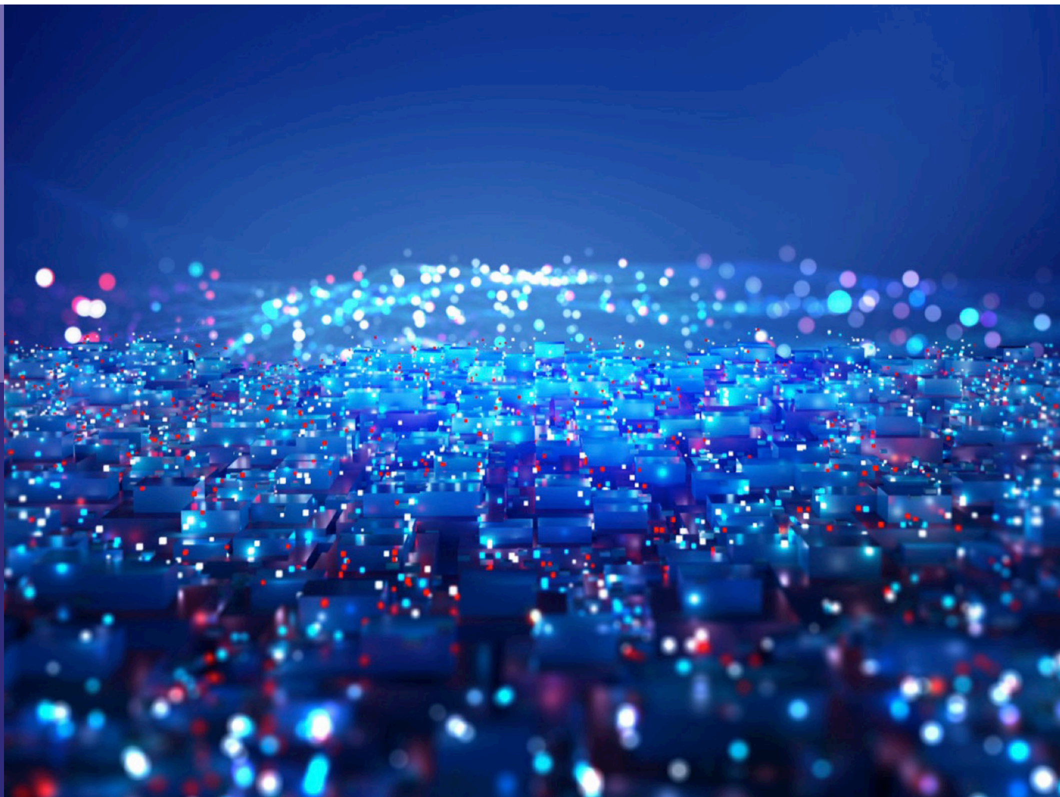
Research Topics

Generative AI in education

Edited by
Ilaria Torre, Diego Zapata-Rivera, Chien-Sing Lee, Antonio Sarasa-Cabezuelo, Ioana Ghergulescu and Paul Libbrecht

Published in
Frontiers in Artificial Intelligence



ISSN 1664-8714
ISBN 978-2-8325-5822-5
DOI 10.3389/978-2-8325-5822-5

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

Generative AI in education

Topic editors
Ilaria Torre — University of Genoa, Italy
Diego Zapata-Rivera — Educational Testing Service, United States
Chien-Sing Lee — Sunway University, Malaysia
Antonio Sarasa-Cabezuelo — Complutense University of Madrid, Spain
Ioana Ghergulescu — Adaptemy, Ireland
Paul Libbrecht — IUBH University of Applied Sciences, Germany

Citation
Torre, I., Zapata-Rivera, D., Lee, C.-S., Sarasa-Cabezuelo, A., Ghergulescu, I., Libbrecht, P., eds. (2024). *Generative AI in education*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-5822-5

Table of contents

TYPE Editorial
PUBLISHED 10 December 2024
DOI 10.3389/frai.2024.1532896

OPEN ACCESS

EDITED AND REVIEWED BY
Julita Vassileva,
University of Saskatchewan, Canada

*CORRESPONDENCE
Diego Zapata-Rivera
dzapata@ets.org

RECEIVED 22 November 2024
ACCEPTED 27 November 2024
PUBLISHED 10 December 2024

CITATION
Zapata-Rivera D, Torre I, Lee C-S, Sarasa-Cabezuelo A, Ghergulescu I and Libbrecht P (2024) Editorial: Generative AI in education. *Front. Artif. Intell.* 7:1532896. doi: 10.3389/frai.2024.1532896

Editorial: Generative AI in education

Diego Zapata-Rivera[1]*, Ilaria Torre[2], Chien-Sing Lee[3], Antonio Sarasa-Cabezuelo[4], Ioana Ghergulescu[5] and Paul Libbrecht[6]

[1]ETS, ETS Research Institute, Princeton, NJ, United States, [2]Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genoa, Genoa, Italy, [3]Department of Computer Systems and Computing, Sunway University, Bandar Sunway, Malaysia, [4]Department of Computer Systems and Computing, Complutense University of Madrid, Madrid, Spain, [5]Adaptemy, Dublin, Ireland, [6]Department of Computer Science and Data Science, IU International University of Applied Sciences, Erfurt, Germany

KEYWORDS

Generative AI, personalized assessment, AI tutors, student emotions, whole learner, interactive listening tasks, AI robots, AI mentors

Editorial on the Research Topic
Generative AI in education

In the field of education, there is a growing interest in the use of Generative Artificial Intelligence (Generative AI) to reshape the educational landscape. This Research Topic investigates the transformative potential of Generative AI in various aspects of education. The papers in this edited volume shed light on the latest discoveries, new insights, novel developments, and future challenges in this rapidly advancing field.

By leveraging machine learning models, these intelligent systems extract useful insights from vast amounts of data, making them capable of delivering highly individualized content. They can analyze a learner's proficiency level, learning style, and pace, and then tailor the study material accordingly. Generative AI can adapt its content generation strategies to meet distinct preferences and learners' needs. This can increase student engagement and comprehension, highlighting its potential to transform traditional teaching methodologies.

This Research Topic also explores the use of Generative AI as part of AI tutors, capable of tailoring instructions and feedback dynamically based on each learner's progress. Acting as an ever-present mentor, Generative AI can offer learning aids beyond class hours, facilitating continuous learning and immediate doubt clarification. This can be crucial for learners encountering obstacles outside the typical school hours or during self-study periods. Anyway, to use Generative AI as a tutor, further research is needed to examine not only the accuracy of its answers but also their emotional content, as emotions play a crucial role in the learning process.

This Research Topic includes 11 papers (Original Research: six; Perspective: two; Opinion: two, and Mini-Review: one). These papers explore areas such as: (a) using Large Language Models (LLMs) to generate feedback, (b) the use and perceived usefulness of a Generative AI chatbot for schoolwork among adolescents, (c) the potential of Generative AI in supporting critical thinking and enhancing human interactions, (d) using ChatGPT to support pre-service mathematics teachers in constructing mathematical proofs, (e) opportunities and challenges of LLMs to model the "whole learner," (f) exploring Generative AI for personalized educational assessment, (g) the use of AI-mentors in career exploration, (h) the responsible integration of AI in education, (i) the use of LLMs to automatically generate interactive listening tasks, (j) the potential of AI-enhanced robots to generate incorrect information and deceive students, and (k) the potential harm when AI-enhanced robots generate. The main contributions of these articles are described below.

Comparing emotions in ChatGPT answers and human answers to the coding questions on stack overflow by Fatahi et al.. This paper presents a study aimed to compare the emotional content in human and AI answers. Specifically, it examines the emotional aspects in answers from ChatGPT and humans to 2,000 questions sourced from Stack Overflow, finding that ChatGPT's answers tend to be more positive, while human responses often express anger and disgust. Additionally, human emotions exhibit a broader spectrum than ChatGPT. The authors suggest that ChatGPT shows promise as a virtual tutor for students by answering queries and fostering collaboration. However, further research is needed on the emotional aspects of responses.

Adolescents' use and perceived usefulness of generative AI for schoolwork: exploring their relationships with executive functioning and academic achievement by Klarin et al.. The article explores adolescents' frequency of use and perceived usefulness of generative AI chatbots for schoolwork, focusing on their relationship with executive functioning (EF) and academic achievement. Two studies were conducted with adolescents. Findings indicate that older students use Generative AI tools as more frequently. Also, students facing more EF challenges perceive Generative AI tools as more useful for completing assignments. However, no significant link was found between the use of Generative AI and academic achievement. Future work involves exploring additional Generative AI issues such as potential gender differences, implications for academic equity and the impact on adolescent cognitive development.

Using Generative AI in education: the case for critical thinking by Lee and Low. This opinion article makes the case for focusing the use of Generative AI in enhancing students' critical thinking and human interactions. The authors describe two case studies: (a) teaching communication skills and (b) teaching data structures and algorithms with AI chatbots. The two cases illustrate the potential use of Generative AI to enhance teaching and learning. The authors discuss the benefits of AI-based personalized feedback in improving student engagement and fostering strategic and critical use of AI tools. The article encourages the ethical and responsible use of generative AI in education with potential implications for the workforce.

Using large language models to support pre-service teachers' mathematical reasoning—an exploratory study on ChatGPT as an instrument for creating mathematical proofs in geometry by Dilling and Herrmann. LLMs can be a great source to extract knowledge. It thus appears natural to expect them to generate the texts of classical mathematical proofs. The authors explore how pre-service teachers employ them to produce proofs. Using the lens of instrumental genesis, their study shows a variety of usage patterns with limited knowledge about the inner workings of the models. It sketches the road to become a teacher support instrument.

Large language models for whole-learner support: opportunities and challenges by Mannekote et al. examines the transformative potential of LLMs in education through the development of personalized learning environments that address both cognitive and non-cognitive dimensions of learners, including motivation and socioemotional needs. The authors underscore the necessity of enhancing the interpretability of LLMs to ensure accurate learner representations, leveraging adaptive technologies for customized pedagogical support, and refining methods for authoring and evaluating educational agents. However, the article also highlights significant challenges, such as model interpretability, ethical considerations, and privacy concerns, which must be resolved.

Opportunities and challenges of using generative AI to personalize educational assessment by Arslan et al.. The article explores the challenges and opportunities of integrating Generative AI in supporting personalized educational assessments. The authors describe potential benefits of Generative AI personalized assessments, such as increasing learner engagement, motivation, performance, and access. Challenges include ensuring validity, reliability, and fairness. Finally, potential solutions include implementing guidelines for the ethical use of AI, aligning the purpose of the assessment with the intended use of Generative AI, and deploying human-in-the-loop approaches.

Navigating STEM careers with AI mentors: a new IDP journey by Chang et al.. The MyIDP, a Web-based STEM career development-mentoring platform, is the synergistic outcome of experts from diverse associations and universities (Hobin et al., 2012). Concerned with time and resource capacities, Chang et al. investigate the efficacy of a comprehensive list of prompts, when students engage with human-Google Gemini mentors. Progress/achievements in the Assessment, Career Exploration, Create Plan and Implement Plan phases, are measured by SMART goals. Findings reveal the emergence of the sequential integration and concurrent collaboration interaction models, and the importance of human mentors in refining and personalizing Gemini's more generic answers.

Shaping integrity: why generative artificial intelligence does not have to undermine education by Tan and Maravilla examines the role of Generative AI in promoting academic integrity. The authors argue that Generative AI can enhance learning by fostering intrinsic motivation, digital literacy, and knowledge construction. Moreover, its responsible integration can support personalized and interactive learning while upholding ethical standards. However, the paper also emphasizes the need for ethical guidelines, transparency, and thoughtful implementation to address challenges such as data privacy and algorithmic bias. Ultimately, the paper concludes that Generative AI is a tool to enrich education, preparing students for the complexities of a technologically advanced world with integrity and ethical awareness.

A generative AI-driven interactive listening assessment task by Runge et al.. This article discusses the development and evaluation of an interactive listening assessment task in the context of a large-scale assessment. LLMs are used to enhance automated item generation. A pilot study with 713 tasks demonstrated the feasibility of this approach, showing that AI-driven item generation can produce high-quality, diverse assessment content. The study highlights the potential of Generative AI and human-in-the loop to improve language testing by interactive assessment tasks.

Deception detection in educational AI: challenges for Japanese middle school students in interacting with Generative AI robots by Salem and Sumi. The authors investigate whether twenty-two Japanese middle school students can detect different types of lies (lying, paltering, pandering, and bullshitting) via an anime face in contrast to a human-like face. Analyses from ten teaching sessions indicate that there are no significant differences in learning effectiveness, and in motivation and encouragement. However, most of the students are deceived. There is also a significant difference with regards to total belief.

Exploring the utilization and deficiencies of generative artificial intelligence in students' cognitive and emotional needs: a systematic mini-review by Ortega-Ochoa et al. examines how Generative AI tools, like ChatGPT, address students' cognitive and emotional needs in educational contexts. The paper reviews four empirical works and notes challenges in scalability and generalizability, emphasizing the need for improved accuracy, personalization, and ethical integration of Generative AI to support meaningful and adaptive learning experiences. Furthermore, the paper highlights Generative AI's effectiveness in fostering engagement, emotional regulation, and instant feedback. However, it also identifies limitations, such as the inability to foster critical thinking, inconsistent response accuracy, and insufficient personalization to individual emotional and cognitive states.

As a whole, this Research Topic provides interesting insights regarding the use of Generative AI in education. The papers collectively explore the multifaceted roles of generative AI in education, examining its impact on emotional engagement, academic achievement, critical thinking, personalized assessment, STEM career guidance, and ethical considerations, while also addressing the challenges and opportunities it presents in shaping the future of learning and assessment. Our contribution represents an early step toward a scientific approach away from the trendy statements. The volume identifies the potential benefits and opportunities for additional work in this area. We hope you find these articles informative and help inspire new work in this active area of research. Finally, we would like to acknowledge the reviewers who participated in this Research Topic as well as Professor Rita Orji (Dalhousie University, Halifax, Canada), who served as the editor for one of the submitted manuscripts.

Author contributions

DZ-R: Writing – original draft, Writing – review & editing. IT: Writing – original draft, Writing – review & editing. C-SL: Writing – original draft, Writing – review & editing. AS-C: Writing – original draft, Writing – review & editing. IG: Writing – original draft, Writing – review & editing. PL: 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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

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.

Reference

Hobin, J. A., Fuhrmann, C. N., Lindstaedt, B. and Clifford, P. S. (2012). *You Need a Game Plan*. Science Adviser. doi: 10.1126/science.caredit.a1200100

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 28 August 2024
DOI 10.3389/frai.2024.1415782

OPEN ACCESS

EDITED BY
Diego Zapata-Rivera,
Educational Testing Service, United States

REVIEWED BY
Ioana Ghergulescu,
Adaptemy, Ireland
Teresa Ober,
Educational Testing Service, United States
Yang Jiang,
Educational Testing Service, United States

*CORRESPONDENCE
Daiva Daukantaitė
✉ daiva.daukantaite@psy.lu.se

RECEIVED 11 April 2024
ACCEPTED 26 June 2024
PUBLISHED 28 August 2024

CITATION
Klarin J, Hoff E, Larsson A and Daukantaitė D (2024) Adolescents' use and perceived usefulness of generative AI for schoolwork: exploring their relationships with executive functioning and academic achievement.
Front. Artif. Intell. 7:1415782.
doi: 10.3389/frai.2024.1415782

Adolescents' use and perceived usefulness of generative AI for schoolwork: exploring their relationships with executive functioning and academic achievement

Johan Klarin, Eva Hoff, Adam Larsson and Daiva Daukantaitė*

Department of Psychology, Lund University, Lund, Sweden

In this study, we aimed to explore the frequency of use and perceived usefulness of LLM generative AI chatbots (e.g., ChatGPT) for schoolwork, particularly in relation to adolescents' executive functioning (EF), which includes critical cognitive processes like planning, inhibition, and cognitive flexibility essential for academic success. Two studies were conducted, encompassing both younger (Study 1: $N = 385$, 46% girls, mean age 14 years) and older (Study 2: $N = 359$, 67% girls, mean age 17 years) adolescents, to comprehensively examine these associations across different age groups. In Study 1, approximately 14.8% of participants reported using generative AI, while in Study 2, the adoption rate among older students was 52.6%, with ChatGPT emerging as the preferred tool among adolescents in both studies. Consistently across both studies, we found that adolescents facing more EF challenges perceived generative AI as more useful for schoolwork, particularly in completing assignments. Notably, academic achievement showed no significant associations with AI usage or usefulness, as revealed in Study 1. This study represents the first exploration into how individual characteristics, such as EF, relate to the frequency and perceived usefulness of LLM generative AI chatbots for schoolwork among adolescents. Given the early stage of generative AI chatbots during the survey, future research should validate these findings and delve deeper into the utilization and integration of generative AI into educational settings. It is crucial to adopt a proactive approach to address the potential challenges and opportunities associated with these emerging technologies in education.

KEYWORDS

generative AI, executive functions, learning, cognition, academic achievement, ChatGPT

1 Introduction

After the release of ChatGPT, a Large Language Model (LLM) generative AI chatbot, a debate emerged in Sweden in spring 2023 regarding its allowance or prohibition in educational settings. At the heart of this debate were concerns about the potential for AI to be exploited for cheating, contrasted with its potential to enhance learning outcomes and promote equality in education by serving as an educational aid for students at risk of falling behind their peers.

The utilization of generative AI tools in higher education is already evident, with a study in Germany revealing that two-thirds of university students employ these tools in their coursework for tasks such as text analysis, creation, problem-solving, and decision-making (von Garrel and Mayer, 2023). A similar trend is observed in Sweden, where a survey among adolescents and young adults aged 15–24 showed that 75% use generative AI for various educational purposes, including structuring presentations and papers, writing texts, studying, and social support (The Nordic Youth Barometer, 2023).[1]

These initial findings regarding AI tool usage suggest their application in strengthening behaviors associated with executive functioning (EF)—a set of cognitive processes critical for planning, concentration, and attention, encompassing working memory, inhibition, and cognitive flexibility, among others (Diamond, 2013). These functions are crucial for goal-directed behavior, especially in tackling tasks requiring significant cognitive effort. In educational settings, such goal-directed activities often involve cognitive assignments like writing papers, necessitating independent initiation, planning, and execution. Lower EFs have been linked to reduced academic achievement across various subjects and stages of life (Pino Muñoz and Arán Filippetti, 2021). Short-term effects of EF deficits can manifest as challenges in planning and problem-solving, difficulties in peer relationships, and unfinished schoolwork (Frazier et al., 2007). In the long term, the skills nurtured by EF are crucial not only for academic and social success in school but also for vital life outcomes such as employment stability and resistance to substance abuse later in life (Bailey, 2007; Miller et al., 2011).

While informative, the aforementioned surveys lack a deeper analysis of the individual conditions that may impact the use and effectiveness of generative AI, as well as the potential relationship between the use of generative AI and EF problems in adolescence. Moreover, there appears to be a scarcity of studies investigating the adoption of these tools among young adolescents aged 12–14. Therefore, in this study we aimed to provide preliminary insights into two key objectives: (1) investigating the frequency of usage and perceived usefulness of generative AI tools in schoolwork, including potential gender differences in both usage and perceived usefulness, and (2) analyzing how these patterns of usage and perceived utility are related to adolescents' EF and academic achievement.

2 Literature review

2.1 Generative AI chatbots

Chatbots are interactive, language-based chat interfaces that automatically respond to user inquiries. They can be categorized into two groups: those utilizing pattern matching and those employing a machine learning approach. Pattern matching chatbots adhere strictly to predefined decision trees and consider only the current dialog turn. In contrast, machine learning-based chatbots can engage more flexibly with users, taking into account the entire context of a conversation (Maroengsit et al., 2019; Adamopoulou and Lefteris, 2020). Both pattern matching and machine learning-based chatbots are forms of artificial intelligence (AI), which simulates human intelligence through machines or systems (Xu et al., 2021). In the realm of chatbot applications, one historical drawback of machine learning-based AI chatbots has been the extensive amount of training data required to yield satisfactory results (Adamopoulou and Moussiades, 2020). LLMs are machine learning models designed to comprehend and generate human language text. The new generation of chatbots, powered by LLMs, exhibits remarkable capabilities due to extensive training datasets and advancements in natural language processing (NLP). These models, such as GPT-3, generate human-like text with high precision (Kasneci et al., 2023) and display emergent abilities in reasoning, planning, decision-making, and in-context learning, primarily due to their vast scale of pre-trained material (Naveed et al., 2023). However, they also carry the risk of perpetuating biases present in their training data (Bender et al., 2021). Recent evidence indicates that LLMs may propagate outdated and harmful race-based content, particularly in healthcare contexts (Omiye et al., 2023). The vast corpora of training data that enable their generative capabilities also give rise to the "black box" issue, wherein the reasoning behind AI outputs remains opaque to both users and developers (Cabitza et al., 2017). Some of these risks can be mitigated by quality filtering of data (Naveed et al., 2023) and using alignment tuning for LLMs, where human feedback is employed to make them helpful, honest, and harmless (HHH) (Askell et al., 2021). While these risks associated with bias from flawed training data and opacity of processes pose significant concerns, addressing these issues could unlock the potential of these tools to enhance cognitive processes (Baido-Anu and Ansah, 2023) and educational experiences (Al Shloul et al., 2024).

2.2 Executive functioning and generative AI

The EF refers to a set of cognitive processes, including working memory, inhibition, and cognitive flexibility (Diamond, 2013), which are essential for planning, concentration, and attention, and are crucial for academic performance in various ways. Although there are definitional disagreements regarding the exact components of EF, there appears to be a consensus that they are involved in goal-directed behavior requiring effortful problem-solving (Diamond, 2013; Gioia et al., 2015). In this study, we use the revised Behavior Rating Inventory of Executive Functioning (BRIEF-2; Gioia et al., 2015), a widely used EF rating scale, to measure various EFs, including inhibition, self-monitoring, flexibility, emotional control, task completion, working memory, and planning/organization.

In an educational setting, students rely on different cognitive processes for goal-directed behavior; for instance, when tackling writing assignments, they must independently initiate, plan, and complete tasks. Hence, EF plays a crucial role in successfully completing school assignments and achieving long-term academic success. Research has consistently linked lower EF to decreased academic achievement across various subjects, from childhood through adulthood (Best et al., 2011; Samuels et al., 2016; Pino Muñoz and Arán Filippetti, 2021). Preliminary evidence suggests that Swedish adolescents utilize generative AI tools as study aids for assignments that challenge their EF (The Nordic Youth Barometer, 2023).

