inter. Lear. Environ.* 2023, 1–14. doi: 10.1080/10494820.2023.2253858
- Ausat, A. M. A., Massang, B., Efendi, M., Nofirman, N., and Riady, Y. (2023). Can chat gpt replace the role of the teacher in the classroom: a fundamental analysis. *J. Educ.* 5, 16100–16106. doi: 10.31004/joe.v5i4.2745
- Ayanwale, M. A., Sanusi, I. T., Adelana, O. P., Aruleba, K. D., and Oyelere, S. S. (2022). Teachers' readiness and intention to teach artificial intelligence in schools. *Comput. Educ.* 3:100099. doi: 10.1016/j.caeai.2022.100099
- Choi, S., Jang, Y., and Kim, H. (2023). Influence of pedagogical beliefs and perceived trust on teachers' acceptance of educational artificial intelligence tools. *Int. J. Hum. Comput. Interact.* 39, 910–922. doi: 10.1080/10447318.2022.2049145
- Ding, L., and Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Phys. Rev. Special Topics-Phys. Educ. Res.* 5:020103. doi: 10.1103/PhysRevSTPER.5.020103
- Eaton, P., and Willoughby, S. D. (2018). Confirmatory factor analysis applied to the force concept inventory. *Phys. Rev. Phys. Educ. Res.* 14:010124. doi: 10.1103/PhysRevPhysEducRes.14.010124
- Eysenbach, G. (2023). The role of chatgpt, generative language models, and artificial intelligence in medical education: a conversation with chatgpt and a call for papers. *JMIR Med. Educ.* 9:e46885. doi: 10.2196/46885
- Heller, P., and Huffman, D. (1995). Interpreting the force concept inventory: a reply to hestenes and halloun. *Phys. Teach.* 33, 503–503. doi: 10.1119/1.2344279
- Hestenes, D., Wells, M., and Swackhamer, G. (1992). Force concept inventory. *Phys. Teach.* 30, 141–158. doi: 10.1119/1.2343497
- Huffman, D., and Heller, P. (1995). What does the force concept inventory actually measure? *Phys. Teach.* 33, 138–143. doi: 10.1119/1.2344171
- Karaman, M. R., and Goksu, I. (2024). Are lesson plans created by chatgpt more effective? An experimental study. *Int. J. Technol. Educ.* 7, 107–127. doi: 10.46328/ijte.607
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learn. Individ. Differ.* 103:102274. doi: 10.1016/j.lindif.2023.102274
- Kieser, F., Wulff, P., Kuhn, J., and Küchemann, S. (2023). Educational data augmentation in physics education research using chatgpt. *Phys. Rev. Phys. Educ. Res.* 19:020150. doi: 10.1103/PhysRevPhysEducRes.19.020150
- Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? *Phys. Rev. Phys. Educ. Res.* 19:010132. doi: 10.1103/PhysRevPhysEducRes.19.010132
- Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., et al. (2023). Challenges and opportunities of moderating usage of large language models in education. *arXiv preprint arXiv:2312.14969*.
- Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., et al. (2024). "Unreflected acceptance-investigating the negative consequences of chatgpt-assisted problem solving in physics education," in *HHA1 2024: Hybrid Human AI Systems for the Social Good* (IOS Press), 199–212. doi: 10.3233/FAIA240195
- Küchemann, S., Avila, K. E., Dinc, Y., Hortmann, C., Revenga, N., Ruf, V., et al. (2024). Are large multimodal foundation models all we need? On opportunities and challenges of these models in education. *EdArXiv*. doi: 10.35542/osf.io/n7dvv
- Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K. E., et al. (2023). Can chatgpt support prospective teachers in physics task development? *Phys. Rev. Phys. Educ. Res.* 19:020128. doi: 10.1103/PhysRevPhysEducRes.19.020128
- Kuroiwa, T., Sarcon, A., Ibara, T., Yamada, E., Yamamoto, A., Tsukamoto, K., et al. (2023). The potential of chatgpt as a self-diagnostic tool in common orthopedic diseases: exploratory study. *J. Med. Internet Res.* 25:e47621. doi: 10.2196/47621
- Lee, G.-G., and Zhai, X. (2024). Using chatgpt for science learning: a study on pre-service teachers' lesson planning. *arXiv preprint arXiv:2402.01674*.
- Liu, X. (2010). *Using and Developing Measurement Instruments in Science Education: A Rasch Modeling Approach*. New York: United States: Information Age Publishing, Inc.
- Moreno, R., Martínez, R. J., and Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology* 2, 65–72. doi: 10.1027/1614-2241.2.2.65
- Planinic, M., Boone, W. J., Susac, A., and Ivanjek, L. (2019). Rasch analysis in physics education research: why measurement matters. *Phys. Rev. Phys. Educ. Res.* 15:020111. doi: 10.1103/PhysRevPhysEducRes.15.020111
- Polak, S., Schiavo, G., and Zancanaro, M. (2022). "Teachers' perspective on artificial intelligence education: an initial investigation," in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–7. doi: 10.1145/3491101.3519866
- Rahman, M. M., and Watanobe, Y. (2023). Chatgpt for education and research: opportunities, threats, and strategies. *Appl. Sci.* 13:5783. doi: 10.3390/app13095783
- Raina, V., and Gales, M. (2022). Multiple-choice question generation: towards an automated assessment framework. *arXiv preprint arXiv:2209.11830*.
- Scott, T. F., Schumayer, D., and Gray, A. R. (2012). Exploratory factor analysis of a force concept inventory data set. *Phys. Rev. Phys. Educ. Res.* 8:020105. doi: 10.1103/PhysRevSTPER.8.020105



Steinert, S., Avila, K. E., Ruzika, S., Kuhn, J., and Küchemann, S. (2023). Harnessing large language models to enhance self-regulated learning via formative feedback. *arXiv preprint arXiv:2311.13984*.

Steinert, S., Krupp, L., Avila, K. E., Janssen, A. S., Ruf, V., Dzsotjan, D., et al. (2024). Lessons learned from designing an open-source automated feedback system for stem education. *arXiv preprint arXiv:2401.10531*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). Llama 2: open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*, 30.

Wright, B. (1997). A history of social science measurement. *Educ. Measur. Issues Pract.* 16, 33–45. doi: 10.1111/j.1745-3992.1997.tb00606.x

Yin, J., Goh, T.-T., Yang, B., and Hu, Y. (2024). Using a chatbot to provide formative feedback: a longitudinal study of intrinsic motivation, cognitive load, and learning performance. *IEEE Trans. Learn. Technol.* 17, 1404–1415. doi: 10.1109/TLT.2024.3364015



## OPEN ACCESS

## EDITED BY

Knut Neumann,  
IPN–Leibniz Institute for Science and  
Mathematics Education, Germany

## REVIEWED BY

Karolina Eszter Kovács,  
University of Debrecen, Hungary  
Stylianios Mystakidis,  
Hellenic Open University, Greece

## \*CORRESPONDENCE

Dominik Thüs  
✉ dominik.thues@uni-saarland.de

RECEIVED 02 August 2024

ACCEPTED 27 September 2024

PUBLISHED 11 October 2024

## CITATION

Thüs D, Malone S and Brünken R (2024)  
Exploring generative AI in higher education: a  
RAG system to enhance student engagement  
with scientific literature.  
*Front. Psychol.* 15:1474892.  
doi: 10.3389/fpsyg.2024.1474892

## COPYRIGHT

© 2024 Thüs, Malone and Brünken. This is an  
open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](#). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic  
practice. No use, distribution or reproduction  
is permitted which does not comply with  
these terms.

# Exploring generative AI in higher education: a RAG system to enhance student engagement with scientific literature

Dominik Thüs\*, Sarah Malone and Roland Brünken

Department of Education, Saarland University, Saarbrücken, Germany

**Introduction:** This study explores the implementation and evaluation of OwlMentor, an AI-powered learning environment designed to assist university students in comprehending scientific texts. OwlMentor was developed participatorily and then integrated into a course, with development and evaluation taking place over two semesters. It offers features like document-based chats, automatic question generation, and quiz creation.

**Methods:** We used the Technology Acceptance Model to assess system acceptance, examined learning outcomes, and explored the influence of general self-efficacy on system acceptance and OwlMentor use.

**Results:** The results indicated complex relationships between perceived ease of use, perceived usefulness, and actual use, suggesting the need for more dynamic models of system acceptance. Although no direct correlation between OwlMentor use and learning gains was found, descriptive results indicated higher gains among users compared to non-users. Additionally, general self-efficacy was strongly related to perceived usefulness, intention to use, and actual use of the system.

**Discussion:** These findings highlight the importance of aligning AI tools with students' needs and existing learning strategies to maximize their educational benefits.

## KEYWORDS

AI in education, technology acceptance model, scientific text comprehension, higher education, AI-powered learning environment, self-efficacy, ChatGPT

## 1 Introduction

In many university courses, for example, in the social sciences, students are expected to study academic texts, such as primary research literature or research reviews in preparation for seminar sessions or exams. These papers, primarily written for seasoned scientists rather than learners, demand a robust level of scientific literacy. This literacy is essential for effectively connecting the central arguments, scientific methods, and data presented, thereby enabling readers to draw accurate and meaningful conclusions. Scientific literacy, as defined by the Program for International Student Assessment (PISA), is “The capacity to use scientific knowledge, to identify questions and to draw evidence-based conclusions in order to understand and help make decisions about the natural world and the changes made to it through human activity” ([Organisation for Economic Co-operation and Development \(OECD\), 2003, p. 15](#)). It involves understanding basic facts, concepts and processes, methods of scientific research, and the connections between science,

technology, and society (Goldman and Bisanz, 2002). It cannot be assumed that all students possess sufficient scientific literacy to understand scientific texts (Sason et al., 2020), and compared to textbooks, engaging with academic papers can be particularly challenging for students. Given these challenges, new technologies such as generative AI hold great promise for improving students' ability to engage with complex scientific literature. AI-powered tools have the potential to provide personalized, adaptive support that helps students navigate difficult concepts and connect new information to prior knowledge. However, as with previous technological advances, such as the introduction of computers in education, there are also risks. Relying too much on AI can bypass critical thinking processes and even introduce misinformation or bias. It is therefore crucial to identify the conditions under which AI tools can truly enhance learning. Research needs to be conducted on how these tools can be effectively integrated into educational environments while ensuring they are based on proven pedagogical strategies to support meaningful learning experiences. One promising approach is to combine generative AI with course-related knowledge bases that provide students with customized support for their academic needs. In the current study, we developed and evaluated an AI-based learning application, OwlMentor, designed to assist students in comprehending scientific texts. We conducted a longitudinal study to not only assess whether using OwlMentor positively affects students' learning outcomes but also to determine the extent to which students voluntarily engage with the platform throughout various learning phases. Additionally, we examined whether this engagement could be explained by the Technology Acceptance Model (Davis, 1989), or if this theoretical model requires further expansion to accommodate changes in students' engagement as they become more proficient in reading scientific texts over the semester.

## 1.1 Scientific text comprehension

Basic text comprehension involves creating preliminary mental representations during reading and refining them by comparing with existing knowledge. Coherence is achieved by resolving inconsistencies between prior knowledge and new information, eliminating contradictions, and connecting text elements and prior knowledge to form a coherent overall representation (Kintsch, 1998; Gernsbacher and Kaschak, 2013). Reading scientific texts is an interactive process where students engage with scientists' ideas and arguments. It requires higher engagement than basic text comprehension, as readers must integrate complex information, critically evaluate the validity of arguments, understand methods, and apply theoretical knowledge. New students and novices often struggle with scientific texts due to unfamiliarity with the discipline's structure and jargon (Goldman and Bisanz, 2002).

When students engage in science text reading, they often encounter significant challenges (Sason et al., 2020) related to their underdeveloped scientific literacy, including:

- **Phenomenon Identification:** The ability to correctly recognize and understand the main concepts, topics, or central scientific events discussed in a text. Obstacles include insufficient prior knowledge, complex presentations, or information overload (McNamara and Kintsch, 1996; O'Reilly and McNamara, 2007).
- **Scientific Explanation:** Being able to explain and understand scientific concepts in texts, requiring an understanding of

complex relationships, familiarity with scientific jargon, the ability to critically analyze texts and the application of theoretical knowledge (Cromley et al., 2010; Norris and Phillips, 2003).

- **Evidence Utilization:** The understanding, interpretation, and evaluation of data to support or disprove conclusions. A Lack of statistical knowledge, for example, can lead to misunderstandings or ignoring crucial information. Studies highlight students' difficulties in evaluating and integrating multiple sources of scientific evidence (Chinn and Brewer, 1998; Duncan et al., 2018).

These challenges underscore the need for strategies to enhance students' understanding and engagement with scientific literature. Effective reading comprehension, especially for scientific texts, is supported by meta-cognitive strategies such as inferring unstated meanings, synthesizing information for cohesive understanding, and linking new information to prior knowledge. Effective methods for promoting text comprehension include:

- **Self-questioning:** Monitors and guides reading comprehension, aiding in identifying phenomena and encouraging active engagement and seeking clarification when needed (Gunn, 2008; Joseph and Ross, 2018; King, 1994).
- **Linking new information to prior knowledge:** Helps students understand complex relationships and scientific concepts by creating a familiar framework for new information (Kendeou and Van Den Broek, 2005, 2007; Sason et al., 2020).
- **Summarizing key passages:** Consolidates understanding by reinforcing main ideas and ensuring critical information is retained (Cromley et al., 2010).
- **Self-explanation techniques:** Students explain the material to themselves, deepening engagement with content by requiring them to process and articulate their understanding of scientific concepts (Chi et al., 1989; Chi et al., 1994).

Research has shown the effectiveness of these strategies. Gunn (2008) found that structured questioning, especially with high domain knowledge, enhances text memory and learning. Joseph and Ross (2018) demonstrated that self-questioning techniques tailored for middle school students with learning disabilities improve comprehension by generating questions before, during, and after reading. King (1994) showed that generating self-questions, particularly those linking new material with prior knowledge, promotes deeper knowledge construction and enhances learning outcomes. Kendeou and van den Broek (2005) found that misconceptions significantly influence text comprehension as they affect memory representation of the text. Their 2007 study showed that prior knowledge and text structure interact to influence cognitive processes during reading, with explicit disconfirmation of misconceptions improving comprehension. Sason et al. (2020) demonstrated that comprehension of science texts improves significantly when students learn to ask questions that connect to the text. Cromley et al. (2010) found that effective reading strategies, such as summarization, are strongly associated with improved comprehension and academic achievement in science, suggesting that supporting students in summarizing key information can significantly enhance their understanding and retention of scientific concepts. Chi et al. (1989) found that good students generate detailed self-explanations, which refine and expand their understanding, leading

to better problem-solving skills and independent knowledge. Chi et al. (1994) demonstrated that self-explanation promotes deeper understanding and integration of new information, as students who self-explained while reading showed greater knowledge gains and constructed more accurate mental models.

However, applying strategies such as self-questioning, linking to prior knowledge, summarizing, and self-explaining is challenging for learners, requiring significant (meta-) cognitive and motivational resources. This can be a barrier to their successful application. Generative artificial intelligence (GAI) and large language models (LLMs) can support these methods by providing interactive engagement, generating practice questions, and offering feedback. These technologies have the potential to enhance students' comprehension of academic texts through targeted support and practice, showing promise for university teaching (Kasneci et al., 2023).

## 1.2 Generative AI assistants to support text comprehension

To effectively develop generative AI assistants that aid students in understanding scientific texts required in university courses, three key theoretical considerations arise. First, it is crucial to translate the capabilities of GAI into pedagogical functionalities of a learning application that can foster meta-cognitive strategies beneficial for comprehending scientific texts. Second, measures must be implemented to ensure that the system produces outputs that are both accurate and relevant. Third, it is important to identify factors that influence whether students will engage actively with the system over an extended period, such as the duration of a semester. In the following sections, we will delve into the theoretical underpinnings of these three aspects.

GAI can be defined as Artificial Intelligence (AI) that generates new data or outputs, using machine learning (Gimpel et al., 2023). LLMs in particular offer a wide range of promising applications in the education sector (Kasneci et al., 2023). They are trained on a large corpus of data to process and generate natural language text (Gimpel et al., 2023). Natural language processing (NLP) aims to enable computers to understand and process human language. Significant progress in this field has been achieved through the introduction of Transformer models (Vaswani et al., 2017), such as BERT and GPT that allow the context of a word to be analyzed in relation to all other words in the text, resulting in improved processing speed and accuracy. Since ChatGPT's launch in 2022, numerous studies have taken a closer look at the benefits and challenges of LLMs and conversational AI in education, such as GPT-3.5 and GPT-4 (Cooper, 2023; Herft, 2023; Kasneci et al., 2023; Pavlik, 2023; Qadir, 2023; Sallam, 2023; Zhai, 2022). For instance, LLMs have been employed to create educational content, including quizzes and flashcards (Bhat et al., 2022; Dijkstra et al., 2022; Gabajiwala et al., 2022), function as pedagogical agents or conversation partners (Abdelghani et al., 2022; Bao, 2019; El Shazly, 2021; Ji et al., 2023), and serve as tools for providing feedback (Jeon, 2021). Based on this research, written initial guidelines as well as recommendations on how to possibly integrate them into educational settings were developed (Gimpel et al., 2023; Kasneci et al., 2023; Mollick and Mollick, 2022). Besides ChatGPT's interface, users can access models like GPT 3.5 and GPT 4 to build their own applications.

In the field of educational applications, it seems particularly promising that conversational AI assistants can be developed that interact with learners in a human-like way and help them for example,

with understanding given scientific texts. A recent meta-analysis (Wu and Yu, 2023) suggests that the use of conversational AI can increase students' performance, motivation, and self-efficacy and reduce anxiety, especially at the university level. Other literature reviews state that conversational AI enhances student skills and motivation (Wollny et al., 2021), significantly impact learning achievement and satisfaction (Kuhail et al., 2022) and facilitate language learning (Huang et al., 2022). Liang et al., 2023 found that GAI interaction can boost self-efficacy and cognitive engagement, both serving as mediators for learning achievement, with GAI interaction also having a direct effect on learning achievement. Although it has not yet been explicitly explored, conversational AI assistance presents a wealth of opportunities for enhancing students' comprehension of scientific texts. Leveraging the advanced capabilities of LLMs, these AI systems excel at inferring meanings, synthesizing information, and connecting concepts within selected scientific texts. Such proficiency suggests that conversational AI provided through pedagogically informed applications could serve as an effective mentor, potentially surpassing student capabilities in these complex cognitive tasks. When considering the use of LLMs in this capacity, it must be taken into account that AI systems are only as good as their training data and are associated with biases and misinformation (Alkaiissi and McFarlane, 2023; Qadir, 2023), limitations in the scope of knowledge, lack of interpretability (Kasneci et al., 2023), the exacerbation of ethical issues, unreliability, toxicity (Zhuo et al., 2023), and risks of technical dependence and misuse (Alshater, 2022; Kasneci et al., 2023). For students new to a topic, these risks are particularly pertinent given that for them it is challenging to verify an LLM's accuracy. Recent studies indicate that the accuracy of ChatGPT responses is around 60% (Kung et al., 2023) and that 52% of the software development responses contained inaccuracies (Kabir et al., 2023). However, it is important to emphasize that such analyses are snapshots and that LLMs such as ChatGPT are continuously evolving. Moreover, there are currently at least two robust methods available to refine LLM responses to prevent learners' misinformation: 1. fine-tuning a pre-trained LLM with one's own data set, which is very cost- and time-intensive, or 2. sending additional information with the initial user prompt (e.g., chain of thought techniques, zero/few-shot prompting or in-context learning; Zhao et al., 2023), which can be used in the short term and with manageable effort. In-Context Learning enhances contextual understanding and aids in mitigating errors like hallucinations, where the LLM generates seemingly credible but inaccurate information (Alkaiissi and McFarlane, 2023). In-Context Learning offers a practical option, especially for teachers who are not experts in computer science. This method makes it also possible to make short-term and minor adjustments to teaching materials (such as the selection of scientific texts) with relatively little effort.

A widely used way to apply contextual learning to an LLM is retrieval augmented generation (RAG), where hallucinations can be reduced by using information retrieval methods to provide additional context to a prompt (Shen et al., 2023). A RAG system involves searching and retrieving documents that semantically match a query and then passing these documents to an LLM. Usually, the documents are retrieved from the database based on the user request and then transmitted to the LLM via a prompt. Such RAG systems aim to reduce the problem of hallucinations, link references to generated responses or remove the need for annotating documents with meta-data (Barnett et al., 2024). It could be shown that RAG Systems can



substantially increase accuracy in some cases (94% improvement over situations where no context is provided) but can still be misled if prompts directly contradict the previously trained understanding of the model (Feldman et al., 2023, 2024). RAG systems provide a great opportunity to equip LLMs with specific knowledge. Especially for educational scenarios where specific scientific literature is provided, RAG-based applications could be valuable to support students.

### 1.3 Technology acceptance model

Even the best learning applications cannot promote learning if the learners use it minimally or fail to utilize all its helpful features. A theory frequently employed to explain how new software or information technologies are adopted by learners is the Technology Acceptance Model (TAM; Davis et al., 1989, Venkatesh and Davis, 1996). The TAM was developed to explain and predict how users accept and implement new technological tools. Over the years, TAM has been frequently studied and extended (for an overview, see Chuttur, 2009; Yousafzai et al., 2007a). It focuses on four central constructs: Perceived Ease of Use, Perceived Usefulness, Intention to Use, and Actual System Use. Perceived Ease of Use evaluates whether potential users perceive the technology as easy to operate. Technologies considered easy to use are more likely to be adopted, as they reduce the effort required for learning and using the system (Venkatesh and Davis, 2000). Perceived Usefulness refers to the belief that the technology will improve one's performance (Opoku and Enu-Kwesi, 2019). Intention to Use describes the extent to which a person has the behavioral Intention to Use the technology. Actual System Use is the actual behavior of users, indicating how often and to what extent the technology is used. According to the extended TAM (Venkatesh and Davis, 1996), both Perceived Ease of Use and Perceived Usefulness directly influence Intention to Use, which in turn is a strong predictor of Actual System Use. External factors such as system experience, educational level, digital Self-Efficacy and age can influence Perceived Ease of Use and Perceived Usefulness. However, there is no consensus on these external factors, as different studies have identified varying influencing variables (Chuttur, 2009; Yousafzai et al., 2007a).

Over the years, the TAM has been frequently studied and extended (Chuttur, 2009; Yousafzai et al., 2007a). These studies show that Perceived Usefulness consistently emerges as a significant predictor of technology acceptance, while the influence of Perceived Ease of Use may vary or decrease over time. For example, Davis et al. (1989) observed that the influence of Perceived Ease of Use on behavioral Intention to Use tends to diminish as users become more familiar with a technology. Further research supports this, showing that the effect is more pronounced in the early stages of technology adoption, but becomes less evident over time (Adams et al., 1992; Chau, 1996; Gefen and Straub, 2000; Igbaria et al., 1996). Subramanian (1994) also found that Perceived Ease of Use has less influence on Actual System Use when the technology is inherently easy to use. The validity of the TAM was mainly confirmed by studies in which data were collected at a single point in time, usually shortly after introducing a new technology. However, Yousafzai et al.'s (2007b) meta-analysis shows a gap in understanding how these relationships change over longer periods. This analysis highlights the need for longitudinal research to capture the evolving nature of technology

adoption, as many TAM studies have not considered longer-term changes or new variables that may become relevant as users continue to engage with the technology. Future TAM research should therefore take more longitudinal and dynamic approaches to understand better how users' intentions evolve (Davis et al., 2024). Recent studies continue to support and expand upon the core constructs of TAM. Yu et al. (2024) demonstrated that Perceived Ease of Use and Perceived Usefulness are strong predictors of continued Intention to Use, with Perceived Usefulness being the strongest predictor. Zou and Huang (2023) highlighted that attitude towards using technology significantly mediates the effects of Perceived Usefulness and Perceived Ease of Use on Intention to Use. These results suggest that while the core constructs of the TAM are still relevant, the use of Perceived Ease of Use and Perceived Usefulness as predictors still needs to be validated in new contexts, including the use of ChatGPT and other LLMs in education.

### 1.4 OwlMentor: an AI-powered learning environment

Our AI-powered learning environment, OwlMentor, was developed using principles from User-Centered Design (UCD) to focus on usability and usefulness. UCD is an iterative process focusing on users' needs at different design stages to ensure the final product meets their preferences (Norman and Draper, 1986). OwlMentor, named after our wise university mascot, includes features to enhance scientific text comprehension. It supports self-questioning through free chat sessions about course literature, allowing students to ask questions and receive immediate answers, fostering active engagement. OwlMentor also generates summaries of scientific texts, helping students understand key passages and main ideas. Additionally, it automatically creates multiple-choice questions from the text, enabling students to test their knowledge and practice regularly. This process involves deciding whether to keep or discard the questions, ensuring students' active involvement. These questions can be compiled into quizzes, promoting a constructive learning environment where students generate outcomes beyond the provided information. At the end of a quiz, OwlMentor provides feedback for each question if requested, enhancing self-explanation by having students articulate why an answer is correct or incorrect. By offering these functions, OwlMentor has the potential to facilitate deeper engagement with course content and improve students' ability to comprehend complex scientific texts.

OwlMentor was developed by the authors as part of the 'Innovation Project OwlMentor,' which was part of the 'Digital Teaching Plug-in' (DaTa-Pin) project funded by the Foundation for Innovation in Higher Education Teaching (Stiftung für Innovation in der Hochschullehre), granted to Saarland University. This project extended over two semesters and took place in two consecutive courses in a master's program in Educational Technology. In the first semester, participatory prototype development was carried out with the involvement of the students, which was afterward revised and offered to the students during the second semester for evaluation. During the first semester, our focus was on the functionality and quality of the OwlMentor responses while also trying to design a good user experience. As part of the revision during the transition between semesters, we applied the results gained from participatory

development with the students to enable a user experience that is both appealing and of high content quality. Given this background, our work can be understood as an exploratory study on the development and integration of AI-based applications in university teaching. The aim of the present study was to develop, implement, and evaluate the integration of OwlMentor into an existing university course, where students must prepare for the course lessons by reading and understanding specific scientific texts provided by the lecturer. Our goal is to improve text comprehension and reinforcement of course content by providing an AI-powered learning environment in which students can interact with a conversational AI and receive support in creating self-assessment questions and quizzes that they can practice and solve.