Thus, the field of AI technology and its utilization among individuals with varying levels of EF is in a nascent stage. Until now studies on the intersection of AI and EF have primarily focused on

1 https://info.ungdomsbarometern.se/publika-rapporter/back2school-2023

exploring the potential of AI to assist clinicians in diagnosing various psychiatric conditions, including Attention Deficit Hyperactivity Disorder (ADHD) (Loh et al., 2022). In addition to this, certain studies have explored the potential of AI chatbots as supportive tools for children with special needs, such as Autism Spectrum Disorder (ASD) and ADHD, conditions which are frequently linked to lower EF (Tervo-Clemmens et al., 2023). This area appears to be in its preliminary stages of development as well (Torrado et al., 2023). While existing studies (e.g., Haleem et al., 2022) offer valuable insights into the impacts of various digital tools on the learning outcomes of young individuals, the introduction of new LLM AI chatbots since 2022 represents a significant advancement in the application of machine learning technologies in education. Their unparalleled accessibility and features signify a notable departure from traditional educational resources, including other digital tools, underscoring the urgent need for further research to comprehensively understand their implications. None of the aforementioned studies have investigated the new LLM AI chatbots and their effects on knowledge acquisition, and there is an apparent dearth of research focusing on non-clinical populations.

In an adult population, preliminary evidence suggests that use of ChatGPT can increase professionals' productivity in writing tasks while maintaining the same quality of output, with participants with lower level of skills benefitting especially from their use (Noy and Zhang, 2023). A similar equalizing effect was observed in a different study (Brynjolfsson et al., 2023), where adult customer support agents with lower skill levels derived greater benefits from their use of the technology compared to their more highly skilled counterparts. Hence, it appears that adult users with lower skill levels seem to have larger relative gains from employing generative AI compared to their more skillful counterparts when solving tasks relevant to their work. It is uncertain if a comparable effect exists among adolescents when addressing tasks related to their academic studies. A different question is if a similar performance raising effect would be as desirable in an educational setting, where learning, not production, is the main goal. Might there be risks for decreases in learning? At the same time, research has indicated that children with poorer EF derive the greatest benefit from activities aimed to improve these functions (Diamond and Lee, 2011). Therefore, before implementing widespread structured use of generative AI in education, it is crucial to thoroughly examine the potential risks and benefits associated with its introduction.

While cheating by submitting work done by generative AI as one's own is an obvious risk and a negative use of these tools, generative AI chatbots such as ChatGPT, if implemented safely, may also have a potential for strengthening individuals' functioning in educational settings (Bai et al., 2023). In a recent study by Jauhiainen and Guerra (2023), ChatGPT demonstrated capabilities of customizing and personalizing learning material to match students at different levels. This was achieved by using ChatGPT to tailor the main text of lessons and attached exercises to three levels: basic, intermediate, and advanced. The level of each student was assessed by collecting grades in four key subjects (Jauhiainen and Guerra, 2023). In a recent quasi-experimental study on older adolescent students, the group with access to ChatGPT showed improvements compared to the non-user group in three subskills: knowing, applying, and reasoning. Additionally, qualitative insights revealed enhanced problem-solving in the experimental condition (Wardat and Alneyadi, 2024). Thus, students, particularly those with EF challenges, might benefit from using generative AI chatbots as aids in their schoolwork—for instance, in initiating or organizing academic tasks to facilitate timely completion of assignments. It can potentially provide an equalizing effect in terms of strengthening problem-solving skills in students with EF challenges for example by integrating AI-support into special education programs. Furthermore, generative AI could assist adolescents in various subjects and topics, providing external feedback on schoolwork and explaining concepts (Kasneci et al., 2023). However, relying excessively on AI chatbots as direct replacements for these functions, especially those involved in completing schoolwork, may diminish the very cognitive abilities they substitute (León Dominguez, 2024). Recent research emphasizes adolescence as a crucial stage for executive function development (Tervo-Clemmens et al., 2023), underscoring the importance of continuing to explore any tools that might either hinder or enhance the natural progression of EF during this life stage.

2.3 Present study

The present study had two primary objectives: (1) to investigate the frequency of usage and perceived usefulness of generative AI tools in schoolwork, including potential gender disparities in both usage and perceived usefulness; and (2) to examine how these usage patterns and perceived utility are associated with adolescents' EF and academic achievement. These objectives were pursued through two separate studies. Study 1 focused on a sample of young adolescents, comprising middle school students, while Study 2 involved older adolescents, including high school students. Given the early stage of research on AI, particularly in educational settings, no specific hypotheses were posited, and the study was approached as an exploratory correlational investigation.

3 Study 1

3.1 Materials and methods

3.1.1 Sample

The analytical sample of the study comprised 385 young adolescents (179 girls, 203 boys, 3 undisclosed or not identifying as either a girl or boy, 24.3% with foreign background).[2] They were enrolled in seventh to ninth grade at four Swedish primary schools, with ages ranging from 12 to 16 years and a mean age of 14 years ($SD_{age} = 0.85$). These schools, located in the Southern part of Sweden, specifically in the Scania region with a population of over 1,340,000, shared similar socioeconomic status according to Statistics Sweden (SCB). All students in the relevant grades were invited to participate in the questionnaire, and the response rate was 80%. Additionally, all middle and high school students in Sweden receive laptops at no cost, regardless of whether they attend a municipal or private school.

2 Defined as the child either being born abroad with at least one parent born abroad as well or being born in Sweden with both parents being born abroad.

3.1.2 Measures

3.1.2.1 The use of generative AI

Information regarding the utilization of generative AI was collected through items assessing its forms, frequency, and usefulness (three items), developed by the authors due to the lack of validated measures for the use of Generative AI for schoolwork. One of these items served as a gate question, permitting access to subsequent related inquiries only for those who responded affirmatively. The gate item was visible to all participants: *"I use AI services such as ChatGPT, My-AI on Snapchat, Youchat, Bing-chat, or Socratic in my schoolwork. For example, when doing homework or solving tasks in school."* Participants responded with either 'Yes' or 'No'.

Participants who answered 'Yes' to the gate item were directed to subsequent follow-up questions inquiring about the specific AI services they used, the frequency of their use of generative AI in schoolwork (on a 4-point scale from 'rarely' to 'often') and their agreement with three statements assessing the tools' perceived usefulness for (1) initiating, (2) planning/organizing, and (3) completing schoolwork (rated on a 5-point scale from 'not correct at all' to 'exactly right'). The scores from these three usability items for ChatGPT in schoolwork were aggregated to create a usability score, demonstrating high internal consistency with a Cronbach's α value of 0.82.

3.1.2.2 The BRIEF 2 self-report form

The revised Behavior Rating Inventory of Executive Functioning (BRIEF-2; Gioia et al., 2015) self report form was used to measure EF. The self-report form targets respondents aged 11–18 years and comprises 55 items divided into seven subscales: Inhibit (the ability to resist or not act on an behavioral impulse), Self-monitor (awareness of one's impact on other people and outcomes), Shift (alteration of attention, flexibility in change, adjustment and problem-solving), Emotional control (the ability to regulate emotions), Task completion (the ability to complete tasks and/or homework on time), Working memory (the ability to hold information in mind during task completion) and Plan/organize (the ability to manage current- and future demands related to tasks). Summing the scores from the Inhibit and Self-monitor subscales composes the Behavioral Regulation Index (BRI), summing the scores from the Shift and Emotional control subscales composes the Emotional Regulation Index (ERI), and summing the scores from the Task completion, Working memory, and Plan/Organize subscales composes the Cognitive Regulation Index (CRI). Summing the scores of all indices collectively contributes to the Global Executive Composite (GEC), which serves as a global measurement of an individual's EF. The BRIEF has demonstrated good validity and reliability across various countries (e.g., Pino Muñoz and Arán Filippetti, 2021; Huizinga et al., 2023). For instance, in the standardization sample for the BRIEF-2 self-report from, Cronbach alpha for the scales ranged between 0.81 and 0.97 (Gioia et al., 2015). In the present study, the BRIEF-2 demonstrated acceptable to excellent Cronbach's alpha values, ranging from 0.71 for self-monitoring to 0.91 for task completion.

3.1.2.3 Academic achievement

The participants' grades in Swedish, Math, and English for the current academic year were collected from the schools and amalgamated to form an aggregate score representing their academic achievement. Grades in Swedish schools are denoted from F to A, with A being the highest. These grades were converted to a scale from 0 to 5 for each subject (Swedish, Math, and English) and were found to be highly correlated: Math correlated with Swedish at 0.70 and with English at 0.63, while the correlation between Swedish and English was 0.67 (all significant at $p < 0.001$). This demonstrates that the grades form a unidimensional construct with a Cronbach's alpha of 0.85. The scores were then combined to create an overall measure of academic achievement.

3.1.3 Procedure

This study was conducted as an exploratory correlation study and is part of the larger project titled "Well-being in School Environment," led by Daiva Daukantaitė at the Department of Psychology, Lund University. Data collection took place at four secondary schools during a designated lecture hour, with both a teacher and a research assistant present. Each student received a personal link to a web-based survey, which they completed digitally using either personal or school-provided laptops. Prior to starting the survey, participants were briefed on the study's purpose and content and were informed about the voluntary nature of their participation, providing consent accordingly. The survey took approximately 45 min to complete. In addition to the scales outlined in the preceding section, other measures related to mental health, emotional regulation, and life satisfaction were collected, although they were not utilized in this particular study. The research has been approved by the Swedish Ethics Committee (reference numbers: 2021–01666 and 2023–01013-02).

3.1.4 Data cleaning

In total, 393 students participated in the survey. Eight participants who did not provide a response (either 'yes' or 'no') regarding their use of generative AI in schoolwork were excluded from the sample, resulting in a reduction from 393 to 385 participants.

Participants who responded "no" to the gate question about the use of generative AI in schoolwork were included in the analytic sample to investigate differences between users and non-users. Despite their response, these participants completed all questions regarding executive functions, allowing for a comparison of non-users with BRIEF-2 scores alongside the user group.

About 5% of participants had single missing items in the BRIEF-2. The handling of missing data adhered to the strict guidelines outlined in the BRIEF-2 manual (Gioia et al., 2015) during the data cleaning process.

3.1.5 Statistical analysis

The statistical analyses were conducted using SPSS 29 (IBM Corp., 2017). Initial exploratory analyses examining gender and background differences among users and non-users of different generative AI tools for schoolwork were conducted using Chi-Square tests. Independent *t*-tests were employed to evaluate gender differences in the perceived usefulness of AI for schoolwork and differences in EFs between users and non-users of AI for schoolwork. Furthermore, Pearson correlations were performed to explore associations between measures of EFs, the frequency of use, perceived usefulness of generative AI for schoolwork, and academic achievement.

Assumptions regarding the homogeneity of group variances and normality of variables were assessed using Levene's test for homogeneity of variance and Shapiro–Wilk tests, respectively. In instances where these assumptions were violated, Mann–Whitney *U* tests and Spearman's correlations were utilized to validate the significance of the independent *t*-tests and Pearson correlations conducted.

TABLE 1 Numbers of users of different generative AI-tools for schoolwork.

Type of tool	All		Girls		Boys		χ^2	*p*
	N	%	*N*	%	*N*	%		
Any tool	57	100	24	100	31	100	0.27	0.612
ChatGPT	40	70	12	50	26	84	3.96	0.047
Bing-Chat	3	5.3	1	4.2	2	6.5		
YouChat	3	5.3	1	4.2	2	6.5		
Socratic	2	3.5	0	0	2	6.5		
Other[a]	25	44	14	58	11	35	0.92	0.341
Snapchat MY-AI[b]	20	35	12	50	8	26	1.46	0.233

[a]If Other was selected as choice, users were asked to specify which generative AI tool they used for schoolwork.
[b]Snapchat MY-AI was the most common answer in the "Other" group.

TABLE 2 Gender differences in perceived usefulness of ChatGPT in initiation, structuring and completion of schoolwork.

Usefulness for schoolwork (ChatGPT)	Boys		Girls		Gender differences		
	M	*SD*	*M*	*SD*	*t*	*p*	Cohen's *d*
Perceived usefulness for							
Initiation	4.13	1.36	4.42	1.38	0.59	0.560	0.21
Structuring	4.32	1.28	4.42	1.00	0.25	0.820	0.08
Completion	4.00	1.44	4.25	1.54	0.46	0.638	0.17
Overall usefulness	12.67	3.62	13.08	3.15	0.35	0.742	0.12

Cohen's *d* was employed as a measure of effect size for *t*-tests, where a value of 0.2 indicates a small effect size, 0.5 represents a medium effect size, and 0.8 signifies a large effect size (Cohen, 2013).

Given the exploratory nature of the study, numerous analyses were conducted, posing an increased risk for type 1 errors (false significance). To mitigate this risk, subscales for EF were only included in analyses if the Global Executive Composite, an overarching measure of EF problems, yielded significant results in preliminary analyses.

4 Results and discussion

4.1 Descriptive statistics

Out of 385 participants, 57 young adolescents (14.8%) indicated that they used some form of generative AI for schoolwork. These usage rates appear to be lower compared to findings among a youth sample aged 15–24 in Sweden (The Nordic Youth Barometer, 2023) and university students in Germany (von Garrel and Mayer, 2023), where approximately two-thirds reported using generative AI in their schoolwork.

Table 1 illustrates the types of generative AI tools utilized by participants in our study for their schoolwork; multiple choices were permitted. Among these tools, ChatGPT emerged as the most favored, being utilized by 40 (70%) adolescents. A Chi-Square test was conducted to examine whether the probability of using different generative AI tools for schoolwork varied by gender. As shown in Table 1, our analysis revealed that boys were more inclined to use ChatGPT for school-related tasks compared to girls, $\chi^2_{(1)} = 3.96$, $p = 0.047$, while no other notable gender differences were identified.

A Chi-Square test assessed the likelihood of using generative AI tools for schoolwork based on foreign background. Results showed no significant differences, $\chi^2_{(1)} = 0.11$, $p = 0.918$.

4.2 Gender differences in frequency of use and perceived usefulness of generative AI for schoolwork

In terms of the frequency of use and perceived usefulness of generative AI for schoolwork among students who reported its usage, the majority indicated infrequent use, with 37.9% using it rather seldom and 46.6% using it very seldom, while only a small percentage reported using it very often (6.9%). These patterns did not differ significantly by gender, $\chi^2_{(3)} = 1.14$, $p = 0.769$.

Regarding perceived usefulness, Table 2 presents descriptive statistics and gender differences in perceived usefulness for schoolwork. While girls tended to rate the usefulness of generative AI for initiating, structuring, and completing various assignments slightly higher, no significant gender differences were observed.

4.3 Differences between users and non-users of generative AI for schoolwork in self-rated EF

To investigate whether users of generative AI for schoolwork differed from non-users in self-rated EF, an independent *t*-test was initially conducted on BRIEF-2 self-report form Global Executive Component (GEC) scores. Users of generative AI for schoolwork

TABLE 3 Differences in self-rated EF problems between students using- and not using generative AI for schoolwork.

Variable	Users (N = 57)		Non-users (N = 328)		Differences		
	M	SD	M	SD	t	p	Cohen's d
Executive functioning							
Inhibit	14.75	4.07	13.53	3.56	2.34	0.020	0.34
Self-monitor	8.88	2.51	8.35	2.30	1.58	0.116	0.23
Shift	14.65	4.00	13.73	3.53	1.77	0.078	0.26
Emotional control	11.12	3.53	9.90	3.09	2.70	0.007	0.39
Task completion	13.74	4.21	12.72	3,79	1.84	0.067	0.26
Working memory	15.44	4.04	14.21	3.92	2.17	0.031	0.31
Plan/ organize	19.28	4.99	17.77	4.38	2.34	0.020	0.34
BRI	23.63	6.10	21.84	5.37	2.28	0.023	0.33
ERI	25.77	6.99	23.63	5.97	2.43	0.016	0.35
CRI	48.46	12.61	44.65	11.35	2.29	0.022	0.33
GEC	97.86	23.78	89.96	21.19	2.55	0.011	0.37

BRI, Behavior Regulation Index; ERI, Emotion Regulation Index; CRI, Cognitive Regulation Index; GEC, Global Executive Composite.

TABLE 4 Pearson correlations between usefulness of ChatGPT for schoolwork, frequency of use of generative AI for schoolwork and global measure of EFs problems and academic achievement.

	Frequency of use	Global measure of EF problems (GEC)	Academic achievement
Perceived usefulness of ChatGPT for			
Initiation	0.48*	0.37*	0.02
Structuring	0.45*	0.28	0.19
Completion	0.36*	0.44*	−0.16
Overall usefulness	0.53*	0.41*	0.03

*$p < 0.05$. GEC, global executive composite.

reported significantly higher GEC scores, $t(376) = 2.55$, $p = 0.011$, Cohen's $d = 0.37$, indicating that those who use AI for schoolwork reported significantly more EF problems. As depicted in Table 3, a notable difference was observed, particularly on the BRIEF-2 Emotional Control subscale, indicating that students with EF deficits may encounter difficulties in the self-regulatory skills required for initiating and completing tasks assigned by others.

4.4 Relationships between frequency of use, perceived usefulness of generative AI for schoolwork and self-rated measures of EFs

The analysis exploring the relationships between the frequency of use, perceived usefulness of generative AI for schoolwork, and self-rated measures of EF was specifically conducted for the ChatGPT user group, given its sufficient sample size for analysis. The results revealed moderate positive significant correlations between the perceived usefulness of ChatGPT, the frequency of use, and EF problems (see Table 4). These findings suggest a potential compensatory relationship, wherein users with more EF problems reported finding ChatGPT more useful for initiating, structuring, and especially completing assignments in their schoolwork. This observation aligns with prior research indicating that individuals with lower skill levels may experience greater productivity gains from using ChatGPT compared to their more proficient counterparts (Brynjolfsson et al., 2023; Noy and Zhang, 2023).

4.5 Relationships between frequency of use and perceived usefulness of generative AI for schoolwork and academic achievement

We explored the associations between the frequency of use and perceived usefulness of generative AI for schoolwork and academic achievement. Although a modest trend suggested that higher grades were associated with finding ChatGPT useful for structuring ($r = 0.19$), conversely, there was a slight inclination for lower grades to be linked with finding the tools helpful for completing school assignments ($r = -0.16$). However, these correlations, including those with frequency of use ($r = -0.05$) and finding tools useful for initiating ($r = 0.02$), were generally weak and not statistically significant (see Table 4).

5 Study 2

To broaden the applicability of our findings from Study 1, we explored the frequency and perceived usefulness of AI chatbots for schoolwork, as well as their association with EF problems, among older adolescents in high school (aged 15–19) in Study 2. Expanding on our insights garnered from schools and recent studies published subsequent to our data collection in Study 1, we incorporated additional dimensions of usefulness based on findings from German and Swedish research (The Nordic Youth Barometer, 2023; von Garrel and Mayer, 2023), as outlined in the Measures section. Additionally, we expanded our assessment by including teacher ratings of adolescents' EF alongside self-rated EF, aiming to compare results obtained from both sources. However, measures of academic achievement were not available for Study 2, as this data for high school students typically becomes accessible only at the end of the school year.

5.1 Materials and methods

5.1.1 Sample

The sample comprised 359 adolescents (239 girls, 111 boys, 9 undisclosed or not identifying as either a girl or boy; 14.8% with foreign background). They were enrolled in the first to third year at one large Swedish high school (gymnasium) with approximately 1,200 students and 5 different main programs. The participants' ages ranged from 15 to 19 years with a mean age of 17 years (SD_{age} = 0.89) representing the various main programs. All students from the selected programs were invited to participate in our survey, resulting in an 81% response rate.

The school is a communal one with relatively high entrance requirements, drawing students from across the Scania region. The region has a unique arrangement between municipalities, allowing students to freely choose their schooling from any part of the larger region, whether private or municipality schools, without tuition fees. As previously mentioned, all high school students in Sweden receive laptops at no cost.

5.1.2 Measures

5.1.2.1 The use of generative AI

The same measure employed in Study 1 to assess the use of generative AI was utilized, with additional items included to evaluate the perceived usefulness of generative AI for schoolwork. In addition to rating the perceived usefulness for initiation, structuring, and task completion, we introduced items regarding perceived usefulness for summarizing, improving texts, explaining concepts, and writing texts. These four additional items were worded similarly to the original questions from Study 1, prompting participants to indicate their level of agreement with each statement on a 5-point scale ranging from 'not correct at all' to 'exactly right.' The Cronbach's alpha for the extended measure for usability of generative AI in schoolwork was 0.74.

Furthermore, we incorporated one item linked to the use of generative AI and avoidance of effortful thinking in schoolwork, which was worded as follows: "I prefer to ask an AI tool for help rather than try on my own when I encounter difficulties in my schoolwork" (rated on a 5-point scale from 'not correct at all' to 'exactly right'). This item will be referred to as 'Avoiding effort' in the forthcoming results section, which will follow the structure of the results section for Study 1.

5.1.2.2 The BRIEF-2 self-report

The BRIEF-2 self-report form, described in more detail in Study 1, was used to measure EF in Study 2 as well. In this study, the BRIEF-2 demonstrated acceptable to excellent Cronbach's alpha values, ranging from 0.73 for self-monitoring to 0.90 for task completion.

5.1.2.3 The BRIEF-2 teacher-report

In addition to the Self-report form, we also utilized the BRIEF-2 Teacher-report form in this study, which is designed for assessing EF in children and adolescents aged 5–18. The teacher form consists of 63 items with nine scales of EF. Seven of these scales (Inhibit, Self-Monitor, Shift, Emotional Control, Working Memory, and Plan/Organize) correspond to the scales in the self-report version. Additionally, two new scales are introduced: Task-Monitor (which assesses difficulties in recognizing minor errors in work output) and Organization of Materials (evaluating the orderliness of workspaces, play areas, and storage spaces) (Gioia et al., 2015). Summing the scores from the nine scales of the teacher form comprise three indices: BRI, ERI, and CRI, as well as the overall score GEC (Gioia et al., 2015). As documented in the BRIEF-2 manual (Gioia et al., 2015), the test–retest reliability for the teacher form was 0.87, and the Cronbach's alpha values for the scales and indices were excellent, ranging from 0.81 to 0.97. In this study, the BRIEF-2 demonstrated excellent Cronbach's alpha values, ranging from 0.86 for organization of materials to 0.95 for emotional control.

5.1.3 Procedure

Data was collected at a large high school. Similar to Study 1, each student received a personalized link to a web-based survey, which they completed digitally using either personal or school-provided laptops. Participants received an information letter outlining the project's aims and their right to withdraw at any time without needing to provide reasons, along with contact information for the project leader in case they had additional questions via email, the week before data collection. Before completing the survey, students were asked to provide consent accordingly. The survey took approximately 30 min to complete. The research has been approved by the Swedish Ethics Committee (reference numbers: 2021–01666 and 2023–01013-02).

5.1.4 Data cleaning

In total, 359 students participated in the survey, and all of them responded to a question regarding their use of generative AI in schoolwork, with options ranging from "never" to "very often." Participants who responded "never" to the gate question about the use of generative AI in schoolwork were included in the analytic sample to investigate differences between users and non-users and also completed all questions regarding executive functions, allowing for a comparison of non-users with BRIEF-2 scores alongside the user group.