## 1.5 Research interest and hypothesis

Our research interest focuses on the use and impact of OwlMentor in a university course. We wanted to explore how students interact with this AI-based learning platform, focusing on their usage behavior, the benefits they derive from this usage, and the overall impact on their learning process. We utilized the TAM as a framework for system evaluation to predict the factors influencing students' engagement with OwlMentor. The TAM posits that two central constructs, Perceived Ease of Use and Perceived Usefulness, directly influence the Intention to Use a technology, which subsequently predicts Actual System Use. Based on this model, we formulated the following hypotheses:

- H1: Perceived Ease of Use is positively related to Intention to Use.
- H2: Perceived Usefulness is positively related to Intention to Use.
- H3: Higher Intention to Use leads to higher Actual System Use.

In addition to evaluating the system's acceptance, we also aimed to assess its effectiveness as an educational tool through the following hypothesis:

- H4: Higher Actual System Use leads to higher learning gains.

These hypotheses aim to capture the relationship between the design and usability factors of OwlMentor and investigate its acceptability and effectiveness as an educational tool. Our objective is not only to evaluate the practical impact of the OwlMentor but also to contribute to a broader understanding of how such technologies can be designed and implemented to improve educational outcomes. Additionally, we included a measure of general Self-Efficacy to explore its potential influence on students' engagement with OwlMentor. General Self-Efficacy, defined by [Bandura \(1982\)](#), refers to an individual's belief in their ability to succeed in various tasks and challenges. This concept is broader than digital Self-Efficacy, which focuses specifically on confidence in using technology. Research has shown that general Self-Efficacy is linked to improved motivation and learning strategies ([Pintrich and De Groot, 1990](#)). Given that previous studies have indicated a positive impact of generative AI on general Self-Efficacy ([Liang et al., 2023](#); [Wu and Yu, 2023](#)), we aimed to investigate whether interacting with OwlMentor could enhance students' overall confidence in managing the course demands. Based on these considerations, we have made an initial attempt to integrate

the AI-based application OwlMentor into teaching, examining both system acceptance and effectiveness in terms of text comprehension and student engagement.

## 2 Materials and methods

### 2.1 Participants

The participants of the present study were international students on the Master's degree program in Educational Technology (EduTech) at a German University. The students attended two consecutive courses in the Learning with Media module, Multimedia Learning I and Multimedia Learning II, which spanned two semesters. In the Multimedia Learning I course in the first semester, a prototype of the OwlMentor was presented and tested together with the students as part of a pilot study. In the second semester, the main study was carried out by implementing and evaluation a further developed version of the OwlMentor in the course. During the courses, OwlMentor was made available to the students for voluntary use in order to support them in their work with course-relevant scientific literature.

Students independently enrolled in the Multimedia Learning courses, which are part of a compulsory module within the master's program, via the university's internal registration system. Their consent to participate in the study was obtained after they had been introduced to OwlMentor. Apart from meeting the general enrollment requirements for the course, there were no other exclusion criteria. The course allowed a maximum of 25 EduTech students per cohort and, although students could register from their first semester, participation was recommended for those who were in their second or third semester.

The course participants were international students with an interdisciplinary background. All participants had at least a bachelor's degree in either a computer science subject (e.g., computer science, data science), education (e.g., teaching degree) or psychology. In terms of their previous knowledge, they could therefore be described as rather heterogeneous. There were two native English speakers in both semesters and all others had at least B2 level (requirement for the EduTech program). In the pilot study (Multimedia Learning I course), the sample consisted of 17 students with an average age of  $M=27.59$  years ( $SD=2.29$ ) and a balanced gender distribution (female:  $n=8$ , male:  $n=9$ ). Regarding their prior knowledge, the students stated in a self-assessment that they had on average moderate knowledge of multimedia learning ( $n=8$ ) and that some already had knowledge of multimedia formats ( $n=3$ ), multimedia principles ( $n=3$ ), and cognitive load theory ( $n=3$ ). In the main study (Multimedia Learning II course) the sample consisted of 16 of the former 17 students with an average age of  $M=27.38$  years ( $SD=2.19$ ) and a balanced gender distribution (female:  $n=8$ , male:  $n=8$ ).

### 2.2 Multimedia learning courses

Both Multimedia Learning Courses were part of the module Learning with Media which has a duration of two semesters and is compulsory in the master program. The Multimedia Learning I course teaches the basics of learning with multimedia

instructions, such as the Cognitive Theory of Multimedia Learning (Mayer, 2014), the Cognitive Load Theory (Sweller, 2011) and basic principles of multimedia learning such as the Multimedia Principle, Modality Principle, or the Redundancy Effect (Mayer and Fiorella, 2021). The seminar was held in summer term 2023 in a blended learning format, comprising eight synchronous classroom sessions, four asynchronous assignment sessions, a presentation session, and a final exam. During this seminar, the OwlMentor prototype was presented, tested, and evaluated. The Multimedia Learning II course teaches further principles of multimedia learning: Expertise-Reversal Principle, Split-Attention Principle, Worked Example Principle, Principles based on Social Cues and Emotional Design Principle (Mayer and Fiorella, 2021). This seminar was also held in a blended learning format in the winter term 2023/24. The course included six on-site content sessions and six asynchronous sessions for preparing scientific literature, held on alternating weeks. The first synchronous session was led by lecturers; subsequent sessions were prepared and conducted by student groups. Additionally, there were an organizational session, a mock exam session, and a final exam session. Students prepared chapters from the Handbook of Multimedia Learning (Mayer and Fiorella, 2021) and relevant research articles. At the seminar's start, OwlMentor and its functions were introduced, with anonymous access, a manual, and instructional videos provided. Throughout the seminar, the advanced OwlMentor version was freely available for dialogue, question generation, practice quizzes, and AI-generated feedback.

## 2.3 Pilot study (first semester)

The main aim of our pilot study was to test the technical functionality of the OwlMentor and the quality of its AI-generated answers and questions. We used the System-Usability-Scale (SUS; Brooke, 1996), a 10-item questionnaire with a reported internal consistency of Cronbach's  $\alpha = 0.91$  (Bangor et al., 2009). In this early version, the dialog function and question generation function were tested for one course topic. At first students were introduced to the Cognitive Theory of Multimedia Learning (Mayer, 2014) and answered nine questions of different educational objective taxonomy levels (remembering, understanding, applying, evaluating; Bloom, 1956). Example questions include "Define the term 'multimedia learning' according to Mayer and Fiorella's handbook," "What Are the Cognitive Processes Involved in Active Learning (SOI model)?" and "Explain the limited capacity assumption of the cognitive theory of multimedia learning." Then the students, divided into groups, asked these questions to the OwlMentor and rated its responses from 1 (very good) to 6 (unsatisfactory). To assess the automatic question generation function, students created six questions per group using selected text passages. They rated the generated questions on difficulty (1 = low, 5 = high) and usefulness (1 = low, 5 = high). Examples of generated questions are "Which of the following is NOT a type of Cognitive Load? Options: A: Intrinsic Load, B: Extrinsic Load, C: Germane Load, D: Emotional Load" and "Which of the following is NOT one of the three cognitive processes essential for active learning according to the SOI model? Options: A: Selecting relevant material, B: Organizing selected material, C: Integrating selected material with

existing knowledge, D: Evaluating the effectiveness of the learning material."

Students rated the response quality with a mean value of 2.19 ( $SD = 0.99$ ), with ratings for Remembering ( $M = 2.29$ ,  $SD = 0.83$ ), Understanding ( $M = 1.88$ ,  $SD = 0.33$ ), Applying ( $M = 2.8$ ,  $SD = 2.21$ ), and Evaluation ( $M = 1.63$ ,  $SD = 0.95$ ). The difficulty of generated questions had a mean value of 1.92 ( $SD = 0.26$ ) and usefulness a mean value of 3.29 ( $SD = 0.46$ ). The students rated the usability of the prototype on the SUS with a value of 56.25, corresponding to the adjective "ok" (Bangor et al., 2009). A short user feedback questionnaire ( $n = 5$ ) and discussions with the students revealed that they found it easy to chat and generate questions with OwlMentor but had mixed opinions about its understanding of user queries, clarity of responses, and overall satisfaction. Specifically, they noted that the AI's responses were sometimes too long and not sufficiently precise, making it challenging to find the exact information they needed. They also mentioned that the user interface was not intuitive and that response times were slow, affecting their overall experience. Moreover, students expressed a desire for additional functionalities, such as the ability to upload or use PDF's in the application. Overall, they believed that while OwlMentor supported their understanding of the course literature, there was room for improvement. Based on their suggestions, we implemented several modifications to OwlMentor, as detailed in Section 2.5.3.

## 2.4 Main study design

The aim of the main study was to integrate and evaluate the extended version of OwlMentor in the Multimedia Learning II course. It was an exploratory study in a pre-/posttest design. Due to the small number of participants and to ensure that all students had equal opportunities in the course, no control group design was used. The dependent variables for the evaluation of OwlMentor were students' learning gains (difference pre/posttest), usability in terms of the TAM model (Perceived Ease of Use, Perceived Usefulness, Intention to Use, Actual System Use), and an expert assessment of OwlMentor's response quality with two independent raters. The log data was analyzed to assess the Actual System Use of the application. The participants' Self-Efficacy was recorded as an additional variable and the interactions with the OwlMentor were also analyzed qualitatively to get an impression of how the students engaged with the application.

## 2.5 OwlMentor

OwlMentor is an AI-powered web application designed to assist students in comprehending scientific texts required for their courses. This section provides an overview of OwlMentor's structure, development, core functionalities, and technical implementation, addressing key aspects of its design.

### 2.5.1 User interface and core functionalities

The user interface comprises four main sections—Dashboard, Chat, Questions, and Quiz—accessible through a navigation bar.

The Dashboard (Figure 1) provides an overview of course details, including topics, session dates, and linked required literature. It also



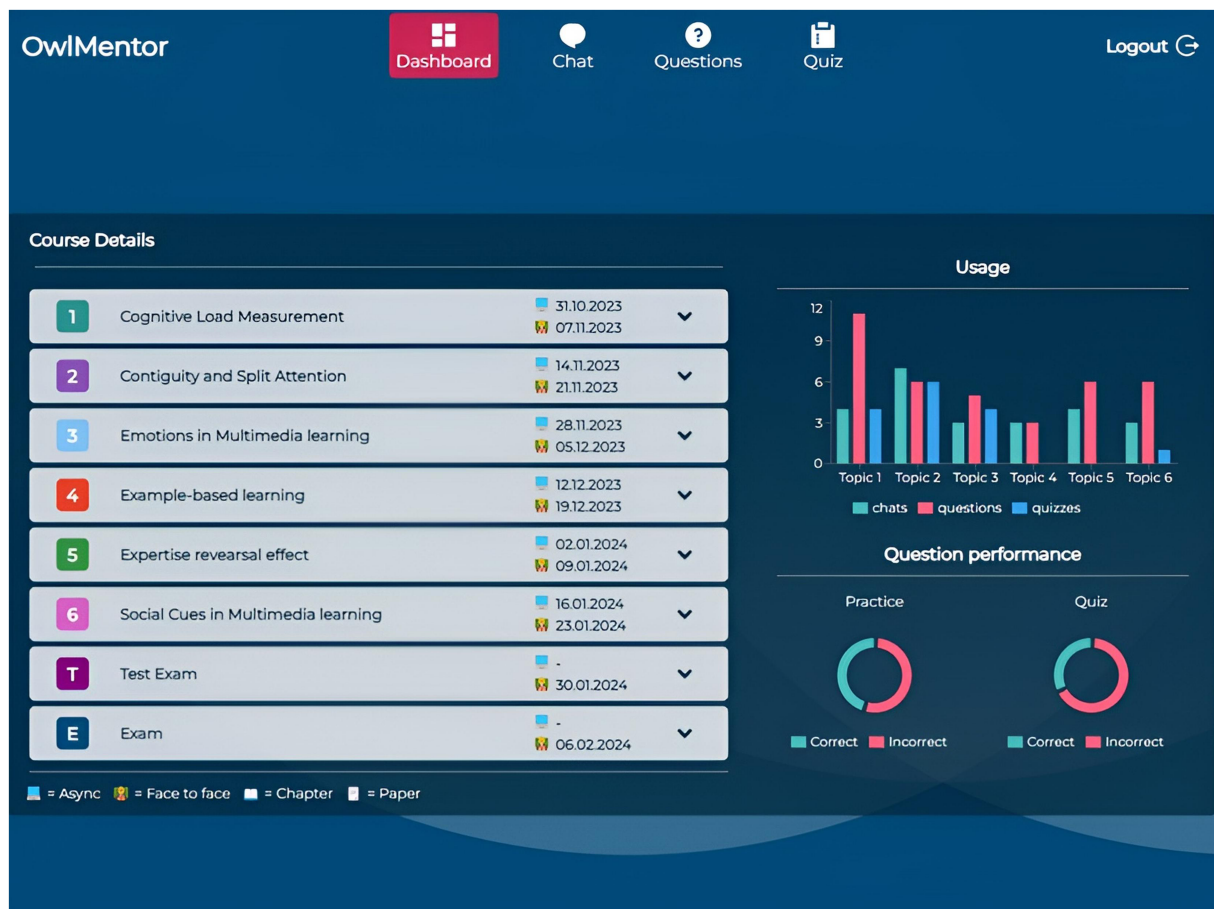


FIGURE 1  
Application screenshot dashboard.

displays usage statistics such as the number of chats initiated, questions generated, and quizzes taken, allowing students to monitor their progress. In the Chat section (Figure 2), students can continue an existing conversation or start a new one about specific course documents. The chat interface allows them to engage in free-form dialogues with OwlMentor. This document-based chat enables students to ask questions, seek clarifications, and explore concepts in depth. By retrieving relevant information from the selected document, OwlMentor provides accurate, context-specific responses based on students' queries - such as summarizing content, explaining concepts, or defining terms - which may assist them in understanding the material.

The Questions section enables students to generate multiple-choice questions automatically by providing text from course documents. They can input specific text excerpts, upload highlighted PDFs, or let the system select random sections. Generated questions appear in an "On Review" section, where students review and validate each question before adding it to their personal question bank. This automatic question generation prompts students to actively engage with the content by critically assessing the quality and relevance of each generated question and its answers. By evaluating correctness and clarity, students deepen their understanding and reinforce key concepts (Figures 3, 4).

The Quiz section allows students to create custom quizzes by selecting a title, the number of questions (3, 6, or 9), and specific questions from their question bank. The quiz interface (Figure 5) presents questions sequentially, enabling navigation and answer submission. After completing a quiz, students receive immediate feedback indicating which questions they answered correctly or incorrectly, along with a brief motivational message based on their performance. They can request detailed, AI-generated feedback on each question to enhance their comprehension (Figure 6). This quiz creation and feedback functionality allows students to assess their knowledge and understand the reasoning behind each answer, as the AI provides explanations that clarify misunderstandings and reinforce learning.

### 2.5.2 Development phases

OwlMentor was developed iteratively over three main phases, incorporating user feedback and technical enhancements. Prior to submitting the proposal for the OwlMentor teaching and research project, the authors conducted a Proof of Concept (Version 1) phase to assess the feasibility of integrating a RAG system. By developing a basic RAG architecture with a simple chat interface, they confirmed its viability and proceeded with further development. During the Prototype Testing with Students (Version 2) phase, a pilot study was conducted (Section 2.3) where students tested the



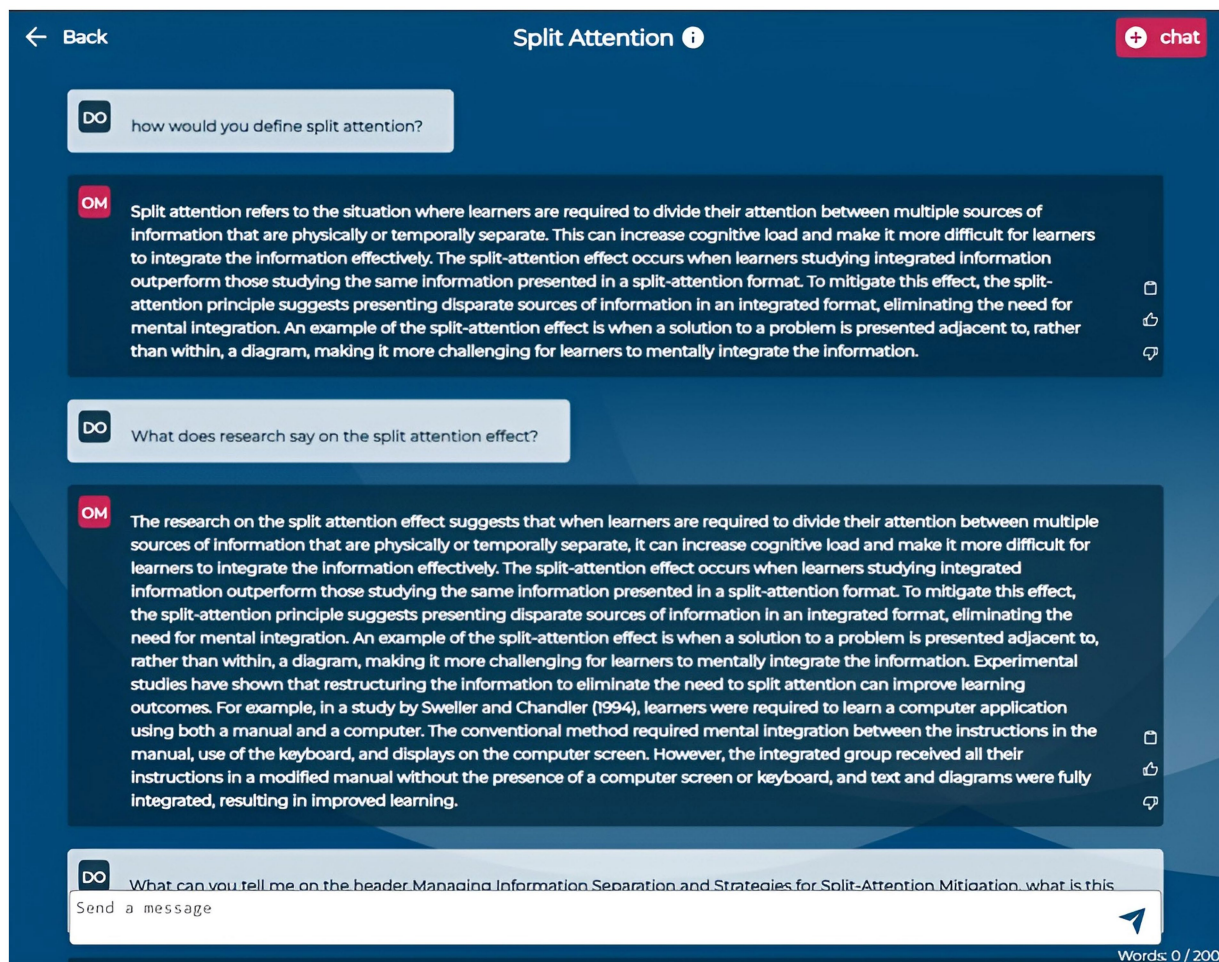


FIGURE 2  
Application screenshot dialog function.

prototype and provided feedback on usability, response quality, and satisfaction. Based on their input, the following modifications were made:

- **User Interface Enhancements:** Students reported that the interface was unintuitive and that response times were slow, affecting usability as reflected in the SUS scores. The interface was redesigned for better clarity and navigation, incorporating a cleaner layout and more intuitive controls. Performance was optimized by implementing streaming responses in the chat function, reducing perceived latency.
- **Improved Response Quality:** The AI's responses were too long and imprecise because the language model searched across all documents that have similar wording (e.g., multimedia learning, cognitive load), making it hard to find correct contexts. To address this, the retrieval strategy was refined to focus on specific documents using logical routed retrieval and adjusted system prompts to produce shorter, more precise answers.
- **Expanded Functionality:** Students requested the ability to generate questions from PDFs, as the initial version only allowed copying and pasting text. A feature was added enabling users to upload highlighted PDFs for automatic question generation and

an option to create random questions from selected documents was introduced.

- **Additional Engagement Features:** To enhance engagement, a quiz function was introduced allowing students to compile questions into custom quizzes where correct answers were not immediately revealed, increasing the challenge. We also added the option to receive elaborative feedback on quiz questions to support deeper learning.

In the Final Version (Version 3), these improvements were integrated, offering enhanced retrieval processes, detailed feedback options, and a streamlined user experience for the main study. The user interface designs for all three versions of OwlMentor can be viewed in [Appendix 6](#).

### 2.5.3 Technical implementation

OwlMentor's AI functionalities rely on OpenAI's GPT-3.5 Turbo model (version 0613) to provide accurate, context-specific responses, generate questions, and offer feedback. Each core function—dialogue, question generation, and feedback—sends tailored requests to the OpenAI API, optimizing performance and precision. These requests are guided by specific system prompt templates ([Appendix C](#)) that

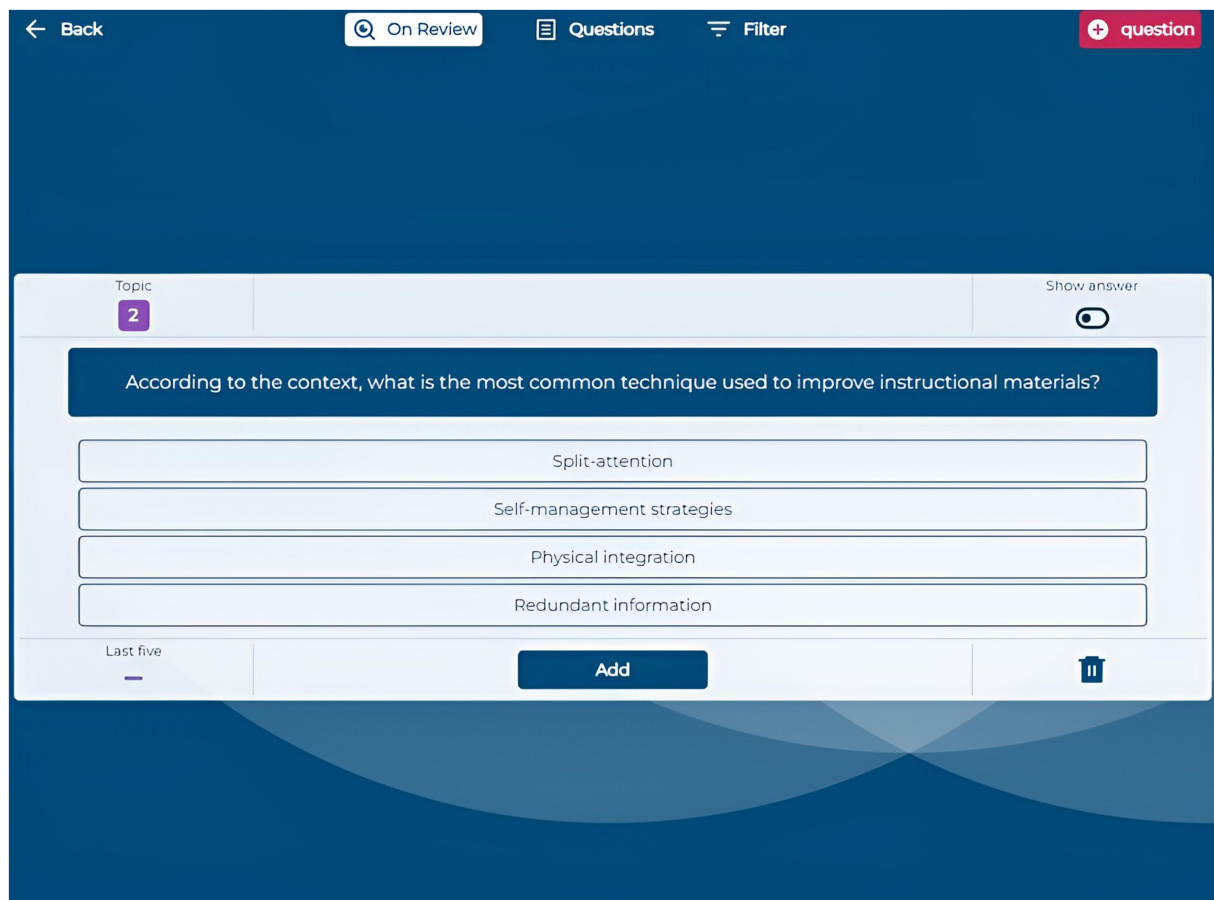


FIGURE 3  
Application screenshot on review area.

instruct different instances of the LLM (e.g., Strategy-, Response Model) on how to process user queries and respond appropriately based on the context.

The backend architecture of OwlMentor is built with Python and FastAPI, using MongoDB for data storage, and the Annoy library to construct a vector database that allows efficient similarity searches for document retrieval. A vector database stores vector representations (or embeddings) of document chunks, enabling fast searches based on the similarity between user queries and the stored vectors.

To generate these vector representations, OwlMentor utilizes OpenAI's text-embedding-ada-002 model. Documents are segmented into chunks of approximately 300 words with 20-word overlaps to maintain context across sections. The Natural Language Toolkit (NLTK) ensures that sentences are not split during this process, preserving the integrity of the text. The embeddings are stored in an Annoy-based vector index with 30 trees and an embedding dimension of 1,536. This vector index serves as a critical component of the RAG system (Figure 7), enabling fast and accurate retrieval of relevant document sections in response to user queries. The RAG system is central to OwlMentor's ability to provide context-specific assistance. When a user submits a query, the system first determines whether it pertains to a relevant document within the vector database. If so, the RAG system applies logical routed retrieval by navigating the document's structure and metadata (e.g., section titles and key topics)

to identify the most pertinent sections. This focused retrieval enhances accuracy by narrowing the search to the most relevant parts of the document. The Strategy Model, a LLM (GPT 3.5-Turbo 0613) with specific instructions, further refines the retrieval process. It assesses whether the user's query is relevant to the document's key topics. If so, it uses logical routed retrieval to direct the query to the most relevant sections. For non-relevant queries, the Strategy Model produces direct outputs like "[EXIT]" or "[INFO]," indicating no retrieval is needed. For relevant queries, it generates refined search queries through query decomposition, breaking down complex queries into simpler, more focused ones, ensuring the most relevant sections are returned for further processing. The retrieved document sections are then passed to the Response Model, which generates responses that are aligned with the user's query and the content of the document. This approach reduces the risk of hallucinations or irrelevant information, ensuring that the system consistently produces responses grounded in the provided course material. The temperature setting of the LLM is configured to 0.1 for the Response Model, and to 0 for the Strategy Model and for feedback generation. This ensures that responses remain precise and grounded in the provided information. The maximum tokens parameter is set to 800 for Response Model, 200 for the Strategy Model, and 250 for feedback. The streaming option is enabled to enhance response times, except for the Strategy Model.

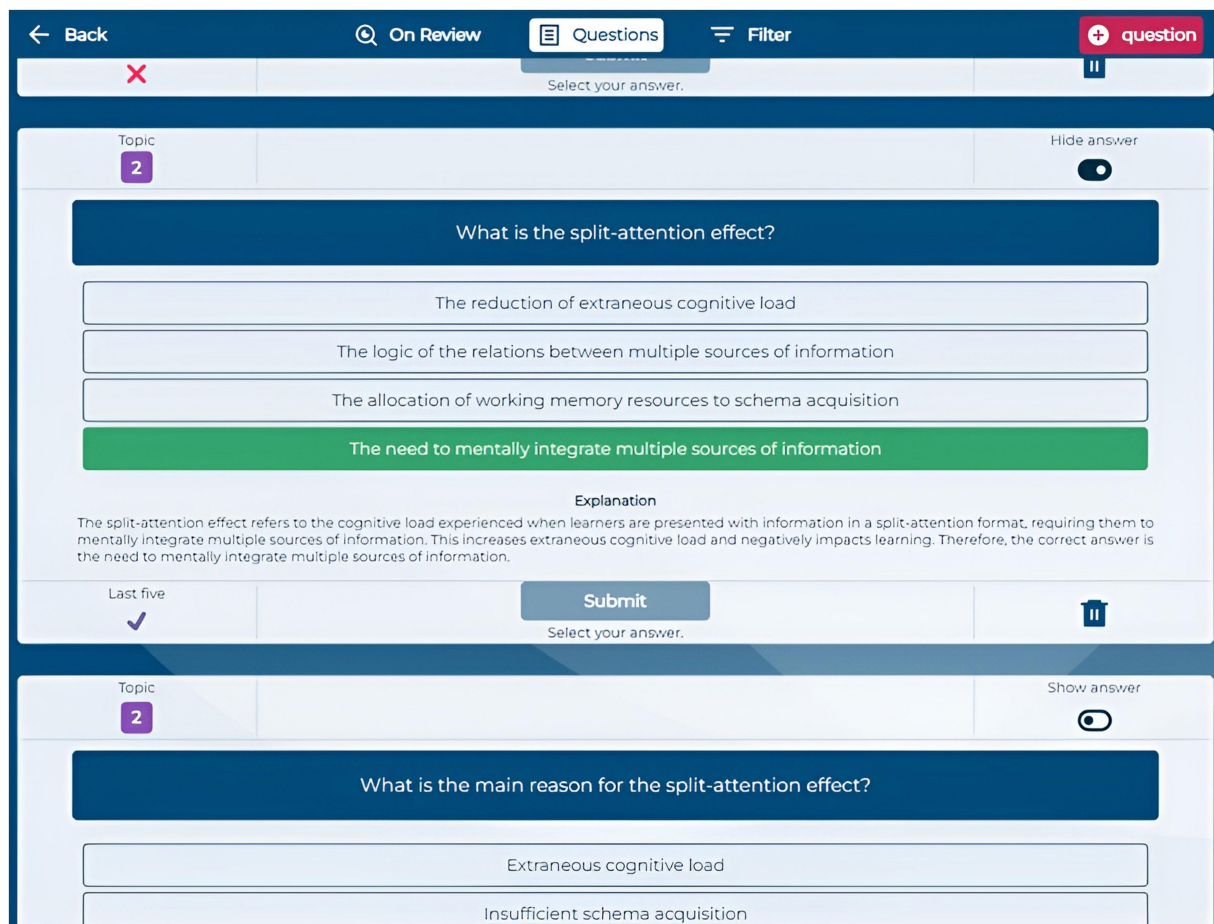


FIGURE 4  
Application screenshot question generation.

OwlMentor's frontend architecture is developed using React JS, ensuring a responsive and interactive user experience. This frontend enables seamless navigation between different functionalities, dynamic content rendering, and real-time updates to track user activity. Detailed lists of backend and frontend dependencies are provided in [Appendices 4, 5](#), offering comprehensive technical insights into OwlMentor's development.

## 2.6 Instruments

### 2.6.1 Pre-and posttest

The pre-and posttest consist of the same 10 items: two items on each of the following five topics of the Multimedia Learning II course: Expertise-Reversal Principle, Split-Attention Principle, Worked Example Principle, Principles based on Social Cues and Emotional Design Principle ([Mayer and Fiorella, 2021](#)). For each topic, two questions were carefully crafted based on specific levels of the education objectives taxonomy ([Bloom, 1956](#)). One question was designed to assess the basic level of 'remembering', while the other was designed to assess the higher levels, specifically 'understanding' or 'applying'. With regard to the question format, the pre-posttest consists of four multiple choice (MC) and six open questions (OP). Example items include:

"What is the split-attention effect? Select the correct answer (Remembering, MC)," "In a study of science learning by [Leslie et al. \(2012\)](#), more experienced students learned better from listening-only texts, while novices benefited more from audiovisual presentations. How would you explain these results using the principle of reversal of subject knowledge? (Understanding, OP)" or "As a primary school teacher who wants to teach the concept of addition using worked examples, how could you improve the effectiveness of your worked examples? (Applying, OP). The pretest was carried out at the beginning of the seminar and the posttest one week before the final exam.

### 2.6.2 Questionnaire TAM (perceive ease of use, perceived usefulness, intention to use) + self-efficacy

A structured self-assessment questionnaire (for complete questionnaire see [Appendix](#)) was used to evaluate the constructs of the technology acceptance model (Perceived Ease of Use, Perceived Usefulness, Intention to Use) and Self-Efficacy. The questionnaire contained proven scales that were selected and modified from the existing literature. Perceived Ease of Use was assessed using a four-item scale derived from [Venkatesh and Davis \(2000\)](#), which showed Cronbach's alpha values between  $\alpha=0.86$  and  $\alpha=0.98$ , as stated by [Venkatesh and Davis \(2000\)](#). Sample items include "I find OwlMentor



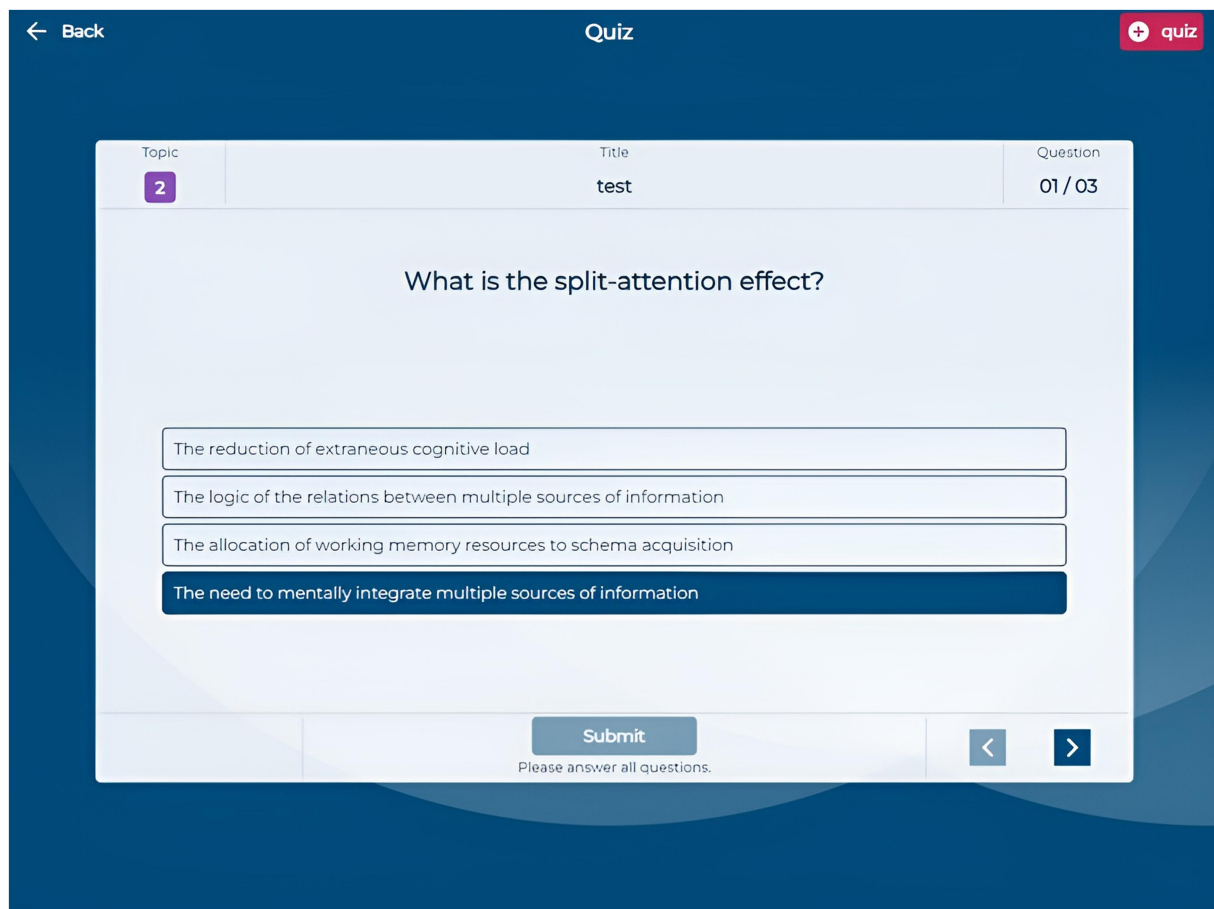


FIGURE 5  
Application screenshot quiz.

easy to use” and “Interacting with OwlMentor does not require much attention.” The scale for Perceived Usefulness was adapted from a three-item scale by [Liaw \(2008\)](#), which had a Cronbach’s alpha reliability of  $\alpha=0.90$ . A sample item for Perceived Usefulness is “I believe OwlMentor is a useful learning tool.” Intention to Use was measured by using a three-item scale from [Liaw \(2008\)](#) with a Cronbach’s alpha of approximately  $\alpha=0.89$ . Sample items are “I intend to use OwlMentor’s content to assist my learning” or “I intend to use OwlMentor to assist my learning in the future.” The responses on all scales were recorded on a seven-point Likert scale ranging from “strongly disagree” to “strongly agree.” Moreover, Self-Efficacy was measured via self-reports. For this purpose, we used the nine-item scale by [Pintrich and De Groot \(1990\)](#), which has an internal consistency of Cronbach’s alpha = 0.89 to 0.92 ([Pintrich and De Groot, 1990](#)). Example items are “I am certain I can understand the ideas taught in this course.” or “I know that I will be able to learn the material for this class.” For the Self-Efficacy scale, we used the original 5-point Likert-scale ranging from “strongly disagree” to “strongly agree.”

### 2.6.3 Reliability

The reliability of the measurement instruments was assessed using Cronbach’s alpha ( $\alpha$ ). The reliability for Perceived Usefulness ranged from  $\alpha=0.92$  to  $\alpha=0.97$  across the three time points. Reliability for Perceived Ease of Use ranged from  $\alpha=0.84$  to  $\alpha=0.96$ . Intention to Use showed a reliability range from  $\alpha=0.92$  to  $\alpha=0.99$ . Self-Efficacy

showed a high reliability with a range of  $\alpha=0.93$  to  $\alpha=0.98$ . For the pretest, the reliability was  $\alpha=0.78$ , for the posttest  $\alpha=0.80$ . These values show that the measurement instruments used in the study were consistently reliable across the various constructs and time points.

## 2.7 Log data/ system use

During use, we saved the conversations, dialogs, and automatically generated questions for qualitative analyses. We also collected the following quantitative log data to quantify the Actual System Use: Number of conversations created, number of dialogs conducted, number of questions generated, number of questions deleted, number of quizzes created, number of quizzes completed, number of feedback received. Each of these quantitative log data represents a user activity. We calculate the Actual System Use by adding up all these user actions to one value. Furthermore, the use of OwlMentor was anonymous and no personal data was collected.

## 2.8 Procedure

First, the revised version of the OwlMentor was presented to the students as part of the Multimedia Learning II seminar and they were



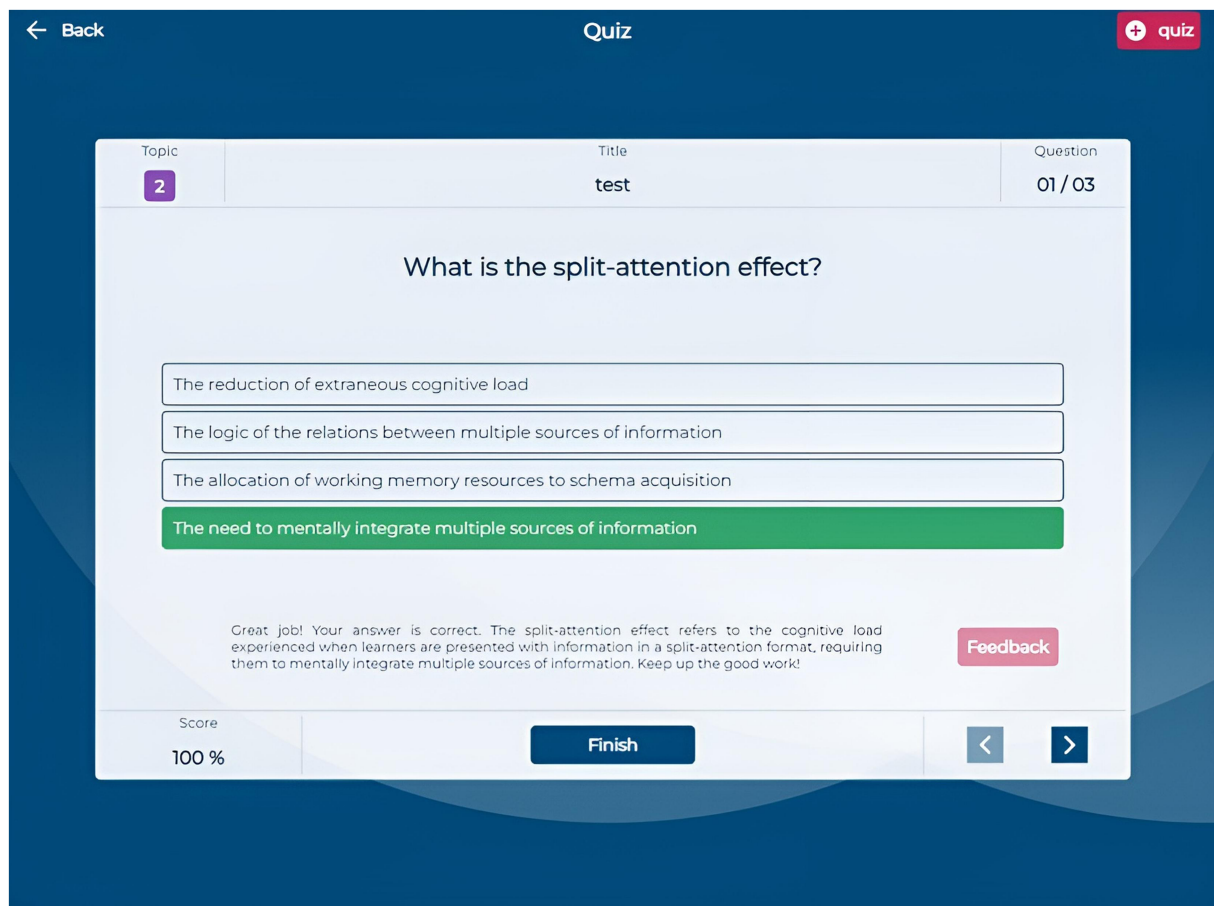


FIGURE 6  
Application screenshot quiz feedback.

asked to use the OwlMentor throughout the course. Figure 8 shows a timeline of the course details and the measurement times. The use of the OwlMentor was voluntary and anonymous. The participants first signed a declaration of consent for the study. They then completed the pretest under the supervision of the course instructors. The pre-test and post-test were designed to assess students' understanding of the key principles covered in the Multimedia Learning II course. The instructors first created a series of questions based on the course literature relevant to the exam and covering the different levels of Bloom's Taxonomy (Bloom, 1956). After a collaborative review, the lecturers selected two questions per topic, which were then formatted into a digital questionnaire using MS Forms. This structured assessment allowed for a sample-based evaluation of students' baseline knowledge and learning gains in key topic areas. Following the pretest, the students were then granted anonymous access to OwlMentor. To facilitate use and ensure effective interaction with the OwlMentor, a short user manual and instructional videos were made available on the digital course platform. The evaluation of the acceptance of the platform was carried out using the TAM questionnaire, which was completed at three points during the course: after completion of the second topic (T2), after the fourth topic (T3) and after the final exam at the end of the seminar (T5). The posttest was carried out at T4 under the supervision of the course instructors as a practice exam.

## 2.9 Analysis

We conducted a comprehensive analysis that included descriptive analysis, hypothesis-testing, dialog functionality evaluation, and explorative analysis. Descriptive analyses were conducted to summarize data on platform usage, questionnaire responses, and pre-and post-test performance.

Hypothesis testing included correlation analyses and linear regression. For H1 (Perceived Ease of Use and Intention to Use) and H2 (Perceived Usefulness and Intention to Use), correlation analyses were performed at three time points (T2, T3, T5). To test if Perceived Usefulness and Perceived Ease of Use are significant predictors of the Intention to Use, a multiple regression analysis was conducted. The dependent variable was Intention to Use, while the independent variables were Perceived Usefulness, Perceived Ease of Use, and time. For H3 (Intention to Use and Actual System Use), We conducted a series of linear regression analyses to investigate the relationship between Intention to Use and subsequent periods of Actual System Use. We measured Actual System Use during three time periods (TP2, TP3, TP4) and Intention to Use at two points in time (T2, T3). Specifically, we examined how Intention to Use at T2 predicted Actual System Use at TP2 and TP3, and how Intention to Use at T3 predicted Actual System Use at TP4.

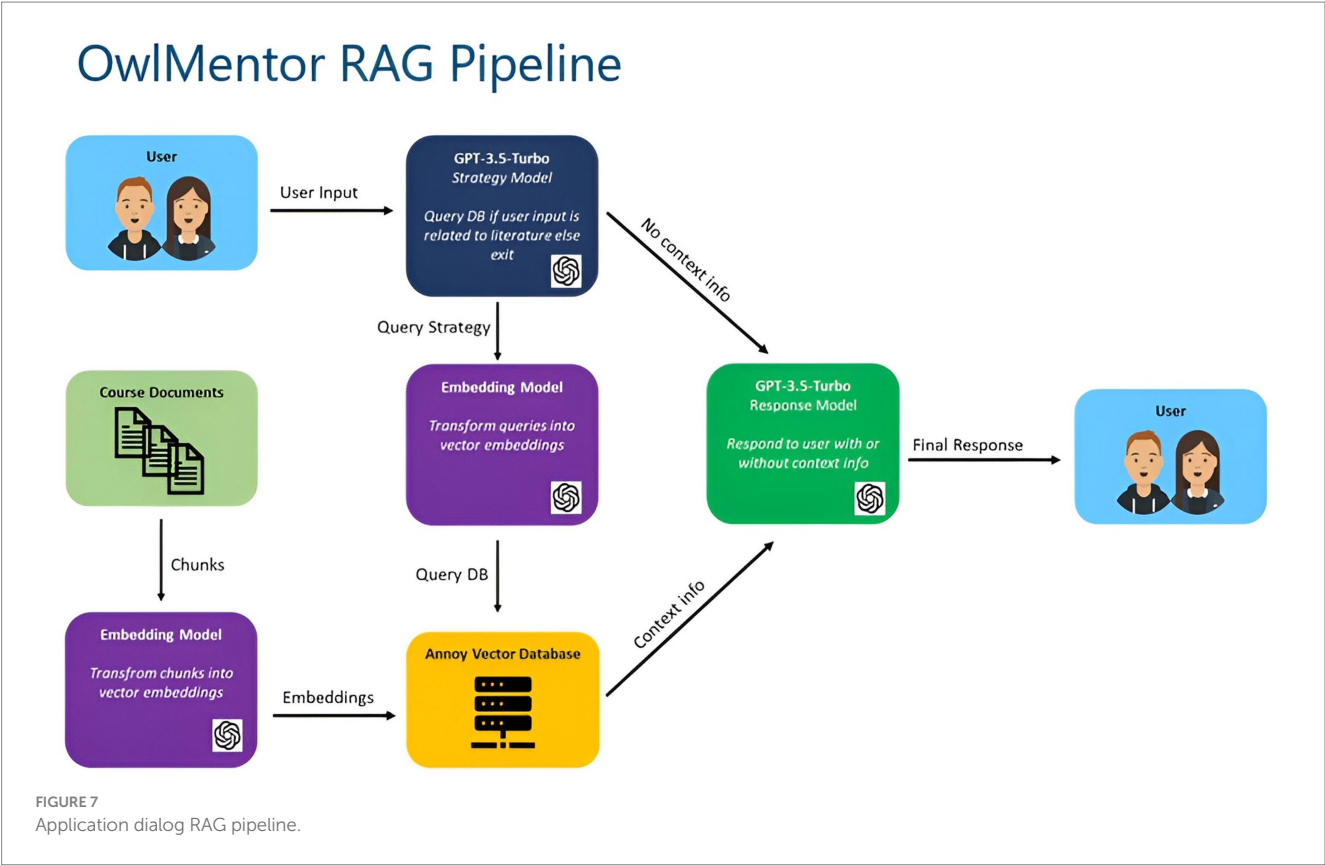


TABLE 1 User request categories for dialog functionality analysis.

Category	Description
Concept explanations and definitions	This category includes requests seeking clear definitions, descriptions, or explanations of specific terms, principles, or concepts. Users wanted basic or detailed information about certain topics to improve their understanding.
Summarization requests	This category includes requests that seek brief summaries, key points, or essential takeaways from a text, chapter, or study. Users wanted a condensed version of the content, often in the form of sentences or bullet points, to get a quick overview.
In-depth explanations	This category includes requests that seek comprehensive explanations or detailed information about specific concepts, principles, or research findings. Users wanted to gain a deeper understanding of complex topics, often aiming to go beyond surface-level definitions.
Relationship and connection	This category includes requests that seek to understand the relationships or connections between different concepts, principles, or effects. Users wanted to comprehend how various elements interact with each other and what impacts they have on one another.
Importance and rationale	This category includes requests that seek the reasons or justification for the importance of a study, paper, or specific information. Users wanted to understand why something is important and what impact or relevance it holds.
Targeted information requests	This category includes requests that seek specific information, concepts, or variables. Users wanted to obtain targeted data or key concepts about a topic, often in the form of lists or brief descriptions.
Practical examples and applications	This category includes requests that seek practical examples or applications of specific concepts, methods, or principles. Users wanted to see concrete examples to better understand abstract ideas or to know how to apply them in practice.
Other	This category includes requests that did not fit into the specific previously defined categories.

- Regression 1: Intention to Use at T2 predicting Actual System Use during TP2
- Regression 2: Intention to Use at T2 predicting Actual System Use during TP3
- Regression 3: Intention to Use at T2 predicting Actual System Use during TP4
- Regression 4: Intention to Use at T3 predicting Actual System Use during TP3

These analyses were conducted to determine whether Intention to Use significantly predicted subsequent Actual System Use. H4 (Actual System Use and learning gain) was tested using a linear regression for Actual System Use in period 1 with the dependent variable learning gain and the independent variable Actual System Use.

The dialog functionality analysis consisted of two parts: categorizing user requests and conducting an expert evaluation of OwlMentor’s responses. When evaluating dialog functionality, user requests were categorized to understand the nature of interactions on the platform. The user requests were categorized into seven main categories, along with an “Other” category for requests that did not fit into the predefined categories. The categories are as follows in Table 1:

Furthermore, two experts reviewed all dialogs between users and OwlMentor and evaluated OwlMentor’s responses. Therefore, the experts rated the responses using a scale where 1 indicates “very good” and 6 indicates “unsatisfactory.” The inter-rater reliability was measured using ICC3 indicating good reliability between the raters with a value of 0.82.

Finally, we examined the influence of Self-Efficacy on other variables measured in the study using correlational analysis. This included analyzing the influence of Self-Efficacy on Perceived Ease of Use, Perceived Usefulness, and Intention to Use at three measurement time points (T2, T3, T5), as well as examining the relationship between Self-Efficacy and Actual System Use throughout the course duration.

### 3 Results

In this section, we present the results of the study, focusing on the reliability of the measurement instruments, descriptive analysis of OwlMentor usage and questionnaire responses, pre-and post-test performance, hypothesis testing, and an exploratory analysis of the dialog functionality and the influence of self-efficacy.

#### 3.1 Descriptive analysis

In this subsection, we provide a detailed analysis of the descriptive statistics for OwlMentor usage, responses from the questionnaire, and pre-and post-test performance. This includes an overview of the engagement patterns and self-reported measures from participants throughout the study period.

##### 3.1.1 OwlMentor usage

Table 2 presents the usage statistics of the OwlMentor application over four different periods (Overall, TP1, TP2, TP3). Overall usage across the entire course duration shows that the most frequently used functions were messages indicating that students primarily engaged with the platform through this interaction. As can be seen from the table, engagement was highest in TP1, while it decreased significantly in TP2. However, there was an increase in engagement from TP2 to TP3, which was closer to the final exam. This pattern can be observed across different measures such as conversations, messages, and questions created. User activity was most frequent in TP1, but despite the initial decrease in TP2, user activity increased again in TP3. In addition, user engagement was higher in the first period with 11 active users compared to only 6 active users in each of the following periods.

##### 3.1.2 Questionnaire

The questionnaire was used to measure Perceived Ease of Use, Perceived Usefulness, Intention to Use, and Self-Efficacy at

TABLE 2 OwlMentor usage statistics.