About 4% of participants had single missing items in the BRIEF-2. The handling of missing data adhered to the strict guidelines outlined in the BRIEF-2 manual (Gioia et al., 2015) during the data cleaning process.

5.1.5 Statistical analysis

As in Study 1, statistical analyses were conducted using SPSS 29 (IBM Corp., 2017). Chi-Square tests examined gender and background differences among AI tool users and non-users. Independent t-tests assessed gender differences in perceived AI usefulness and EFs differences. Pearson correlations explored associations between EFs, AI use and perceived usefulness.

TABLE 5 Numbers of users of different generative AI-tools for schoolwork.

Type of tool	All		Girls		Boys		χ^2	p
	N	%	N	%	N	%		
Any tool	189	100	126	100	61	100	3.17	0.574
ChatGPT	168	88.9	106	84	60	98	8.04	0.004
Bing-Chat	4	2.1	2	1.6	2	3.3		
Bard	3	1.6	1	0.7	2	3.3		
Snapchat MY-AI	61	100	51	40	9	15	12.48	<0.001
Other	7	3.7	6	4.7	1	1.6		

TABLE 6 Gender differences in perceived usefulness of generative AI in schoolwork and for avoiding effort.

	Boys		Girls		Gender differences		
	M	SD	M	SD	t	p	Cohen's d
Perceived usefulness for							
Initiation	3.32	1.37	3.13	1.36	−0.91	0.365	−0.14
Structuring	2.80	1.18	2.44	1.32	−1.85	0.066	−0.28
Completion	2.28	1.2	2.57	1.32	1.44	0.153	0.23
Summarization	3.77	1.13	3.69	1.20	−0.4	0.693	−0.06
Improving texts	2.45	1.27	2.46	1.30	0.05	0.962	0.01
Explaining concepts	3.38	1.35	3.7	1.44	1.42	0.156	0.22
Writing texts	2.37	1.27	2.21	1.29	−0.82	0.413	−0.13
Overall usefulness	20.41	4.96	20.19	5.34	−0.27	0.787	−0.04
Avoiding effort	2.02	1.03	2.21	1.08	1.13	0.259	0.18

Assumptions of variance homogeneity and normality were checked with Levene's and Shapiro–Wilk tests. Violations were addressed with Mann–Whitney U tests and Spearman's correlations. Cohen's d measured effect sizes for t-tests (0.2: small, 0.5: medium, 0.8: large).

To reduce type 1 errors, EF subscales were analyzed only if the Global Executive Composite showed significant results.

6 Results study 2 and discussion

6.1 Descriptive statistics

Out of 357 participants, 189 adolescents (52.6%) indicated that they used some form of generative AI for schoolwork, a percentage that was still lower than those found in published Swedish (The Nordic Youth Barometer, 2023) and German studies (von Garrel and Mayer, 2023). Table 5 presents which types of generative AI tools they use in their schoolwork; multiple choices were allowed. ChatGPT was the most popular generative AI for schoolwork, used by 168 (88.9%) adolescents.

A Chi-Square test was conducted to determine whether the likelihood of being a user of different generative AI tools for schoolwork differed based on gender. As presented in Table 5, we found that boys were more likely to use ChatGPT, $\chi^2_{(1)} = 8.04$, $p = 0.004$, while girls preferred Snapchat MY-AI for schoolwork, $\chi^2_{(1)} = 12.48$, $p < 0.001$.

A Chi-Square test assessed the likelihood of using generative AI tools for schoolwork based on foreign background. A trend indicated that participants with a foreign background were somewhat more likely to use generative AI, though this was not statistically significant, $\chi^2(1) = 3.78$, $p = 0.052$.

6.2 Gender differences in frequency of use and perceived usefulness of generative AI for schoolwork

Regarding the frequency of generative AI use for schoolwork among students who reported using AI, the majority used it rather seldom (50%) or sometimes (28%), while a smaller percentage used it rather often (17.4%) and often (4.2%). No gender differences were observed in these patterns, $\chi^2_{(3)} = 1.09$, $p = 0.780$.

As for perceived usefulness, Table 6 shows that no significant gender differences were found in the perceived usefulness of generative AI for schoolwork.

6.3 Differences between users and non-users of generative AI for schoolwork in self- and teacher-rated EFs

To investigate whether users of generative AI for schoolwork differed from non-users in self-rated and teacher-rated EF, two

 10.3389/frai.2024.1415782

TABLE 7 Pearson correlations between perceived usefulness of generative AI for schoolwork and frequency of use of generative AI for schoolwork and measures of EF.

		Global measure of EF problems (GEC)	
	Frequency of use	Self-rated	Teacher-rated
Perceived usefulness for			
Initiation	0.37***	0.01	0.01
Structuring	0.39***	0.10	0.02
Completion	0.38***	0.29***	0.29**
Summarizing	0.28***	−0.07	0.04
Improving texts	0.26***	0.18*	0.17*
Explaining concepts	0.19*	0.11	0.00
Writing texts	0.25***	0.19*	0.10
Overall usefulness	0.52***	0.26***	0.14
Effort avoidance	0.50***	0.21**	0.19*

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.
GEC, global executive composite.

independent t-tests were conducted. Although both self-reported ($M = 94.31$, $SD = 19.66$) and teacher-reported EF problems ($M = 71.00$, $SD = 13.85$) were somewhat higher among users compared to non-users ($M = 91.26/69.15$, $SD = 18.45/13.93$ for self/teacher ratings), no significant differences were found between the two groups in either self-reported or teacher-rated EF problems, as measured with the BRIEF-2 Global Executive Component (GEC).

6.4 Relationships between frequency of use, perceived usefulness of generative AI for schoolwork and self-and teacher-rated EFs

To explore the relationships between the frequency of use and perceived usefulness of generative AI for schoolwork and self- and teacher-rated EF problems, Pearson correlations were conducted. The frequency of use of generative AI for schoolwork showed positive and significant correlations with all aspects of perceived usefulness. Regarding EF problems, the most prominent and moderately strong correlations were observed between self- and teacher-rated EF and perceived usefulness for completing schoolwork ($r = 0.29$, $p < 0.01$ for both self and teacher ratings), followed by perceived usefulness for improving texts ($r = 0.18$ and 0.17, $p < 0.05$) and effort avoidance ($r = 0.21$ and 0.19, $p < 0.05$) (see Table 7). For informational purposes, the overall scores of self- and teacher-rated EFs were moderately positively correlated ($r = 0.39$, $p < 0.001$).

7 General discussion

In this study, our primary aim was to investigate the frequency and perceived usefulness of LLM generative AI chatbots for schoolwork, focusing specifically on their relationships with adolescents' executive functioning (EF), which encompasses cognitive processes, including planning, inhibition, and cognitive flexibility, all critical for academic success. To accomplish this aim and ensure the validity of the findings we conducted two studies involving both younger and older adolescents, aiming to comprehensively examine these relationships across different age groups. As far as we are aware, this study represents the first attempt to explore these aspects among adolescents, providing valuable insights into their use of LLM generative AI chatbots and their implications for academic performance and cognitive development.

7.1 Frequency of use of generative AI in schoolwork

In Study 1, approximately 14.8% of participants reported using generative AI, while in Study 2, this figure was 52.6%. The notable disparity in usage rates between the two studies may be attributed to several factors. Firstly, the temporal aspect likely played a role, with Study 2 conducted nearly a year later than Study 1, suggesting a potential rise in the adoption of generative AI over time. Additionally, differences in participant age between the two studies could have influenced usage rates, with older adolescents possibly demonstrating a higher inclination to utilize such technology. This is supported by findings from other studies, such as The Nordic Youth Barometer (2023) and research by von Garrel and Mayer (2023), which also indicate an upward trajectory in generative AI usage with age. Moreover, as older adolescents often encounter more complex assignments in their schoolwork, they might rely more on generative AI tools to aid them. Lastly, variations in school environments, such as differences in technological infrastructure or educational practices, might have contributed to the observed differences in usage rates.

Although our findings reveal variability in the adoption of generative AI among different adolescent populations, the usage rates observed in our studies were notably lower compared to those reported in studies involving older youth populations, such as university students in Germany (von Garrel and Mayer, 2023) and a broader age range of youths (15–24 years) in Sweden (The Nordic Youth Barometer, 2023), where usage rates reached as high as two-thirds. One potential explanation for this disparity, as mentioned above, is the age difference among participants. Additionally, variations in the formulation of our survey questions, albeit slight, might have influenced participant responses. Finally, while data collection in previously published studies likely occurred anonymously and remotely, in our study we collected data in school settings using pseudoanonymization (assigning each participant a numerical code, with their names registered on a single master list). Despite assuring participants that their information would be treated confidentially and that individual data would not be analyzed or presented, this approach may have led to underreporting of usage rates. Therefore, further research involving data collection in school settings and utilizing diverse methodologies is warranted to delve deeper into this phenomenon.

7.2 The perceived usefulness of generative AI in schoolwork and its relationship to executive functioning and academic achievement

Only in Study 1 did we observe that users of AI for schoolwork reported significantly more EF problems compared to non-users. However, in both studies, we found a consistent pattern indicating that users of generative AI in schoolwork with more EF problems perceived AI as more useful for schoolwork, particularly regarding its perceived usefulness for completing school-related assignments. This result may be linked to previous research indicating that individuals with lower work-related skill levels (e.g., difficulties in writing tasks relevant to their work) derive greater productivity improvements from using AI tools compared to their more proficient counterparts (Brynjolfsson et al., 2023; Noy and Zhang, 2023). However, our findings also raise important questions about the manner in which AI is utilized in completing school assignments. Specifically, it remains unclear whether AI is primarily used to assist with tasks already initiated or to independently complete entire assignments. This distinction is crucial as it not only informs discussions surrounding the ethical implications of AI usage in education but also highlights concerns about academic integrity and the uncritical reliance on AI-generated material. For instance, Noy and Zhang (2023) found that a majority of individuals who reported productivity gains simply copied and pasted output from ChatGPT. If AI is used uncritically, with a reliance on AI-generated material, it poses risks beyond academic integrity. Such risks include potential biases in the output (Bender et al., 2021) and the dissemination of outdated or harmful content, as evidenced by the discovery of race-based content in healthcare applications (Omiye et al., 2023). Therefore, it is imperative to address not only the ethical concerns but also the broader implications of uncritical AI usage in academic settings.

In our study, we observed that adolescents with more EF problems predominantly use ChatGPT and other generative AI tools for completing assignments rather than for initiating or structuring them, which raises significant concerns. While utilizing these tools for initiation or structuring, especially in the initial stages of structured AI usage in a school setting, could potentially enhance or augment existing abilities and foster educational attainment, their direct substitution—where these tools replace rather than enhance existing abilities—may exacerbate academic disparities in the long term. Recent research underscores adolescence as a pivotal period for EF maturation (Tervo-Clemmens et al., 2023), emphasizing the necessity of continued investigation into any tools that may interfere with or amplify the natural development of EFs during this life stage. Additionally, León-Domínguez (2024) proposes theoretical scenarios, informed by neuroscience, to explore the potential outcomes of increased generative AI usage among adolescents. A primary concern is the possibility of certain groups developing an excessive dependence on these technologies, which might encourage the evasion of challenging cognitive tasks, potentially resulting in a stagnation or deterioration of cognitive capabilities in the long term (León-Domínguez, 2024). This highlights the importance of considering the broader implications of generative AI usage among adolescents and the need for proactive measures to mitigate potential negative impacts on cognitive development and academic achievement.

7.3 Strengths and limitations

The study has several strengths, including the utilization of two relatively large samples of adolescents and a high response rate, which enhances the study's external validity by providing a diverse representation. Moreover, the inclusion of teacher ratings of EF in Study 2 adds depth and validation to the results, offering a more comprehensive understanding of the relationship between generative AI use and EF among adolescents.

However, there are also some limitations to consider. Firstly, the survey for Study 1 was conducted in spring 2023, during the early stages of introducing generative AI chatbots. Consequently, the results may not reflect current trends or usage patterns, as technology adoption and usage habits may have evolved since then and the timing of the study could be a potential confound. Furthermore, more nuanced research is needed to clarify how AI chatbots were used for different school-related tasks, such as whether they assisted with tasks already initiated by students or were used to independently complete entire assignments. Understanding these distinctions could provide deeper insights into the specific ways AI chatbots are utilized. Additionally, the relatively low number of AI users for schoolwork observed in Study 1 may have impacted the statistical power of our analyses, potentially leading to Type II errors.

Another limitation lies in the reliance solely on self-report measures for EF in Study 1. To overcome this, future research could integrate teacher and parent ratings for younger individuals and explore performance-based assessments in schoolwork with and without generative AI. This approach would offer a more comprehensive understanding of the impact of these tools on EF among adolescents.

Additionally, the absence of validated scales for evaluating the usefulness of ChatGPT and other generative AI tools in schoolwork poses a challenge. Custom questions and scales were developed for this study, but they may not fully capture the relevant factors. While the scale for the usefulness of ChatGPT demonstrated good internal consistency in both studies, further validation in future research is necessary Moreover, a more precise inquiry—such as quantifying how frequently students use AI tools for schoolwork in terms of times per week or day—might yield more insightful results compared to the more ambiguous response options like "rarely" or "often" used in the current study.

Furthermore, in Study 2, data was collected from a single, albeit large, school, which may introduce selection bias. Therefore, conducting multi-site studies would enhance the generalizability of the results. However, the classes from the school were chosen randomly, covering a broad range of specialties, and the response rate of 81% may provide a good representation of that age group of adolescents. While non-significant differences regarding foreign background were found in both studies, a clear tendency observed in Study 2 suggests that this should be studied further in a larger sample to examine the nuances of these relationships as well as their connections to EF. Given the novelty of this research field, further exploration of other potential covariates related to the relationships between AI use, perceived usefulness, and EF/academic achievement would be beneficial.

Lastly, the extensive number of analyses conducted in this study increases the risk of type 1 errors, which should be taken into account when interpreting the findings. Given the exploratory nature of the research and the limited existing literature on the topic, further validation through additional research studies is warranted.

7.4 Practical and theoretical implications

The utilization of generative AI for schoolwork appears to be relatively uncommon among Swedish youths aged 12–16, as evidenced by the findings of Study 1. However, in a slightly older population, comprising high school students, more than half reported using these tools for their academic tasks. Although the discrepancy could be partially attributed to age differences, in line with current research (The Nordic Youth Barometer, 2023; von Garrel and Mayer, 2023), the almost one-year difference between data collections is likely to be a significant confound. The results may not fully reflect current trends or usage patterns, given that technology adoption and usage habits could have evolved since the data were collected. Nevertheless, the introduction of such tools, with the potential to amplify, impact, or substitute critical abilities necessary for academic success, without thorough scientific investigation, raises significant concerns.

Educators should take note of the implications highlighted by this study, particularly regarding the potential risks associated with generative AI use among students with EF difficulties. Evidence from both Study 1 and Study 2 suggests that these individuals may be more inclined to use AI tools and may prefer utilizing them in a manner that does not enhance existing abilities. This could have negative implications for their long-term learning outcomes and the natural development of EF.

However, if thoughtfully implemented, AI chatbots might have the potential to aid students, especially those with EF deficits, by enhancing their ability to plan and manage tasks effectively. Given the link between EF deficits and academic struggles, which can lead to a vicious cycle resulting in early school dropout (Esch et al., 2014), these tools could potentially bridge the gap between students with varying levels of EF, offering opportunities for educational equity.

8 Conclusion

This study represents the first exploration of how individual characteristics, such as EF, relate to the frequency and perceived usefulness of LLM generative AI chatbots for schoolwork among adolescents. By conducting two studies involving both younger and older adolescents, we gained valuable insights into these relationships across different age groups. Our findings illuminate the usage patterns of generative AI among the studied adolescents, although it is important to note that these patterns could have evolved since the data were collected. The observed disparities between studies underscore the necessity for further investigation into the factors influencing generative AI usage rates among adolescent populations. Future studies should also evaluate whether there are preferences for different tools tied to gender, as the use of various tools may impact users in different ways. The potential gender-based differences in the likelihood of using tools like ChatGPT and Snapchat MY-AI for schoolwork could be a significant topic for future research. If these tools demonstrate varying effectiveness when used for schoolwork, differences in usage between genders could have negative implications from an equality standpoint.

Our findings reveal that adolescents with more EF problems tend to perceive generative AI tools as more useful for schoolwork, especially for completing assignments. This association prompts questions about the role of AI in education and its potential impact on academic integrity and ethical considerations.

This study emphasizes the urgency for policymakers, researchers, and educators to carefully evaluate the integration of generative AI into school environments. While our findings suggest potential risks associated with generative AI use, particularly among students with EF difficulties, these tools also may offer the possibility of aiding students in managing academic tasks more effectively. Further research and proactive measures are essential to ensure the safe and effective use of these technologies, while also considering their implications for academic equity and cognitive development among adolescents.

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 Swedish Ethics Committee (reference numbers: 2021–01666 and 2023–01013-02). The studies were conducted in accordance with the local legislation and institutional requirements. The ethics committee/institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because they deemed that passive informed consent from the participants' legal guardians/next of kin was sufficient to mitigate the risk of selection bias. However, digital informed consent was obtained from all participants.

Author contributions

JK: Conceptualization, Formal analysis, Investigation, Writing – original draft. EH: Conceptualization, Methodology, Supervision, Writing – review & editing. AL: Writing – review & editing. DD: Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study received partial funding from Hogrefe Publishing for the data collection. Importantly, the funding source had no influence on the data analysis, interpretation, or conclusions drawn from the study. This ensures the integrity and independence of the research findings.

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2024.1415782/full#supplementary-material

References

Adamopoulou, E., and Lefteris, M. (2020). Chatbots: history, technology, and applications. *Mach. Learn. Appl.* 2:100006. doi: 10.1016/j.mlwa.2020.100006

Adamopoulou, E., and Moussiades, L. (2020). An Overview of Chatbot Technology. In: I. Maglogiannis, L. Iliadis and E. Pimenidis (eds.) *Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology*, Springer, Cham. 584. doi: 10.1007/978-3-030-49186-4_31

Al Shloul, T., Mazhar, T., Abbas, Q., Iqbal, M., Ghadi, Y. Y., Shahzad, T., et al. (2024). Role of activity-based learning and ChatGPT on students' performance in education, computers and education. *Artif. Intell.* 6. doi: 10.1016/j.caeai.2024.100219

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., et al. (2021). A general language assistant as a laboratory for alignment. [Epubh ahead of print]. doi: 10.48550/arXiv.2112.00861

Bai, L., Liu, X., and Su, J. (2023). ChatGPT: the cognitive effects on learning and memory. *Brain X* 1:e30. doi: 10.1002/brx2.30

Bailey, C. E. (2007). Cognitive accuracy and intelligent executive function in the brain and in business. *Ann. N. Y. Acad. Sci.* 1118, 122–141. doi: 10.1196/annals.1412.011

Baido-Anu, D., and Ansah, 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

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, 610–23. FAccT '21. New York, NY, USA: Association for Computing Machinery.

Best, J. R., Miller, P. H., and Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative National Sample. *Learn. Individ. Differ.* 21, 327–336. doi: 10.1016/j.lindif.2011.01.007

Brynjolfsson, E., Li, D., and Raymond, L. (2023). Generative AI at work. ArXiv [econ. GN]. arXiv. Available at: http://arxiv.org/abs/2304.11771

Cabitza, F., Rasoini, R., and Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA* 318, 517–518. doi: 10.1001/jama.2017.7797

Cohen, J. (2013). Statistical power analysis for the behavioral sciences. Cambridge: Academic Press. doi: 10.4324/9780203771587

Diamond, A. (2013). Executive functions. *Annu. Rev. Psychol.* 64, 135–168. doi: 10.1146/annurev-psych-113011-143750

Diamond, A., and Lee, K. (2011). Interventions shown to aid executive function development in children 4 to 12 years old. *Science* 333, 959–964. doi: 10.1126/science.1204529

Esch, P., Bocquet, V., Pull, C., Couffignal, S., Lehnert, T., Graas, M., et al. (2014). The downward spiral of mental disorders and educational attainment: a systematic review on early school leaving. *BMC Psychiatry* 14:237. doi: 10.1186/s12888-014-0237-4

Frazier, T. W., Youngstrom, E. A., Glutting, J. J., and Watkins, M. W. (2007). ADHD and achievement: meta-analysis of the child, adolescent, and adult literatures and a concomitant study with college students. *J. Learn. Disabil.* 40, 49–65. doi: 10.1177/00222194070400010401

Gioia, G. A., Isquith, P. K., Guy, S. C., and Kenworthy, L. (2015). BRIEF2: behavior rating inventory of executive function. Lutz, FL: Second Psychological Assessment Resources.

Haleem, A., Javaid, M., Qadri, M., and Suman, R. (2022). Understanding the role of digital Technologies in Education: a review. *Sustain. Operat. Comput.* 3:4. doi: 10.1016/j.susoc.2022.05.004

Huizinga, M., Smidts, D. P., Baeyens, D., and Kan, K.-J. (2023). The Dutch version of the behavior rating inventory of executive Function-2 (BRIEF-2). *Psychol. Test Adapt. Dev.* 4, 97–115. doi: 10.1027/2698-1866/a000038

IBM Corp (2017). IBM SPSS statistics for windows, version 25.0. Armonk, NY: IBM Corp.

Jauhiainen, J. S., and Guerra, A. G. (2023). Generative AI and ChatGPT in school Children's education: evidence from a school lesson. *Sustain. Sci. Pract. Policy* 15:14025. doi: 10.3390/su151814025

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

León-Domínguez, U. (2024). Potential cognitive risks of generative transformer-based AI chatbots on higher order executive functions. *Neuropsychol.*, 38, 293–308. doi: 10.1037/neu0000948

Loh, H. W., Ooi, C. P., Barua, P. D., Palmer, E. E., Molinari, F., and Acharya, U. R. (2022). Automated detection of ADHD: current trends and future perspective. *Comput. Biol. Med.* 146:105525. doi: 10.1016/j.compbiomed.2022.105525

Maroengsit, W., Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P., and Theeramunkong, T. A.Survey on Evaluation Methods for Chatbots. (2019). A survey on evaluation methods for Chatbots. In Proceedings of the 2019 7th International Conference on Information and Education Technology, 111–119. ICIET 2019. New York, NY, USA: Association for Computing Machinery.