N = 16	Time period	Overall	TP1	TP2	TP3
Actual usage	M	21.38	9.38	4.44	7.56
	SD	32.64	10.82	9.20	15.21
	Sum	342	150	71	121
Conversations	M	2.25	1.00	0.56	0.69
	SD	2.84	1.10	1.03	1.25
	Sum	36	16	9	11
Messages	M	15.13	6.25	3.00	5.88
	SD	24.21	7.11	6.15	13.79
	Sum	242	100	48	94
Message likes	M	0.38	0.25	0.06	0.06
	SD	1.03	1.00	0.25	0.25
	Sum	6	4	1	1
Questions	M	1.88	1.00	0.5	0.38
	SD	3.07	2.03	1.41	0.89
	Sum	30	16	8	6
Temporary questions	M	0.56	0.19	0.13	0.25
	SD	1.55	0.75	0.50	0.58
	Sum	9	3	2	4
Deleted questions	M	0.31	0.25	0	0.06
	SD	0.70	0.58	0	0.25
	Sum	5	4	0	1
Practiced questions	M	0.88	0.44	0.19	0.25
	SD	2.63	1.75	0.75	1.00
	Sum	14	7	3	4

No quizzes were created or completed and no feedback was received.

TABLE 3 Questionnaire data.

Group	Variable	T2 (n = 15) M (SD)	T3 (n = 13) M (SD)	T5 (n = 10) M (SD)
All users	PU	3.66 (1.64)	3.36 (1.79)	2.73 (1.34)
	PE	4.27 (1.73)	5.04 (1.07)	4.25 (1.49)
	IU	2.82 (1.86)	2.72 (1.77)	2.10 (1.56)
	SE	3.66 (1.19)	3.58 (0.78)	3.57 (0.80)
Users	PU	3.80 (1.41) (n = 10)	3.85 (1.88) (n = 9)	3.19 (1.36) (n = 7)
	PE	4.05 (1.73) (n = 10)	4.97 (1.23) (n = 9)	4.46 (1.70) (n = 7)
	IU	2.87 (1.55) (n = 10)	3.00 (1.95) (n = 9)	2.33 (1.76) (n = 7)
	SE	3.40 (1.29) (n = 10)	3.57 (0.88) (n = 9)	3.43 (0.92) (n = 7)
None users	PU	3.40 (2.20) (n = 5)	2.25 (1.00) (n = 4)	1.67 (0.33) (n = 3)
	PE	4.70 (1.86) (n = 5)	5.19 (0.66) (n = 4)	3.75 (0.87) (n = 3)
	IU	2.73 (2.58) (n = 5)	2.08 (1.26) (n = 4)	1.56 (0.96) (n = 3)
	SE	4.18 (0.85) (n = 5)	3.61 (0.58) (n = 4)	3.89 (0.30) (n = 3)

Perceived usefulness (PU), perceived ease of use (PE), intention to use (IU): 1 = strongly disagree, 7 = strongly agree, self-efficacy (SE): 1 = strongly disagree, 5 = strongly agree.

three time points (T2, T3, T5). As shown in the questionnaire data in [Table 3](#), mean Perceived Usefulness decreased from T2 to T3 and continued to decrease from T3 to T5, moving from somewhat disagree and neutral at T2 and T3 to somewhat disagree and disagree at T5. Perceived Ease of Use increased from T2 to T3, but then decreased from T3 to T5. Intention to Use



remained stable from T2 to T3 but decreased from T3 to T5. Self-Efficacy remained relatively stable across all three time points and showed only minimal changes. For the users of the AI-based learning platform, the mean Perceived Usefulness initially increased from T2 to T3 and then decreased from T3 to T5. Perceived Ease of Use showed an increase from T2 to T3 and a slight decrease from T3 to T5. Intention to Use remained relatively stable from T2 to T3, with a slight decrease from T3 to T5. Self-Efficacy remained relatively stable across all three time points with minimal changes. For the non-users, mean Perceived Usefulness decreased consistently from T2 to T5. Perceived Ease of Use increased from T2 to T3 and then decreased at T5. Intention to Use showed a decreasing trend from T2 to T5. Self-Efficacy remained relatively stable, with a decrease from T2 to T3 and slight increase from T3 to T5. Comparing users and non-users, users generally reported higher Perceived Usefulness and Intention to Use at all three time points, while non-users showed a more pronounced decrease in Perceived Usefulness over time. Perceived Ease of Use increased similarly for both groups from T2 to T3, but users maintained higher Perceived Ease of Use at T5 compared to non-users. Self-Efficacy was slightly higher for non-users at all time points, although the differences were minimal.

### 3.1.3 Pre- posttest performance

The results before and after the test show that users' overall performance improved significantly after taking part in the course,  $t(13) = -3.56, p < 0.001$ , *Cohen's d* = 1.93. There were improvements in all areas of the individual topics, with users scoring higher in the post-test than in the pre-test. The increases were more pronounced for some topics than for others, which is mainly due to lower values in the pretest. The detailed statistics in the Table 4 provide a comprehensive overview of these improvements and illustrate the participants' overall learning progress in the different subject areas during the course.

Table 5 shows the results of the pre-and post-tests as well as the difference for users. For users of the platform, overall performance improved from the pre-test to the post-test, with substantial gains across all topics. The learning gains for users were consistent across the different subject areas.

Non-users also exhibited performance improvements from the pre-test to the post-test, though the overall gains were slightly lower compared to users (Table 6). Non-users showed improvements on all topics with varying degrees of progress in each topic. Comparing users and non-users, users generally demonstrated higher overall learning gains across all topics. Both groups showed improvement, but users had more pronounced gains in most topics. Non-users also improved, but their performance increases were generally lower than those of the users.

TABLE 4 Pre- posttest performance all users.

All users <i>M</i> ( <i>SD</i> )	Pre ( <i>n</i> = 14)	Post ( <i>n</i> = 14)	Diff ( <i>n</i> = 14)
Overall	4.68 (3.07)	11.57 (3.62)	6.89 (3.57)
Topic 1	1.04 (0.41)	2.14 (0.74)	1.11 (0.81)
Topic 2	1.14 (1.03)	2.21 (1.03)	1.07 (1.40)
Topic 3	1.29 (1.07)	2.18 (0.91)	0.89 (1.30)
Topic 4	0.68 (1.01)	2.71 (1.01)	2.04 (1.03)
Topic 5	0.54 (0.50)	2.32 (0.99)	1.79 (1.07)

TABLE 5 Pre- Posttest Performance users.

Users <i>M</i> ( <i>SD</i> )	Pre ( <i>n</i> = 10)	Post ( <i>n</i> = 10)	Diff ( <i>n</i> = 10)
Overall	4.95 (3.46)	12.30 (3.34)	7.35 (3.85)
Topic 1	1.00 (0.47)	2.25 (0.79)	1.25 (0.86)
Topic 2	1.30 (1.06)	2.45 (0.90)	1.15 (1.56)
Topic 3	1.20 (1.23)	2.30 (0.95)	1.10 (1.45)
Topic 4	0.90 (1.13)	3.00 (0.58)	2.10 (0.94)
Topic 5	0.55 (0.50)	2.30 (1.01)	1.75 (1.03)

TABLE 6 Pre- posttest performance none users.

None users <i>M</i> ( <i>SD</i> )	Pre ( <i>n</i> = 4)	Post ( <i>n</i> = 4)	Diff ( <i>n</i> = 4)
Overall	4.00 (2.00)	9.75 (4.13)	5.75 (2.90)
Topic 1	1.13 (0.25)	1.88 (0.63)	0.75 (0.65)
Topic 2	0.75 (0.96)	1.63 (1.25)	0.88 (1.03)
Topic 3	1.50 (0.58)	1.88 (0.85)	0.38 (0.75)
Topic 4	0.13 (0.25)	2.00 (1.58)	1.88 (1.38)
Topic 5	0.50 (0.58)	2.38 (1.11)	1.88 (1.32)

## 3.2 Hypothesis testing

In this section, we present the results of our hypothesis testing. We examined four hypotheses related to the relationships between Perceived Ease of Use, Perceived Usefulness, Intention to Use, Actual System Use, and learning gains. The findings for each hypothesis are detailed below.

### 3.2.1 H1 and H2

H1 stated that Perceived Ease of Use is positively related to the Intention to Use. Perceived Ease of Use and Intention to Use were measured at three time points, and the correlations between them were analyzed. At T2, there was a significant positive correlation between Perceived Ease of Use and Intention to Use [ $r(15)=0.66$ ,  $p=0.007$ ], indicating that higher Perceived Ease of Use was associated with a higher Intention to Use. This correlation suggests a strong, positive relationship between the two variables at this time point. However, at T3, no significant correlation was found [ $r(13)=0.01$ ,  $p=0.984$ ], indicating no relationship between Perceived Ease of Use and Intention to Use at this time point. Similarly, at T5, there was no significant correlation [ $r(10)=0.22$ ,  $p=0.533$ ], suggesting that Perceived Ease of Use did not significantly relate to Intention to Use at this later time point. Based on these results, the hypothesis is partly confirmed as a significant positive association between Perceived Ease of Use and Intention to Use was only observed at T2.

H2 stated that Perceived Usefulness is positively related to Intention to Use. Significant positive correlations between Perceived Usefulness and Intention to Use were found for all three time points. At T2, there was a significant positive correlation between Perceived Usefulness and Intention to Use [ $r(15)=0.94$ ,  $p<0.001$ ], indicating that higher Perceived Usefulness was strongly associated with a higher Intention to Use. At T3, a significant positive correlation was found [ $r(13)=0.79$ ,  $p<0.001$ ], indicating a strong relationship between Perceived Usefulness and Intention to Use at this time point. Similarly, at T5, there was a significant positive correlation [ $r(10)=0.87$ ,  $p<0.001$ ], suggesting that higher Perceived Usefulness continued to be strongly associated with higher Intention to Use. Based on these results, the hypothesis is confirmed.

A multiple regression analysis was conducted to examine whether Perceived Usefulness and Perceived Ease of Use significantly predict the Intention to Use (Intention to Use), controlling for time (Time). The overall model was significant,  $F(3, 34)=37.541$ ,  $p<0.001$ , and accounted for approximately 76.8% of the variance in Intention to Use, with  $R^2=0.768$ . The regression coefficients indicated that Perceived Usefulness was a significant predictor of Intention to Use ( $\beta=0.950$ ,  $t=9.336$ ,  $p<0.001$ ), suggesting that higher Perceived Usefulness is associated with higher Intention to Use. However, Perceived Ease of Use ( $\beta=-0.014$ ,  $t=-0.133$ ,  $p=0.895$ ) and Time ( $\beta=0.090$ ,  $t=0.493$ ,  $p=0.625$ ) were not significant predictors of Intention to Use.

### 3.2.2 H3

H3 stated that higher Intention to Use leads to higher Actual Usage. A series of linear regression analyses were conducted to examine the predictive relationship between Intention to Use and subsequent Actual System Use periods. For the first regression, Intention to Use at T2 did not significantly predict Actual System Use during TP2,  $F(1, 13)=0.487$ ,  $p=0.497$ . The model explained only 3.6% of the variance ( $R^2=0.036$ , adjusted  $R^2=-0.038$ ). The second

regression analysis indicated that Intention to Use at T2 did not significantly predict Actual System Use during TP3,  $F(1, 13)=1.477$ ,  $p=0.246$ . This model explained 10.2% of the variance ( $R^2=0.102$ , adjusted  $R^2=0.033$ ). In the third regression, Intention to Use at T2 was not a significant predictor of Actual System Use during TP4,  $F(1, 13)=1.136$ ,  $p=0.306$ . The model explained 8% of the variance ( $R^2=0.080$ , adjusted  $R^2=0.010$ ). However, the fourth regression analysis revealed that Intention to Use at T3 significantly predicted Actual System Use during TP3,  $F(1, 13)=10.730$ ,  $p=0.007$ . This model accounted for 49.4% of the variance ( $R^2=0.494$ , adjusted  $R^2=0.448$ ). Based on these results, Hypothesis 3 is partially confirmed. While Intention to Use at T3 significantly predicted Actual System Use during TP3, Intention to Use at T2 did not significantly predict Actual System Use at TP2, TP3, or TP4.

### 3.2.3 H4

H4 stated that higher usage of the OwlMentor leads to higher learning gains. The difference between pre-and post-test scores was calculated as the learning gain, and the Actual System Use for the complete period of the course was analyzed. A simple linear regression analysis was conducted to examine whether the Actual System Use significantly predicts the learning gain for the users of the application. The results of the regression indicated that Actual System Use was not a significant predictor of learning gain,  $F(1, 8)=0.330$ ,  $p=0.581$ , and explained only 4.0% of the variance ( $R^2=0.040$ , adjusted  $R^2=-0.080$ ). Based on these results, the hypothesis is not confirmed.

## 3.3 Dialog functionality analysis

106 user requests were categorized into seven main categories, along with an “Other” category for requests that did not fit into the predefined categories. The frequencies of these categories are as shown in Table 7. Furthermore, an expert evaluation was conducted to assess the quality of OwlMentor’s responses. The mean rating given to 106 OwlMentor responses was 1.59 with a standard deviation of 0.94, suggesting that the overall quality of the responses was rated between “very good” and “good.”

## 3.4 Explorative analysis

Due to the overall low usage of our application, the negative correlation between Perceived Usefulness and Intention to Use and

TABLE 7 Distribution of user requests.

Category	Number of requests
Concept explanations and definitions	40
Summarization requests	23
In-depth explanations	14
Relationship and connection	8
Importance and rationale	7
Targeted information requests	6
Practical examples and applications	6
Other	2

the low Perceived Usefulness values combined with stable and high Self-Efficacy values over time, an exploratory analysis was conducted. The explorative assumption was that Self-Efficacy has a significant influence on how useful the application is perceived, as well as on its intended and Actual System Use. In this exploratory analysis, we investigated the influence of Self-Efficacy on Perceived Usefulness, Perceived Ease of Use, Intention to Use, and Actual System Use.

### 3.4.1 Self-efficacy and perceived usefulness

At T2, the correlation between Self-Efficacy and Perceived Usefulness was positive but not significant [ $r(15) = 0.33$ ,  $p = 0.235$ ], indicating that higher Self-Efficacy was weakly positive associated with Perceived Usefulness at this time point. At T3, there was a negative correlation between Self-Efficacy and Perceived Usefulness [ $r(13) = -0.53$ ,  $p = 0.061$ ], suggesting a moderate association of higher Self-Efficacy and lower Perceived Usefulness, although this relationship was not statistically significant. By T5, the negative correlation between Self-Efficacy and Perceived Usefulness was significant [ $r(10) = -0.85$ ,  $p = 0.002$ ], indicating that higher Self-Efficacy was strongly associated with lower Perceived Usefulness at this later time point.

### 3.4.2 Self-efficacy and perceived ease of use

At T2, there was a significant positive correlation between Self-Efficacy and Perceived Ease of Use [ $r(15) = 0.61$ ,  $p = 0.016$ ], indicating that higher Self-Efficacy was associated with higher Perceived Ease of Use at this time point. At T3, the correlation between Self-Efficacy and Perceived Ease of Use became negative but was not significant [ $r(13) = -0.17$ ,  $p = 0.587$ ], suggesting that higher Self-Efficacy was not significantly associated with lower Perceived Ease of Use at this time point. At T5, the negative correlation remained but was not significant [ $r(10) = -0.21$ ,  $p = 0.566$ ], indicating that higher Self-Efficacy continued to show a non-significant tendency towards lower Perceived Ease of Use.

### 3.4.3 Self-efficacy and intention to use

At T2, the correlation between Self-Efficacy and Intention to Use was positive but not significant [ $r(15) = 0.23$ ,  $p = 0.414$ ], indicating that higher Self-Efficacy was not significantly associated with higher Intention to Use at this time point. At T3, a significant negative correlation was observed between Self-Efficacy and Intention to Use [ $r(13) = -0.59$ ,  $p = 0.035$ ], suggesting that higher Self-Efficacy was associated with lower Intention to Use at this time point. By T5, this negative correlation became even more pronounced and significant [ $r(10) = -0.81$ ,  $p = 0.005$ ], indicating that higher Self-Efficacy was strongly associated with lower Intention to Use at this later time point.

### 3.4.4 Self-efficacy and actual system use

There was a negative correlation between Self-Efficacy at T2 and Actual System Use during both subsequent time periods (TP4, TP2). The correlation between Self-Efficacy at T2 and Actual System Use during TP2 was significant [ $r(15) = -0.52$ ,  $p = 0.047$ ], suggesting that higher Self-Efficacy was significantly associated with lower actual usage at TP2. However, the correlation between Self-Efficacy at T2 and Actual System Use for TP4 was negative but not significant [ $r(15) = -0.48$ ,  $p = 0.071$ ], indicating that higher Self-Efficacy was weakly associated with lower Actual System Use at TP4. There was a significant negative correlation between Self-Efficacy at T3 and Actual

System Use during TP3 [ $r(13) = -0.65$ ,  $p = 0.016$ ], indicating that higher Self-Efficacy was significantly associated with lower actual usage during TP3.

Additionally, there was a negative correlation between Self-Efficacy at all three time points (T2, T3, T5) and the Actual System Use for the overall TP. Specifically, there was a significant negative correlation between Self-Efficacy at T2 and Actual System Use for the overall TP [ $r(15) = -0.54$ ,  $p = 0.037$ ], between Self-Efficacy at T3 and Actual System Use for the overall TP [ $r(13) = -0.67$ ,  $p = 0.012$ ], and between Self-Efficacy at T5 and Actual System Use for the overall TP [ $r(10) = -0.76$ ,  $p = 0.010$ ], indicating that higher Self-Efficacy at each of these time points was significantly associated with lower overall Actual System Use.

## 4 Discussion

In the following section, we provide a detailed discussion of the key findings from this study, examine its limitations, offer an outlook on potential directions for future research and development, and present our conclusions.

### 4.1 Discussion of key findings

In this discussion, we analyze the results of OwlMentor's use and impact in a university course, where it was utilized to help students understand scientific texts. We examine the relationship between the individual variables of the TAM model across three measurement points, showing that the assumptions of the TAM model do not always hold true in every case and suggesting the need for a more flexible approach in the future. Specifically, the relationships between Perceived Ease of Use, Intention to Use, and Actual System Use are more complex as the TAM model suggests. Additionally, we consider the role of general Self-Efficacy when analyzing the TAM model. We also clarify the extent to which the use of OwlMentor is associated with learning gains. Finally, we consider user interactions with OwlMentor and highlight the quality of AI responses, underscoring the effectiveness of the RAG approach.

Based on the TAM (Davis et al., 1989; Venkatesh and Davis, 1996), our first hypothesis (H1) posited that Perceived Ease of Use would positively correlate with the Intention to Use OwlMentor. Our findings partially confirmed this: at the beginning of the course, there was a significant positive correlation, indicating that students who found the platform easy to use were more likely to intend to use it. However, this correlation was not significant later. Descriptive statistics showed that while Perceived Ease of Use initially increased, it slightly decreased over time, and Intention to Use remained stable initially but declined later. This aligns with Davis et al. (1989), who noted that the effect of Perceived Ease of Use on behavioral intention subsided over time. Researchers like Adams et al. (1992), Chau (1996), Gefen and Straub (2000), and Igarria et al. (1996) suggest that the influence of Perceived Ease of Use on Intention to Use is stronger in the early stages but diminishes over time. Subramanian (1994) also found that Perceived Ease of Use has less impact on usage over time if the technology is inherently easy to use. Our results show that while Perceived Ease of Use initially influenced Intention to Use, this relationship diminished, suggesting ease of use alone is insufficient for sustained engagement.

Factors like Perceived Usefulness and Self-Efficacy, related to students' growing domain knowledge, seem to become more influential over time. Initially, students may have found OwlMentor easy to use but later realized it did not offer as much benefit as expected, or they could meet course demands without the tool. The strong, stable correlation between Perceived Usefulness and Intention to Use supports that perceived added value is crucial for sustained use. Additionally, the shift from a positive to a negative correlation between Self-Efficacy and both Perceived Usefulness and Intention to Use suggests that students with higher Self-Efficacy felt less need for the tool as the course progressed. This aligns with Davis (1985), who noted that Perceived Usefulness has more influence than Perceived Ease of Use on system acceptance. Chuttur (2009) also pointed out that external factors like system experience, level of education, and age may influence system usage, while Yousafzai et al. (2007a) proposed that moderators affecting Perceived Usefulness and Perceived Ease of Use include Self-Efficacy, experience, educational level, skills, and knowledge.

The second hypothesis (H2) posited that Perceived Usefulness would positively correlate with the Intention to Use. This hypothesis was fully confirmed, with significant positive correlations at all three time points, consistent with the TAM (Davis, 1989; Chuttur, 2009). This finding aligns with previous studies and meta-analyses showing that Perceived Usefulness is a stronger predictor of technology adoption than Perceived Ease of Use (Amoako-Gyampah and Salam, 2004; Sharp, 2006; Yousafzai et al., 2007b; Yu et al., 2024). Our results confirmed this strong correlation, indicating that students are more likely to use OwlMentor if they find it useful, underscoring its importance for technology adoption (Opoku and Enu-Kwesi, 2019; Zou and Huang, 2023). Descriptive statistics show a decline in average Perceived Usefulness over the study, suggesting students became less positive about the platform's usefulness, contributing to the decline in Intention to Use, even though Perceived Ease of Use initially increased. This could be addressed in future by integrating regular knowledge assessments to adapt the support provided by the tool according to the students' evolving needs. The relationship between Self-Efficacy and Perceived Usefulness further supports this interpretation. Initially, there was no significant correlation between Self-Efficacy and Perceived Usefulness, but over time, a significant negative correlation emerged. This delayed emergence could be attributed to the time students needed to familiarize themselves with the course content, the difficulty of the texts, and how well they could cope with these demands. Students could only validly assess the platform's usefulness after thoroughly testing its capabilities and understanding the challenges posed by the learning tasks. Additionally, as students became more familiar with the course material and better understood the content over time, they may have found OwlMentor less necessary. This suggests that students with high confidence in their abilities found the course requirements manageable without OwlMentor. Consequently, they perceived the platform as less useful. These confident students likely believed they could succeed in the course without additional help, leading to a lower perceived necessity for the platform. Additionally, using OwlMentor required extra effort, and given their high Self-Efficacy, students might have decided that the time and effort needed to use the tool were not justified by its perceived benefits.

The third hypothesis (H3) proposed that Intention to Use would lead to Actual System Use. This hypothesis was only partially

confirmed. During the initial phase of the study, the analysis showed that students' Intention to Use OwlMentor did not significantly predict their Actual System Use. However, in the subsequent phase, as students became more familiar with OwlMentor, a significant positive relationship emerged. This indicates that the expected relationship between Intention to Use and Actual System Use became more evident over time, as students gained more experience with the platform. According to TAM, Perceived Ease of Use and Perceived Usefulness are important determinants of Intention to Use and Actual System Use of technologies (Davis, 1989; Venkatesh and Davis, 2000). However, our results suggest that the initial interest triggered by Perceived Usefulness did not immediately translate into Actual System Use, possibly due to students' exploratory approach or their existing confidence in mastering the course requirements without additional tools. The negative relation of Self-Efficacy on Actual System Use also provides important context. A negative correlation between Self-Efficacy and Actual System Use was observed from the early stages, which strengthened over time. This reflects the relation of Self-Efficacy on Intention to Use, where higher Self-Efficacy was associated with lower Intention to Use and subsequently lower Actual System Use. Students with high Self-Efficacy, who were confident that they could manage the demands of the course independently, saw less need for the OwlMentor and therefore used it less frequently. To encourage use by students with high Self-Efficacy, the tool could offer low-effort, high-benefit features such as automatic checking of progress in understanding science texts. Overall, it can be said that Intention to Use did not have a strong impact on Actual System Use initially, but its influence increased as students recognized the relevance of the platform. However, students with high Self-Efficacy consistently used the platform less, probably because they felt able to succeed without its help.

The fourth hypothesis (H4) stated that higher use of the OwlMentor would lead to higher learning gains as this platform provides different functions designed to support scientific text comprehension. However, our analysis revealed no significant correlation between Actual System Use and learning gains, suggesting that the frequency of platform use alone is not sufficient to achieve higher learning gains. Another possible explanation is that students with high Self-Efficacy performed better in the post-test, which might have obscured the correlation. Future research should examine the relationship between Self-Efficacy and post-test performance more closely. This could provide further insights into how Self-Efficacy influences learning outcomes. The descriptive statistics show that both users and non-users of the OwlMentor platform experienced significant learning gains between pre- and posttest, but these gains were not directly related to the extent of platform use. This indicates that the seminar's quality of instruction and peer presentations likely contributed primarily to the students' learning gains. The decreasing use of OwlMentor over the course suggests that students may have relied more on other learning strategies and resources to prepare for the exam. This aligns with previous findings that high Self-Efficacy led to lower Perceived Usefulness and Intention to Use the platform. Students with high confidence in their abilities may have felt able to succeed without additional support from the AI tool, perceiving it as offering no added value, which further discouraged its use. In summary, while the course facilitated learning, OwlMentor showed promise as users demonstrated higher learning gains than non-users. Although no direct correlation between Actual System Use and



learning gains was found, the descriptive results suggest the need for further experimental studies to confirm the effectiveness of AI-based tools in education. This underscores the importance of integrating AI tools into existing learning strategies and ensuring they offer clear, tangible benefits to students. The results emphasize that matching technological tools to students' needs is crucial for their usefulness.

The analysis of OwlMentor's dialog features offers insights into student interactions and response quality. Categorizing user requests showed a need for explanations, definitions, summaries, detailed explanations, and understanding relationships between concepts. These categories are consistent with strategies known to improve scientific text comprehension, such as self-questioning, summarizing key passages, and linking new information to prior knowledge (Gunn, 2008; Joseph and Ross, 2018; Kendeou and Van Den Broek, 2005, 2007; Sason et al., 2020). For instance, requests for concept explanations and definitions correspond to the need for understanding basic facts, concepts, and processes, which are essential for scientific literacy (Organisation for Economic Co-operation and Development (OECD), 2003; Goldman and Bisanz, 2002). Summarization requests reflect the strategy of consolidating understanding by condensing key information, aiding in the retention and comprehension of complex texts (Cromley, 2010). Requests for on-depth explanations and understanding relationships between concepts highlight students need to engage more deeply with content, similar to self-explanation methods that promote deeper learning and integration of text material (Chi et al., 1989; Chi et al., 1994; King, 1994). By supporting these strategies through its dialog function, OwlMentor helps students handle the challenges of understanding academic texts. This fits with the idea that generative AI, like large LLMs, can offer interactive and adaptable help (Gimpel et al., 2023; Kasneci et al., 2023). Such AI can make it easier for students to use complex learning strategies and improve their engagement and understanding of academic material. However, OwlMentor takes over these tasks completely, meaning that users only need to be generative in their prompting. Whether this approach is as beneficial as performing these tasks entirely by themselves is not proven and requires further experimental studies.

The expert evaluation of OwlMentor's responses, which received positive ratings, shows that the RAG approach can effectively provide accurate and relevant responses generated by the AI. By providing specific contextual information from the course literature, the AI was able to generate accurate responses to student queries, supporting the idea that RAG systems can improve LLM performance in educational environments (Shen et al., 2023; Barnett et al., 2024). This suggests that RAG systems can support students in working with scientific texts by providing contextually informed answers, thereby addressing some of the challenges associated with understanding complex academic material. Our findings align with the theoretical promise of generative AI in education, highlighting the potential of RAG-enhanced LLMs to support students' learning processes, and facilitate deeper engagement with course materials (Gimpel et al., 2023; Kasneci et al., 2023). However, despite the positive expert evaluations, many students did not find the system helpful. This suggests that providing accurate and relevant answers, similar to ensuring ease of use, is likely a necessary but not sufficient condition for the system to be perceived as helpful and actively used. Person variables, such as students' expertise and Self-Efficacy, seem to play a significant role in this perception. High Self-Efficacy and advanced knowledge might reduce the perceived necessity for additional support from the AI tool,

influencing its overall acceptance and usage. Therefore, while the technical performance of OwlMentor is crucial, its integration and Perceived Usefulness must also consider individual differences among students to enhance its effectiveness. Our findings illustrate how generative AI can be integrated into university courses and highlight the potential for improving the performance of such systems by providing contextual information. These insights emphasize the need for ongoing development and refinement of AI technologies like OwlMentor in educational settings to enhance their effectiveness and address their technical and pedagogical limitations (Alkaissi and McFarlane, 2023; Feldman et al., 2024).

## 4.2 Limitations

Our research question of how an AI-based learning platform to support scientific text comprehension, such as OwlMentor, can be integrated into university teaching and whether RAG systems are suitable for supporting students in their work with academic texts was partially answered. In principle, it should be noted that the results found here can provide initial indications for the integration of AI in university teaching but are severely limited in their generalizability due to methodological weaknesses. Although this study provides valuable insights into the usage behavior, perceptions, and learning outcomes associated with the AI-based learning platform OwlMentor, the following limitations must be considered when interpreting the results. The sample size of 16 participants is relatively small and limits the generalizability of the results. Furthermore, the lack of a control group makes it difficult to attribute the observed learning gains solely to the use of the AI-based learning platform. Moreover, the different features of OwlMentor cannot be considered in isolation from each other or from the overall course activities. Our pre-post intervention analysis shows the progress over time for the entire course, but it does not distinguish the impact of individual features of the platform. The study relied primarily on correlation analyses to examine relationships between variables that do not establish causal relationships. Moreover, data collected through self-report may contain response bias, as participants may give socially desirable answers or inaccurately recall their perceptions. In addition, OwlMentor usage fluctuated significantly throughout the study period, particularly at exam times, complicating the interpretation of results and suggesting that platform usage may be heavily influenced by immediate academic demands and schedules. OwlMentor was designed to enhance scientific text comprehension. If this goal was achieved is not clear, as there were no specific measures of scientific text comprehension in this study. Future research should include specific assessments of scientific text comprehension to better understand the platform's impact on this crucial aspect of academic learning.

## 4.3 Future research and development

Future research should focus on controlled studies isolating the effects of OwlMentor's individual features, like the dialog function and question generation, on student learning and engagement. Enhancing LLM response quality using the RAG approach and

assessing the accuracy and pedagogical value of generated questions are critical areas. Investigating the role of Self-Efficacy in the use and perception of AI-powered platforms will shed light on how these tools can be tailored to students' varying levels of confidence. It's important to address the specific needs of students at various educational levels; undergraduates might need more support than graduates. Our study suggests that students with higher Self-Efficacy and knowledge may need different support, such as adaptive prompts or progress checks in text comprehension. Incorporating ongoing student feedback into the development process will keep the platform relevant and effective. Future development efforts for OwlMentor will focus on integrating advanced language models, such as GPT-4 or GPT-4o, to enhance the retrieval process. These models, with their larger context windows, will allow the system to provide more detailed contextual information, further improving response accuracy. Additionally, function calling will be leveraged to enable the LLM to interact more effectively with external tools. To support learning success, OwlMentor will be extended to offer targeted strategies such as summarizing sections, highlighting key points, and guided navigation through the text, offering students specific strategies for text comprehension. These features will allow students to quickly access the information they need without extensive interaction. Another improvement will simplify question generation, enabling students to create and refine questions directly within the chat interface and save them seamlessly. New question formats, including true/false and open-ended questions, will also be introduced. On the instructor side, future iterations will allow educators to integrate their own courses and materials into OwlMentor, further personalizing the platform to specific educational needs. Additionally, research should explore the impact of AI-based platforms on different student groups to ensure equitable benefits and identify necessary changes. Integrating various learning forms, such as peer or instructor collaboration, could enhance learning experiences by combining AI tools with traditional methods. To improve the TAM, future adaptations should consider the dynamic nature of user experience and expertise development. Incorporating regular assessments of domain knowledge and Self-Efficacy could enhance the Perceived Usefulness and sustained engagement with AI-based educational tools like OwlMentor. Understanding how these factors evolve over time can help in designing more adaptive and supportive learning environments. Finally, evaluating AI tools across academic disciplines will identify beneficial features for each area, informing targeted enhancements to address the unique challenges of different fields of study.

## 5 Conclusion

This study explored the integration and impact of OwlMentor, an AI-based learning platform, within a university course. Our findings indicate that the static nature of the TAM does not adequately account for the evolving influences of Perceived Ease of Use, Perceived Usefulness, and Self-Efficacy over time. While TAM partially explained initial adoption behaviors, the dynamic changes in these factors suggest the need for a more flexible approach that considers temporal shifts and the development of domain knowledge. The

effectiveness of the RAG approach was demonstrated through positive expert evaluations of OwlMentor's responses, which were accurate and relevant. However, student perceptions did not always align with these evaluations, indicating that technical performance alone is insufficient for sustained engagement. The integration of person-specific factors and adaptive functionalities is crucial for enhancing Perceived Usefulness and continued use. In conclusion, OwlMentor shows potential in supporting scientific text comprehension, but its success hinges on dynamic adaptation to user needs, continuous value addition, and integration with existing learning strategies. Future AI-enhanced educational tools must adopt a flexible, student-centered approach, emphasizing the importance of regular assessments and iterative development to remain relevant and effective in higher education.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The data was collected anonymously, all participants were of legal age, fully informed and had given their consent. The use of the application was part of a university course with no additional risks for the participants. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## Author contributions

DT: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. SM: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing – review & editing. RB: Conceptualization, Funding acquisition, Methodology, Resources, Supervision, Validation, Writing – review & editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The "Innovation Project OwlMentor" was part of the "Digital Teaching Plug-in" (DaTa-Pin) project funded by the Foundation for Innovation in Higher Education Teaching (Stiftung für Innovation in der Hochschullehre) granted to Saarland University (grant number: FBM2020-EA-2350-01363). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1474892/full#supplementary-material>

## References

- Abdelghani, R., Wang, Y. H., Yuan, X., Wang, T., Sauzéon, H., and Oudeyer, P. Y. (2022). GPT-3-driven pedagogical agents for training children's curious question-asking skills. *arXiv preprint*, arXiv:2211-14228v6. doi: 10.48550/arXiv.2211.14228
- Adams, D. A., Nelson, R. R., and Todd, P. A. (1992). Perceived usefulness, ease of use, and usage of information technology: a replication. *MIS Q.* 16, 227–247. doi: 10.2307/249577
- Alkaissi, H., and McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 15:e35383. doi: 10.7759/cureus.35383
- Alshater, M. M. (2022). Exploring the role of artificial intelligence in enhancing academic performance: a case study of ChatGPT. *SSRN Electron. J.* doi: 10.2139/ssrn.4312358
- Amoako-Gyampah, K., and Salam, A. F. (2004). An extension of the technology acceptance model in an ERP implementation environment. *Inf. Manag.* 41, 731–745. doi: 10.1016/j.im.2003.08.010
- Bandura, A. (1982). Self-efficacy mechanism in human agency. *Am. Psychol.* 37, 122–147. doi: 10.1037/0003-066X.37.2.122
- Bangor, A., Kortum, P., and Miller, J. T. (2009). Determining what individual SUS scores mean: adding an adjective rating scale. *J. Usability Stud.* 4, 114–123.
- Bao, M. (2019). Can home use of speech-enabled artificial intelligence mitigate foreign language anxiety—investigation of a concept. *Arab World English J.* 5, 41–50. doi: 10.24093/awej/call5.4
- Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., and Abdelrazek, M. (2024). Seven failure points when engineering a retrieval augmented generation system. *arXiv preprint*. doi: 10.48550/arXiv.2401.05856
- Bhat, S., Nguyen, H. A., Moore, S., Stamper, J., Sakr, M., and Nyberg, E. (2022). Towards automated generation and evaluation of questions in educational domains. In *Proceedings of the 15th international conference on educational data mining* 701.
- Bloom, B. S. (Ed.) (1956). *Taxonomy of education objectives book 1-cognitive domain*. David McKay Company.
- Brooke, J. (1996). SUS: A “quick and dirty” usability scale. In *Usability evaluation in industry*. (Eds.) P. W. Jordan, B. Thomas, B. A. Weerdmeester and I. L. McClelland (London: Taylor & Francis), 189–194.
- Chau, P. Y. (1996). An empirical assessment of a modified technology acceptance model. *J. Manag. Inf. Syst.* 13, 185–204. doi: 10.1080/07421222.1996.11518128
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., and Glaser, R. (1989). Self-explanations: how students study and use examples in learning to solve problems. *Cogn. Sci.* 13, 145–182.
- Chi, M. T. H., De Leeuw, N., Chiu, M. H., and LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cogn. Sci.* 18, 439–477.
- Chinn, C. A., and Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *J. Res. Sci. Teach.* 35, 623–654. doi: 10.1002/(SICI)1098-2736(199808)35
- Chuttur, M. (2009). Overview of the Technology Acceptance Model: Origins, developments, and future directions. *Sprouts: Working Papers on Information Systems*, 9, 290. Available at: [https://aisel.aisnet.org/sprouts\\_all/290](https://aisel.aisnet.org/sprouts_all/290)
- Cooper, G. (2023). Examining science education in ChatGPT: an exploratory study of generative artificial intelligence. *J. Sci. Educ. Technol.* 32, 444–452. doi: 10.1007/s10956-023-10039-y
- Cromley, J. G., Snyder-Hogan, L. E., and Luciw-Dubas, U. A. (2010). Reading comprehension of scientific text: a domain-specific test of the direct and inferential mediation model of reading comprehension. *J. Educ. Psychol.* 102, 687–700. doi: 10.1037/a0019452
- Davis, F. D. (1985). A technology acceptance model for empirically testing new end-user information systems: theory and results. Doctoral dissertation, Massachusetts Institute of Technology.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 13, 319–340. doi: 10.2307/249008
- Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Manag. Sci.* 35, 982–1003. doi: 10.1287/mnsc.35.8.982
- Davis, F. D., and Granić, A. (2024). *The technology acceptance model: 30 years of TAM* (1st ed.). Springer Cham, Switzerland. doi: 10.1007/978-3-030-45274-2
- Dijkstra, R., Genç, Z., Kayal, S., and Kamps, J. (2022). Reading comprehension quiz generation using generative pre-trained transformers. In *iTextbooks@AIED pp.* 4–17.
- Duncan, R. G., Chinn, C. A., and Barzilai, S. (2018). Grasp of evidence: problematizing and expanding the next generation science standards' conceptualization of evidence. *J. Res. Sci. Teach.* 55, 907–937. doi: 10.1002/tea.21468
- El Shazly, R. (2021). Effects of artificial intelligence on English speaking anxiety and speaking performance: a case study. *Expert. Syst.* 38:e12667. doi: 10.1111/exsy.12667
- Feldman, P., Foulds, J. R., and Pan, S. (2023). Trapping LLM hallucinations using tagged context prompts. *arXiv preprint*. doi: 10.48550/arXiv.2306.06085
- Feldman, P., Foulds, J. R., and Pan, S. (2024). RAGged edges: the double-edged sword of retrieval-augmented chatbots. *arXiv preprint*. doi: 10.48550/arXiv.2403.01193
- Gabajiwal, E., Mehta, P., Singh, R., and Koshy, R. (2022). Quiz maker: automatic quiz generation from text using NLP. In *Futuristic trends in networks and computing technologies: Select proceedings of fourth international conference on FTNCT 2021* (523–533)
- Gefen, D., and Straub, D. (2000). The relative importance of perceived ease-of-use in IS adoption: a study of e-commerce adoption. *J. AIS* 1, 1–30. doi: 10.17705/1jais.00008
- Gernsbacher, M. A., and Kaschak, M. P. (2013). “Text comprehension” in *The Oxford handbook of cognitive psychology*. ed. D. Reisberg (London: Oxford University Press), 462–474.
- Gimpel, H., Hall, K., Decker, S., Eymann, T., Lämmermann, L., Mädche, A., et al. (2023). *Unlocking the power of generative AI models and systems such as GPT-4 and ChatGPT for higher education: A guide for students and lecturers*. University of Hohenheim, Stuttgart.
- Goldman, S. R., and Bisanz, G. L. (2002). Toward a functional analysis of scientific genres: Implications for understanding and learning processes. in *The psychology of science text comprehension*. (Eds.) J. Otero, J. A. León and A. C. Graesser (Erlbaum, Mahwah, NJ), 19–50.
- Gunn, T. M. (2008). The effects of questioning on text processing. *Read. Psychol.* 29, 405–442. doi: 10.1080/02702710802165374
- Herft, A. (2023). A teacher's prompt guide to ChatGPT: aligned with “what works best”. CESE NSW “what works best in practice. Available at: <https://drive.google.com/file/d/15qAxnUzOwAPwH> (Accessed February, 2024).
- Huang, W., Hew, K. F., and Fryer, L. K. (2022). Chatbots for language learning—are they really useful? A systematic review of chatbot-supported language learning. *J. Comput. Assist. Learn.* 38, 237–257. doi: 10.1111/jcal.12610
- Igbaria, M., Parasuraman, S., and Baroudi, J. (1996). A motivational model of microcomputer usage. *Journal of MIS* 13, 127–143. doi: 10.1080/07421222.1996.11518115
- Jeon, J. A. (2021). Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis. *Computer Assisted Language Learning*, 36, 1338–1364. doi: 10.1080/09588221.2021.1987272
- Ji, H., Han, I., and Ko, Y. (2023). A systematic review of conversational AI in language education: focusing on the collaboration with human teachers. *J. Res. Technol. Educ.* 55, 48–63. doi: 10.1080/15391523.2022.2142873
- Joseph, L. M., and Ross, K. M. (2018). Teaching middle school students with learning disabilities to comprehend text using self-questioning. *Interv. Sch. Clin.* 53, 276–282. doi: 10.1177/1053451217736866



- Kabir, S., Udo-Imeh, D. N., Kou, B., and Zhang, T. (2023). Who answers it better? An in-depth analysis of ChatGPT and stack overflow answers to software engineering questions. *arXiv preprint*. doi: 10.48550/arXiv.2308.02312
- Kasneci, E., Seifler, K., Küchemann, S., Bannert, M., Dementieva, D., and Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.*, 103:102274. doi: 10.1016/j.lindif.2023.102274
- Kendeou, P., and Van Den Broek, P. (2005). The effects of readers' misconceptions on comprehension of scientific text. *J. Educ. Psychol.* 97, 235–245. doi: 10.1037/0022-0663.97.2.235
- Kendeou, P., and Van Den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Mem. Cogn.* 35, 1567–1577. doi: 10.3758/BF03193491
- King, A. (1994). Guiding knowledge construction in the classroom: effects of teaching children how to question and how to explain. *Am. Educ. Res. J.* 31, 338–368. doi: 10.3102/00028312031002338
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press, Cambridge.
- Kuhail, M. A., Alturki, N., Alramlawi, S., and Alhejori, K. (2022). Interacting with educational chatbots: a systematic review. *Educ. Inf. Technol.* 28, 973–1018. doi: 10.1007/s10639-022-10747-1
- Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., et al. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digital Health* 2:e0000198. doi: 10.1371/journal.pdig.0000198
- Leslie, K. C., Low, R., Jin, P., and Sweller, J. (2012). Redundancy and expertise reversal effects when using educational technology to learn primary school science. *Educ. Technol. Res. Dev.* 60, 1–13. doi: 10.1007/s11423-011-9199-0
- Liang, J., Wang, L., Luo, J., Yan, Y., and Fan, C. (2023). The relationship between student interaction with generative artificial intelligence and learning achievement: serial mediating roles of self-efficacy and cognitive engagement. *Front. Psychol.* 14:1285392. doi: 10.3389/fpsyg.2023.1285392
- Liaw, S. S. (2008). Investigating students' perceived satisfaction, behavioral intention, and effectiveness of e-learning: a case study of the blackboard system. *Comput. Educ.* 51, 864–873. doi: 10.1016/j.compedu.2007.09.005
- Mayer, R. E. (2014). Cognitive theory of multimedia learning. In *The Cambridge handbook of multimedia learning*. (Ed.), R. E. Mayer. 2nd ed (Cambridge University Press, Cambridge), 43–71.
- Mayer, R. E., and Fiorella, L. (2021). *The Cambridge handbook of multimedia learning* (3rd ed.). Cambridge University Press, Cambridge.
- McNamara, D. S., and Kintsch, W. (1996). Learning from texts: effects of prior knowledge and text coherence. *Discourse Process.* 22, 247–288. doi: 10.1080/01638539609544975
- Mollick, E. R., and Mollick, L. (2022). New modes of learning enabled by AI chatbots: three methods and assignments. *SSRN Electron. J.* doi: 10.2139/ssrn.4300783
- Norman, D. A., and Draper, S. W. (1986). *User centered system design: New perspectives on human-computer interaction*. Lawrence Erlbaum Associates, Mahwah, NJ.
- Norris, S. P., and Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Sci. Educ.* 87, 224–240. doi: 10.1002/sce.10066
- Opoku, M. O., and Enu-Kwesi, F. (2019). Relevance of the technology acceptance model (TAM) in information management research: a review of selected empirical evidence. *Res. J. Bus. Manag.* 6, 55–62. doi: 10.17261/Pressademia.2019.1028
- O'Reilly, T., and McNamara, D. S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *Am. Educ. Res. J.* 44, 161–196. doi: 10.3102/0002831206298171
- Organisation for Economic Co-operation and Development (OECD) (2003). *PISA literacy skills for the world of tomorrow: Further results from PISA 2000*. OECD Publications, Paris, France.
- Pavlik, J. V. (2023). Collaborating with ChatGPT: considering the implications of generative artificial intelligence for journalism and media education. *J. Mass Commun. Educ.* 78, 84–93. doi: 10.1177/10776958221149577
- Pintrich, P. R., and De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *J. Educ. Psychol.* 82, 33–40. doi: 10.1037/0022-0663.82.1.33
- Qadir, J. (2023). Engineering education in the era of ChatGPT: promise and pitfalls of generative AI for education. In 2023 IEEE global engineering education conference (EDUCON). (1–9)
- Sallam, M. (2023). The utility of ChatGPT as an example of large language models in healthcare education, research and practice: systematic review on the future perspectives and potential limitations. *medRxiv*. doi: 10.1101/2023.02.19.23286155
- Sason, H., Michalsky, T., and Mevarech, Z. (2020). Promoting middle school students' science text comprehension via two self-generated "linking" questioning methods. *Front. Psychol.* 11:595745. doi: 10.3389/fpsyg.2020.595745
- Sharp, J. H. (2006). Development, extension, and application: a review of the technology acceptance model. *Inform. Syst. Educ. J.* 5, 1–11.
- Shen, T., Long, G., Geng, X., Tao, C., Zhou, T., and Jiang, D. (2023). Large language models are strong zero-shot retrievers. *arXiv preprint*. doi: 10.48550/arXiv.2304.14233
- Subramanian, G. H. (1994). A replication of perceived usefulness and perceived ease of use measurement. *Decis. Sci.* 25, 541–553. doi: 10.1111/j.1540-5915.1994.tb01859.x
- Sweller, J. (2011). "Cognitive load theory" in *The psychology of learning and motivation: cognition in education*. eds. J. P. Mestre and B. H. Ross (Elsevier Academic Press), 37–76.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems*, 30:5998–6008. doi: 10.48550/arXiv.1706.03762
- Venkatesh, V., and Davis, F. D. (1996). A model of the antecedents of perceived ease of use: development and test. *Decis. Sci.* 27, 451–481. doi: 10.1111/j.1540-5915.1996.tb01822.x
- Venkatesh, V., and Davis, F. D. (2000). A theoretical extension of the technology acceptance model: four longitudinal field studies. *Manag. Sci.* 46, 186–204. doi: 10.1287/mnsc.46.2.186.11926
- Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., and Drachler, H. (2021). Are we there yet? A systematic literature review on chatbots in education. *Frontiers. Artif. Intell.* 4:6454924. doi: 10.3389/frai.2021.654924
- Wu, R., and Yu, Z. (2023). Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. *Br. J. Educ. Technol.* 54, 667–685. doi: 10.1111/bjet.13278
- Yousafzai, S. Y., Foxall, G. R., and Pallister, J. G. (2007a). Technology acceptance: a meta-analysis of the TAM: part 1. *J. Model. Manag.* 2, 251–280. doi: 10.1108/17465660710834453
- Yousafzai, S. Y., Foxall, G. R., and Pallister, J. G. (2007b). Technology acceptance: a meta-analysis of the TAM: part 2. *J. Model. Manag.* 2, 281–304. doi: 10.1108/17465660710834462
- Yu, C., Yan, J., and Cai, N. (2024). ChatGPT in higher education: factors influencing ChatGPT user satisfaction and continued use intention. *Front. Educ.* 9:1354929. doi: 10.3389/feduc.2024.1354929
- Zhai, X. (2022). ChatGPT user experience: implications for education. *SSRN Electron. J.* doi: 10.2139/ssrn.4312418
- Zhao, L., Zhang, L., Wu, Z., Chen, Y., Dai, H., Yu, X., et al. (2023). When brain-inspired AI meets AGI. *Meta Radiol.* 1:100005. doi: 10.1016/j.metrad.2023.100005
- Zhuo, T. Y., Huang, Y., Chen, C., and Xing, Z. (2023). Red teaming ChatGPT via jailbreaking: Bias, robustness, reliability and toxicity. *arXiv preprint*. doi: 10.48550/arXiv.2301.12867
- Zou, M., and Huang, L. (2023). To use or not to use? Understanding doctoral students' acceptance of ChatGPT in writing through technology acceptance model. *Front. Psychol.* 14:1259531. doi: 10.3389/fpsyg.2023.1259531





## OPEN ACCESS

## EDITED BY

Knut Neumann,  
IPN–Leibniz Institute for Science and  
Mathematics Education, Germany

## REVIEWED BY

Hongzhi (Veronica) Yang,  
The University of Sydney, Australia  
Barry Lee Reynolds,  
University of Macau, China

## \*CORRESPONDENCE

Denis Federiakin  
✉ denis.federiakin@uni-mainz.de

RECEIVED 06 January 2024

ACCEPTED 08 October 2024

PUBLISHED 29 November 2024

## CITATION

Federiakin D, Molerov D,  
Zlatkin-Troitschanskaia O and Maur A (2024)  
Prompt engineering as a new 21st century  
skill.  
*Front. Educ.* 9:1366434.  
doi: 10.3389/feduc.2024.1366434

## COPYRIGHT

© 2024 Federiakin, Molerov,  
Zlatkin-Troitschanskaia and Maur. This is an  
open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](#). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic  
practice. No use, distribution or reproduction  
is permitted which does not comply with  
these terms.

# Prompt engineering as a new 21st century skill

Denis Federiakin\*, Dimitri Molerov, Olga Zlatkin-Troitschanskaia  
and Andreas Maur

Department of Business and Economics Education, Johannes Gutenberg University Mainz, Mainz,  
Germany

Artificial Intelligence (AI) promises to revolutionize nearly every aspect of human learning. However, users have observed that the efficacy of AI assistants hinges crucially on the quality of the prompts supplied to them. A slight alteration in wording can make the difference between an assistant misinterpreting an instruction and exceeding expectations. The skill of precisely communicating the essence of a problem to an AI assistant is as crucial as the assistant itself. This paper aims to introduce Prompt Engineering (PE) as an emerging skill essential for personal and professional learning and development in the 21st century. We define PE as the skill of articulating a problem, its context, and the constraints of the desired solution to an AI assistant, ensuring a swift and accurate response. We show that no existing related frameworks on 21st skills and others cover PE to the extent that allows for its valid assessment and targeted promotion in school and university education. Thus, we propose a conceptual framework for this skill set including (1) comprehension of the basic prompt structure, (2) prompt literacy, (3) the method of prompting, and (4) critical online reasoning. We also discuss the implications and challenges for the assessment framework of this skill set and highlight current PE-related recommendations for researchers and educators.

## KEYWORDS

prompt engineering, artificial intelligence, 21st century skills, ChatGPT, digital skills, critical online reasoning, LLM

## 1 Introduction

The development of assisting Artificial Intelligence (AI) tools promises to revolutionize almost all fields of human learning. The widespread adoption of emerging digital technologies has accelerated the development and the speed of information exchange. It has become obvious that learners require a specific competence to be able to process various forms of information to successfully undertake tasks in disciplinary and cross-disciplinary contexts. As part of this transformative trend, the cultivation of 21st century skills has been deemed essential to preparing a global workforce to succeed in an increasingly data-centric and information-driven society.

While a universal definition of 21st century skills is hardly possible due to numerous different frameworks, their common features can be determined. These skills are generic, not specifically tied to any particular professional domain, and essential for personal development in the ever-changing 21st century (Foster and Piacentini, 2023). These skills include online information problem-solving (Goldman and Brand-Gruwel, 2018) and other abilities required to evaluate and process new information and competently use it in various settings (Foster, 2023a; Pellegrino, 2023).