Miller, G. E., Chen, E., and Parker, K. J. (2011). Psychological stress in childhood and susceptibility to the chronic diseases of aging: moving toward a model of behavioral and biological mechanisms. *Psychol. Bull.* 137, 959–997. doi: 10.1037/a0024768

Naveed, H., Khan, A., Qiu, S., Saqib, M., Anwar, S., Usman, M., et al. (2023). A comprehensive overview of large language models. arXiv:2307.06435v9 [cs.CL]. doi: 10.48550/arXiv.2307.06435

Noy, S., and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 187–192. doi: 10.1126/science.adh2586

Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V. and Daneshjou, R. (2023). Large language models propagate race-based medicine. *npj Digit. Med.* 6:195. doi: 10.1038/s41746-023-00939-z

Pino Muñoz, M., and Arán Filippetti, V. (2021). Confirmatory factor analysis of the BRIEF-2 parent and teacher form: relationship to performance-based measures of executive functions and academic achievement. *Appl. Neuropsychol. Child* 10, 219–233. doi: 10.1080/21622965.2019.1660984

Samuels, W., Tournaki, N., Blackman, S., and Zilinski, C. (2016). Executive functioning predicts academic achievement in middle school: a four-year longitudinal study. *J. Educ. Res.* 109, 478–490. doi: 10.1080/00220671.2014.979913

Tervo-Clemmens, B., Calabro, F. J., Parr, A. C., Fedor, J., Foran, W., and Luna, B. (2023). A canonical trajectory of executive function maturation from adolescence to adulthood. *Nat. Commun.* 14:6922. doi: 10.1038/s41467-023-42540-8

The Nordic Youth Barometer, (2023), Back2School. Available at: https://info.ungdomsbarometern.se/publika-rapporter/back2school-2023

Torrado, J. C., Bakke, C., and Gabarron, E. (2023). Chatbots and Children with Special Educational Needs Interaction. In: M. Antona and C. Stephanidis (eds.) *Universal Access in Human-Computer Interaction. HCII 2023.* Lecture Notes in Computer Science, 14021. Cham: Springer. doi: 10.1007/978-3-031-35897-5_32

von Garrel, J., and Mayer, J. (2023). Artificial intelligence in studies—use of ChatGPT and AI-based tools among students in Germany. *Humanit. Soc. Sci. Commun.* 10, 1–9. doi: 10.1057/s41599-023-02304-7

Wardat, Y., and Alneyadi, S. (2024). Integrating ChatGPT in grade 12 quantum theory education: An exploratory study at emirate school (UAE). *Int. J. Inform. Educ. Technol.* 14:389. doi: 10.18178/ijiet.2024.14.3.2061

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., et al. (2021). Artificial intelligence: a powerful paradigm for scientific research. *Innovation* 2:100179. doi: 10.1016/j.xinn.2021.100179

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 16 September 2024
DOI 10.3389/frai.2024.1393903

Check for updates

OPEN ACCESS

EDITED BY
Ilaria Torre,
University of Genoa, Italy

REVIEWED BY
Mirela Popa,
Maastricht University, Netherlands
Ifigeneia Mavridou,
Tilburg University, Netherlands

*CORRESPONDENCE
Somayeh Fatahi
✉ somayeh.fatahi@usask.ca

RECEIVED 29 February 2024
ACCEPTED 22 August 2024
PUBLISHED 16 September 2024

CITATION
Fatahi S, Vassileva J and Roy CK (2024) Comparing emotions in ChatGPT answers and human answers to the coding questions on Stack Overflow.
Front. Artif. Intell. 7:1393903.
doi: 10.3389/frai.2024.1393903

Comparing emotions in ChatGPT answers and human answers to the coding questions on Stack Overflow

Somayeh Fatahi*, Julita Vassileva and Chanchal K. Roy

Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

Introduction: Recent advances in generative Artificial Intelligence (AI) and Natural Language Processing (NLP) have led to the development of Large Language Models (LLMs) and AI-powered chatbots like ChatGPT, which have numerous practical applications. Notably, these models assist programmers with coding queries, debugging, solution suggestions, and providing guidance on software development tasks. Despite known issues with the accuracy of ChatGPT's responses, its comprehensive and articulate language continues to attract frequent use. This indicates potential for ChatGPT to support educators and serve as a virtual tutor for students.

Methods: To explore this potential, we conducted a comprehensive analysis comparing the emotional content in responses from ChatGPT and human answers to 2000 questions sourced from Stack Overflow (SO). The emotional aspects of the answers were examined to understand how the emotional tone of AI responses compares to that of human responses.

Results: Our analysis revealed that ChatGPT's answers are generally more positive compared to human responses. In contrast, human answers often exhibit emotions such as anger and disgust. Significant differences were observed in emotional expressions between ChatGPT and human responses, particularly in the emotions of anger, disgust, and joy. Human responses displayed a broader emotional spectrum compared to ChatGPT, suggesting greater emotional variability among humans.

Discussion: The findings highlight a distinct emotional divergence between ChatGPT and human responses, with ChatGPT exhibiting a more uniformly positive tone and humans displaying a wider range of emotions. This variance underscores the need for further research into the role of emotional content in AI and human interactions, particularly in educational contexts where emotional nuances can impact learning and communication.

KEYWORDS

generative AI, large language models, natural language processing, emotion analysis, Stack Overflow, ChatGPT, programming assistance

1 Introduction

With the advancement of technology, especially in artificial intelligence (AI), we are witnessing the emergence of novel tools. Over the past decade, text-based chatbots have gained widespread popularity across diverse application domains. This surge in adoption has been described as a 'chatbot tsunami' (Grudin and Jacques, 2019), enabling human interaction with machines through natural written language. In November of 2022, OpenAI introduced

ChatGPT-3.5 (Open AI, 2023), a chatbot AI built on top of existing Large Language Models (LLMs) to facilitate interactive communication through a conversational interface. OpenAI achieved this interactive capability by employing reinforcement learning from human feedback, building upon prior work from InstructGPT (Ouyang et al., 2022). ChatGPT rapidly gained popularity and attained a milestone by amassing 100 million users by January 2023 (Saini, 2023), reaching 1.5 billion monthly visitors as of the time of writing this paper. ChatGPT can generate diverse text forms, encompassing scientific abstracts, domain-specific question answers, programming code, lifelike conversational exchanges, text summarization, language translation, and providing suggestions and recommendations. However, ChatGPT also carries potential risks, such as enabling copyright violations, plagiarism, over-dependence, and possibly reduced creativity. Also, studies indicate that people are concerned about cybersecurity threats posed by malicious entities using ChatGPT to create harmful code, hack, gather information, and trick people into revealing private or sensitive information (Okey et al., 2023; Khoury et al., 2023; Poremba, 2023; Malwarebytes, 2023). The risk of AI-generated content being passed off as human-written may lead to potential harm, such as the spread of fake content on social media. AI-generated content is riskier than human-written posts because AI can produce vast amounts of tailored misinformation quickly, making it hard to detect and flag. Its ability to personalize messages increases their persuasiveness, and the autonomous, adaptable nature of AI allows it to continuously evolve and evade detection. It could also cause significant problems in various areas, such as information security and digital forensics. In addition to its ability to provide specific answers to user questions, ChatGPT can be utilized for completing written assignments and examinations on behalf of students, raising concerns about AI-assisted cheating (Susnjak and McIntosh, 2024). In response, some schools have implemented bans on access to ChatGPT on campus (Dibble, 2023). The implications of ChatGPT in the field of education were explored in a review and the findings revealed that educators expressed concerns about the use of ChatGPT in education, fearing that students might outsource their work to ChatGPT due to its capability to rapidly generate acceptable texts (Mhlanga, 2023). Although the development of ChatGPT can be challenging, it may simplify the application of AI in teaching and learning, making it more accessible for instructors and helping students increase their knowledge in a proper way. Despite the drawbacks associated with misusing AI Chatbots like ChatGPT, there are numerous advantages in its application in education. These include personalized tutoring: ChatGPT facilitates personalized tutoring, leading to enhanced learning outcomes (Alshahrani, 2023). Automated essay grading: It streamlines the essay grading process, saving valuable time for teachers (Parker et al., 2023). ChatGPT aids in language translation, making educational materials more accessible to a broader audience (Zhao et al., 2024). Interactive learning: It promotes interactive learning, offering effective support for students (Murad et al., 2023). Adaptive learning: ChatGPT could potentially adapt teaching methods based on a student's progress and performance, especially in the context of learning programming languages.

Given the significant influence of emotions on perception, the learning process (Tyng et al., 2017), and communication in humans, there is substantial evidence indicating that learning is intertwined with emotions (Hökkä et al., 2020; Bohn-Gettler and Kaakinen, 2022; Um et al., 2012). Consequently, if ChatGPT interacts with humans in a way that aligns with the positive aspects such as optimism and positive emotions, it could prove more beneficial in the field of education by effectively assuming the role of a teacher, as previous research has shown the impact of AI on learning (Wang et al., 2024; Kasneci et al., 2023). However, there is still a lack of research in education comparing the emotional content in human and AI responses, which warrants further investigation for a comprehensive understanding.

The current trend in chatbot development is toward empathetic and emotionally intelligent bots, capable of recognizing user feelings and generating fitting answers (Adamopoulou and Moussiades, 2020). However, to what extent can AI chatbots understand human emotions and respond with a human level of empathy, and to what extent can they mimic believable emotional responses to a situation? Despite notable advancements in chatbot development (Adam et al., 2021; Rapp et al., 2021; Adamopoulou and Moussiades, 2020), accurately capturing and expressing the right emotions within chatbot interactions remains a persistent challenge. ChatGPT has shifted researchers' perspectives to some extent. For instance, one study (Elyoseph et al., 2023) showcased ChatGPT's capacity to produce suitable Emotional Awareness (EA) responses, with the potential for significant improvement over time.

However, there is a gap in research when it comes to comparing the emotional aspects of human and ChatGPT responses in different areas. Our study focuses on understanding these differences in how questions related to programming language are answered. Learning programming is tough for many in education today, and ChatGPT can assist students and programmers with problem-solving. That is why we have chosen this field to look into the emotional differences in responses between humans and ChatGPT. The findings from this research may offer useful insights that could inform future developments of ChatGPT, enhancing its utility for education and learning.

To accomplish this goal, we address the following research questions:

Research Questions:

RQ1: What is the distribution of emotions in questions, human-generated answers and ChatGPT answers?

RQ2: What are the dominant emotions exhibited in ChatGPT answers, and what are the dominant emotions in human-generated answers?

RQ3: How does the range of expressed emotions differ between human generated and ChatGPT answers?

To answer our research questions, we conducted a comprehensive analysis, comparing emotional aspects in responses from both ChatGPT and humans to 2000 questions sourced from Stack Overflow (SO).

The subsequent sections of this paper are structured as follows: Section 2 delves into the literature review of our study, providing an overview of recent research related to ChatGPT. Section 3 outlines our methodology. The results are presented in Section 4, followed by the discussion in Section 5 and the conclusion in Section 6.

2 Literature review

Since November 2022, ChatGPT has attracted significant attention, resulting in numerous applications and extensive research

(Kalla et al., 2023). This section aims to provide a clearer introduction to the applications of ChatGPT within the context of software engineering. Specifically, we explore its utility within the Stack Overflow dataset, focusing on studies that examine ChatGPT's role and effectiveness in this domain.

Jalil et al. (2023), the researchers assess the performance of ChatGPT in addressing common queries from a used software testing curriculum. Their investigation shows that ChatGPT's current capabilities effectively address 77.5% of the questions examined. Among these questions, it provides fully or partially accurate answers in 55.6% and correct or partially correct explanations in 53% of the cases. The findings of this research diverge from the outcomes of implementing ChatGPT in other fields, such as medicine (Kung et al., 2023) or law (Choi et al., 2021), where ChatGPT has demonstrated success in passing specific portions of their examinations. This discrepancy suggests that although ChatGPT exhibits capability, it also possesses limitations, including a lack of comprehensive knowledge and the tendency to make incorrect assumptions, contributing to potential response inaccuracies.

Surameery and Shakor (2023), researchers investigated the role of ChatGPT in solving programming bugs. They found that ChatGPT is superior to other tools in cost, speed, customization, ease of use, and scalability. However, when it comes to fitting into existing systems, traditional debugging tools are more effective due to their integration capabilities. Additionally, the accuracy of ChatGPT depends on the quality of its training data, while traditional debugging tools generally provide higher accuracy levels.

An empirical investigation (Nascimento et al., 2023) compared the performance of software engineers and AI systems, such as ChatGPT, across various evaluation metrics. ChatGPT-generated code was evaluated against code created by developers and submitted on Leetcode. The study revealed that automated systems like ChatGPT can, in certain instances, surpass the performance of novice software engineers in specific tasks. This superiority was particularly apparent in resolving easy and medium-level problems, where ChatGPT consistently outperformed novice contest programmers.

Kabir et al. (2023), a comprehensive analysis was conducted on ChatGPT's answers to 517 questions sourced from Stack Overflow (SO). The assessment encompassed the correctness, consistency, comprehensiveness, and conciseness of these answers. The manual analysis indicated that 52% of the answers provided by ChatGPT contained inaccuracies, while 77% were found to be excessively verbose. Nevertheless, users preferred ChatGPT's answers 39.34% of the time due to their thoroughness and articulate language style. The results of the linguistic analysis demonstrated the formal nature of ChatGPT's answers, which rarely expressed negative sentiments. Although the user study showed that users had a higher preference and quality rating for SO, they occasionally erred by favoring incorrect ChatGPT answers due to the model's well-articulated language style and seemingly plausible logic presented with positive assertions. These findings highlight the requirement for meticulous error correction within ChatGPT while also emphasizing the need to make users aware of the potential risks associated with answers that appear accurate.

Liu et al. (2023) conducted a study to examine the comparative efficacy of ChatGPT and SO in assisting programmers. Two groups of students with similar programming abilities were instructed to use the two platforms to solve three programming tasks: algorithmic challenges, library usage, and debugging. The findings reveal that, in terms of code quality, ChatGPT exhibits significantly better performance than SO when aiding in the completion of algorithmic and library-related tasks. However, Stack Overflow proves more beneficial for debugging tasks. Concerning task completion speed, the ChatGPT group demonstrates notably faster results than the SO group, specifically in algorithmic challenges, while displaying similar performance in the other two tasks.

Delile et al. (2023) explore the privacy issues encountered by developers. They compare the responses accepted on SO with those generated by ChatGPT for these queries to evaluate if ChatGPT could be a helpful alternative. The results reveal that most privacy-related questions center on choice/consent, aggregation, and identification. Additionally, ChatGPT provides roughly 56% of responses that match the accuracy level of SO.

Following Stack Overflow's decision to ban ChatGPT, Borwankar et al. (2023) examined how the users of SO responded to this change. They studied the quality of content using natural language processing (NLP) techniques and voting patterns across SO and the AskProgramming subreddit on Reddit. The results indicate that SO users adjusted their answer style after the limitation, leading to more positive, longer responses than AskProgramming subreddit users. This study shows that there has been an improvement in content quality post-limitation, reflected in increased upvotes for answers.

The research discussed in this section sheds light on various aspects of ChatGPT's capabilities and limitations in software engineering. A prevailing consensus suggests that the ability to differentiate between ChatGPT and human-generated text is crucial. As mentioned, emotion is a distinguishing factor for identifying the human and ChatGPT (Pamungkas, 2019). However, it is noteworthy that relatively few studies have investigated emotion in communication with ChatGPT. Elyoseph et al. (2023) focused on assessing the emotional awareness (EA) capabilities of ChatGPT. Using the Levels of Emotional Awareness Scale (LEAS), researchers conducted two examinations involving 20 scenarios to evaluate ChatGPT's EA performance, comparing it to norms established by a previous study. ChatGPT's emotional awareness (EA) scores were compared to those in a previous study (Nandrino et al., 2013). To check how accurate ChatGPT's answers were, two professional psychologists independently rated each answer based on how well it fits the situation. They used a scale from 0, meaning "the feelings described do not match the scenario at all," to 10, meaning "the emotions described fit the scenario perfectly." The results illustrated that ChatGPT can produce suitable emotional awareness (EA) answers, potentially enhancing performance over time.

To the best of our knowledge, no existing research compares emotions in answers generated by ChatGPT with those created by humans in the software engineering domain. Additionally, there has yet to be a study focused on using emotion as a fingerprint to differentiate between human and ChatGPT-generated answers. This paper addresses these gaps and delves into this unexplored research area.

3 Methodology

This study uses a comparative analysis approach to assess the answers generated by ChatGPT in contrast to those provided by humans on Stack Overflow.

TABLE 1 A sample of Stack Overflow questions with answers from human and ChatGPT.

<table>
<tr><td rowspan="2">Question</td><td rowspan="2">I need to grab the base64-encoded representation of the ViewState. Obviously, this would not be available until fairly late in the request lifecycle, which is OK. For example, if the output of the page includes: <input type="hidden"name="__VIEWSTATE" id="__VIEWSTATE" value="/wEPDwUJODU0Njc5MD...==" />
I need a way on the server side to get the value "/wEPDwUJODU0Njc5MD...=="
To clarify, I need this value when the page is being rendered, not on PostBack. e.g. I need to know the ViewState value that is being sent to the client, not the ViewState I'm getting back from them.</td><td>Human answer</td><td>Rex, I suspect a good place to start looking is solutions that compress the ViewState—grabbing ViewState on the server before it is sent down to the client and gzipping it. That is exactly where you want to be.
Scott Hanselman on ViewState Compression (2005) ViewState Compression with System.IO.Compression (2007)</td></tr>
<tr><td>ChatGPT answer</td><td>To get the base64-encoded representation of the ViewState on the server side, you can use the following code in your ASP.NET page:
"csharp
string viewStateValue = Page.ClientScript.GetHiddenField('__VIEWSTATE').Value
This code uses the 'ClientScript' object to access the hidden '__VIEWSTATE' field and retrieve its value".
Note that this code should be executed after the ViewState has been populated, which typically happens during the Page's "PreRender" event. You can add the code in the appropriate event handler of your page, such as the "Page_Load" or "Page_PreRender" event.
Keep in mind that the ViewState value may not be available until the page has been fully rendered, so you might need to experiment with different events to find the right timing for retrieving it</td></tr>
</table>

3.1 Data

We chose the subject matter, Software Engineering, because it is less inclined toward emotional interactions, as it is primarily about resolving programming problems. This makes it a more 'neutral' domain for analysis. For instance, datasets containing reviews of products, hotels, or restaurants are likely to feature numerous strongly opinionated positive and negative reviews. This could skew the comparison, placing a heavier emphasis on human emotions.

As a source of data, we chose Stack Overflow because it is a popular free question-and-answer community, extensively used for many years by programmers and software engineers (Spolsky and Atwood, 2008), and the data is readily accessible and abundant. We used the open dataset[1] provided in Kabir et al., 2023, comprising answers from ChatGPT and human experts to a randomly selected set of 2000 Stack Overflow questions. As an example, one of the data points is presented in Table 1.

3.2 Model

To extract emotions from the questions, ChatGPT answers, and human answers, we utilized an emotion-multilabel model from Hugging Face. This model is a fine-tuned version of cardiffnlp/twitter-roberta-base-2022-154 m (Mohammad et al., 2018) and is based on EmoBERT, a novel emotion-based variant of the BERT (Devlin et al., 2018) transformer model. Multi-label EmoBERT comprises three main components. The first component involves two encoders: one for all tokens (Word Encoder) and another designed explicitly for emotion-word hashtag tokens (Hashtag Encoder). The second component is a compositional layer that represents sentiment semantic composition. The final component is a label correlation layer that learns the correlation between co-existing emotions.

The emotions were selected for this model based on the basic emotion model (Ekman, 1992; Plutchik, 1980; Parrott, 2001; Frijda, 2017) and the valence–arousal–dominance (VAD) model (Russell, 1980). Finally, researchers considered an emotion classification task encompassing 11 emotions commonly expressed in tweets: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust for this model.

Its performance demonstrated close approximation to published results in extracting emotions from the Stack Overflow dataset, achieving a Micro-F1 score of 83.36 (Li and Xiao, 2023). The schematic diagram of our study is shown in Figure 1.

4 Results

When the Multi-label EmoBERT model is applied to the dataset, it produces a vector of emotions as an output (Table 2). This vector comprehensively represents the emotional content embedded within the questions and the sets of humans and ChatGPT answers. In the initial phase, we used the Chi-squared test to examine the relationship between the emotion in the question and the emotional response from both humans and ChatGPT. As depicted in Table 3, our findings indicate a notable correlation between the emotions conveyed in questions and the emotional responses from both human participants and ChatGPT. It should be mentioned that we opted to use the Chi-squared test instead of the Pearson correlation for the following reasons. The Chi-squared test is particularly suitable for categorical data, in contrast, the Pearson correlation coefficient is designed to measure the linear relationship between two continuous variables. Our data consisted of categorical variables in this step, making the Chi-squared test the appropriate choice. Also, the Chi-squared test is used to determine whether there is a significant association between two categorical variables. Our objective was to assess the independence or association between these variables. The Pearson correlation assumes a linear relationship between variables and can provide misleading results when applied to non-linear or non-continuous data. The Chi-squared test does not require any assumptions about the

1 https://github.com/SamiaKabir/ChatGPT-Answers-to-SO-questions

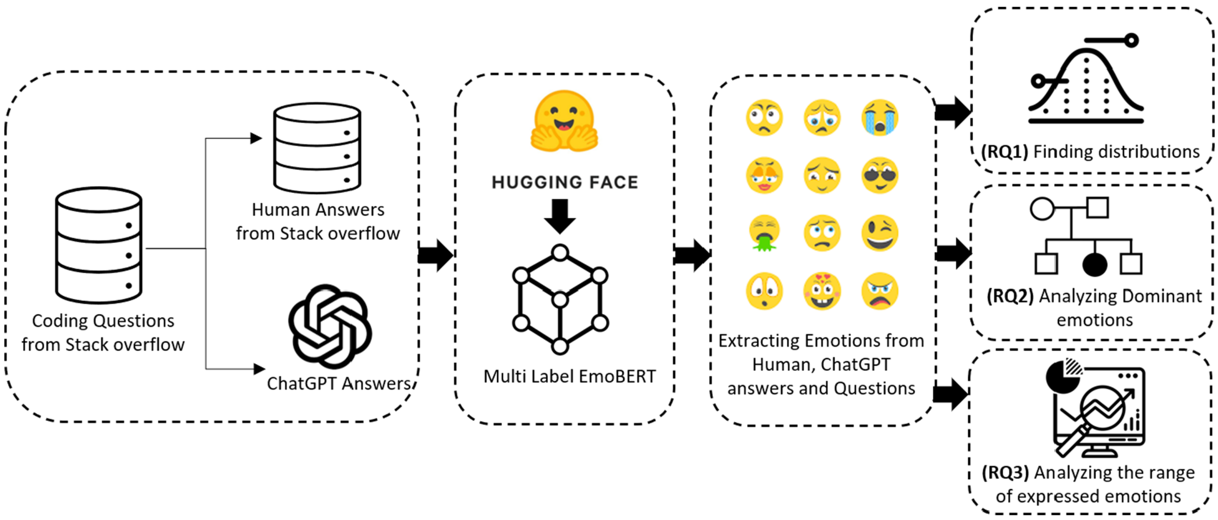


FIGURE 1
Schematic diagram of our study.

TABLE 2 Values of emotions for a sample of Stack Overflow questions and answers.