Not only has ChatGPT become a pervasive presence within the computer-reliant programming and technology sector (Chen et al., 2023a; Ridnik et al., 2024) and the research

community (Kasneci et al., 2023; Giray, 2023), it has also established itself in various service industries (Opara et al., 2023). Consequently, this new tool has infiltrated the learning and workflow of students, transcending the boundaries of technological focus. The impact of such AI tools on society has already been so immense that some researchers have claimed that some fields, such as education, are significantly *disrupted* by them (Cain, 2024).

Hosseini et al. (2023) surveyed students beginning university who reported using ChatGPT (very) often while learning at school and/or in professional training. Proficient utilization can surmount inhibition thresholds associated with familiarizing oneself with a particular topic, expedite information processing through summarization, visually and systematically process information, validate writing, and serve various other functions (Mohr et al., 2023). Some have claimed that AI tools like ChatGPT can promote “unlearning,” resulting in students acquiring less knowledge and underperforming due to less intensive cognitive learning processes (Abbas et al., 2024). Another aspect of this negative impact is the blind trust in ChatGPT’s responses, causing users to accept the outputs of Large Language Models (LLMs) without critical evaluation (Krupp et al., 2023).

The malleability and adaptability inherent in LLMs render tools like ChatGPT capable of fundamentally altering virtually all processes to which they are applied. Nonetheless, LLMs are not the only type of AI assistance on the agenda. Text-to-Image models, Speech-to-Text, polymodal AI tools, and other tools have already been in the practice for quite a while, sufficiently expanding the societal and learning impact of AI.

While LLMs have been in development for years, the release of ChatGPT to the general public by OpenAI in autumn 2022 has marked a shift in use affordances of digital and Internet-based tools even compared to the seemingly ubiquitous search engines. In contrast to such engines, LLMs provide full-text responses to longer inputs by users, are more friendly to further inquiry and chatbot communication, but typically include fewer references or direct hyperlinks that would guide users to leave their interface. Still, the usefulness of open-access ChatGPT for learning (not least in higher education) has been quickly noted.

Although ChatGPT erupted onto the scene very quickly, users have just as quickly noticed that the performance of many types of AI-assisting tools highly depends on the quality of prompts supplied to them (Ekin, 2023). Changing just a couple of words in the prompt can split the difference between the AI tool failing to understand the instruction and outperforming the request. From a technical standpoint, the importance of prompt accuracy is not particularly surprising, since LLMs (the engine of tools such as ChatGPT) are focused on predicting the next language token. Tokens are essentially building blocks of written language—punctuation, specific forms of words, word endings (such as -s or -ed), and so on. They combine in a sequence to produce the written text. Correspondingly, the fundamental task of such LLMs is just to use probability to predict the next token conditional on the previous tokens. Given this, it is expected that the model performance will depend on the quality of the prompt. Typically, the more detailed and explicit the prompt is, the more precise the model is in its response.

Moreover, users might experience difficulties evaluating the quality of LLM output. Recent research has already registered that LLMs (including ChatGPT) can hallucinate (Alkaissi and McFarlane, 2023). LLMs can invent facts and references that are non-existent or factually

incorrect. This degrades the quality of model output even to a degree of rendering it unusable. Users might overlook this, which poses an additional challenge in the use of LLMs. This challenge is compounded by the users’ concurrent adoption of conflicting roles, serving both as the processor and the supervisor of the task because ChatGPT does not indicate how certain it is about the given answer or whether the prompt needs to provide more information. The amalgamation of these dual responsibilities contributes to the heightened complexity and intricacy of the communication process within the context of utilizing LLMs, and ipso facto requires meta-awareness and ambiguity tolerance on the users’ part.

Additionally, some research has suggested that the correct prompting of an LLM can enhance its performance to the point that special fine-tuning of a foundational model (trained on a generic corpus of texts without any particular specialization) might be unnecessary. For example, Nori et al. (2023) and Maharajan et al. (2024) have shown that the correct prompting technique can improve LLM performance to the extent that that foundational models outperform specially fine-tuned LLMs in medical knowledge. This demonstrates that prompting is an immensely powerful phenomenon that holds a dramatic influence on LLM performance.

Recently, Microsoft has released BingGPT and Google introduced Gemini as the preliminary merged search engines with LLM capabilities. LLMs with increased capabilities have been continuously released over the past months. This wild universality of LLMs and their capacity to quickly work with unstructured information renders their application increasingly and continuously important and popular across many fields. Hence, the necessity of exact prompting skills may vary by application and are expected to change, however, the general insights on LLMs apply, as long as the types of interfaces, training, and output quality prevail.

While some AI tools themselves can help to reformulate and improve the prompts (Zhou et al., 2022) via dialoguing with a user, it takes time and still does not guarantee the desired result. Moreover, some tools (i.e., Text-to-Image models) might experience difficulties in improving the prompt in the situations where users apply too many constraints on the desired solution. These constraints might mislead the model, forcing it to focus on the insufficient details. In the end, the only way left to communicate with the tool is through trial-and-error until the user is satisfied with the solution. The iterative nature of trial-and-error can be time-consuming, inefficient, economically burdensome, fallible, and may introduce security risks into critical decision-making processes, potentially leaving errors unobserved and further impacting corporate success and the well-being of users.

Thus, being able to concisely communicate the nature of the problem to the AI tool is as valuable as the tool itself. Without this skill, users may fail to receive an acceptable and correct solution. In this respect, we disagree with those researchers in the data science community who argue that PE is merely a facet of general communication skills (Morton, 2024). Merely speaking a language does not assume good communication skills, and similarly, a good communicator may not inherently possess the skills necessary to effectively interact with AI. Therefore, we aim to define PE as a distinct skill which warrants investigation within the educational and psychological sciences.

## 2 Research objectives

Numerous higher-order (meta)cognitive skill concepts have been developed through research, which theoretically define the proficient

utilization of digital information, communication, and learning tools. These include skills related to Information and Communication Technology (ICT; Kaarakainen et al., 2018), digital skills as measured in the PISA assessment (OECD, 2023), and Critical Online Reasoning (COR) skills (Molero et al., 2020; Nagel et al., 2020, 2022; Schmidt et al., 2020). The previous concepts at least do not explicitly address and conceptualize the skills for the competent use of AI-supported tools, but only elaborate on the necessity of creativity and higher-order, metacognitive skills without a specific relation to AI (PISA 2025 framework; Hu et al., 2023). As transversal skills like information problem-solving gain prominence in education as essential 21st century skills (Foster, 2023a; Pellegrino, 2023), the cultivation of PE skills becomes particularly crucial. Competent ChatGPT use heavily relies on well-formulated and elaborated prompts. The proficiency in crafting prompts is essential to anticipate and minimize the risk of inaccurate answers, necessitating a thoughtful process of reflection and rehearsal.

This paper aims to address this desideratum and to conceptualize PE as a specific skill in the 21st century. We claim that defining it in a manner akin to generic competencies (which are universal across many professions; Shavelson et al., 2019) is beneficial, as the variety of tasks that LLMs can solve or assist in solving are virtually infinite. Based on a systematic, structured analysis and synthesis of previous relevant concepts as well as the elaboration of specific requirements for dealing with AI tools competently, we aim to develop a new conceptual framework and discuss implications for a corresponding PE assessment framework, which holds particular research and practical significance. Studies have already assessed PE as a skill without explicitly defining its components and indicators (Knoth et al., 2024b) and related it to collateral literacies such as AI literacy (Knoth et al., 2024a). Such preliminary work indicates that educational researchers recognize the importance of investigating PE as a distinct skill. A comprehensive PE definition and conceptual framework have not yet been developed.

Following the conceptual analyses, we conclude that this skill is necessary for learning and working with such AI-supported tools like ChatGPT, and as such requires separate and specific investigation from the educational science perspective as a new 21<sup>st</sup> century skill.

### 3 Research focus

In this paper, we commence by elucidating the concept and theory surrounding PE and online reasoning skills. This strategic, analytic approach aligns with the notion that assessment, fundamentally perceived as a process of reasoning from evidence, necessitates a thoughtful design (Mislevy and Haertel, 2007). We utilize the assessment-targeted approach because it is exactly the field of educational assessment that links together theoretical ideas about the construct nature and the rigorous orientation to the data (Pellegrino et al., 2001). Therefore, this paper seeks to establish a foundation that integrates both theoretical frameworks on online reasoning skills and practical insights from PE. This dual approach aims to inform the design of assessments, ensuring they are not only rooted in sound theoretical principles but are also practically applicable to the specific context in which they are employed.

Hence, to provide a necessary foundation for developing a PE assessment framework, this paper takes one of the first steps, aiming to spark a discussion on the conceptual framework of PE skills in

educational research. Taking the inspiration from the Evidence-Centered Design (ECD; Mislevy and Haertel, 2007), we start by defining claims on how students are supposed to understand and use AI tools in the context of online reasoning. The insights from this paper will serve to inform the development of the ECD-based model of PE in future research.

Regarding PE, we make a distinction by the type of model it is applied to. In this paper, we focus on the LLMs, and not Text-to-Image Models, since they have their own specific manner of engineering prompts (Liu and Chilton, 2022; Oppenlaender, 2022). LLMs (or polymodal models) might be applied to a significantly wider variety of tasks, making them more flexible.

Moreover, we focus on the application of ChatGPT as the main and one of the most general AI-assisted tools. We also focus on the user side of PE, and not on the technical side of improving the model performance by specifically training it for the task. This machine learning subfield is also called PE, but it focuses on technicalities, like text embedding optimizations (Gu et al., 2023), or training on specific outputs indicating the nature of reasoning of a larger model (Mukherjee et al., 2023). Hence, for the sake of this paper, we exclude any procedure that implies re-estimation or optimization of the LLM parameters from the scope of PE and focus it exclusively on the user-side of LLM applications.

In addition, we make a distinction between PE as a (composite) skill and PE as a practice. PE as a practice has been described to some extent by other researchers (Cain, 2024; Wang et al., 2023b), and some showcase examples aimed at learning PE (Google, 2024). The description of PE practice is focused on unsystematized hints, tricks, and examples intended to help users achieve the desired result from an LLM. In the description of PE as a practice, many researchers emphasize that PE is often a continuous process that unfolds over several iterations of interactions between a human user and an LLM, much like many other information processing-related practices. This makes the description of the PE skill, like descriptions of many other information processing-related skills (Goldman and Brand-Gruwel, 2018) challenging because this structure needs to be able to incorporate the sequential aspect of the process. In such processes, many distinct cognitive components might activate in different orders or simultaneously, complicating their untangling for research investigations. However, the structure of the skills utilized in PE practice has been scarcely addressed, which serves as the motivation for this paper.

The (online) information literacy concept, regardless of the exact framework or definition, typically splits into passive (user) and active (developer) use (Koltay, 2011). In this paper, we discuss PE in the context of only passive use by (higher education) students for learning and knowledge acquisition, which corresponds to engineering prompts and evaluating LLM output. PE itself, however, can be considered under the frameworks of computer-assisted text production or creative writing, but these frameworks also lie beyond the focus of this paper as these aspects are more closely related to linguistics, media and communication science, rather than educational science.

While skill descriptions in Internet use and information acquisition might also apply to LLMs in general, our conceptual analysis illustrates that skills for competent use of LLMs (including PE) for learning differ from most skills in frameworks. Given the enormous interactivity of LLMs, their dependency on user input, their

virtually unlimited knowledge, and their tendency to hallucinate while remaining very convincing, we conclude that PE requires distinct skills not covered by traditional (online) information literacy frameworks. To conceptualize a specific PE skillset, we review related prior skills frameworks, illustrating what we can learn from them, but also how they fall short in modeling specific PE (sub)skills, revealing a gap in the research.

## 4 Review of online skills frameworks and their distinctions from prompt engineering

When designing conceptual (and assessment) frameworks for 21<sup>st</sup> century skills related to AI, one first needs to identify the knowledge and skills students need while engaging with tasks. We, therefore, relate PE to a selection of prominent skill frameworks and show that they are too global, referring primarily to search-engine-based Internet inquiry, and do not cover LLMs, even lacking entire PE components (see also [Zlatkin-Troitschanskaia et al., 2021](#)). This section aims to illustrate that, although these skill frameworks are relevant for contextualizing PE in a pre-2023 Internet, they are currently insufficient for describing PE itself or in the context of the Internet in 2024.

### 4.1 Prompt engineering as part of (exploratory) technology use and targeted inquiry for information acquisition

Prompt Engineering (PE) for learning in (higher) education can be considered under at least two broader kinds of activities that relate to research on 'literacies': digital technology use and manipulation (as part of digital literacy) and information acquisition (as covered in information literacy) (for a differentiation, [Koltay, 2011](#); for a synthesis for assessment of digital information literacy in higher education, [Sparks et al., 2016](#)).

In using digital technology, (higher) education students' understanding and use of LLMs (and corresponding tools like ChatGPT) can be examined as the ability to use a specific class of platforms for adequate purposes in ways to achieve desired results, e.g., to obtain textual output with specific qualities from an LLM. This paradigm highlights that the user

1. decides to consult an LLM (vs. other information resources) to acquire specific types of information,
2. interacts with the selected LLM,
3. ends the interaction when a compelling mental or emotional state is reached (e.g., satisfaction, frustration, tiredness, boredom), having either completed their inquiry or not.

Here, the motivation for selecting a specific LLM is important, and LLM(s) might or might not be the only source of information for the user. Regardless, PE in this context refers to the sub-phase of inputting and refining prompts and marks the user's main active input to the LLM to obtain desired information. This perspective helps frame students' general tool exploration and experimentation (among platform novices), their versatility in interacting through inputs, troubleshooting, ludic use, and unintended or original technology uses.

In the context of exploratory technology use, such as the inquiry of a new topic to assess a resource's usefulness, users may not necessarily aim for efficient prompting. Instead, they may seek to test the capabilities of an LLM within their domain of interest, evaluating factors such as breadth and depth of answers, as well as information quality. This testing may involve pushing the LLM to its limits to understand its full potential. By contrast, for specific inquiry, more skillful goal-directed users can be expected to seek to obtain only the types of information the present LLM can indeed produce (above their desired quality threshold). Thus, part of the PE skill set includes knowledge and understanding of what the LLM system can and cannot do to judge its suitability to a given task and use it only for as long as it is helpful (section 5.1). Advanced users can benefit from understanding the capabilities and limitations, ethics and privacy tradeoffs of different LLMs (including their multimodal data capabilities), the coverage of their training databases, reasoning capabilities, speed, energy use, cost, and other metadata. This knowledge enables them to select the most suitable set of tools for their inquiries.

### 4.2 Prompt engineering and online information problem-solving skills

In the tradition of research on information literacy, conceptualizations have combined general tool use for active production and (passive) interpretations of information, but have still been deficient in conceptualizing Internet-based skills ([Foster and Piacentini, 2023](#)). More applied conceptual approaches have sought to narrow this gap, from Multiple Document Literacy and Multiple Source Use/Comprehension/Understanding to Information Problem Solving on the Internet (IPS-I) (for an overview, [Goldman and Brand-Gruwel, 2018](#); [List and Alexander, 2019](#)).

IPS-I ([Brand-Gruwel et al., 2009](#)) was derived as a descriptive model of (five) phases that users go through when solving tasks that require the acquisition of information expanded for the Internet. These phases cover analysis (problem analysis and prior knowledge activation, searching the Internet for information, preliminary and deeper processing of information along with evaluations and reflection) and synthesis (text response drafting, and process and product evaluation), accompanied by regulation (on task, time, test content). Despite the good description of the online inquiry phases in the IPS-I model, its explanation of sufficient reasons to conclude an online inquiry can be improved ([Goldman and Brand-Gruwel, 2018](#)). Moreover, IPS-I employs task covering typical online platforms but does not model their affordances explicitly, and does not yet include a differentiation of single-or multi-platform inquiry, or search engine vs. LLM querying.

In general, the IPS-I framework, although promising, needs to be adapted to LLMs. For instance, the search component is entirely geared toward search engines, and evaluation does not account for specific LLM cues or the pages-long machine-generated text (although it does highlight checking page ownership). Moreover, reasoning in IPS-I does not cover how the typical responses of LLMs can veil a lack of specificity or contain factually incorrect information.

The weighting of facets should differ for PE, as well. In evaluation, there are fewer website cues to be considered, while specific LLM language cues can become more important. Knowledge of LLM



production becomes more crucial for discernment of output quality; for instance, domain knowledge which serves as a critical reference is typically still underdeveloped among learners. Regarding syntheses and reasoning, LLMs can effectively carry out part of the thinking for users. The key question becomes how satisfied, if at all, are users with the respective LLM output (when to cross-check with further sources or not), and what are possible pitfalls of LLMs to be hedged against. As LLMs can now forward queries to other LLMs and deliver results for several of the IPS-I inquiry subphases for the user, while augmenting the requirements for others such as prompt formulation, it remains to be seen in future descriptive studies if the IPS-I phase structure will hold for this new platform type.

### 4.3 Prompt engineering and PISA's framework 'learning in the digital world'

One of the flagships of the educational assessment practices is OECD's Programme for International Student Assessment (PISA). While it focuses on school students, PISA showcases assessment innovations and key 21<sup>st</sup> century skills targeted in each wave. With the advent of assessing broader skills, such as critical thinking and problem-solving in PISA (Pellegrino, 2023), the necessity of competent use of technology and fostering PE skills becomes increasingly evident. Assessing and unveiling educational needs to prepare students in leveraging technology effectively in an increasingly information-driven society, including the adept formulation of prompts for AI tools like ChatGPT, becomes a pivotal aspect in ensuring comprehensive educational outcomes.

The PISA 2025 model responds in part to recent technological trends with the Learning in the Digital World (LDW) framework. The authors break the skills of learning with and from software down into

- (1) computational and scientific inquiry practices (analyzing problems and recognizing patterns, working with software outputs, conducting experiments and analyzing data),
- (2) metacognitive monitoring and cognitive regulation (progress monitoring and adaptation, performance and knowledge evaluation), and.
- (3) noncognitive regulation processes (maintaining task engagement and affective states).

While PISA's LDW framework is limited to offline administration to school students (thereby limiting its use for open search), a chatbot has been implemented, and exploratory tool use is designed intuitively so as to tap into (secondary school level) technology-based data manipulation and examination principles. On-task learning is included as part of the assessment and is measured through logged indicators constructed from situational inference rules. Inputs are limited to presented stimuli. However, the LDW framework (2025) also presupposes a closed information environment, which in this instance does not seem to include LLM-like tools (i.e., a highly versatile chatbot).

### 4.4 Prompt engineering and open (web) search assessments

The development of AI tools goes toe to toe with the assessment innovations, making the current educational assessment practices

more authentic and complex (Sabatini et al., 2023). Assessment innovations have rapidly evolved in the recent past, also paving the way for more complex conceptualizations of skills (with the prospects of feasible operationalization). Particularly, there is a shift from closed information pools (including in multiple source use) to open (web) search assessment environments (Wineburg et al., 2022). This places much more emphasis on subskills such as targeted search, rigorous selection, and cross-referencing to obtain useful information, under the uncertainty connected to never seeing the entire information pool. This contrasts with the careful evaluation of every source and deduction of information from a limited information pool, as in assessments with a document library (Shavelson et al., 2019). Thus, open web search assessments, compared to closed ones, differentially tap and weight-assessed subskills in a more ecologically valid setup. They place importance on design features such as abundance vs. scarcity of acceptable quality alternatives, noise vs. no noise, access, familiarity, and affordances for information search and organization. However, they also raise the need to account for cheating opportunities and changes in the information pool.

LLMs can further increase these differences, while also synthesizing and filtering out some of the interim complexity of open web search assessments. LLMs are generative models, meaning that they can generate new texts that never existed before. Hence, many of the open web search characteristics also apply to their use. An LLM's capability to recombine, structure, preselect, and synthesize information (while leaving the undisplayed bits and even sources opaque) differentiates it from search engines and results in a corresponding weighting of required skills. Users need to put less thinking effort into drawing inferences and compiling an (initial) draft that summarizes their inquiry, as the system can do this step for them. Corrections are also facilitated. Compared to search engines, LLMs are more dialogical, e.g., in explicitly restating, reaffirming, specifying, correcting user's prompts (seemingly in their own terms), and are able to autonomously apply suggested changes to entire text blocks of results. Therefore, studying and assessing PE as a new education-relevant skill requires the application of sufficiently complex forms of assessment practices.

### 4.5 Prompt engineering and critical online reasoning skills

Critical Online Reasoning (COR) is a recent conceptualization of the skillset necessary to acquire, evaluate, and reason with and about sources and information from the Internet, developed for the setting of learning in higher education (Molero et al., 2020; Nagel et al., 2020, 2022). COR provides a convenient conceptual adaptation and development of IPS-I phases to the process of solving complex, open-ended information problems in a mixed information quality environment (Molero et al., 2020). In particular, COR includes three interconnecting facets: "Online Information Acquisition" (OIA), "Critical Information Evaluation" (CIE), and "Reasoning Using Evidence, Argumentation, and Synthesis" (REAS), as well as a meta-cognitive facet (MCA) for the situation-specific activation and regulation of the COR skills.

COR was developed to implement advances in assessment, specifically to include ecologically valid open web search in assessment (with associated web behavior tracking). COR has aspired to capture

competent behaviors regarding differences in the credibility of online sources and information, focusing on students' discernment of Internet sources and content in the face of realistic challenges online, such as a multitude of information, low-quality information, and/or misinformation (Molerov et al., 2020).

COR, too, was conceptualized before the popularization of LLMs. Today, students may find satisfying answers on ChatGPT and avoid further search or synthesis. LLMs can offer recommendations on general evaluation criteria. By loading or copying texts and sources into the LLM as part of a prompt, students may also obtain full machine evaluations.

## 4.6 Prompt engineering and artificial intelligence literacy

One of the more recent concepts, Artificial Intelligence (AI) literacy, is also relevant to the discussion of PE. In response to the rapidly developing AI field, the educational science community has already begun attempting to define the skills needed for competent AI use. AI literacy concerns AI in general, which is far broader than the topic of using LLMs. The concept of AI literacy distinguishes between generic and domain-specific use of AI (Knoth et al., 2024a), which includes numerous AI tools and machine learning models developed for narrow use in specific professional fields. Moreover, the explicit inclusion of attitudes in AI literacy (Wang et al., 2023a) binds at least part of the operational indicators of the construct to the self-report format, which becomes troublesome in the case of developing educational assessments.

Attempts to include more objective measures of AI literacy, however, tend to focus on the general knowledge of respondents about the structure, nature, and functioning of AI (Hornberger et al., 2023; Weber et al., 2023), which is not the same as defining *what allows a person to use AI successfully*, albeit complementary to this ability. Moreover, the employed items have been of multiple-choice format, which increases standardization but can threaten the authenticity of the assessment and ecological validity of claims about the respondents.

Given that our definition of PE aims to isolate the cognitive nature of the skill specifically tailored for the use of LLMs, we conclude that this is a distinct skill, which might be considered a part of AI literacy but is not defined by it. Within AI literacy (i.e., among all available types of AIs), LLMs and their use comprise a portion of a specific type of AI. Within LLM use, PE comprises a significant portion of the skills needed. Moreover, AI literacy frameworks are still in the early stages and evolving, requiring more specificity in various subareas. While general considerations about the functioning of AIs apply to PE as well, most parts are still underspecified.

In summary, the abovementioned skill frameworks focus on the perspective of students as agentic learners who actively regulate their own learning processes. Therefore, PE and LLM use are compatible with the above frameworks, as they essentially relate to the consistently emerging cognitive, metacognitive, and self-regulatory skills (Foster, 2023a; Roll and Barhak-Rabinowitz, 2023). They can be attributed to the respective search phases as well as partly to the evaluation and reasoning phases, if the assessment's operationalization grants students access to LLMs. However, the frameworks are not specific enough in addressing how users can skillfully interact with the AI-supported tools like ChatGPT to obtain desired information,

leaving a number of conceptual questions open. As with any new technology, novel affordances call for new (sub)skills, too.

## 5 Conceptualization of prompt engineering skills

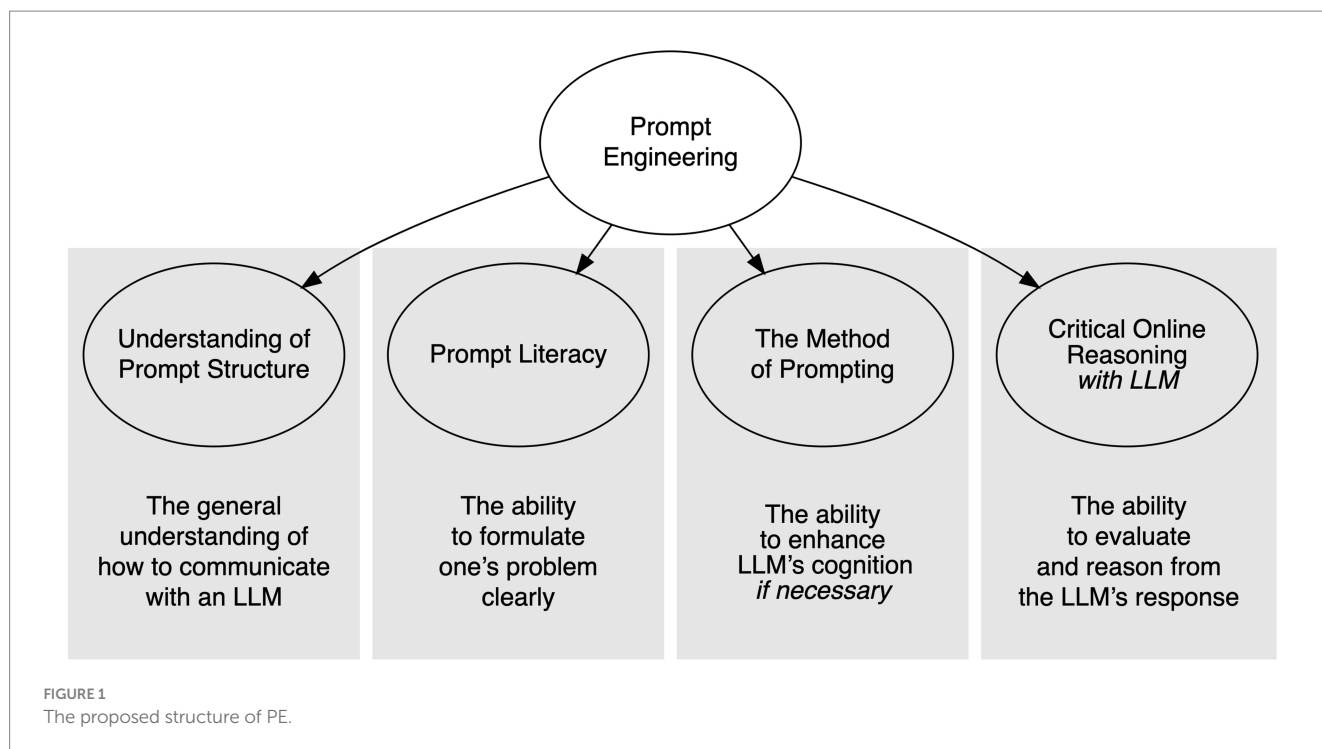
For the sake of this paper, we define PE as the skill of communicating the problem, its context, and the constraints imposed on the desirable solution to an LLM to solve it correctly as fast as possible (Lo, 2023a). Thus, this skill conveys the user's needs to an LLM in a manner that the model can understand. However, since the tasks to which the LLMs can be applied vary, the prompts for their application vary as well (White et al., 2023), so there is no general "best" prompt structure. Instead, it makes sense to describe the PE as a composite skill consisting of a combination of several subskills involved in the communication with the LLM. We describe PE as a composite multidimensional skill consisting of four skills, intertwining in the practice of using an LLM (Figure 1).

Since the purpose of this paper is to develop an operationalization of PE that can be flexibly used in the development of various assessments, either in parts or entirely, we do not provide a taxonomy of behavioral indicators (e.g., Bloom's taxonomy; Anderson and Krathwohl, 2000). Such taxonomies are closely linked to the exact type of claims about respondents that the assessment aims to make and are therefore defined by the purpose of the assessment (Mislevy and Haertel, 2007). Describing any taxonomy of behavioral indicators in the context of PE would reduce the possible scope of such applications and purposes. Instead, we aim to broadly describe the structure of the PE construct, which can be developed and used further. The application of an incorrect taxonomy can result in the misspecification of the assessment framework, a decrease in the authenticity of the assessment, and a degradation of the validity of the final claims. To further enrich the understanding of PE, we juxtapose online skills frameworks to integrate and adapt their essence to PE skills.

## 5.1 Understanding the basic prompt structure

Giray (2023, p. 2630), following the DAIR.AI (2023) frameworks, lists four elements of a prompt. These elements, combined together, substitute a prompt, which formulates the problem, gives the model the necessary information to solve it, and contains output in the desired form:

- Instruction – a specific task that guides the model's behavior (e.g., "Proofread the text");
- Context – external information or additional context that provides background knowledge to the model, helping it generate relevant responses (e.g., "The text is an email that needs to follow an official corporate style");
- Input data – the content of the prompt that the model needs to solve, might vary given the instruction (e.g., contents of different emails);
- Output indicator – specifies the type or format of the desired output (e.g., "Do not rewrite the text, only correct grammar, spelling, and punctuation").



Incorporating “output indicators” in the prompt structure implies that a user needs to have a projected image of the result. This means that PE inherently requires a user to know what they want. This will allow an AI tool to shape the output according to the user’s expectations. This allows the user to have representational benchmarks, against which the output of an AI tool is judged.

In sum, understanding the necessity to provide all these elements, as well as the ability to optimize them when needed, can be considered the necessary part of PE. The quality of these components is particularly important, since imprecise formulations or irrelevant information can derail the LLM’s response to the prompt.

## 5.2 Prompt literacy

Prompt literacy addresses the user’s ability to be precise in their formulations. None of the research to date has come up with any exhaustive lists of requirements for being precise in prompts. Still, [Hwang et al. \(2023\)](#) define prompt literacy as the ability to generate precise prompts as input for AI tools, interpret the outputs, and iteratively refine prompts to achieve desired results. Others vaguely address this literacy in terms of avoidance of pitfalls and common mistakes that learners make while engineering prompts ([Busch et al., 2023](#); [Lo, 2023b](#)). Nonetheless, in practice, avoidance of the aforementioned pitfalls might not be necessary, as the improvement after an initial input or the general awareness of the limitations of one’s suboptimal prompts allows for the avoidance of incorrect conclusions. Mostly, researchers highlight such aspects of prompt literacy as ([Giray, 2023](#)):

- Ambiguity or lack of specificity – without a concrete focused input, an LLM might wander away from a desired solution.

- Bias reinforcement – an ill-formulated prompt might provoke an LLM to give an answer which can be interpreted as biased.
- Overfitting and unrealistic dependency on model limitations – an LLM might not know all the specific details of a certain field or area, and as a result might be not the best consultant on an overly specific topic.
- The correct context – LLM needs *necessary and sufficient* context to work with (e.g., “write an email” is not a specific enough prompt to solve a problem correctly).
- Overly complex prompts – supplying too much information might trigger the LLM to focus on an irrelevant part of the prompt.
- Ethical considerations – ethical and ecological use of an AI that does not inherently have values or an ethical system remains the responsibility of the user.

The aspect of ethical considerations has received significant attention in PE literature. It relates to the fact that many LLMs have pre-built system prompts (hidden from the user) that explicitly prohibit them from discussing unethical topics (e.g., crimes or violence). Because of this, so-called “jailbreaking” has gained special attention in PE literature as a way to relax this limitation ([Zhou et al., 2024](#); [Yu et al., 2024](#)). This topic is closely related to the general ethics of AI usage and is a specific field of AI research in general ([Jobin et al., 2019](#)) and AI research in education specifically ([Borenstein and Howard, 2021](#); [Burststein, 2024](#)) for the Duolingo Standards on Responsible AI). While a rigorous discussion of this topic is beyond the scope of this paper, the ethics of PE, as well as the ethics of a major part of general AI use, boils down to keeping a human in the loop of the process involving an AI tool and holding the person accountable for the decisions made ([Shah, 2024](#)).

Interestingly, in some cases, the context of the problem needs to be reduced rather than explicated. For instance, in the study by

Krupp et al. (2023), a physics problem was phrased along the following lines: “Tarzan swings hanging on a liana of a given length with a given speed from a given height. He picks up Jane (who has a given mass) standing still on the ground. Calculate Tarzan’s speed right after he has picked Jane up.” This phrasing can derail ChatGPT into discussing the Tarzan story, rendering the output useless. However, rephrasing the problem in terms of pendulums and loads will result in ChatGPT giving a correct response (or, in analogous cases, at least providing the user with the correct formulas for further calculations). This example illustrates that LLMs, like humans, can struggle to discern details from the core of the problem if the context is too unexpected. This behavior makes LLMs similar to humans who might experience the same difficulties (Carnoy et al., 2015). Therefore, a proficient level of PE requires the user to carefully measure the necessary and sufficient context for solving the given problem, not only expanding it (as is typically highlighted in the literature) but also reducing it in some cases.

In general, there is a striking similarity to item writing principles from test development and prompt literacy. For example, Haladyna and Rodriguez (2013) highlight 23 features that test items should have, among which, for example, are:

- Test important content, avoid overly specific and overly general content;
- Avoid opinions unless qualified;
- Avoid trick items;
- Edit and proof items;
- Keep linguistic complexity appropriate to the desired output;
- State the central idea clearly and concisely.

These similarities make sense conceptually, since, in both cases, the prompt/item writer is trying to be as precise, unambiguous, and economical as possible to achieve the purpose of prompt or assessment. This is attributed to the fact that, in both situations, a higher number of brief items can yield more reliable data, as the information can be amassed across a greater number of instances (Piacentini et al., 2023).

## 5.3 The method of prompting

The method of prompting is an aspect of PE that up until now has almost exclusively been studied from the technical perspective, or just anecdotally described by the users. The method of prompting is an inherent component of PE that includes using special verbal ways of organizing the prompt information to help an AI tool solve the posed problem. Although not all methods of prompting are suitable for every problem, it is crucial for users to understand these methods and identify when they are applicable. This knowledge can significantly improve the performance of LLMs.

In the following, we provide a relatively detailed discussion of the methods of prompting, as it offers practice-related insights into the functioning of LLMs. We suggest that understanding these aspects might be more important than the technically oriented knowledge of internal AI machinery, which tends to attract researchers’ attention when they attempt to measure AI literacy. Hence, these methods of prompting might serve as a basis for the assessment of PE.

In general, there are many different methods of prompting. For example, Sahoo et al. (2024) describe 29 distinct techniques, most of which can be considered separate methods of prompting. However, some of them can be seen as variants of each other, and others are prompting strategies individually tailored to specific tasks. Moreover, new methods of prompting that require increasingly more skills than information processing (e.g., collective prompting that includes communication between human users and thus requires social communication skills; Wang et al., 2024) are continuously being created, making it impossible to exhaustively describe and systematize all recent prompting methods. Therefore, we limit ourselves to a brief description of some of the most prominent and important “families” of prompting methods.

### 5.3.1 Few-shots prompting

Few-Shots (FS) prompting<sup>1</sup> (Brown et al., 2020) may be important when an LLM is required to reason by analogy. FS prompting refers to the idea of providing the model with several examples of similar tasks and their solutions before the actual task. This idea bears a heavy similarity to Bandura and Jeffrey’s (1973) observational learning concept, suggesting that people might learn something from observing other people doing it. The mechanics of LLMs’ reasoning in this approach to prompting (the input-label mapping, the distribution of the input, the label space, and the output format; Min et al., 2022) details the “motor reproduction” process – one of four processes that account for learning according to Bandura and Jeffrey (1973), except, it unfolds in the verbal space.

Another process, retention, also has a reflection in the prompting literature since if the total length of the prompt exceeds the context memory of an LLM (the number of input tokens that the model uses to condition its responses), LLM’s performance decreases (Mosbach et al., 2023; Kuratov et al., 2024). This exact phenomenon is fundamental to some of the “jailbreaking” techniques, which aim to overwhelm the context memory of an LLM to make it “forget” the ethical constraints contained in the latent system prompt (Jiang et al., 2024). However, with the recent chase after an exponential increase in the number of context tokens that an LLM can remember (up to millions of tokens; Reid et al., 2024; Zhang et al., 2024), these “jailbreaking” techniques become obsolete, and the problem of insufficient LLM context memory decreases. This chase unlocks other features of LLM use, such as the so-called mega-prompts (a couple of pages long) and the use of dozens of examples for few-shot (FS) learning.

Other processes from Bandura and Jeffrey (1973), attention and motivation, are barely covered in the prompting literature. However, while motivation is non-existent in AI literature in general on account of LLMs lacking it, the attention mechanism is almost solely responsible for LLMs’ existence (Vaswani et al., 2017). This mechanism allows LLMs to find the dependencies between language tokens from different parts of a token sequence in a computationally efficient

<sup>1</sup> Originally, this method of prompting has been termed *Few-shots (AI) learning* (Brown et al., 2020), and later the term has become *In-context (AI) learning* (Dong et al., 2022) but since we refer to AI learning as to the process of optimizing the model parameters, we have re-labelled this method of prompting to better reflect its nature.



manner. Still, this mechanism is an architectural feature and has no impact on prompting strategy.

FS prompting also reflects a somewhat traditional insight from psychological research in intelligence (Gentner et al., 2001) and higher-order reasoning (Alexander et al., 2016) which states that analogy is the fundamental concept of these processes. In the context of AI, an LLM having few examples of what is required from it can focus on the key aspects of the task better, generalize from them, and repeat the required information processing on the actual task.

FS prompting has had such an immense impact on the model performance that it has become a general practice in evaluating the model performance to reflect in the reporting documents what method of prompting exactly (e.g., 5-shots or zero-shots) has been used when several competing LLMs are measured against benchmark tasks (Bragg et al., 2021).

### 5.3.2 Chain-of-thought prompting

When an AI tool, for example, is required to perform some complex informational tasks (e.g., to formulate the implications of a text, or to reason from it in regards to a specific context), it needs to be allowed to spell out its reasoning steps (Kojima et al., 2022). Since LLMs do not have implicit higher-order reasoning skills (as they only *predict the next language token*), their reasoning can only occur in the form of “thinking aloud.” This method of prompting in particular resembles concurrent thinking aloud (Fuchs et al., 2019). If the core of a request to an LLM includes several complex operations on the textual information, requiring the model to explicitly describe the steps that lead to its conclusion invokes the higher-order reasoning skills in the model and sufficiently improves the quality of the output (Wei et al., 2022). Such method of prompting is called *Chain-of-Thought* (CoT) prompting.

However, not all tasks require CoT prompting as, for example, some requests may just require creating a simple overview of a topic or rewriting a text. Hence, knowing about this method of prompting and recognizing when and how to have a model “think aloud” is also required from a user. This prompting method also has implications for the machine learning community, since training the model on the datasets that explicitly describe those reasoning steps can sufficiently boost its intelligence, even if the size of the model is relatively small. In such cases, the model trains to mimic the reasoning steps described in the training corpus, which is sufficient to exhibit impressive reasoning skills in the model evaluation (Mukherjee et al., 2023; Mitra et al., 2023).

Currently, CoT prompting has sparked a separate area of research in PE (Sahoo et al., 2024). For example, CoT promoting has been generalized to Graph-of-Thought (Yao Y. et al., 2023) and X-of-Thought (Ding et al., 2023) reasoning strategies that force LLMs to learn to reason internally, without spelling the solution process out. A significant portion of such research is dedicated to “interiorizing” this higher-order reasoning in LLM. The purpose of this “interiorizing” is essentially to make LLMs automatically (in a hidden manner, “internally”) apply the reasoning steps without spelling the reasoning steps out (“externally”). This appears to be the key to unlocking the extremely complex cognitive performance of AI (Chu et al., 2023). This research direction bears similarities to Vygotsky’s concept of interiorization, which states that higher psychological functions initially develop with external support in the real world and then become executed internally within the human mind without requiring

this support (Bertau and Karsten, 2018). Importantly, these similarities are only superficial, since Vygotsky described the development of human psychological phenomena.

### 5.3.3 Tree-of-thought prompting

While this is a generalization of CoT (Yao S. et al., 2023; Yao Y. et al., 2023; Long, 2023), it has gained particular prominence. While some technical implementations of ToT require coding applications, a non-technical prompting variant has been suggested. It requires a prompt that emulates a collaborative brainstorming session among experts (Al-Samarraie and Hurmuzan, 2018). Hulbert (2023) uses the following prompt: “Imagine three different experts are answering this question. All experts will write down 1 step in their thinking, then share it with the group. Then all experts will move to the next step, etc. If any expert realizes at any point that they are wrong, then they leave. The question is...” This method enables the model to fulfill multiple roles and potentially enhances its performance.

*Self-Consistency* (Wang et al., 2022) is another technique used to enhance model performance. Essentially, this method poses the same query to the model multiple times and determines the most frequently occurring response. While there are various sophisticated methods to refine this procedure (Wang et al., 2022), it is also possible to apply it in a straightforward, manual fashion, ensuring that the model does not retain memory of previous responses (e.g., by initiating new chat sessions). This concept is akin to the *wisdom of crowds* (Surowiecki, 2005), which posits that a collective group of individuals often makes more accurate judgments than individual members of the group. In the context of LLMs, the model is treated as if it were a crowd of people, with each new attempt at answering the question acting as an independent opinion from the group. It is crucial, however, for the user to ensure that each response remains separate from the previous one – repeating the question in a continuous chat thread may lead to biased reasoning due to influence from the previous attempts.

*Self-fact-checking* is another strategy (Semnani et al., 2023), which helps to mitigate LLM hallucinations. While originally designed as a chat-bot function, users can manually adopt this technique. Here, the response is divided into individual claims, verifying their accuracy separately, and constructing a final response only from those which are correct. Although the self-fact-checking chat-bot has shown superior performance compared to other LLMs, it operates more slowly due to the additional steps involved. Nevertheless, users can incorporate elements of this method by inquiring about the veracity of sources or specific facts. This practice draws strong parallels with retrospective thinking aloud (Prokop et al., 2020), focusing on the evaluation of information rather than its generation.

### 5.3.4 Role-model

Another strategy relating to the method of prompting are special role-model hints that a user can have an LLM consider when answering the request. Such approaches have been described only anecdotally to date (Ivanovs, 2023). For example, some users have noticed that ChatGPT can perform better if it has been offered money for the successful solution.<sup>2</sup> Although this is an obviously nonsensical statement, adding this suggestion to the prompt

<sup>2</sup> <https://x.com/voooooogel/status/1730726749854663093?s=20>

evidently increases the meticulousness of ChatGPT's response. It has been suggested that this is an artifact of the dataset that was used for the ChatGPT training. Since it included some Internet forums where users ask for help, some of such requests included the promise of monetary prizes for those that would help to overcome the problem. Correspondingly, the solutions provided to such requests were more verbally rich, rigorous and meticulous, and better overall. Hence, once tokens meaning the promise of money for the solution (or similar ones—for example, saying that the user's work or life depends on the success of the solution) are used in the prompt, ChatGPT imitates the responses that were given to similar requests on the Internet. This is consistent with the intuition of neural networks learning the data features, which can be overlooked by the creators of the training datasets but are still present (Buolamwini, 2017). In terms of superficial psychological analogies, this calls for observational learning (Greer et al., 2007) to be externally motivated (Hendijani et al., 2016).

The list of similar role-model hints is constantly increasing, as users discover new saddle features of ChatGPT's behavior. Some recommendations to date include:

- Asking LLM to “take a deep breath” (because, apparently, this combination of tokens is used when people describe the successful solutions of the problem after a long and frustrating chain of attempts; Yang et al., 2023),
- Asking ChatGPT to imagine, that it is now May (that is related to the fact that ChatGPT also receives a latent timestamp of the prompt as well as the training dataset also having timestamps; apparently, close to holidays (especially in December), the length of human responses which were contained in the training dataset decreased, resulting in ChatGPT giving more concise responses leading up to and after wide-spread holidays<sup>3</sup>),
- Stating that a user “unfortunately has no fingers,” so “they cannot type” (apparently, it is especially successful in the request of writing programming code; it makes ChatGPT provide a final solution to the problem, incorporating all small changes to the final code at the same time; Ivanovs, 2023).
- Additionally, several other tricks, such as making the LLM repeat the question before answering or stressing human-relevant motivation factors (Bsharat et al., 2023), appear to have a positive impact on LLM performance.

## 5.4 Toward specifying prompt engineering in relation to critical online reasoning

Walter (2024) has suggested that the ability to critically evaluate the output of an LLM is a crucial part of successfully integrating AI into educational processes, alongside prompting skills. Additionally, Krupp et al. (2023) found that one of the major problems in students' use of LLMs is the lack of critical evaluation of LLM outputs. Given that the prompting structure (section 5.1) includes output indicators, PE implicitly requires the user to conceptualize the desired LLM output and evaluate the actual output against it. We suggest that this

set of skills is necessary at the stage of evaluating LLM output and deciding on further actions.

With regard to the COR facets (section 4.5), particularly the evaluation (CIE) and reasoning (REAS) facets are necessary for concluding whether or not the LLM output satisfies the necessary criteria of the desired solution. The information acquisition facet (OIA) is relevant at the stages of selecting a platform such as an LLM for a (sub) inquiry, choosing between several available LLMs, and formulating prompts. When formulating a relevant prompt, the user is responsible for correctly phrasing and articulating the prompt in relation to the actual problem they are trying to solve.

The meta-cognitive facet (MCA) is related to the motivation of the user to critically evaluate the LLM output. This facet is the most elusive in the COR structure since it is hard to operationally disentangle *low motivation to use COR abilities* from *low COR abilities*. This problem is one of the most important in the field of assessment of higher-order cognitive skills in general and 21st century skills in particular, as they by definition include this meta-cognitive component. Current assessments presume motivation and awareness as given within the test-taking window, thanks to extrinsic motivators (i.e., test-taking incentives) and explicit task instructions.

The features making COR an important component of PE include its interactivity, high emphasis on ecological validity, focus on information quality and web behavior tracking capabilities. Given that LLMs are by definition interactive, if a user finds the output of an LLM unsatisfactory, they might change the prompt on the spot or correct it in natural language immediately after evaluating the output. This significantly alters traditional understandings and conceptualizations of critical reasoning because they are often defined and measured in much less interactive environments.<sup>4</sup>

Moreover, high authenticity and ecological validity have been crucial parts of the COR skills from the very beginning of their development, demanding innovative assessment formats that do not restrict the natural unfolding of these processes by utilizing traditional standardized and well-studied response formats (such as multiple choice). Additional features of COR skills, such as their connection to the online environment and their utility for filtering out fake information (Molero et al., 2020), strengthen the tie of this skillset to PE. Given LLMs' propensity to hallucinate and invent non-existent information (such as imaginary sources), the focus in COR skills on source quality evaluation is highly relevant.

## 6 Implications and challenges for developing a prompt engineering assessment framework

The purpose of this paper is not to develop an assessment framework of PE but to provide an initial conceptualization of the PE construct as a skill set that enables a person to use LLMs successfully. This has implications for the corresponding assessment framework to be developed in the future. Section 5 might serve as a PE construct

<sup>3</sup> <https://x.com/RobLynch99/status/1734278713762549970?s=20>

<sup>4</sup> For an exception, see Jahn and Kenner (2018), whose 4 phases model synthesizes critical thinking into both a receptive and an interactive half arch; the latter includes hypothesis formation and testing.

model within the terminology of ECD (Mislevy and Haertel, 2007), as it lists the proposed conceptual components of PE and broadly describes their content. However, not all components might be used in an assessment if one is willing to accept the limitations of the claims about the respondents that a selective construct model may entail. Moreover, some components of the constructed model can be added to the proposed structure if this is justified for a given assessment. For example, it is expected that different LLMs perform better when solving different kinds of problems. Hence, the general awareness of specific proficiencies of different LLMs or ethical considerations in LLM use can be designated as separate components of PE if necessary.

We intentionally refrain, however, from providing an ECD evidence model in this paper. Given that PE is a 21st century skill, its assessments can be developed for an enormously wide range of situations. From formative assessments in high schools to high-stakes assessments in recruiting, different aspects of PE might be more or less relevant for different contexts. Given the inherent connection of the evidence model with the assessment context, we do not provide any limitations on the types of evidence that can be used to assess PE as a skill.

However, when it comes to the ECD task model, one issue becomes abundantly clear: it is nearly impossible to assess PE with the traditional multiple-choice response format. The inherent property of PE—interactivity—demands innovative response formats for any PE assessment. The necessity for a highly authentic task model, in turn, impacts the scoring procedure. While psychometricians are understandably comfortable with traditional response formats, utilizing them for the assessment of such highly complex skills appears to be a misspecification of the assessment framework.

The use of restricted virtual chats similar to PISA's collaborative problem-solving assessment (OECD, 2017) is a possible approach here. In such assessments, the multiple-choice items are masked by students selecting pre-formulated replies suggested by a test developer, along with a coherent storyline and rescue points mimicking learning progression toward the relevant outcomes (Piacentini et al., 2023). However, while the limited variability and flexibility of suggested responses in such virtual chats can be advantageous for assessment standardization, the assessment's authenticity is still decreased in this option. Creating choice-rich environments is a very complex task on its own (Piacentini et al., 2023).

Instead, the use of open-ended response formats is appealing for PE assessment. They appear to be highly effective in capturing the shifting assessment purpose from summatively evaluating the presence of static knowledge to evaluating students' ability to acquire and scrutinize knowledge in different contexts (Roll and Barhak-Rabinowitz, 2023). In the face of this paradigm shift, it has been asserted that generating, assessing, and processing such complex data streams from these interactive tasks is feasible on a large scale only through the utilization of advanced digital technologies (Hu et al., 2023). Additionally, more robust claims can be made on students' (differential) skills if they work on invention activities in an unconstrained (or less constrained) environment (Piacentini et al., 2023).

The scoring of open-ended items, however, is usually done with human raters, prohibiting interactivity and making such assessments very expensive. Here, the utilization of other language models – for automated scoring of open-ended responses (LaFlair et al., 2023), as well as other innovations from the field of automated LLM evaluation

with specific evaluator language models (Kim et al., 2024) – can improve the economic feasibility of such assessments. Still, such scoring procedures will be based on uninterpretable statistical models scoring the responses, which can be considered a threat to validity (Lottridge et al., 2023).

Importantly, this also impacts procedural aspects of the assessment structure, such as the ECD delivery model. Until Small Language Models (Zhu et al., 2024) can be utilized as efficiently as LLMs, such technologically enhanced assessments will predictably require not only a computer but also a stable online connection. In general, as with any assessment, designing a PE assessment appears to involve a complex network of trade-offs between multiple aspects, with no solution fitting all situations.

Given all these implications, one of the most significant potential advantages of PE assessment is its possible orientation for learning, which goes much deeper than traditional formative assessment (Hu and Wang, 2024). The provision of learning resources through ChatGPT in assessment tasks can synergistically serve multiple purposes: enabling non-linear learning trajectories, facilitating interactivity for meaning-making, capturing digital traces to unveil intermittent cognitive processes, etc. Only tasks that trigger deeper learning can unveil misconceptions and faulty strategies, identifying further needs for support in follow-up training (Piacentini et al., 2023). This approach provides limitless opportunities for assessing self-regulated learning (Roll and Barhak-Rabinowitz, 2023). The interactivity and adaptability of LLMs can tailor challenges to different abilities, improving measurement quality and the authenticity of assessments (Piacentini et al., 2023) while maintaining student engagement in the assessment (Foster, 2023b). Overall, the thoughtful implementation of a PE assessment has the potential to grow into an unprecedented assessment-for-learning tool with capabilities previously unseen.

## 7 Conclusion

With the development of AI assisting tools, multiple areas of human learning are experiencing rapid changes. AI promises to revolutionize nearly all fields of information processing – from high-stakes decision making to education. This is especially evident in the discussions around AI-based chatbots like ChatGPT, which are based upon the use of LLMs – machine learning engines which create a sequence of language tokens (words, letters, and punctuation) as an output in response to an initial prompt from the user. Still, multiple reports have emerged stating that incorrect phrasing or an inappropriate context of the problem is capable of degrading the output of an LLM beyond any use. These reports highlight that the skill of communicating with an LLM—PE—might be as important as the AI-assisted tool itself. Moreover, since LLMs can be applied universally across nearly all areas of learning and professional activity, this skill should be conceptualized as a universal skill, similarly to the widely-recognized 21<sup>st</sup> century skills. Given that the rise of AI and its applications is expected to be increasingly wide-spread, the necessity for studying this skill in the field of educational science becomes evident (Gattupalli et al., 2023). This paper constitutes one of the first approaches to this topic, attempting to justify such investigation of this skill in the tradition of educational assessment.