Emotions	Anger	Anticipation	Disgust	Fear	Joy	Love	Optimism	Pessimism	Sadness	Surprise	Trust
SO Question	0.03	0.72	0.04	0.01	0.05	0	0.06	0.02	0.02	0.04	0.02
Human Answer	0.11	0.45	0.12	0.01	0.13	0.01	0.05	0.02	0.04	0.06	0.02
ChatGPT Answer	0.02	0.49	0.02	0	0.24	0.01	0.15	0.01	0.01	0.02	0.03

TABLE 3 Relationship between emotions in Stack Overflow questions and answers from human and ChatGPT.

Emotion in Stack Overflow and Human answer	X^2 (df = 10, N = 2000) = 123.74, p < 0.001
Emotion in Stack Overflow and ChatGPT answer	X^2 (df = 10, N = 2000) = 311.32, p < 0.001

nature of the relationship between the variables other than their categorical nature, making it more flexible and appropriate for our analysis. In the next step, we calculate the average value of emotions in SO questions, human answers, and ChatGPT answers. As Figure 2 shows, questions frequently exhibit emotions such as anticipation, anger, and disgust. There is a notable similarity in emotional patterns between humans and ChatGPT.

Upon a detailed comparison of human and ChatGPT answers, shown in Figure 2, it becomes evident that ChatGPT answers tend to be more optimistic and joyful, while human answers often contain more expressions of anger and disgust.

In the next step, our focus shifts to identifying the predominant emotion in both the questions and their corresponding answers. We determine the maximum emotion intensity for each answer rather than considering a range of different emotions. As shown in Table 2, a spectrum of emotions is present, but Anticipation emerges as the dominant emotion for both the question and the answers from humans and ChatGPT. Now we focus our attention on the dominant emotions, excluding those detected marginally, e.g., Love, Pessimism, and Trust (see Figure 3). Notably, when we focus solely on the dominant emotion for each question/answer and calculate the mean emotion value, it differs from the previous calculation.

To understand the data distribution, we examined how the data is dispersed using a box plot (Figure 4). The results show that for some emotions, such as Anger, Disgust, and Joy, there are considerable differences between ChatGPT and human responses. Notably, human emotions exhibited a wide dispersion. Conversely, for emotions like Anticipation, Optimism, and Sadness, the distributions appeared consistent between human and ChatGPT-generated answers.

We performed a t-test to compare the data represented in the boxplots shown in Figure 4. For the "Anger" variable, the results indicate that the t-statistic is 6.27 and the p-value is 1.28e-6. These results suggest a highly significant difference between the groups being compared. The large t-statistic indicates a substantial difference in the means of the two groups, while the extremely small p-value indicates that this difference is statistically significant, far below the conventional significance level of 0.05. This provides strong evidence that the observed differences in "Anger" levels between the groups are not due to random chance.

For "Disgust," the results show that the t-statistic is 4.36 and the p-value is 1.89e-5. Since the p-value (0.0000189) is far below the common significance level of 0.05, we can reject the null hypothesis. This means there is strong evidence to suggest a significant difference between the means of the two groups.

For "Joy," the p-value is 4.41e-7 (0.00000044), which is much smaller than the conventional significance level of 0.05. This indicates that there is a statistically significant difference between the two groups.

10.3389/frai.2024.1393903

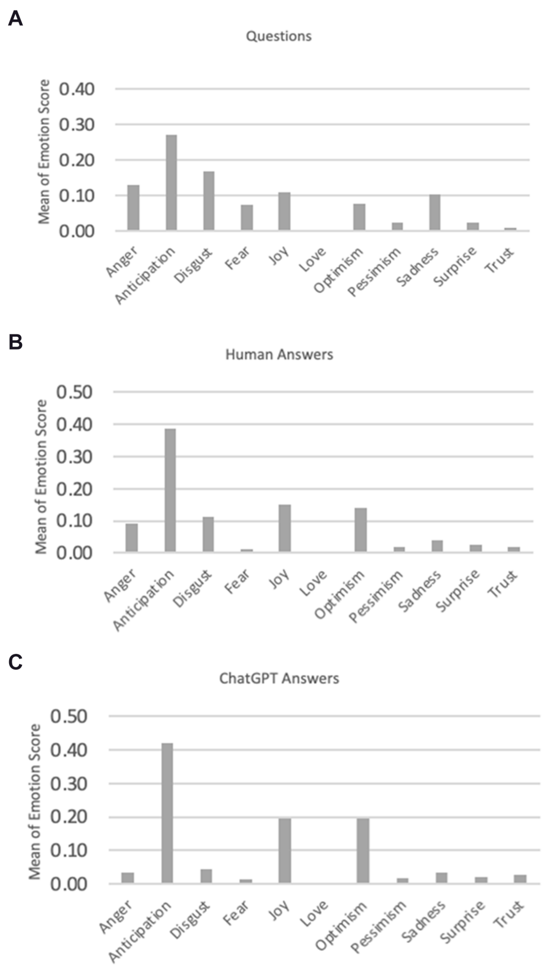


FIGURE 2
Means of emotion scores for SO questions, human answers, and ChatGPT answers.

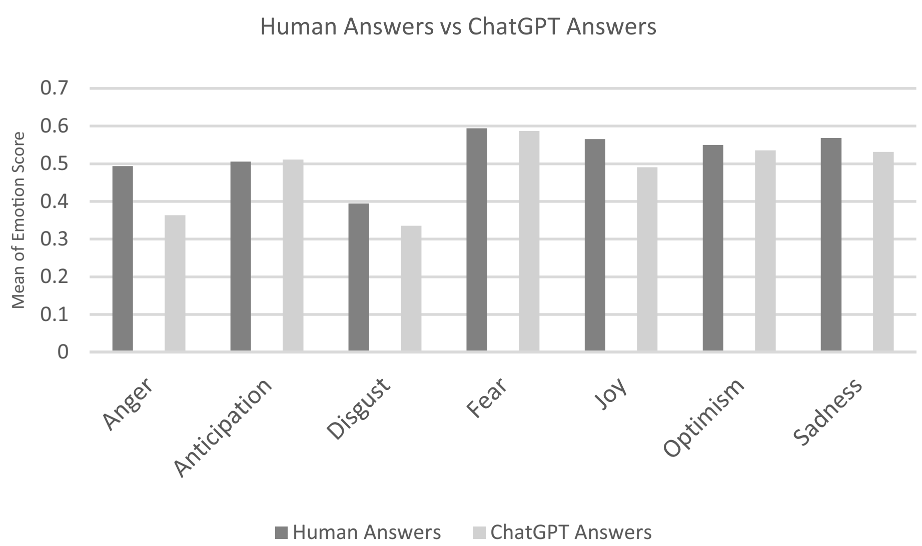


FIGURE 3
Comparison of dominant emotions in human answers and ChatGPT answers.

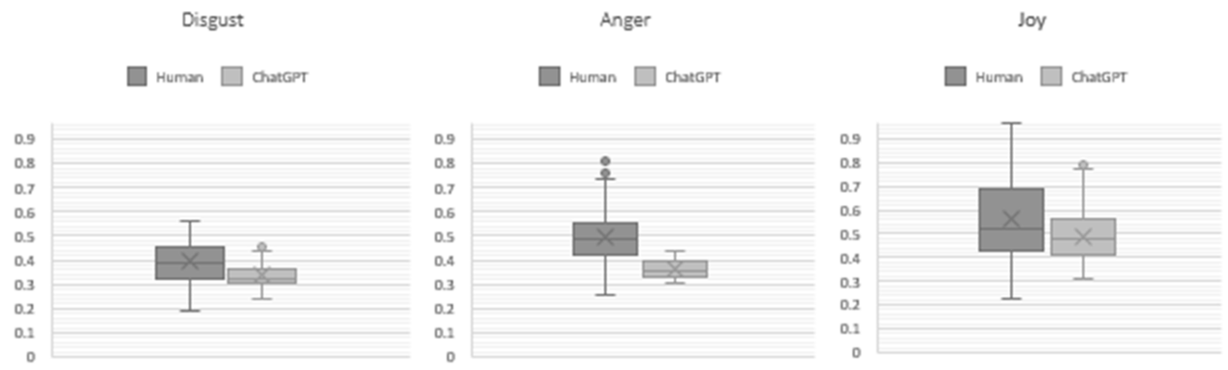


FIGURE 4
Box plot comparison of emotional responses between ChatGPT and human.

TABLE 4 Topic modeling results.

Topic Label	Top 10 words in the topic
Text Processing and Data Structures	emacs, dictionary, date, printf, echo, hello, struct, sizeof, iphone, margin
File and Project Management	file, path, files, server, folder, directory, project, command, version, windows
Application Development and Performance	code, performance, like, tools, application, http, provides, data, specific, good
Database Operations and User Management	table, data, query, code, thread, database, user, session, server, page
Visual Design and Libraries	color, myclass, bean, boost, colors, datetime, signal, iterator, serialization, vector
Object-Oriented Programming and Methods	string, class, function, value, public, code, method, return, event, object

In the next step, we applied topic modeling (Alghamdi and Alfalqi, 2015) to uncover the relationships between topics and emotions (Table 4). Specifically, we utilized Latent Dirichlet Allocation (LDA; Blei et al., 2003), a probabilistic generative model widely employed in natural language processing (NLP) and machine learning. LDA is a technique adept at identifying underlying topics within a set of documents. In our analysis, we employed the LDA model on the questions, human answers, and ChatGPT-generated answers to extract topics.

When examining Figure 5, it becomes evident that each topic is associated with a predominant emotion. For instance, in the case of Text Processing and Data Structures topic, Disgust holds a prominent proportion compared to other emotions in the human answers.

We conducted a statistical analysis, and the results are shown in Table 5. The results confirm that in each topic, the Chi-square test results show significant differences in the proportions of emotions between ChatGPT and human responses. The very low p-values across all topics confirm that the observed differences are statistically significant, indicating real and substantial differences in how emotions are represented by ChatGPT versus humans in each topic (Tables 6, 7).

5 Discussion

We conducted a comprehensive analysis comparing emotional aspects in answers from ChatGPT and humans to 2000 questions sourced from Stack Overflow. Our findings indicate notable differences in emotional expression between humans and ChatGPT across various topics. Humans tend to have more negative responses with higher variance compared to ChatGPT. ChatGPT's responses tend to lean toward optimism, whereas humans are more inclined toward expressing anger and disgust. This difference may be one of reasons people prefer ChatGPT's answers 40% of the time due to their thoroughness and articulate language style (Liu et al., 2023).

When humans express an emotion, the variance is larger than in ChatGPT. It appears that ChatGPT provides responses based on patterns in training data and aims to be helpful. Human responders on Stack Overflow are real individuals with their own feelings, thoughts, and experiences.

Upon investigating the results comparing emotions in different topics, it is evident that humans tend to exhibit consistent emotional responses, encompassing feelings of Disgust or Anger. In contrast, ChatGPT demonstrates a discrepancy in emotional expression across various topics, expressing a range of emotions including Optimism, Sadness, Joy, Fear, and Anticipation. This suggests that human responses may be more authentic and natural, stemming from the inherent frustration of the searching process.

In exploring **RQ1**, the initial analysis shows that ChatGPT's answers tend to be more optimistic and joyful, while human answers often contain more expressions of anger and disgust.

In addressing **RQ2**, when we focus on the predominant emotion and determine the maximum emotion intensity for each answer, Anticipation emerges as the dominant emotion for both the questions and the answers from humans and ChatGPT. Additionally, the results show considerable differences between ChatGPT and human responses for some emotions, such as Anger, Disgust, and Joy. Human emotions exhibit higher variance compared to ChatGPT.

Regarding **RQ3**, investigating emotions across different subjects reveals substantial differences in how emotions are represented by ChatGPT versus humans in each topic. Specifically, our analysis highlights that humans tend to exhibit a narrower emotional range within each topic, predominantly showing negative emotions like Disgust and Anger. This consistency in human emotional expression

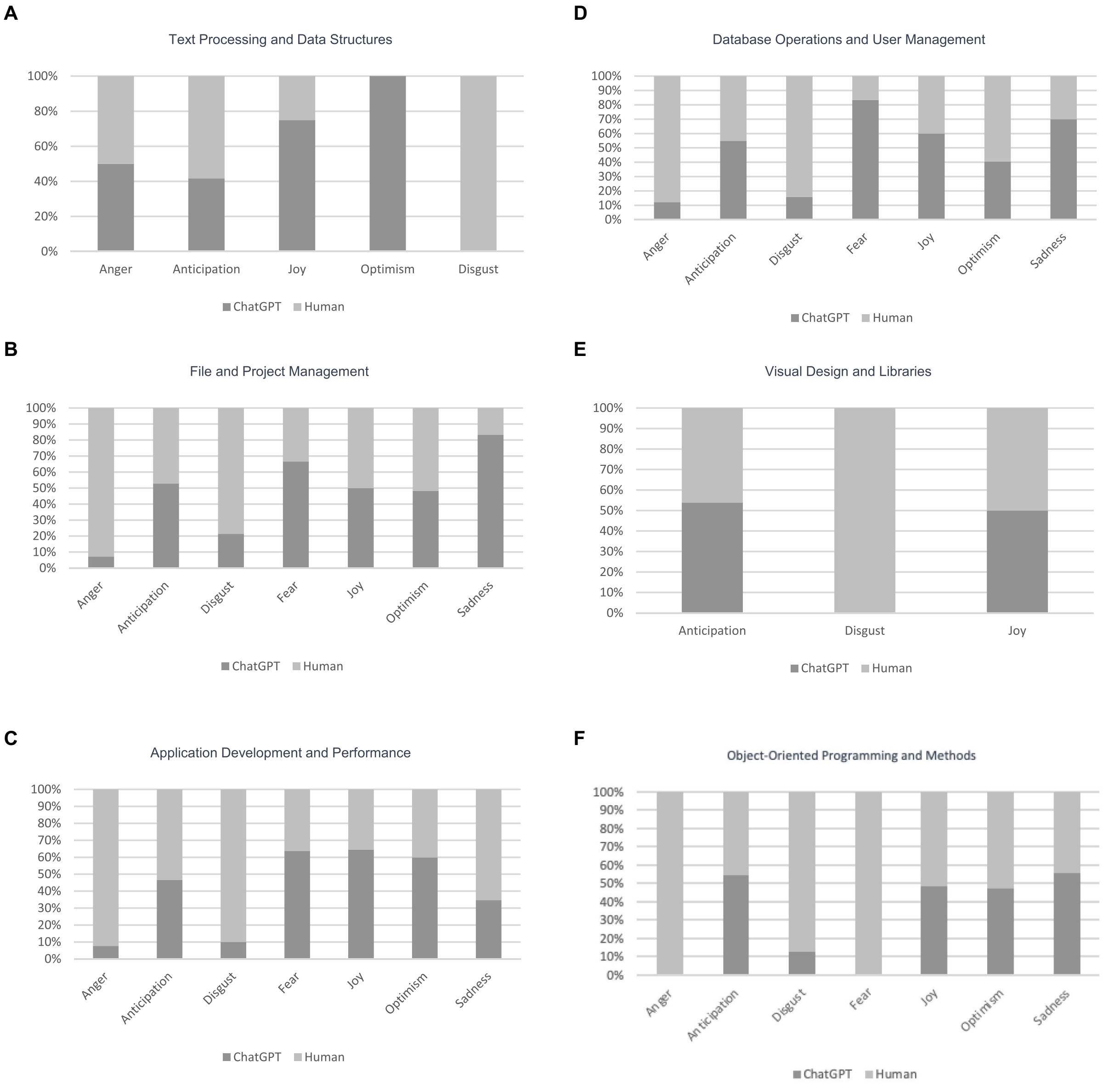


FIGURE 5
Comparison between dominant emotions of humans and ChatGPT answers for different topics.

TABLE 5 Chi-square test results for emotional proportions in ChatGPT vs. human responses across topics.

Topic	Chi
Text Processing and Data Structures	X^2 (df = 6) = 324.70, $p < 0.001$
File and Project Management	X^2 (df = 6) = 161.30, $p < 0.001$
Application Development and Performance	X^2 (df = 6) = 146.39, $p < 0.001$
Database Operations and User Management	X^2 (df = 6) = 173.86, $p < 0.001$
Visual Design and Libraries	X^2 (df = 6) = 78.62, $p < 0.001$
Object Oriented Programming and Methods	X^2 (df = 6) = 184.33, $p < 0.001$

across topics could be attributed to the personal and situational frustrations users encounter when seeking help on technical issues.

Understanding these differences in emotional expression is crucial for improving AI systems like ChatGPT and holds importance from three perspectives.

Detection of AI-generated Text: The emotional differences can serve as unique fingerprints, enabling the detection of text generated by ChatGPT when compared to human-generated content.

Application in Supportive Roles: Generative AI and LLM-based chatbots may be particularly suited for roles where being a patient,

TABLE 6 Examples of ChatGPT answers.

Answer	Extracted emotion
Exposing the `eval()` function in JavaScript to execute dynamically generated JavaScript can indeed pose a security risk. This is because any JavaScript code passed to the `doScript()` function will be executed without any validation or sandboxing. By allowing the execution of raw JavaScript code, you are essentially opening up your application to potential script injection attacks, where an attacker could execute harmful or unauthorized actions on the client-side. To mitigate this risk, it is recommended to carefully validate and sanitize any user-generated JavaScript code before executing it. One approach is to leverage server-side validation and only allow specific, pre-defined operations to be executed. Alternatively, you could use a more secure method such as using ExternalInterface to call pre-defined, specific JavaScript functions that perform the necessary actions. This way, you can have more control over the execution and ensure that only trusted operations are performed. In summary, executing raw JavaScript code from Flex can be a security risk if not properly validated and sanitized. It is important to carefully consider the potential risks and implement appropriate security measures to protect your application and users.	Fear
Unfortunately, I could not find any free Vi add-ins specifically for MS Visual Studio 2005. The Google Summer of Code project you mentioned may have been promising, but it's unclear if it has been completed or if the add-in is available. ViEmu is a popular commercial option for using Vi/Vim keybindings in Visual Studio. However, since you mentioned that you cannot test it at work and do not have a Windows computer at home, it may not be a viable solution for you. In situations like this, where you have specific requirements or restrictions, it's always best to consult with your IT department or adhere to the company's policy on software installation. They may provide alternative solutions or guidance on how to proceed.	Sadness
While Adhearsion is a popular choice for Ruby developers working with Asterisk, there are alternatives available for developing AGI-based IVR applications. One such alternative is the AsteriskRuby framework, which is well-documented and supported by the community. However, if you are open to considering non-Ruby frameworks, some other options include: 1. FastAGI: FastAGI is an interface for writing AGI scripts in any programming language. It provides a simple socket-based protocol to communicate with Asterisk. This allows you to choose the programming language you are most comfortable with. 2. PHPAGI: PHPAGI is a PHP framework specifically designed for building AGI scripts. It provides a high-level object-oriented API for working with the various features of Asterisk. 3. Java AGI: Java AGI is a Java library for building AGI scripts. It provides a programming interface that abstracts the complexities of the AGI protocol, making it easier to develop robust IVR applications in Java. Ultimately, the choice of framework depends on your specific requirements and the programming language you are most comfortable with. It is recommended to evaluate each option, considering factors like documentation, community support, and the ease of integration with your existing infrastructure.	Optimism

TABLE 7 Examples of human answers.

Answer	Extracted emotion
SipX is really the wrong answer. I've written some extremely complicated VoiceXML on SipX 3.10.2 and it's been all for naught since SipX 4 is dropping SipXVXML for an interface that requires IVRs to be compiled JARs. Top that off with Nortel filing bankruptcy, extremely poor documentation on the open-source version, poor compliance with VXML 2.0 (as of 3.10.2) and SIP standards (as of 3.10.2, does not trunk well with ITSPs). I will applaud it for a bangup job doing what it was designed to do, be a PBX. But as an IVR, if I had it to do all over again, I'd do something different. I do not know what for sure, but something different. I'm toying with Trixbox CE now and working on tying it into JVoiceXML or VoiceGlue. Also, do not read that SipX wiki crap. It compares SipX 3.10 to AsteriskNOW 1 to Trixbox 1. Come on. It's like comparing Mac OS X to Win95! A more realistic comparison would be SipX 4 (due out 1Q 2009) to Asterisk 1.6 and Trixbox 2.6, which would show that they accomplish near identical results except in the arena of scalibility and high availability; SipX wins at that. But, for maturity and stability, I'd advocate Asterisk. Also, my real-world performance results with SipXVXML: Dell PowerEdge R200, Xeon Dual Core 3.2GHz, handles 17 calls before jitters. HP DL380 G4, Dual Xeon HT 3.2 GHz, handles 30 calles before long pauses. I'll post my findings when I finish evaluating VoiceGlue and JVoiceXML but I think I'm going to end up writing a custom PHP called from AGI since all the tools are native to Asterisk.	Anger
ViEmu works great with Visual Studio. I used Vi(m) strictly in Linux, but I was turned on to bring the Vi(m) editing process into the Windows world by JP Boodhoo. JP praises about it also.	Joy

optimistic, and joyful partner in dialog is beneficial, such as in education, online help, and customer service applications.

Enhancing AI Emotional Diversity: For developing more believable and empathetic AI chatbots, it is important to enhance their emotional diversity and variance. This can lead to more realistic emotional responses, making AI interactions more engaging and relatable. Future research could explore personalized and emotionally adaptive AI chatbots that reflect and respond to the user's emotional tone. Such chatbots could be invaluable in areas like mental health counseling, child, and elderly care.

6 Conclusion

In this research, we conducted a comprehensive analysis comparing the emotional aspects of answers from ChatGPT and humans to 2,000 questions sourced from Stack Overflow. Our findings indicate notable differences in emotional expression between humans and ChatGPT across various topics. Humans tend to have more negative responses with higher variance compared to ChatGPT. ChatGPT's responses tend to lean toward optimism, whereas humans are more inclined toward expressing anger and disgust.

Theses analysis highlights distinct emotional patterns across different topics. These insights underscore the need for improving AI systems to enhance their believability and user engagement, particularly in roles requiring supportive and patient interaction.

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

SF: Writing – original draft, Writing – review & editing. JV: Writing – review & editing. CR: 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 by the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant Program, grant RGPIN-2021-03521, awarded to Dr. Julita Vassileva; the Living Skies Postdoctoral Fellowship, awarded to Dr. Fatahi; and the NSERC CREATE in Software Analytics program, with fund number 121277, awarded to Dr. Chanchal Roy.

Acknowledgments

We would like to acknowledge the assistance of ChatGPT, version 3.5, for its support in checking the English grammar and enhancing the clarity of this paper. Its valuable contribution helped ensure the precision and quality of the written content.

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

Adam, M., Wessel, M., and Benlian, A. (2021). AI-based chatbots in customer service and their effects on user compliance. *Electron. Mark.* 31, 427–445. doi: 10.1007/s12525-020-00414-7

Adamopoulou, E., and Moussiades, L., (2020). "An overview of chatbot technology." In *IFIP international conference on artificial intelligence applications and innovations* (pp. 373–383). Springer, Cham.

Alghamdi, R., and Alfalqi, K. (2015). A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)* 6, 147–153. doi: 10.14569/IJACSA.2015.060121

Alshahrani, A. (2023). The impact of ChatGPT on blended learning: current trends and future research directions. *Int J Data and Network Sci* 7, 2029–2040. doi: 10.5267/j.ijdns.2023.6.010

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022.

Bohn-Gettler, C. M., and Kaakinen, J. K. (2022). Introduction to the special issue on emotions in reading, learning, and communication. *Discourse Process.* 59, 1–12.

Borwankar, S., Khern-am-nuai, W., and Kannan, K.N., (2023). Unraveling the impact: An empirical investigation of ChatGPT's exclusion from stack overflow.

Choi, J. H., Hickman, K. E., Monahan, A. B., and Schwarcz, D. (2021). ChatGPT goes to law school. *J. Leg. Educ.* 71:387.

Delile, Z., Radel, S., Godinez, J., Engstrom, G., Brucker, T., Young, K., et al. (2023). "Evaluating privacy questions from stack overflow: can ChatGPT compete?." In *2023 IEEE 31st international requirements engineering conference workshops (REW)* (pp. 239–244). IEEE.

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.

Dibble, M. (2023). Schools ban ChatGPT amid fears of artificial intelligence-assisted cheating. Washington, DC: VOA News.

Ekman, P. (1992). An argument for basic emotions. *Cognit. Emot.* 6, 169–200.

Elyoseph, Z., Hadar-Shoval, D., Asraf, K., and Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. *Front. Psychol.* 14:1199058.

Frijda, N. H. (2017). The laws of emotion. New York, NY: Psychology Press.

Grudin, J., and Jacques, R., (2019). "Chatbots, humbots, and the quest for artificial general intelligence." In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–11).

Hökkä, P., Vähäsantanen, K., and Paloniemi, S. (2020). Emotions in learning at work: a literature review. *Vocat. Learn.* 13, 1–25. doi: 10.1007/s12186-019-09226-z

Jalil, S., Rafi, S., LaToza, T. D., Moran, K., and Lam, W. (2023). "Chatgpt and software testing education: promises & perils." In *2023 IEEE international conference on software testing, verification and validation workshops (ICSTW)* (pp. 4130–4137). IEEE.

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 arXiv* 2308:02312.

Kalla, D., Smith, N., Samaah, F., and Kuraku, S. (2023). Study and analysis of chat GPT and its impact on different fields of study. *Int J Innovative Sci Res Technol* 8, 1–12.

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.

Khoury, R., Avila, A.R., Brunelle, J., and Camara, B.M., (2023). "How secure is code generated by chatgpt?." In *2023 IEEE international conference on systems, man, and cybernetics (SMC)* (pp. 2445–2451). IEEE.

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

Li, J., and Xiao, L., (2023). "Multi-emotion recognition using multi-EmoBERT and emotion analysis in fake news." In *Proceedings of the 15th ACM web science conference 2023* (pp. 128–135).

Liu, J., Tang, X., Li, L., Chen, P., and Liu, Y. (2023). Which is a better programming assistant? A comparative study between chatgpt and stack overflow. *arXiv preprint arXiv* 2308:13851.

Mhlanga, D. (2023). "Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning" in FinTech and artificial intelligence for sustainable development: The role of smart technologies in achieving development goals (Springer Nature Switzerland: Cham), 387–409.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). "Semeval-2018 task 1: affect in tweets." In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 1–17).

Murad, I. A., Surameery, N. M. S., and Shakor, M. Y. (2023). Adopting ChatGPT to enhance educational experiences. *Int J Info Technol Comp Eng* 3, 20–25. doi: 10.55529/ijitc.35.20.25

Nandrino, J. L., Baracca, M., Antoine, P., Paget, V., Bydlowski, S., and Carton, S. (2013). Level of emotional awareness in the general French population: effects of gender, age, and education level. *Int. J. Psychol.* 48, 1072–1079. doi: 10.1080/00207594.2012.753149

Nascimento, N., Alencar, P., and Cowan, D. (2023). Comparing software developers with chatgpt: an empirical investigation. *arXiv preprint arXiv* 2305:11837.

Okey, O. D., Udo, E. U., Rosa, R. L., Rodríguez, D. Z., and Kleinschmidt, J. H. (2023). Investigating ChatGPT and cybersecurity: a perspective on topic modeling and sentiment analysis. *Comput. Secur.* 135:103476. doi: 10.1016/j.cose.2023.103476

Open AI. (2023). "ChatGPT." Available at: https://openai.com/blog/chatgpt

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *Adv. Neural Inf. Proces. Syst.* 35, 27730–27744.

Pamungkas, E. W. (2019). Emotionally-aware chatbots: a survey. *arXiv preprint arXiv* 1906:09774.

Parker, J. L., Becker, K., and Carroca, C. (2023). ChatGPT for automated writing evaluation in scholarly writing instruction. *J. Nurs. Educ.* 62, 721–727. doi: 10.3928/01484834-20231006-02

Parrott, W. G. (2001). Emotions in social psychology: Essential readings. New York, NY, US: psychology press.

Plutchik, R. (1980). "A general psychoevolutionary theory of emotion" in Theories of emotion (New York: Academic press), 3–33.

Poremba, S., (2023). ChatGPT confirms data breach, raising security concerns. Retrieved from security intelligence website: https://securityintelligence.com/articles/chatgpt-confirms-data-breach.

Rapp, A., Curti, L., and Boldi, A. (2021). The human side of human-chatbot interaction: a systematic literature review of ten years of research on text-based chatbots. *Int J Human-Computer Stud* 151:102630. doi: 10.1016/j.ijhcs.2021.102630

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39:1161. doi: 10.1037/h0077714

Saini, N. (2023). ChatGPT becomes fastest growing app in the world, records 100mn users in 2 month. *LiveMint.*

Spolsky, J., and Atwood, J., (2008). Introducing stack overflow.

Surameery, N. M. S., and Shakor, M. Y. (2023). Use chat GPT to solve programming bugs. *Int J Info Technol Comp Eng (IJITC)* 3, 17–22. doi: 10.55529/ijitc.31.17.22

Susnjak, T., and McIntosh, T. R. (2024). ChatGPT: the end of online exam integrity? *Educ. Sci.* 14:656. doi: 10.3390/educsci14060656

Tyng, C. M., Amin, H. U., Saad, M. N., and Malik, A. S. (2017). The influences of emotion on learning and memory. *Front. Psychol.* 8:235933. doi: 10.3389/fpsyg.2017.01454

Um, E., Plass, J. L., Hayward, E. O., and Homer, B. D. (2012). Emotional design in multimedia learning. *J. Educ. Psychol.* 104:485. doi: 10.1037/a0026609

Wang, X., Pang, H., Wallace, M. P., Wang, Q., and Chen, W. (2024). Learners' perceived AI presences in AI-supported language learning: a study of AI as a humanized agent from community of inquiry. *Comput. Assist. Lang. Learn.* 37, 814–840. doi: 10.1080/09588221.2022.2056203

Malwarebytes (2023). What is ChatGPT? ChatGPT Security Risks. Available at: https://www.malwarebytes.com/cybersecurity/basics/chatgpt-ai-security (Accessed on Jun. 26, 2023)

Zhao, W., Huang, S., and Yan, L., (2024). "ChatGPT and the future of translators: overview of the application of interactive AI in English translation teaching." In *2024 4th international conference on computer communication and artificial intelligence (CCAI)* (pp. 303–307). IEEE.

frontiers | Frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 07 October 2024
DOI 10.3389/frai.2024.1460651

Check for updates

OPEN ACCESS

EDITED BY
Rita Orji,
Dalhousie University, Canada

REVIEWED BY
Leonardo Brandão Marques,
Federal University of Alagoas, Brazil
Grace Ataguba,
Academic City University College, Ghana

*CORRESPONDENCE
Burcu Arslan
✉ barslan@ets.org

RECEIVED 06 July 2024
ACCEPTED 04 September 2024
PUBLISHED 07 October 2024

CITATION
Arslan B, Lehman B, Tenison C, Sparks JR, López AA, Gu L and Zapata-Rivera D (2024) Opportunities and challenges of using generative AI to personalize educational assessment.
Front. Artif. Intell. 7:1460651.
doi: 10.3389/frai.2024.1460651

Opportunities and challenges of using generative AI to personalize educational assessment

Burcu Arslan*, Blair Lehman, Caitlin Tenison, Jesse R. Sparks, Alexis A. López, Lin Gu and Diego Zapata-Rivera

ETS Research Institute, Princeton, NJ, United States

In line with the positive effects of personalized learning, personalized assessments are expected to maximize learner motivation and engagement, allowing learners to show what they truly know and can do. Considering the advances in Generative Artificial Intelligence (GenAI), in this perspective article, we elaborate on the opportunities of integrating GenAI into personalized educational assessments to maximize learner engagement, performance, and access. We also draw attention to the challenges of integrating GenAI into personalized educational assessments regarding its potential risks to the assessment's core values of validity, reliability, and fairness. Finally, we discuss possible solutions and future directions.

KEYWORDS

personalization, educational assessment, generative artificial intelligence, validity, reliability, fairness

1 Introduction

Personalized learning has been shown to enhance learner motivation, engagement, and performance (Bernacki et al., 2021; Walkington, 2013; Walkington and Bernacki, 2018, 2019). Personalization can be delivered via humans (e.g., students or teachers), digital assessment systems (e.g., via a virtual agent embedded in a digital platform), or a combination (e.g., recommender systems). In educational assessment, standardization has been one of the most essential requirements for fair and valid measurement (Sireci, 2020). However, more recent discussions put the learners in front and expect that personalized assessments yield similar benefits to personalized learning (Bennett, 2023; Buzick et al., 2023; Sireci, 2020). The transition from standardized to more personalized assessment of learning (i.e., summative assessment) and assessment *for* learning (i.e., formative assessment) comes with inherent challenges in ensuring the validity, reliability, and fairness of more tailored, individualized assessments.

Artificial intelligence (AI) in education dates back more than four decades (see Holmes and Tuomi, 2022; Williamson and Eynon, 2020, for reviews). However, recent technological advancements and generative AI (GenAI) have broadened AI's scale and potential applications in education due to its ability to create human-like text, being generic enough to be employed for different tasks, and real-time personalization capabilities. GenAI is a subcategory of AI designed to generate content, including images, videos, and text. Large language models (LLMs) are specifically trained on vast amounts of text data. When powered by LLMs, GenAI models can have a contextual understanding, enhanced memory, and create content based on natural language input (Hadi et al., 2023).

Recent research has increasingly focused on the integration of GenAI and LLMs in educational settings, examining their potential (e.g., Barany et al., 2024; Gökoğlu, 2024; Hu, 2023; Kasneci et al., 2023; Mazzullo et al., 2023; Nguyen et al., 2023; Olney, 2023; Pankiewicz and Baker, 2023; Pardos and Bhandari, 2024; Wang et al., 2022). Similarly, some studies focus on the application of GenAI and LLMs in educational assessment, exploring their impact and implications (e.g., Bulut et al., 2024; Hao et al., 2024; Jiang et al., 2024; von Davier, 2023; Swiecki et al., 2022).

Despite these advancements, to our knowledge, the potential opportunities and challenges of using GenAI to personalize educational assessment have not been explored. As we mentioned above, the shift from one-size-fits-all assessments to more culturally relevant and responsive approaches is becoming more critical, especially as stakeholders recognize the limitations of traditional assessments in responding to the needs of diverse populations. Thus, personalized educational assessments are increasingly viewed as a means to enhance learner engagement, performance, and access (Bennett, 2023; Buzick et al., 2023; Randall et al., 2022; Sireci, 2020).

Similar to the other application areas, advances in GenAI offer opportunities and challenges to personalized educational assessment (see Kirk et al., 2024, for benefits and risks of personalization in general with LLMs). GenAI can be integrated with the existing frameworks for including personalization, adaptation, or responsiveness in assessments, such as caring assessments (Lehman et al., 2024; Zapata-Rivera et al., 2020), socioculturally responsive assessments (Bennett, 2023), formative assessments (Bennett, 2011; Black and Wiliam, 2009), and intelligent tutoring systems (Corbett et al., 1997; Graesser et al., 2012). For example, in line with the caring assessment framework, GenAI may be leveraged to tailor content to the learner's emotional, motivational, and cognitive state. Similarly, in line with socioculturally responsive assessments, GenAI may adapt assessment content to reflect diverse perspectives and contexts, considering the learner's cultural background. Moreover, GenAI may enhance formative assessments and intelligent tutors by providing real-time, personalized feedback in a conversational style that helps learners improve continuously (e.g., Cheng et al., 2024). By leveraging these established frameworks, GenAI can offer robust personalized assessments that are not only effective but also responsive to the diverse needs of learners.

Integrating GenAI into the existing frameworks may play a crucial role in efficiently personalizing educational assessments by automatically generating images, videos, scenarios, and metadata and evaluating and scoring assessment items. Moreover, GenAI has the potential to generate or modify assessment items in real-time (Arslan, 2024), adapt to the learner's responses, performance, interests, or cultural background, and provide personalized feedback and reporting dashboards. Additionally, GenAI can be used to create personalized conversations about the construct that can be used for assessment purposes to create assessment content at varying levels of language complexity or translate it into multiple languages. These potential uses of GenAI can help to achieve the previous efforts of enhancing the assessment experience through maximizing learner performance and engagement, activating existing *funds of knowledge* (González et al., 2005), and making assessments more relevant and accessible to learners, including neurodiverse and multilingual learners (Sireci, 2020).

However, using GenAI to personalize educational assessment also introduces significant challenges, such as ensuring fairness and maintaining validity and reliability. Research increasingly highlights the challenges and risks associated with GenAI, including issues such as bias, copyright infringement, the potential for harmful content, minimal control over its output, security concerns, and lack of interpretability and explainability (Bender et al., 2021; Kasneci et al., 2023). Table 1 shows the potential opportunities and challenges of using GenAI for personalized assessments, potential solutions, and future directions.

2 Potential opportunities for applying GenAI in personalized assessments

GenAI may offer significant opportunities to enhance the personalization of assessments to maximize learner motivation and engagement, performance, and access.

2.1 Personalization for maximizing motivation and engagement

Increased motivation during test-taking leads to cognitive engagement, resulting in learners giving their best effort when answering assessment items (Finn, 2015; Wise and Kong, 2005; Wise, 2017). Cognitive engagement improves the likelihood that test scores will accurately represent what learners know and can do, as the interpretation of scores relies on the assumption that learners are trying their best (Finn, 2015; Wise, 2017). An effective way of maximizing engagement for learners with diverse interests and sociocultural is to personalize the context of assessment items (Bennett, 2023; Bernacki and Walkington, 2018; Sireci, 2020; Walkington and Bernacki, 2018). Context personalization can significantly enhance learner motivation and engagement by allowing learners to bring their cultural identity to the learning environment, leading to better learning outcomes (Walkington, 2013; Walkington and Bernacki, 2018, 2019).

LLMs have made it possible to personalize the context of assessment items *during* assessment based on each learner's input about their interests embedded in their cultural identities, thus maximizing engagement through situational interest (see Hidi and Renninger, 2006) and has the potential to allow learners to show what they know and can do (Arslan, 2024). Unlike personalization approaches that leverage background variables (e.g., race/ethnicity) to create culturally relevant forms and assign each form to a group of learners based on their demographic information (e.g., Sinharay and Johnson, 2024), using LLMs offers real-time tailoring of content to individual interests and cultural background by providing learners agency and relevance that is often missing in standardized assessments. This approach acknowledges the diversity among learners, avoiding the pitfall of assuming homogeneity within groups (Arslan, 2024).

Integrating conversational virtual agents into assessment platforms is another way of making assessments more engaging and user-friendly. The virtual agents, powered by LLMs, can respond to queries in natural language, providing real-time, contextual support that assists both students and teachers during their interactions with

TABLE 1 Potential opportunities and challenges of using GenAI for personalized assessments, potential solutions, and future directions.

Opportunities	Challenges	Consequences	Potential solutions and future directions
Maximizing Engagement • On-the-fly context personalization (e.g., Arslan, 2024; Bennett, 2023; Bernacki and Walkington, 2018; Hidi and Renninger, 2006; Sireci, 2020; Walkington, 2013; Walkington and Bernacki, 2018, 2019). • Conversational agents to enhance usability (Zapata-Rivera, 2012; Bull and Kay, 2016; Zapata-Rivera and Greer, 2002). Maximizing Performance • Conversational agents to gather additional evidence, to provide just-in-time feedback (e.g., Kochmar et al., 2020; Ma et al., 2014; Mazzullo et al., 2023; Meyer et al., 2024; Matelsky et al., 2023; Pardos and Bhandari, 2024; Wang and Han, 2021), and to enhance dashboards for reporting (e.g., Forsyth et al., 2017; Xhakaj et al., 2017). Increasing Access • Conversational agents, scaffolds, and language supports for neurodiverse and multilingual learners (e.g., Ali et al., 2020; López, 2023; Yang, 2024).	• Lack of control over the quality and content of the output (e.g., Bender et al., 2021; Jurenka et al., 2024). • Hallucinations, potential bias and fairness issues (e.g., Jurenka et al., 2024; Hu, 2023; Jurenka et al., 2024; Jurenka et al., 2024; Jurenka et al., 2024; Jurenka et al., 2024; Jurenka et al., 2024; Ye et al., 2023; Jurenka et al., 2024). • Lack of interpretability and explainability of the system's underlying decision-making (see Jurenka et al., 2024 for a survey).	Jeopardizing assessment's core values of: • Validity. • Reliability. • Fairness (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (US), 2014; Jurenka et al., 2024).	• Developing guidelines and standards for the ethical use of GenAI in personalized assessment (e.g., Hu, 2023; Jurenka et al., 2024). • Aligning the purpose and goals of the assessment with how GenAI is being leveraged (Jurenka et al., 2024). • Developing methodologies to evaluate the quality of the output of GenAI (e.g., human-in-the loop approaches; Jurenka et al., 2024; Jurenka et al., 2024; Jurenka et al., 2024). • Working with practitioners and students to co-design solutions (Penuel, 2019). • Identifying and implementing guardrails (Rai et al., 2024). • Combining neuro-symbolic approaches and/or using computational cognitive architectures to create decision-making systems that leverage knowledge of human cognition (e.g., Jurenka et al., 2024; Jurenka et al., 2024).

10.3389/frai.2024.1460651

the platform (Zapata-Rivera, 2012; Bull and Kay, 2016; Zapata-Rivera and Greer, 2002).

2.2 Personalization for maximizing performance

Unlike traditional summative assessments, personalized formative assessments can significantly enhance performance by providing feedback tailored to each learner's needs (e.g., Kochmar et al., 2020; Ma et al., 2014; Mazzullo et al., 2023; Hu, 2023; Wang and Han, 2021). LLMs can generate hints and adaptive feedback during assessments at scale in an efficient way, helping learners understand their mistakes and learn from them in real-time (e.g., Meyer et al., 2024; Matelsky et al., 2023; Pardos and Bhandari, 2024) and facilitate adaptive conversations that guide learners through their thought processes (Hu, 2023; Forsyth et al., 2024; Zapata-Rivera et al., 2024).

LLMs can also enhance reporting by providing detailed, narrative insights for learners, teachers, and other interest holders. These insights can help interest holders understand assessment information more deeply and make informed decisions. For example, it can influence what teachers know about their student and their decision-making through dashboards (e.g., Forsyth et al., 2017; Xhakaj et al., 2017).

2.3 Personalization for increasing access

Personalized assessments can be crucial in increasing access for diverse learner populations, including neurodiverse and/or multilingual learners. LLMs offer various options for making assessments more linguistically responsive to the needs of multilingual learners (Yang, 2024). A significant way LLMs can enhance accessibility for multilingual learners is by providing support and scaffolds, such as translations to the learner's preferred language, language simplification, glossaries, and read-aloud features. These tools allow multilingual learners—who comprise 10.6% of the student population in US public schools (National Center for Education Statistics, 2024a)—to utilize all available linguistic resources without compromising the construct being measured. These tools offer multilingual learners alternative ways to access and engage with assessment content, ensuring that language barriers do not block learners' ability to fully demonstrate what they know and can do (Bennett, 2023; Sireci, 2020). In this context, LLMs are leveraged to provide enriched, inclusive means for all learners to access the assessment content and showcase their conceptual understanding using multiple modes of communication (e.g., linguistic, visual, aural, spatial, gestural) to reflect the diversity of needs and abilities in U.S. public schools (National Center for Education Statistics, 2024b). In essence, LLMs allow learners to use their entire linguistic repertoire, enabling them to express their KSAs through multiple forms of representation, including oral and written language and drawings (García and Wei, 2014; López, 2023). This approach, associated with translanguaging, supports providing multiple forms of expression, making assessments more inclusive and reflective of learners' diverse backgrounds (Bennett, 2023).

Conversational virtual agents powered by LLMs can also be used to further support usability for neurodiverse and/or multilingual learners by enabling interactive, natural language-based supports with a choice of spoken and written communication in understanding and navigating the assessment platform, interpreting assessment items, and providing real-time, context-sensitive assistance. (e.g., see Ali et al., 2020). This potential application makes the platform more user-friendly, as discussed in the above section, and may ensure that all learners, regardless of their language proficiency, can fully participate in the assessment process.

3 Challenges, potential solutions, and future directions

Despite its potential, GenAI introduces significant challenges for personalized assessments. In this section, we first mention GenAI's challenges in this context. Subsequently, we provide an overview of potential solutions to these challenges and future directions for research.

3.1 Challenges

Alongside research applying GenAI to new problems and domains, a growing body of work highlights the limitations and risks associated with its use. These discussions address potential biases, copywriting infringement, and the harmful content that can be introduced by large training datasets over which users have little control (Bender et al., 2021). Additionally, concerns about data privacy and security, particularly in educational contexts, are increasingly relevant when using these models (Kasneci et al., 2023). These general issues pose specific challenges when considering how GenAI can be used responsibly to support the design, administration, and reporting of personalized assessments while upholding the core values of validity, reliability, and fairness (see Johnson, 2024). Although these challenges may vary depending on the type and purpose of the assessment (e.g., formative vs. summative), we discuss several overarching challenges that are likely to shape the future development of personalized assessments using GenAI.