We demonstrate that this emerging skillset is not covered by the existing frameworks for the 21<sup>st</sup> century skills, although, it fits within them nicely. Therefore, we suggest understanding PE as a composite skill reflecting people's ability to communicate with an LLM to solve informational problems and/or more complex disciplinary and cross-disciplinary tasks. This skill includes components reflecting the understanding of the basic prompt structure that is required for an LLM to understand the request, as well as the ability to navigate through the pitfalls of inappropriate formulation of the request. We show that the latter component, prompt literacy, bears a striking similarity to the item writing guidelines from the test development field, meaning that test developers already have a head start in understanding the art of PE. Moreover, PE requires an alternation between formulating the request and evaluating the output of the model to improve, reformulate, or stop and use the current solution provided by an LLM. This component is covered via critical online reasoning skills. Additionally, in prompting methods, we discuss different tricks in information organization and phrasing that can significantly increase LLMs' performance.

While we discuss some implications and challenges for AI assessment framework based on the initial conceptual framework of PE, concrete recommendations on the practices of the PE assessment or attempts to assess this construct lie far beyond the scope of this conceptual paper. Such a task, as well as the specific suggestion for the assessment framework of PT, would require numerous further theoretical and methodological investigations. This paper aims to contribute to the initial milestone of such a challenging endeavor and to justify the approach to the analysis of PE as a new 21<sup>st</sup> century skill, to outline its possible conceptual structure, and to call for further research.

We must critically recognize, however, that the current LLM development boom is outpacing many peer-reviewed academic research processes, assessment development cycles, and likely our abilities to maintain a sufficiently up-to-date overview. By the time a paper is completed, results and recommendations may become outdated and will certainly be incomplete. This paper does not attempt to be exhaustive; instead, it aims to initiate the discussion on PE as a skill relevant to professionals in the 21<sup>st</sup> century.

Moreover, the size of challenges in obtaining useful and credible content from LLMs keeps shifting as single facets of the inquiry process are augmented by new features. Thus, PE advice and skill components are bound to a specific time and system version (Chen et al., 2023b). They can become socially differentiated as experienced users may perform information quality or utility-enhancing services,

such as offering ready prompts, prompt generation recommendations, or tools.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

DF: Conceptualization, Writing – original draft, Writing – review & editing. DM: Investigation, Writing – original draft, Writing – review & editing. OZ-T: Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Writing – original draft, Writing – review & editing. AM: Investigation, Writing – original draft, Writing – review & editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research work was conducted in the context of the research Unit CORE (Critical Online Reasoning in Higher Education) funded, by the German Research Foundation (Funding number: 5404).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

- Abbas, M., Jam, F. A., and Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *Int. J. Educ. Technol. High. Educ.* 21:10. doi: 10.1186/s41239-024-00444-7
- Alexander, P. A., Singer, L. M., Jablansky, S., and Hattan, C. (2016). Relational reasoning in word and in figure. *J. Educ. Psychol.* 108, 1140–1152. doi: 10.1037/edu0000110
- Alkaissi, H., and McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 15:e35179. doi: 10.7759/cureus.35179
- Al-Samarraie, H., and Hurmuzan, S. (2018). A review of brainstorming techniques in higher education. *Think. Skills Creat.* 27, 78–91. doi: 10.1016/j.tsc.2017.12.002
- Anderson, L. W., and Krathwohl, D. R. (2000). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Upper Saddle River: Pearson Education (US).
- Bandura, A., and Jeffrey, R. W. (1973). Role of symbolic coding and rehearsal processes in observational learning. *J. Pers. Soc. Psychol.* 26, 122–130. doi: 10.1037/h0034205
- Bertau, M. C., and Karsten, A. (2018). Reconsidering interiorization: self moving across language spacetimes. *New Ideas Psychol.* 49, 7–17. doi: 10.1016/j.newideapsych.2017.12.001
- Borenstein, J., and Howard, A. (2021). Emerging challenges in AI and the need for AI ethics education. *AI Ethics* 1, 61–65. doi: 10.1007/s43681-020-00002-7
- Bragg, J., Cohan, A., Lo, K., and Beltagy, I. (2021). Flex: unifying evaluation for few-shot nlp. *arXiv*, 1–20. doi: 10.48550/arXiv.2107.07170
- Brand-Gruwel, S., Wopereis, I., and Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Comput. Educ.* 53, 1207–1217. doi: 10.1016/j.compedu.2009.06.004



- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33, 1877–1901. doi: 10.48550/arXiv.2005.14165
- Bsharat, S. M., Myrzakhan, A., and Shen, Z. (2023). Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4. *arXiv*, 1–26. doi: 10.48550/arXiv.2312.16171
- Buolamwini, J. A. (2017). Gender shades: Intersectional phenotypic and demographic evaluation of face datasets and gender classifiers. Master dissertation. Cambridge (MA), Massachusetts Institute of Technology
- Burstein, J. (2024). Responsible AI standards. Duolingo. Available at: <https://duolingo-papers.s3.amazonaws.com/other/DET+Responsible+AI+Standards+-+040824.pdf>
- Busch, K., Rochlitz, A., Sola, D., and Leopold, H. (2023). “Just tell me: prompt engineering in business process management,” in Enterprise, business-process and information systems modeling, eds. AaH. van der, D. Bork, H. A. Proper and R. Schmidt Cham: Springer Nature Switzerland, 3–11
- Cain, W. (2024). Prompting change: exploring prompt engineering in large language model AI and its potential to transform education. *TechTrends* 68, 47–57. doi: 10.1007/s11528-023-00896-0
- Carnoy, M., Khavenson, T., and Ivanova, A. (2015). Using TIMSS and PISA results to inform educational policy: a study of Russia and its neighbours. *J. Compar. Int. Educ.* 45, 248–271. doi: 10.1080/03057925.2013.855002
- Chen, E., Huang, R., Chen, H. S., Tseng, Y. H., and Li, L. Y. (2023a). GPTutor: a ChatGPT-powered programming tool for code explanation. *arXiv*, 1–26. doi: 10.48550/arXiv.2305.01863
- Chen, L., Zaharia, M., and Zou, J. (2023b). How is ChatGPT's behavior changing over time? *Harvard Data Science Rev.* 6, 1–47. doi: 10.1162/996088923.5317da47
- Chu, Z., Chen, J., Chen, Q., Yu, W., He, T., Wang, H., et al. (2023). A survey of chain of thought reasoning: advances, frontiers and future. *arXiv*, 1–31. doi: 10.48550/arXiv.2309.15402
- DAIR.AI. (2023). Elements of a prompt. Available at: <https://www.promptingguide.ai/introduction/elements> (Accessed December 22, 2023)
- Ding, R., Zhang, C., Wang, L., Xu, Y., Ma, M., Zhang, W., et al. (2023). Everything of thoughts: defying the law of penrose triangle for thought generation. *arXiv*, 1–34. doi: 10.48550/arXiv.2311.04254
- Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., et al. (2022). A survey for in-context learning. *arXiv*, 1–22. doi: 10.48550/arXiv.2301.00234
- Ekin, S. (2023). Prompt engineering for ChatGPT: a quick guide to techniques, tips, and best practices. *TechRxiv*, 1–11. doi: 10.36227/techrxiv.22683919.v2
- Foster, N. (2023a). “21st century competencies: challenges in education and assessment” in Innovating assessments to measure and support complex skills. eds. N. Foster and M. Piacentini (Paris: OECD Publishing), 30–44.
- Foster, N. (2023b). “Exploiting technology to innovate assessment” in Innovating assessments to measure and support complex skills. eds. N. Foster and M. Piacentini (Paris: OECD Publishing), 98–109.
- Foster, N., and Piacentini, M. (2023). Innovating assessments to measure and support complex skills. OECD Publishing. doi: 10.1787/e5f3e341-en
- Fuchs, L. S., Äikäs, A., Björn, P. M., Kytälä, M., and Hakkarainen, A. (2019). Accelerating mathematics word problem solving performance and efficacy with think-aloud strategies. *South Afr. J. Childhood Educ.* 9, 1–10. doi: 10.4102/sajce.v9i1.716
- Gattupalli, S., Maloy, R. W., and Edwards, S. A. (2023). Prompt Literacy: A Pivotal Educational Skill in the Age of AI. *College of Education Working Papers and Reports Series* 6. University of Massachusetts Amherst. doi: 10.7275/3498-wx48
- Gentner, D., Holyoak, K. J., and Kokinov, B. N. (2001). The analogical mind. Cambridge: The MIT Press.
- Giray, L. (2023). Prompt engineering with ChatGPT: a guide for academic writers. *Ann. Biomed. Eng.* 51, 2629–2633. doi: 10.1007/s10439-023-03272-4
- Goldman, S. R., and Brand-Gruwel, S. (2018). “Learning from multiple sources in a digital society” in International handbook of the learning sciences. eds. F. Fischer, C. E. Hmelo-Silver, S. R. Goldman and P. Reimann (New York: Routledge), 86–95.
- Google. (2024). Prompting guide 101: A quick-start handbook for effective prompts. Available at: <https://services.google.com/fh/files/misc/gemini-for-google-workspace-prompting-guide-101.pdf> (Accessed December 22, 2023).
- Greer, R. D., Dudek-Singer, J., and Gautreaux, G. (2007). “Observational learning” in Behavior analysis around the world: A special issue of the international journal of psychology. ed. C. Dalbert (London: Psychology Press), 486–499.
- Gu, J., Han, Z., Chen, S., Beirami, A., He, B., Zhang, G., et al. (2023). A systematic survey of prompt engineering on vision-language foundation models. *arXiv*, 1–21. doi: 10.48550/arXiv.2307.12980
- Haladyna, T. M., and Rodriguez, M. C. (2013). Developing and validating test items. New York: Routledge.
- Hendijani, R., Bischak, D. P., Arvai, J., and Dugar, S. (2016). Intrinsic motivation, external reward, and their effect on overall motivation and performance. *Hum. Perform.* 29, 251–274. doi: 10.1080/08959285.2016.1157595
- Hornberger, M., Bewersdorff, A., and Nerdel, C. (2023). What do university students know about artificial intelligence? Development and validation of an AI literacy test. *Comput. Educ.* 5:100165. doi: 10.1016/j.caeai.2023.100165
- Hosseini, M., Gao, C. A., Liebovitz, D. M., Carvalho, A. M., Ahmad, F. S., Luo, Y., et al. (2023). An exploratory survey about using ChatGPT in education, healthcare, and research. *medRxiv*, 1–21. doi: 10.1101/2023.03.31.23287979
- Hu, X., Shubeck, K., and Sabatini, J. (2023). “Artificial intelligence-enabled adaptive assessments with intelligent tutors” in Innovating assessments to measure and support complex skills. eds. N. Foster and M. Piacentini (Paris: OECD Publishing), 173–187.
- Hu, S., and Wang, X. (2024). FOKe: a personalized and explainable education framework integrating foundation models, knowledge graphs, and prompt engineering. *arXiv*, 1–17. doi: 10.48550/arXiv.2405.03734
- Hulbert, D. (2023). Using tree-of-thought prompting to boost ChatGPT's reasoning. Available at: <https://medium.com/@dave1010/using-tree-of-thought-prompting-to-boost-chatgpts-reasoning-318914eb0e76> (Accessed January 4, 2024).
- Hwang, Y., Lee, J. H., and Shin, D. (2023). What is prompt literacy? An exploratory study of language learners' development of new literacy skill using generative AI. *arXiv*. doi: 10.48550/arXiv.2311.05373
- Ivanovs, A. (2023). Users are turning to reinforcement prompts to fix ChatGPT laziness. Available at: <https://stackdiary.com/users-are-turning-to-reinforcement-prompts-to-fix-chatgpt-laziness/> (Accessed January 4, 2024)
- Jahn, D., and Kenner, A. (2018). “Critical thinking in higher education: how to foster it using digital media” in The digital turn in higher education. eds. D. Kergel, B. Heidkamp, P. K. Telléus, T. Rachwal and S. Nowakowski (Wiesbaden: Springer Fachmedien Wiesbaden), 81–109.
- Jiang, F., Xu, Z., Niu, L., Xiang, Z., Ramasubramanian, B., Li, B., et al. (2024). ArtPrompt: ASCII art-based jailbreak attacks against aligned LLMs. *arXiv*, 1–17. doi: 10.48550/arXiv.2402.11753
- Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat Machine Intelligence* 1, 389–399. doi: 10.1038/s42256-019-0088-2
- Karakainen, M. T., Kivinen, O., and Vainio, T. (2018). Performance-based testing for ICT skills assessing: a case study of students and teachers' ICT skills in Finnish schools. *Univ. Access Inf. Soc.* 17, 349–360. doi: 10.1007/s10209-017-0553-9
- Kasneji, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* 103:102274. doi: 10.1016/j.lindif.2023.102274
- Kim, S., Suk, J., Longpre, S., Lin, B. Y., Shin, J., Welleck, S., et al. (2024). Prometheus 2: an open source language model specialized in evaluating other language models. *arXiv*, 1–16. doi: 10.48550/arXiv.2405.01535
- Knoth, N., Decker, M., Laupichler, M. C., Pinski, M., Buchholtz, N., Bata, K., et al. (2024a). Developing a holistic AI literacy assessment matrix—bridging generic, domain-specific, and ethical competencies. *Comput. Educ. Open* 6:100177. doi: 10.1016/j.caeo.2024.100177
- Knoth, N., Tolzin, A., Janson, A., and Leimeister, J. M. (2024b). AI literacy and its implications for prompt engineering strategies. *Comput. Educ.* 6:100225. doi: 10.1016/j.caeai.2024.100225
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *arXiv*, 1–42. doi: 10.48550/arXiv.2205.11916
- Koltay, T. (2011). The media and the literacies: media literacy, information literacy, digital literacy. *Media Cult. Soc.* 33, 211–221. doi: 10.1177/0163443710393382
- Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., et al. (2023). Unreflected acceptance—investigating the negative consequences of ChatGPT-assisted problem solving in physics education. *arXiv*, 1–9. doi: 10.48550/arXiv.2309.03087
- Kuravov, Y., Bulatov, A., Anokhin, P., Sorokin, D., Sorokin, A., and Burtsev, M. (2024). In Search of needles in a 10M haystack: Recurrent memory finds what LLMs Miss. *arXiv*. doi: 10.48550/arXiv.2402.10790
- LaFlair, G., Yancey, K., Settles, B., and von Davier, A. A. (2023). “Computational psychometrics for digital-first assessments: a blend of ML and psychometrics for item generation and scoring”, eds. V. Yaneva and M. von Davier *Advancing natural language processing in educational assessment*. (Routledge), 107–123.
- List, A., and Alexander, P. A. (2019). Toward an integrated framework of multiple text use. *Educ. Psychol.* 54, 20–39. doi: 10.1080/00461520.2018.1505514
- Liu, V., and Chilton, L. B. (2022). Design guidelines for prompt engineering text-to-image generative models. *arXiv*, 1–26. doi: 10.48550/arXiv.2109.06977
- Lo, L. S. (2023a). The art and science of prompt engineering: a new literacy in the information age. *Internet Ref. Serv. Q.* 27, 203–210. doi: 10.1080/10875301.2023.2227621
- Lo, L. S. (2023b). The CLEAR path: a framework for enhancing information literacy through prompt engineering. *J. Acad. Librariansh.* 49:102720. doi: 10.1016/j.acalib.2023.102720
- Long, J. (2023). Large language model guided tree-of-thought. *arXiv*, 1–11. doi: 10.48550/arXiv.2305.08291
- Lottridge, S., Ormerod, C., and Jafari, A. (2023). “Psychometric considerations when using deep learning for automated scoring”, eds. V. Yaneva and M. von Davier *Advancing natural language processing in educational assessment*. (Routledge), 15–30.

- Maharajan, J., Garikipati, A., Singh, N. P., Cyrus, L., Sharma, M., Ciobanu, M., et al. (2024). OpenMedLM: Prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports*. 14:14156. doi: 10.1038/s41598-024-64827-6
- Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., et al. (2022). Rethinking the role of demonstrations: what makes in-context learning work? *arXiv*, 1–9. doi: 10.48550/arXiv.2202.12837
- Mislevy, R., and Haertel, G. (2007). Implications of evidence-centered design for educational testing. *Educ. Meas. Issues Pract.* 25, 6–20. doi: 10.1111/j.1745-3992.2006.00075.x
- Mitra, A., Del Corro, L., Mahajan, S., Coda, A., Simoes, C., Agarwal, S., et al. (2023). Orca 2: teaching small language models how to reason. *arXiv*, 1–53. doi: 10.48550/arXiv.2311.11045
- Mohr, G., Reinmann, G., Blüthmann, N., Lübcke, E., and Kreinsen, M. (2023). Übersicht zu Chat-GPT im Kontext Hochschullehre. Hamburg: Hamburger Zentrum für Universitäres Lehren und Lernen. Hamburg University.
- Molero, D., Zlatkin-Troitschanskaia, O., Nagel, M.-T., Brückner, S., Schmidt, S., and Shavelson, R. J. (2020). Assessing University Students' Critical Online Reasoning Ability: A Conceptual and Assessment Framework With Preliminary Evidence. 5:1102. doi: 10.3389/feduc.2020.577843
- Morton, J. (2024). Using prompt engineering to better communicate with people. Available at: <https://hbr.org/2024/01/using-prompt-engineering-to-better-communicate-with-people> (Accessed February 15, 2024)
- Mosbach, M., Pimentel, T., Ravfogel, S., Klakow, D., and Elazar, Y. (2023). Few-shot fine-tuning vs. in-context learning: a fair comparison and evaluation. *arXiv*, 1–29. doi: 10.48550/arXiv.2305.16938
- Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., and Awadallah, A. (2023). Orca: progressive learning from complex explanation traces of gpt-4. *arXiv*, 1–51. doi: 10.48550/arXiv.2306.02707
- Nagel, M.-T., Schäfer, S., Zlatkin-Troitschanskaia, O., Schemer, C., Maurer, M., Molero, D., et al. (2020). How Do University Students' Web Search Behavior, Website Characteristics, and the Interaction of Both Influence Students' Critical Online Reasoning? *Frontiers in Education*, 5:565062. doi: 10.3389/feduc.2020.565062
- Nagel, M.-T., Zlatkin-Troitschanskaia, O., and Molero, D. (2022). Validation of newly developed tasks for the assessment of generic Critical Online Reasoning (COR) of university students and graduates. *Frontiers in Education*, 7:914857. doi: 10.3389/feduc.2022.914857
- Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., et al. (2023). Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv*, 1–21. doi: 10.48550/arXiv.2311.16452
- OECD (2017). PISA 2015 results (volume V): collaborative problem solving. Paris: OECD Publishing.
- OECD (2023). PISA 2025 learning in the digital world framework (second draft). OECD. Available at: <https://www.oecd.org/media/oecdorg/satellitesites/pisa/PISA%2025%20Learning%20in%20the%20Digital%20World%20Assessment%20Framework%20-%20Second%20Draft.pdf> (Accessed January 4, 2024).
- Opara, E., Mfon-Ette Theresa, A., and Aduke, T. C. (2023). ChatGPT for teaching, learning and research: prospects and challenges. *Global Acad. J. Human. Soc. Sci.* 5, 33–40. doi: 10.36348/gajhss.2023.v05i02.001
- Oppenlaender, J. (2022). A taxonomy of prompt modifiers for text-to-image generation. *Behav. Inform. Technol.*, 1–14. doi: 10.1080/0144929X.2023.2286532
- Pellegrino, J. W. (2023). "Introduction: arguments in support of innovating assessments" in *Innovating assessments to measure and support complex skills*. eds. N. Foster and M. Piacentini (Paris: OECD Publishing), 15–28.
- Pellegrino, J. W., Chudowsky, N., and Glaser, R. (2001). *Knowing what students know: The science and Design of Educational Assessment*. Washington, DC: National Academy Press.
- Piacentini, M., Foster, N., and Nunes, C. A. A. (2023). "Next-generation assessments of 21st century competencies: insights from the learning sciences" in *Innovating assessments to measure and support complex skills*. eds. N. Foster and M. Piacentini (Paris: OECD Publishing), 45–60.
- Prokop, M., Pilař, L., and Tichá, I. (2020). Impact of think-aloud on eye-tracking: a comparison of concurrent and retrospective think-aloud for research on decision-making in the game environment. *Sensors* 20:2750. doi: 10.3390/s20102750
- Reid, M. Gemini Team Google (2024). Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. *arXiv*, 1–154. doi: 10.48550/arXiv.2403.05530
- Ridnik, T., Kreda, D., and Friedman, I. (2024). Code generation with AlphaCodium: from prompt engineering to flow engineering. *arXiv*, 1–10. doi: 10.48550/arXiv.2401.08500
- Roll, I., and Barhak-Rabinowitz, M. (2023). "Measuring self-regulated learning using feedback and resources" in *Innovating assessments to measure and support complex skills*. eds. N. Foster and M. Piacentini (Paris: OECD Publishing), 159–171.
- Sabatini, J., Hu, X., Piacentini, M., and Foster, N. (2023). "Designing innovative tasks and test environments" in *Innovating assessments to measure and support complex skills*. eds. N. Foster and M. Piacentini (Paris: OECD Publishing), 131–146.
- Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. (2024). A systematic survey of prompt engineering in large language models: techniques and applications. *arXiv*, 1–9. doi: 10.48550/arXiv.2402.07927
- Schmidt, S., Zlatkin-Troitschanskaia, O., Roeper, J., Klose, C., Weber, M., Bültmann, A.-K., et al. (2020). Undergraduate Students' Critical Online Reasoning—Process Mining Analysis. *Frontiers in Psychology*, 11:576273. doi: 10.3389/fpsyg.2020.576273
- Semnani, S., Yao, V., Zhang, H., and Lam, M. (2023). WikiChat: stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia. *arXiv*, 1–27. doi: 10.48550/arXiv.2305.14292
- Shah, C. (2024). From prompt engineering to prompt science with human in the loop. *arXiv*, 1–7. doi: 10.48550/arXiv.2401.04122
- Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., and Marino, J. (2019). Assessment of university students' critical thinking: next generation performance assessment. *Intern. J. Testing*. doi: 10.1080/15305058.2018.1543309
- Sparks, J. R., Katz, I. R., and Beile, P. M. (2016). Assessing digital information literacy in higher education: a review of existing frameworks and assessments with recommendations for next-generation assessment. *ETS Res. Rep Series* 2016, 1–33. doi: 10.1002/ets2.12118
- Surowiecki, J. (2005). The wisdom of crowds. *Anchor*. doi: 10.5555/1095645
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *arXiv*. doi: 10.48550/arXiv.1706.03762
- Walter, Y. (2024). Embracing the future of artificial intelligence in the classroom: the relevance of AI literacy, prompt engineering, and critical thinking in modern education. *Int. J. Educ. Technol. High. Educ.* 21:15. doi: 10.1186/s41239-024-00448-3
- Wang, Z. J., Chakravarthy, A., Munechika, D., and Chau, D. H. (2024). Workflow: social prompt engineering for large language models. *arXiv*, 1–8. doi: 10.48550/arXiv.2401.14447
- Wang, B., Rau, P. L. P., and Yuan, T. (2023a). Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. *Behav. Inform. Technol.* 42, 1324–1337. doi: 10.1080/0144929X.2022.2072768
- Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., et al. (2023b). Prompt engineering for healthcare: methodologies and applications. *arXiv*, 1–33. doi: 10.48550/arXiv.2304.14670
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., et al. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv*, 1–24. doi: 10.48550/arXiv.2203.11171
- Weber, P., Pinski, M., and Baum, L. (2023). Toward an objective measurement of AI literacy. PACIS 2023 Proceedings, 60. Available at: <https://aisel.aisnet.org/pacis2023/60> (Accessed December 22, 2023).
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., and Chi, E. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv*, 1–43. doi: 10.48550/arXiv.2201.11903V
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., et al. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv*, 1–19. doi: 10.48550/arXiv.2302.11382
- Wineburg, S., Breakstone, J., McGrew, S., Smith, M. D., and Ortega, T. (2022). Lateral reading on the open internet: a district-wide field study in high school government classes. *J. Educ. Psychol.* 114, 893–909. doi: 10.1037/edu0000740
- Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., et al. (2023). Large language models as optimizers. *arXiv*, 1–42. doi: 10.48550/arXiv.2309.03409
- Yao, Y., Li, Z., and Zhao, H. (2023). Beyond chain-of-thought, effective graph-of-thought reasoning in large language models. *arXiv*, 1–21. doi: 10.48550/arXiv.2305.16582
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., et al. (2023). Tree of thoughts: deliberate problem solving with large language models. *arXiv*, 1–14. doi: 10.48550/arXiv.2305.10601
- Yu, Z., Liu, X., Liang, S., Cameron, Z., Xiao, C., and Zhang, N. (2024). Don't listen to me: understanding and exploring jailbreak prompts of large language models. *arXiv*, 1–18. doi: 10.48550/arXiv.2403.17336
- Zlatkin-Troitschanskaia, O., Hartig, J., Goldhammer, F., and Krstev, J. (2021). Students' online information use and learning progress in higher education—A critical literature review. 46, 1996–2021.
- Zhang, P., Shao, N., Liu, Z., Xiao, S., Qian, H., Ye, Q., et al. (2024). Extending Llama-3's context ten-fold overnight. *arXiv*, 1–5. doi: 10.48550/arXiv.2404.19553
- Zhou, A., Li, B., and Wang, H. (2024). Robust prompt optimization for defending language models against jailbreaking attacks. *arXiv*, 1–28. doi: 10.48550/arXiv.2401.17263
- Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., et al. (2022). Large language models are human-level prompt engineers. *arXiv*, 1–40. doi: 10.48550/arXiv.2211.01910
- Zhu, Y., Zhu, M., Liu, N., Ou, Z., Mou, X., and Tang, J. (2024). LLaVA-phi: efficient multi-modal assistant with small language model. *arXiv*, 1–6. doi: 10.48550/arXiv.2401.02330

# Frontiers in Psychology

Paving the way for a greater understanding of human behavior

The most cited journal in its field, exploring psychological sciences - from clinical research to cognitive science, from imaging studies to human factors, and from animal cognition to social psychology.

## Discover the latest Research Topics

[See more →](#)

### Frontiers

Avenue du Tribunal-Fédéral 34  
1005 Lausanne, Switzerland  
[frontiersin.org](https://frontiersin.org)

### Contact us

+41 (0)21 510 17 00  
[frontiersin.org/about/contact](https://frontiersin.org/about/contact)