Personalizing assessments with GenAI offers benefits such as reducing construct-irrelevant variance indirectly by maximizing engagement (e.g., see Section 2.1) or directly by maximizing access (e.g., Section 2.3). However, without careful use, GenAI is just as likely to introduce new sources of construct-irrelevant variance. Approaches like Evidence-Centered Design (ECD; Mislevy et al., 2003) systematically align every aspect of the assessment process with theoretical and empirical evidence needed to support the claims made based on test scores. Part of the strength of this approach for generating valid assessments is the transparency at each step of the assessment development process and mapping decisions made to the intended interpretations and uses of the test. When using GenAI for on-the-fly content generation (e.g., see Section 2.1) or as a conversational virtual agent (e.g., see Sections 2.1 and 2.3), the lack of control over the output of GenAI makes it harder to ensure that the assessment content is measuring what we intend to measure (e.g., see Hong et al., 2024). With less control over the content, risks span from the introjection of inappropriate (e.g., see Greshake et al., 2023), non-sensical (Ye et al., 2023), incorrect (Hicks et al., 2024), or biased content and representations that these models have been known to exhibit (Cheung et al., 2024; Jiang et al., 2024; UNESCO, IRCAI, 2024; Schleifer et al., 2024; Zhou et al., 2024). Moreover, LLMs perform complex computations, complicating the interpretation of their decision-making processes. This 'black-box' nature of GenAI makes it harder to detect the sources of problematic output and to create explanations for interest holders (see Zhao et al., 2024 for a survey of the explainability of LLMs).

One of the cornerstone principles of standardized summative assessments is the consistency of test forms and the comparability of scores, which ensures the reliability and validity of scores across different test administrations (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (US), 2014). For example, in standardized summative assessments, on-the-fly personalization with GenAI, which generates uniquely tailored items during the assessment, may introduce construct-irrelevant variance into the measurement. This poses significant challenges to critical tenets of reliability and validity and complicates the currently established process for evaluating and documenting the reliability or precision of a given assessment. These challenges add a new dimension to ongoing discussions of the need for an expanded psychometric toolbox (such as computational psychometrics; von Davier et al., 2021), as well as more explicit guidance on valid score inferences when incorporating AI (Huggins-Manley et al., 2022) and personalization in assessments (Buzick et al., 2023).

Finally, developing and maintaining GenAI models specialized for personalized assessments involves numerous technical challenges. While prompt engineering is a popular method for adjusting a GenAI model's behavior, its ability to alter the model's actions is limited to what the model has already learned during pre-training (Bozkurt and Sharma, 2023; Jurenka et al., 2024). Alternative approaches like fine-tuning are much more expensive, requiring both quality data and expertise (e.g., Chevalier et al., 2024). Lastly, concerns regarding the hosting and management of GenAI models highlight critical data privacy and security issues. These are especially pertinent in educational contexts, where the sensitivity of learner data requires stringent security measures and ethical considerations (see Johnson, 2024).

3.2 Potential solutions and future directions

There are several key areas for future research to understand better how GenAI can be leveraged for personalized assessments. The first area is identifying, developing, and distributing guidelines and standards for the ethical use of GenAI in personalized assessments. A set of guidelines and standards helps guide future research and development and facilitates clear expectations for interest holders. Several emerging efforts exist to establish responsible AI standards in educational assessment (Burstein, 2023; Johnson, 2024). However, continued work is needed to establish guidelines and standards that encompass the full potential uses of GenAI in assessment design and development (e.g., content development to be evaluated by humans vs. on-the-fly personalization). To this end, as we briefly mentioned in the Introduction, existing frameworks for personalization, adaptation, and responsiveness in assessments may help identify these potential uses and important use cases.

The second area for future research is identifying how to best leverage GenAI in different testing contexts. As is typically the case in education, there is unlikely to be a one-size-fits-all solution for leveraging GenAI for personalized assessments (Bennett, 2023). For example, on-the-fly item generation may not be appropriate for a summative, high-stakes assessment in which there are high demands for score comparability. However, it may be appropriate for a formative, low-stakes assessment in a classroom context. Moreover, when additional approaches are taken to mitigate the inherent challenges of GenAI (e.g., nonfactual information and bias), it may be appropriate to leverage it to provide learners with conversational support during the assessment. Thus, it is essential to align the purpose and goals of the assessment with how GenAI is being leveraged and to develop methodologies to evaluate the quality of the output of GenAI before operationalizing the personalized assessments (e.g., see Zapata-Rivera et al., 2024 for leveraging ECD). It will be critical to regularly evaluate the impact of using GenAI-developed content for assessments on the perceptions of various interest holders when applied in different manners to different testing contexts. (e.g., teachers, learners, policymakers). It will also be essential to leverage GenAI to address the current needs of practitioners (e.g., teachers, assessment developers) and learners to improve the experience of developing, administering, and completing assessments. For example, teachers may struggle to provide all aspects of students' Individualized Education Programs (IEPs) during an assessment due to tools without the appropriate nuance and/or resource limitations, such as one teacher in a class of 30 students (Lehman et al., submitted).[1] Researchers can work with practitioners and students to co-design solutions to these real-world problems that utilize GenAI (Penuel, 2019).

When establishing how best to leverage GenAI in different testing contexts, a third area of research is needed to identify the guardrails that must be implemented to address some of the abovementioned challenges. While it may be tempting to let GenAI run free to maximize its potential benefits fully, key guardrails can be implemented to limit unintended negative consequences and maintain rigorous, appropriate content for personalized assessments. For example, implementing a 'human-in-the-loop' approach allows for human inspection and evaluation before GenAI-generated content is presented to learners (Amirizaniani et al., 2024; Drori and Te'eni, 2024; Park, 2024). However, this type of human review can limit some potential uses of GenAI, such as on-the-fly personalization. Moreover, rigorous research is essential to narrow the decision space for GenAI and mitigate the 'black-box' nature of LLMs. This can be achieved by integrating neuro-symbolic approaches or using computational cognitive architectures to develop decision-making systems that leverage an understanding of human cognition (e.g., Sumers et al., 2023; Sun, 2024). Additionally, combining these approaches with insights from key interest holders—such as teachers, students, and assessment developers—can help identify effective ways to utilize GenAI while minimizing unintended negative consequences.

1 Lehman, B., Gooch, R., Tenison, C., & Sparks, J. R. (submitted). The role of teachers in digital personalized assessments. Paper submitted to the annual meeting of the American Educational Research Association. Denver, CO.

The previous areas for future research have focused on the content generation process via GenAI. However, there is also a need for rigorous research to evaluate the personalized assessments that are developed with GenAI. This research will need to evaluate the quality of the content developed with or by GenAI and how the use of GenAI impacts the broader uptake of personalized assessments. When appropriate, it will be necessary to evaluate the utility of GenAI content within the current assessment development process. For example, it will be necessary to document if GenAI content results in more efficient content development processes that still maintain high levels of quality (e.g., see Park, 2024). Another area for future research is how GenAI could be leveraged to support response scoring, which could support personalized assessment and more efficient reporting (e.g., Section 2.2.).

Lastly, the full potential of utilizing GenAI for personalized assessments can only be realized if interest holders (e.g., teachers, learners, curriculum experts, and policymakers) view those assessments as valid, reliable, and fair, thus trustworthy and helpful in supporting learning.

4 Conclusion

Overall, there is a significant opportunity to enhance the deployment and effectiveness of personalized assessments, which could offer learners more relevant test materials, leading to greater engagement, improved performance, and broader access. This, in turn, has the potential to produce more valid test outcomes. However, while the potential of GenAI to create more valuable assessments is promising, it is crucial to proceed with caution. The field must continue to explore how GenAI can be effectively harnessed, but this exploration should be grounded in a rigorous evaluation of its utility.

As we move forward, it is essential not to abandon the potential for future advancements in assessments in favor of holding onto outdated development and evaluation processes (Huggins-Manley et al., 2022; Sireci, 2020). While embracing the possibilities offered by AI, we must ensure that these new tools are evaluated against criteria that recognize the affordances of both current and future technologies. However, this should never come at the expense of the core values of assessments—validity, reliability, fairness, and alignment with valued educational goals. By balancing innovation with caution, we can strive to create assessments that are both cutting-edge and trustworthy.

Author contributions

BA: Conceptualization, Writing – original draft, Writing – review & editing. BL: Conceptualization, Writing – original draft, Writing – review & editing. CT: Conceptualization, Writing – original draft, Writing – review & editing. JS: Conceptualization, Writing – original draft, Writing – review & editing. AL: Conceptualization, Writing – original draft, Writing – review & editing. LG: Conceptualization, Writing – original draft, Writing – review & editing. DZ-R: Conceptualization, Funding acquisition, 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 was funded by ETS Research Institute. This material is based upon work supported by the National Science Foundation and the Institute of Education Sciences under Grant #2229612. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Education.

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

References

Ali, M. R., Razavi, S. Z., Langevin, R., Al Mamun, A., and Kane, B. (2020). "A virtual conversational agent for teens with autism spectrum disorder: experimental results and design lessons." In *Proceedings of the 20th ACM international conference on intelligent virtual agents* (pp. 1–8).

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (US) (2014). Standards for educational and psychological testing. Washington, DC: AERA.

Amirizaniani, M., Yao, J., Lavergne, A., Okada, E. S., and Chadha, A. (2024). Developing a framework for auditing large language models using human-in-the-loop. arXiv preprint at: https://arxiv.org/pdf/2403.16809

Arslan, B. (2024). Personalized, adaptive, and inclusive digital assessment and learning environments. [conference presentation]. E-ADAPT conference, Potsdam, Germany. Available at: https://osf.io/82p5f/?view_only=cba3f410bc1e462fb086e3361ffed0bc (Accessed September 04, 2024).

Barany, A., Nasiar, N., Porter, C., Zambrano, A. F., and Andres, A. L. (2024). "ChatGPT for education research: exploring the potential of large language models for qualitative codebook development." In *International conference on artificial intelligence in education* (pp. 134–149). Cham: springer Nature Switzerland.

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* (pp. 610–623).

Bennett, R. E. (2011). Formative assessment: a critical review. *Assess. Educ. Principles, Policy & Prac.* 18, 5–25. doi: 10.1080/0969594X.2010.513678

Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educ. Assess.* 28, 83–104. doi: 10.1080/10627197.2023.2202312

Bernacki, M. L., Greene, M. J., and Lobczowski, N. G. (2021). A systematic review of research on personalized learning: personalized by whom, to what, how, and for what purpose (s)? *Educ. Psychol. Rev.* 33, 1675–1715. doi: 10.1007/s10648-021-09615-8

Bernacki, M. L., and Walkington, C. (2018). The role of situational interest in personalized learning. *J. Educ. Psychol.* 110, 864–881. doi: 10.1037/edu0000250

Black, P., and Wiliam, D. (2009). Developing a theory of formative assessment. *Educ. Assess. Eval. Account.* 21, 5–31. doi: 10.1007/s11092-008-9068-5

Bozkurt, A., and Sharma, R. C. (2023). Generative AI and prompt engineering: the art of whispering to let the genie out of the algorithmic world. *Asian J. Distance Educ.* 18, i–vii.

Bull, S., and Kay, J. (2016). SMILI☺: a framework for interfaces to learning data in open learner models, learning analytics and related fields. *Int. J. Artif. Intell. Educ.* 26, 293–331. doi: 10.1007/s40593-015-0090-8

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., et al. (2024). The rise of artificial intelligence in educational measurement: opportunities and ethical challenges. *arXiv preprint arXiv* 2406:18900. doi: 10.48550/arXiv.2406.18900

Burstein, J. (2023). The Duolingo English Test Responsible AI Standards. Retrieved from https://go.duolingo.com/ResponsibleAI (Accessed July 7, 2024).

Buzick, H. M., Casabianca, J. M., and Gholson, M. L. (2023). Personalizing large-scale assessment in practice. *Educ. Meas. Issues Pract.* 42, 5–11. doi: 10.1111/emip.12551

Cheng, L., Croteau, E., Baral, S., Heffernan, C., and Heffernan, N. (2024). Facilitating student learning with a chatbot in an online math learning platform. *J. Educ. Comput. Res.* 62, 907–937. doi: 10.1177/07356331241226592

Cheung, V., Maier, M., and Lieder, F. (2024). Large language models amplify human biases in moral decision-making. *Psyarxiv preprint*. doi: 10.31234/osf.io/aj46b

Chevalier, A., Geng, J., Wettig, A., Chen, H., Mizera, S., Annala, T., et al. (2024). Language models as science tutors. *arXiv preprint* arXiv. doi: 10.48550/arXiv.2402.11111

Corbett, A. T., Koedinger, K. R., and Anderson, J. R. (1997). "Intelligent tutoring systems" in *Handbook of human-computer interaction, second, completely revised edition*. eds. M. Helander, T. K. Landauer and P. Prabhu (North-Holland: Elsevier Science B. V.), 849–874.

Drori, I., and Te'eni, D. (2024). Human-in-the-loop AI reviewing: feasibility, opportunities, and risks. *J. Assoc. Inf. Syst.* 25, 98–109. doi: 10.17705/1jais.00867

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Res. Report Series* 2015, 1–17. doi: 10.1002/ets2.12067

Forsyth, C. M., Peters, S., Zapata-Rivera, D., Lentini, J., Graesser, A. C., and Cai, Z. (2017). Interactive score reporting: an AutoTutor-based system for teachers. In R. Baker, E. Andre, X. Hu, T. Rodrigo and BoulayB. du (Eds.), Proceedings of the international conference on artificial intelligence in education, LNCS. Switzerland: springer Verlag. pp. 506–509.

Forsyth, C.M., Zapata-Rivera, D., Graf, A., and Jiang, Y. (2024). "Complex conversations: LLMs vs. knowledge engineered conversation-based assessment." In *Proceedings of the international conference on educational data mining*.

García, O., and Wei, L. (2014). Translanguaging: Language, Bilingualism and Education. US: Palgrave Macmillan.

Gökoğlu, S. (2024). "Challenges and limitations of generative AI in education," in *Transforming education with generative AI*. ed. N. Gunsel (GI Global), 158–181.

González, N., Moll, L. C., and Amanti, C. (2005). Funds of knowledge: Theorizing practices in households, communities, and classrooms. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Graesser, A. C., Conley, M. W., and Olney, A. (2012). "Intelligent tutoring systems" in APA Educational Psychology handbook, Vol. 3. Application to learning and teaching. eds. K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major and H. L. Swanson (American Psychological Association), 451–473.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. (2023). "Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection." In *Proceedings of the 16th ACM workshop on artificial intelligence and security* (pp. 79–90).

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., et al. (2023). Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *TechRxiv*. doi: 10.36227/techrxiv.23589741.v2

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., and Harris, D. J. (2024). Transforming assessment: the impacts and implications of large language models and generative ai. *Educ. Meas. Issues Pract.* 43, 16–29. doi: 10.1111/emip.12602

Hicks, M. T., Humphries, J., and Slater, J. (2024). ChatGPT is bullshit. *Ethics Inf. Technol.* 26:38. doi: 10.1007/s10676-024-09775-5

Hidi, S., and Renninger, K. (2006). The four-phase model of interest development. *Educ. Psychol.* 41, 111–127. doi: 10.1207/s15326985ep4102_4

Holmes, W., and Tuomi, I. (2022). State of the art and practice in AI in education. *Eur. J. Educ.* 57, 542–570. doi: 10.1111/ejed.12533

Hong, P., Ghosal, D., Majumder, N., Aditya, S., Mihalcea, R., and Poria, S. (2024). Stuck in the quicksand of numeracy, far from AGI summit: evaluating LLMs' mathematical competency through ontology-guided perturbations. *arXiv preprint arXiv* 2401:09395. doi: 10.48550/arXiv.2401.09395

Hu, X. (2023). "Empowering education with LLMs - the next gen interface and content generation workshop [demo]." Presented at the *international conference on artificial intelligence in education* (Tokyo, Japan, July 03-07, 2023). AIED 2023.

Huggins-Manley, A. C., Booth, B. M., and D'Mello, S. K. (2022). Toward argument-based fairness with an application to AI-enhanced educational assessments. *J. Educ. Meas.* 59, 362–388. doi: 10.1111/jedm.12334

Jiang, Y., Hao, J., Fauss, M., and Li, C. (2024). Detecting ChatGPT-generated essays in a large-scale writing assessment: is there a bias against non-native English speakers? *Comput. Educ.* 217:105070a. doi: 10.1016/j.compedu.2024.105070

Johnson, M. (2024). ETS principles for responsible use of AI in assessments. ETS Highlights. Available at: https://www.ets.org/Rebrand/pdf/ETS_Convening_executive_summary_for_the_AI_Guidelines.pdf (Accessed July 7, 2024).

Jurenka, I., Kunesch, M., McKee, K. R., Gillick, D, et al. (2024). Towards responsible development of generative AI for education: an evaluation-driven approach. Retrieved from: https://storage.googleapis.com/deepmind-media/LearnLM/LearnLM_paper.pdf (Accessed September 2, 2024).

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

Kirk, H. R., Vidgen, B., Röttger, P., and Hale, S. A. (2024). The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intell.* 6, 383–392. doi: 10.1038/s42256-024-00820-y

Kochmar, E., Vu, D. D., Belfer, R., Gupta, V., Serban, I. V., and Pineau, J. (2020). "Automated personalized feedback improves learning gains in an intelligent tutoring system." *Artificial intelligence in education: 21st international conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II*, 12164, 140–146.

Lehman, B., Sparks, J. R., Zapata-Rivera, D., Steinberg, J., and Forsyth, C. (2024). A culturally enhanced framework of caring assessments for diverse learners. *Pract. Assess. Res. Eval.* 29. doi: 10.7275/pare.2102

López, A. A. (2023). Examining how Spanish-speaking English language learners use their linguistic resources and language modes in a dual language mathematics assessment task. *J. Latinos Educ.* 22, 198–210. doi: 10.1080/15348431.2020.1731693

Ma, W., Adesope, O. O., Nesbit, J. C., and Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: a meta-analysis. *J. Educ. Psychol.* 106, 901–918. doi: 10.1037/a0037123

Matelsky, J. K., Parodi, F., Liu, T., Lange, R. D., and Kording, K. P. (2023). A large language model-assisted education tool to provide feedback on open-ended responses. *arXiv preprint arXiv* 2308:02439. doi: 10.48550/arXiv.2308.02439

Mazzullo, E., Bulut, O., Wongvorachan, T., and Tan, B. (2023). Learning analytics in the era of large language models. *Analytics* 2, 877–898. doi: 10.3390/analytics2040046

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., et al. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Comput. Educ.: Artificial Intelligence* 6:100199. doi: 10.1016/j.caeai.2023.100199

Mislevy, R. J., Almond, R. G., and Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Res. Report Series* 2003, 1–29. doi: 10.1002/j.2333-8504.2003.tb01908.x

National Center for Education Statistics. (2024a). English learners in public schools. Condition of education. U.S. Department of Education, Institute of Education Sciences. Retrieved from https://nces.ed.gov/programs/coe/indicator/cgf (Accessed July 7, 2024).

National Center for Education Statistics. (2024b). Students with disabilities. Condition of education. U.S. Department of Education, Institute of Education Sciences. Retrieved from https://nces.ed.gov/programs/coe/indicator/cgg (Accessed July 7, 2024).

Nguyen, H. A., Stec, H., Hou, X., Di, S., and McLaren, B. M. (2023). "Evaluating ChatGPT's decimal skills and feedback generation in a digital learning game" in *Responsive and Sustainable Educational Futures. EC-TEL 2023*. eds. O. Viberg, I. Jivet, P. Muñoz-Merino, M. Perifanou and T. Papathoma (Cham, Switzerland: Springer Nature), 278–293.

Olney, A. (2023). "Generating multiple choice questions from a textbook: LLMs match human performance on most metrics." Paper presented at *the empowering education with LLMs - the next gen interface and content generation workshop at AIED 2023* (Tokyo, Japan, July 03–07).

Pankiewicz, M., and Baker, R. S. (2023). Large language models (GPT) for automating feedback on programming assignments. *arXiv preprint arXiv* 2307:00150. doi: 10.48550/arXiv.2307.00150

Pardos, Z. A., and Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLoS One* 19:e0304013. doi: 10.1371/journal.pone.0304013

Park, Y. (2024). "Digital-first content development for test-taker delight and fairness." Paper presented at *the 2024 annual meeting of the National Council on measurement in education*. Philadelphia, PA.

Penuel, W. R. (2019). "Co-design as infrastructuring with attention to power: building collective capacity for equitable teaching and learning through design-based implementation research" in Collaborative curriculum Design for Sustainable Innovation and Teacher Learning. eds. J. Pieters, J. Voogt and N. P. Roblin (Cham, Switzerland: springerOpen), 387–401.

Rai, P., Sood, S., Madisetti, V. K., and Bahga, A. (2024). Guardian: A multi-tiered defense architecture for thwarting prompt injection attacks on llms. *J. Softw. Eng. Appl.* 17, 43–68. doi: 10.4236/jsea.2024.171003

Randall, J., Slomp, D., Poe, M., and Oliveri, M. E. (2022). Disrupting white supremacy in assessment: toward a justice-oriented, antiracist validity framework. *Educ. Assess.* 27, 170–178. doi: 10.1080/10627197.2022.2042682

Schleifer, A. G., Klebanov, B. B., Ariely, M., and Alexandron, G. (2024). Anna Karenina strikes again: pre-trained LLM embeddings may favor high-performing learners. *arXiv preprint arXiv* 2406:06599. doi: 10.48550/arXiv.2406.06599

Sinharay, S., and Johnson, M. S. (2024). Computation and accuracy evaluation of comparable scores on culturally responsive assessments. *J. Educ. Meas.* 61, 5–46. doi: 10.1111/jedm.12381

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educ. Meas. Issues Pract.* 39, 100–105. doi: 10.1111/emip.12377

Sumers, T. R., Yao, S., Narasimhan, K., and Griffiths, T. L. (2023). Cognitive architectures for language agents. *arXiv preprint arXiv* 2309:02427. doi: 10.48550/arXiv.2309.02427

Sun, R. (2024). Can a cognitive architecture fundamentally enhance LLMs? Or vice versa? *arXiv preprint arXiv* 2401:10444. doi: 10.48550/arXiv.2401.10444

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., et al. (2022). Assessment in the age of artificial intelligence. *Comput. Educ.: Artificial Intelligence* 3:100075. doi: 10.1016/j.caeai.2022.100075

UNESCO, IRCAI (2024). Challenging Systematic Prejudices: An Investigation into Gender Bias in Large Language Models. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000388971 (Accessed July 7, 2024).

von Davier, M. (2023). "Training Optimus prime, MD: a case study of automated item generation using artificial intelligence–from fine-tuned GPT2 to GPT3 and beyond" in *Advancing natural language processing in educational assessment*. eds. V. Yaneva and M. von Davier (New York, NY: Routledge), 90–106.

von Davier, A. A., DiCerbo, K., and Verhagen, J. (2021). "Computational psychometrics: A framework for estimating learners' knowledge, skills and abilities from learning and assessments systems," in *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python*. eds. A. A. von Davier, R. J. Mislevy, and J. Hao (Springer), 25–43.

Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: the impact of relevant contexts on performance and learning outcomes. *J. Educ. Psychol.* 105, 932–945. doi: 10.1037/a0031882

Walkington, C., and Bernacki, M. L. (2018). Personalization of instruction: design dimensions and implications for cognition. *J. Exp. Educ.* 86, 50–68. doi: 10.1080/00220973.2017.1380590

Walkington, C., and Bernacki, M. L. (2019). Personalizing algebra to students' individual interests in an intelligent tutoring system: how moderators of impact. *J. Artif. Intell. Educ.* 29, 58–88. doi: 10.1007/s40593-018-0168-1

Wang, D., and Han, H. (2021). Applying learning analytics dashboards based on process-oriented feedback to improve students' learning effectiveness. *J. Comput. Assist. Learn.* 37, 487–499. doi: 10.1111/jcal.12502

Wang, Z., Valdez, J., Basu Mallick, D., and Baraniuk, R. G. (2022). "Towards human-like educational question generation with large language models" in International conference on artificial intelligence in education. eds. M. M. Rodrigo, N. Matsuda, A. I. Cristea and V. Dimitrova (Cham: springer International Publishing), 153–166.

Williamson, B., and Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learn. Media Technol.* 45, 223–235. doi: 10.1080/17439884.2020.1798995

Wise, S. L. (2017). Rapid-guessing behavior: its identification, interpretation, and implications. *Educ. Meas. Issues Pract.* 36, 52–61. doi: 10.1111/emip.12165

Wise, S. L., and Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Appl. Meas. Educ.* 18, 163–183. doi: 10.1207/s15324818ame1802_2

Xhakaj, F., Aleven, V., and McLaren, B. M. (2017). "Effects of a dashboard for an intelligent tutoring system on teacher knowledge, lesson plans and class sessions." In *Artificial intelligence in education: 18th international conference, AIED 2017, Wuhan, China, June 28–July 1, 2017, Proceedings 18* (pp. 582–585). Springer International Publishing.

Yang, X. (2024). Linguistically responsive formative assessment for emergent bilinguals: exploration of an elementary teacher's practice in a math classroom. *Int. Multilingual Res. J.* 1–24, 1–24. doi: 10.1080/19313152.2024.2339757

Ye, H., Liu, T., Zhang, A., Hua, W., and Jia, W. (2023). Cognitive mirage: a review of hallucinations in large language models. *arXiv preprint arXiv* 2309:06794. doi: 10.48550/arXiv.2309.06794

Zapata-Rivera, D. (2012). "Adaptive score reports," in *Proceedings of the user modeling, adaptation, and personalization conference*. eds. J. Masthoff, B. Mobasher, M. Desmarais, and kambou (Berlin/Heidelberg: Springer), 340–345.

Zapata-Rivera, D., Forsyth, C. M., Graf, A., and Jiang, Y. (2024). "Designing and evaluating evidence-centered-design-based conversations for assessment with LLMs." *Proceedings of EDM 2024 workshop: Leveraging large language models for next generation educational technologies*.

Zapata-Rivera, J. D., and Greer, J. (2002). Exploring various guidance mechanisms to support interaction with inspectable learner models. *Proceed. Intell. Tutoring Syst. ITS* 2363, 442–452. doi: 10.1007/3-540-47987-2_47

Zapata-Rivera, D., Lehman, B., and Sparks, J. R. (2020). "Learner modeling in the context of caring assessments." In *Adaptive instructional systems: Second international conference, AIS 2020, held as part of the 22nd HCI international conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22* (pp. 422–431). Springer International Publishing.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., et al. (2024). Explainability for large language models: a survey. *ACM Trans. Intell. Syst. Technol.* 15, 1–38. doi: 10.1145/3639372

Zhou, M., Abhishek, V., Derdenger, T., Kim, J., and Srinivasan, K. (2024). Bias in generative AI. *arXiv preprint arXiv* 2403:02726. doi: 10.48550/arXiv.2403.02726

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 08 October 2024
DOI 10.3389/frai.2024.1461137

Check for updates

OPEN ACCESS

EDITED BY
Chien-Sing Lee,
Sunway University, Malaysia

REVIEWED BY
Antonio Sarasa-Cabezuelo,
Complutense University of Madrid, Spain
Melody Tan,
Sunway University, Malaysia

*CORRESPONDENCE
Chi-Ning Chang
changc10@vcu.edu

†These authors have contributed equally to this work

RECEIVED 07 July 2024
ACCEPTED 16 August 2024
PUBLISHED 08 October 2024

CITATION
Chang C-N, Hui J, Justus-Smith C and Wang T-W (2024) Navigating STEM careers with AI mentors: a new IDP journey. *Front. Artif. Intell.* 7:1461137. doi: 10.3389/frai.2024.1461137

Navigating STEM careers with AI mentors: a new IDP journey

Chi-Ning Chang*, John Hui†, Cammie Justus-Smith† and Tzu-Wei Wang†

School of Education, Virginia Commonwealth University, Richmond, VA, United States

Introduction: Mentoring is crucial to the success of STEM higher education. The Individual Development Plan (IDP) is a common career development tool in STEM graduate education that facilitates structured mentor-mentee interactions and goal setting. This study examined the integration of AI mentors into the myIDP framework to provide real-time support and career insights.

Methods: Using Google Gemini as an AI mentor, this study developed and assessed AI prompts within the myIDP framework. Eighteen STEM graduate students, primarily from underrepresented groups, were trained to engage with the AI mentor. Their interactions, feedback, and comments were analyzed using sentiment and thematic analysis.

Results: Participants reported positive experiences with AI mentors, noting benefits, such as immediate responses, up-to-date information, access to multiple AI mentors, enhanced ownership of career development, and time savings. However, concerns about misinformation, bias, privacy, equity, and algorithmic influences have also been raised. The study identified two hybrid human-AI mentoring models—Sequential Integration and Concurrent Collaboration—that combine the unique strengths of human and AI mentors to enhance the mentoring process.

Discussion: This study underscores the potential of AI mentors to enhance IDP practices by providing timely feedback and career information, thereby empowering students in their STEM career development. The proposed human-AI mentoring models show promise in supporting underrepresented minorities, potentially broadening participation in STEM fields.

KEYWORDS

individual development plan, myIDP, career development, career planning, STEM mentoring, human-AI mentoring, mentorship, large language model

1 Introduction

Mentoring plays a critical role in advancing success in higher education (U.S. Department of Education, 2012). With the growing emphasis on career readiness and global competitiveness in the fields of science, technology, engineering, and mathematics (STEM) (National Science Foundation, 2022), career development has become a fundamental component of STEM mentoring. This form of mentoring goes beyond basic guidance, establishing a strategic framework for navigating educational and professional challenges. By accelerating skill development, offering industry insights, fostering networking, and providing essential career and psychosocial support, effective mentorship is crucial for thriving in the competitive STEM landscape (National Academies of Sciences, Engineering, and Medicine, 2020).

Recognizing the value of mentorship, the U.S. CHIPS and Science Act of 2022, along with various federal funding agencies [e.g., National Science Foundation (NSF)], has emphasized the integration of mentorship into workforce development, particularly in sectors like semiconductor manufacturing, where aligning academic preparation with industry needs is critical. The act mandates that all NSF-supported graduate students utilize Individual Development Plans (IDPs), a widely recognized tool in STEM graduate education, to map educational goals, facilitate career exploration, and guide professional development in collaboration with principal investigators or mentors (CHIPS and Science Act, 2022; National Science Foundation, 2024). By fostering structured, two-way communication within the mentor-mentee relationship, IDPs help students achieve their career aspirations (Chang et al., 2021; Hobin et al., 2012), thereby supporting the development of a diverse and skilled workforce essential for maintaining U.S. leadership in global science and technology.

myIDP is a commonly used web-based IDP platform[1] developed in 2012 by the American Association for the Advancement of Science (AAAS), the Federation of American Societies for Experimental Biology (FASEB), and experts from multiple universities (Fuhrmann et al., n.d.). It has gained significant popularity in U.S. universities, enabling graduate students to self-assess, explore different career paths, and set SMART (Specific, Measurable, Achievable, Relevant, and Time-Bound) goals with the assistance of mentoring teams. During the self-assessment stage, myIDP offers assessments for scientific skills, career interests, and work-related values. Based on the assessment results, the platform provides users with a variety of career options, each with a matching score to help them consider career fit. Additionally, myIDP offers resources including articles, books, and professional societies to help users understand each career path. To further explore their target career, myIDP provides tips for attending relevant events and networking with professionals. Once a career path is selected, users are prompted to create SMART goals for career advancement, skill improvement, and project development. The platform also offers tips on identifying a mentoring team to discuss these goals and support the career development process. Upon completion, a certificate can be generated, which is often part of the paperwork required to meet degree or program requirements at some universities (Chang et al., 2023).

Due to the rapid evolution of STEM professions over the last decade, some resources about current career options on myIDP are limited and outdated. For example, as of July 1, 2024, the suggested books for each career path were mostly published before 2010, with publication dates ranging from 1993 to 2015. Consequently, student users working with myIDP rely heavily on human mentors for support. The drawback is that mentees cannot receive immediate feedback and up-to-date career information from human mentors due to time limitations and knowledge blind spots. In light of the advancements and applications of Large Language Models (LLMs), such as those used by Google Gemini and ChatGPT, this study explores the feasibility of using artificial intelligence (AI) mentors to enhance myIDP practice. We hypothesize that graduate students can receive real-time support throughout the process from AI mentors, while acknowledging that this integration might also face challenges (Köbis and Mehner, 2021).

Recent advancements in Generative Pre-trained Transformer (GPT) models have significantly influenced various industries around the world (Dehouche, 2021; ChatGPT Generative Pre-trained Transformer and Zhavoronkov, 2022; Li, 2020). In educational research, notable examples of these applications are the language processing AI systems ChatGPT and Gemini. These systems are built on Large Language Models (LLMs), which are so-called because of the substantial memory required for their training, maintenance, and optimization. LLMs operate using algorithms that analyze extensive text data to identify patterns and relationships within text. Through this training, LLMs develop the ability to generate outputs that align with the patterns observed in the training data. The more data the model is trained on, the more precise it becomes in producing accurate outputs. The models developed by OpenAI (ChatGPT) and Google (Gemini) were released to the public to gather additional data through organic usage. In addition to these efforts of sourcing data from the public by natural user experiences, OpenAI and Google both gathered enormous amounts of data for their models to assimilate. OpenAI had roughly ten billion dollars of funding from Microsoft to train and develop their model. Google has access to even larger amounts of user data through their search engine, website ads, and many other data sources they own. This distinction in data-as-a-resource allowed OpenAI and Google to develop some of the leading models in terms of accuracy and performance. The potential creates possibilities for higher education mentoring and career guidance in STEM.

Human-centered approaches to mentoring in STEM fields play a pivotal role in guiding students from academic learning to professional careers. This form of mentorship is critical as it facilitates the practical application of theoretical knowledge, enabling mentees to acquire essential skills, attitudes, and professional networks essential for success in STEM (National Academies of Sciences, Engineering, and Medicine, 2020). According to National Academies of Sciences, Engineering, and Medicine (2020), effective mentorship not only fosters career and psychosocial development but also cultivates deep, impactful relationships that contribute to the holistic development of STEM professionals. Further, research by Atkins et al. (2020) highlighted that mentoring significantly aids the career planning process, especially for underrepresented minority students, by fostering a scientific identity and providing them with role models and opportunities for growth in research contexts. This support strengthens students' self-identification as scientists and encourages diverse pathways into further STEM leadership, underscoring the need for tailored and research-focused mentoring approaches. These traditional mentoring methods are instrumental in developing the next generation of STEM leaders, making them an indispensable part of educational strategies aimed at enhancing career trajectories in these fields. However, as valuable as these types of mentoring relationships are, they are time and resource intensive.

While traditional human-centered mentoring in STEM fields has proven invaluable, generative AI offers a unique opportunity to bridge resource gaps, performing personalized tasks for students that can supplement human mentor availability (Neumann et al., 2021; Wollny et al., 2021). Scholars have highlighted the potential of AI to offer students more personalized career guidance, aligning with a changing labor market (Duan and Wu, 2024) as well as enhancing professional

1 https://myidp.sciencecareers.org/

development in specific fields such as preservice teacher education (Lu et al., 2024). AI mentoring also has the potential to provide equitable access to resources, particularly through virtual mentoring and AI-driven platforms (Akiba and Fraboni, 2023). These platforms not only facilitate personalized career advising but also significantly elevate engagement and satisfaction among diverse and underserved student populations (Okado et al., 2023). As AI continues to change mentoring and career development, it has the potential to be a powerful tool for inclusivity and adaptability in learning environments.

However, despite these promising developments, there remains a critical gap in understanding how exactly mentors and mentees might integrate their current mentoring practices to make the most of generative AI. A critical area that remains underexplored is the integration of generative AI into the development planning aspects of IDPs. Students could utilize aspects of AI to tailor development plans to their own needs and career paths, while leaving time for more nuanced and human-directed activities. Addressing these gaps could enhance the functionality of platforms like myIDP and help mentors and mentees envision new ways to utilize the IDP process and make effective use of their mentoring time.

The purpose of this study is to investigate the potential of AI mentors to enhance myIDP practices for STEM graduate students. The research methodology comprises three key components: (1) development of a comprehensive set of AI-based prompts aligned with the current myIDP framework, (2) evaluation of these prompts by a diverse cohort of STEM graduate students, predominantly from underrepresented groups, and (3) collection of participant feedback on the efficacy of AI mentors and strategies for optimal integration with human mentorship. Participants tested the AI prompts while engaging with myIDP and provided insights on the strengths and limitations of AI mentorship. Additionally, they offered perspectives on effectively leveraging both AI and human mentors throughout the IDP process. The findings contribute to the emerging field of AI integration in STEM mentoring, highlighting potential benefits, identifying areas for improvement, and exploring avenues to empower underrepresented minorities in STEM.

2 Materials and methods

2.1 Materials

This study initially evaluated the LLM-supported technologies of Google Gemini and ChatGPT. To streamline the process, we decided to utilize only one platform. During the data collection phase of our study, Gemini rapidly enhanced its features, a development anticipated given Google's extensive data collection, amassing over 7 billion data points daily for more than a decade (Sagiroglu and Sinanc, 2013). The vast resources available to Gemini were evident in its evolution during our prompt testing period. In our comparative tests, Gemini demonstrated more accurate real-time results and current responses compared to ChatGPT, pre-trained on data up to 2022. Furthermore, Gemini offered integration with Google products, such as Google Scholar, providing peer-reviewed literature to support its responses. Based on our architectural knowledge of LLMs and prompt testing outcomes, we concluded that Gemini was better suited for producing supportive results for myIDP guidance. The data for this study was generated using prompts created for myIDP, chosen due to its significant impact and extensive user base.

There are four major components in the myIDP framework: Assessment, Career Exploration, Create Plan, and Implement Plan. The research team, with diverse expertise in computer science, mentoring, myIDP, STEM career development, and research methods, collaboratively generated meaningful prompts by leveraging their varied knowledge. After initial testing of these prompts, the team refined the prompts to help graduate students use myIDP more effectively. The final prompts are shown in Table 1. The Assessment component included prompts to clarify unclear or unfamiliar assessment items for Skills (e.g., demonstrating workplace etiquette), Interests (e.g., negotiating agreements), and Values (e.g., independence vs. working alone). The Career Exploration component covered prompts for topics such as considering career fit, reading about careers, attending events and workshops, talking to people, and choosing a career path. The Create Plan component included prompts for creating goals related to career advancement, skill improvement, and project development. For the final component, Implement Plan, prompts were designed for building a mentoring team.

To integrate Google Gemini as an AI-mentor within the myIDP framework, participants were asked to use the prompts we generated (see Table 1) while navigating through the myIDP process. The prompts we generated were designed to help mentees acquire critical, up-to-date information, regarding their career development, which the myIDP portal cannot efficiently provide. For instance, the prompt "What is a typical workday as a [job title]?" could be used when mentees were navigating through the process of "Read about Careers" in myIDP. By using the generated prompts to ask Google Gemini questions aligned with the different aspects in myIDP, the AI mentor can be effectively integrated with the myIDP framework. The process of providing feedback on their experience with the AI-mentor and prompts will be addressed in the next section.

2.2 Methods

2.2.1 Data collection and sample

After the study was approved by the Institutional Review Board at the authors' university (Protocol #HM20028450), we recruited participants via the university's daily events newsletter as well as directed communication via the School of Engineering and Career Services. Starting with the myIDP website, participants were asked to integrate the provided prompts into their myIDP process, then provide feedback through a survey about their experiences using Gemini as their AI mentor. To help the participants with the process, we provided a tutorial that walked them through the myIDP process as well as how to use Google Gemini as an AI mentor with the prompts. Participants provided their comments on the prompts they tested and the responses from the AI mentor along with the Google Gemini public shared links. After completing the prompts testing, participants were asked to complete open-ended questions of their thoughts on the strengths and concerns of using the AI mentor. Participants each received a \$100 gift card to compensate for their participation.

This study involved 18 participants, all of whom were from STEM fields, as defined by the U.S. Department of Homeland Security's STEM Designated Degree Program List (U.S. Department of Homeland Security, 2023), which includes health-related disciplines. Of these, 15 were full-time graduate students in the United States, and three were visiting students from other countries. Half of the

TABLE 1 Prompts with the sentiment analysis results.

myIDP	Examples of prompts	Number of uses	Number of comments	Average sentiment (−1 to + 1)
Assessment				0.37
Skills assessment (Optional)	What is [item name]? (e.g., What is [demonstrating workplace etiquette]?)	8	6	0.42
Interests assessment (Optional)	What is [item name]? (e.g., What is [negotiating agreements]?)	8	6	0.43
Values assessment (Optional)	What is [item name]? (e.g., What is [independence]? What is [working alone]?)	10	5	0.23
Career exploration				0.26
Consider career fit (Choose at least one prompt)	What are some ways to know if I would be a good [job title]?	17	12	0.32
	[Talk about your top values first]. Could I pursue a career in [career path]?	5	3	0.78
Read about careers (Choose at least four prompts)	What is a typical workday as a [job title]?	17	12	0.03
	What is the average salary for a [job title] in [career path] this year?	16	11	0.13
	What job searching websites can I use to find job postings in [career path]?	10	11	0.24
	What skills and qualifications are typically required for careers in [career path]?	10	4	0.19
	What is the demand like for jobs in [career path]?	8	4	0.23
	What are some potential career growth opportunities or advancement prospects in [career path]?	4	2	0.5
	What are the main challenges or drawbacks associated with careers in [career path]?	7	4	0.36
	Are there any specific certifications or additional training that can enhance job prospects in [career path]?	5	3	0.42
	Can you suggest any resources or websites to explore for further information on careers in [career path]?	4	2	0.6
	What does the future hold for jobs in [career path]?	1	1	0.63
	Are there any notable trends or emerging areas within [career path] that might impact future job prospects?	1	0	NA
Attend events and workshops (Choose at least three prompts)	What are some annual or regional events in [city/locale] to assist in [given goal]?	6	2	0.21
	What are some upcoming events or workshops related to [career path] in [city/locale]?	13	9	0.3
	Can you recommend any conferences or industry-specific events that are beneficial for someone interested in [career path]?	11	6	0.15
	Are there any scholarships, grants, or funding opportunities available for attending career-related events or workshops in [career path]?	10	6	0.36
	How can attending events or workshops help me gain insights and network within [career path]?	6	2	0.33
	How can I make the most out of attending events or workshops to enhance my career prospects in [career path]?	3	3	0.33
	Can you suggest any resources or websites where I can find information about career-focused events and workshops in [career path]?	6	3	0.38

(Continued)

10.3389/frai.2024.1461137

TABLE 1 (Continued)

myIDP	Examples of prompts	Number of uses	Number of comments	Average sentiment (−1 to + 1)
Talk to people (Choose at least four prompts)	Where can I find people to ask questions about [career path] topics?	6	3	0.15
	What are some effective strategies for networking in [career path]?	12	5	0.46
	How can I identify and approach professionals in [career path] for informational interviews?	11	5	0.21
	What are some key questions to ask during an informational interview?	12	5	0.4
	How can I make a positive impression and build connections through networking events or online platforms?	8	4	0.22
	Can you provide tips on following up and maintaining relationships after networking events or informational interviews?	9	5	0.32
	How important is it to establish a personal brand or online presence for networking purposes?	6	4	0.09
	Are there any common networking mistakes to avoid?	7	3	0.21
	How can I leverage social media platforms for professional networking in [career paths]?	1	0	NA
	Are there any specific resources or websites to help me find networking opportunities in [career path]?	3	1	0.44
	Can you suggest some professionals in [career path] in [city/institution], with whom I can connect to gain insights into a career in [career path]?	5	2	0.36
Choose a career path (Required)	[Talk about your long-term goal first]. What transition experience do I need to reach this long-term goal? [You may have follow-up questions about the suggestions from AI]	18	8	0.26
Create plan				0.37
Career advancement goals (Choose at least one prompt)	As a [role], what are some SMART career advancement goals [in the next X years]?	11	7	0.23
	[Describe your current status]. This year, I want to [list the career advancement areas you want to improve this year]. Do you think my plan is feasible?	7	4	0.47
	[Describe your current status]. What are my SMART goals to improve [your target career advancement area this year]? [You may have follow-up questions about the suggestions from AI]	2	1	0.42
Skill goals (Choose at least two prompts)	How [SMART metric] is this goal for [identity] on a scale of 1 to 10 with 10 being unlikely? [Goal]	2	0	NA
	What are some SMART skill goals for [identity] in [concentration/major]?	6	4	0.43
	What are some SMART skill goals for [career path]?	8	4	0.48
	What are some SMART skill goals for becoming a [career path]?	9	2	0.57
	[Describe your current status]. This year, I want to [list the skills you want to improve this year]. Do you think my plan is feasible?	5	1	0.46
	[Describe your current status]. What are my SMART goals to improve [your target skill this year]? [You may have follow-up questions about the suggestions from AI]	8	3	0.35

(Continued)

10.3389/frai.2024.1461137

TABLE 1 (Continued)

myIDP	Examples of prompts	Number of uses	Number of comments	Average sentiment (−1 to + 1)
Project goals (Choose at least two prompts)	How [SMART metric] is this goal for [context] on a scale of 1 to 10 with 10 being unlikely? [Goal]	2	2	0.31
	What are some SMART project goals for [context] in [concentration/major]?	7	3	0.27
	What are some SMART project goals for [career path]?	10	6	0.32
	What are some SMART project goals for becoming a [career path]?	10	5	0.32
	[Describe your current status]. This year, I want to [list the project areas you want to improve this year]. Do you think my plan is feasible?	3	1	0.53
	[Describe your current status]. What are my SMART goals to improve [your project area this year]? [You may have follow-up questions about the suggestions from AI]	4	1	0.42
Implement plan				0.52
Mentoring team (Choose at least three prompts)	What are the key qualities or characteristics to look for in a mentor?	14	5	0.59
	How can I identify potential mentors who align with my career goals and interests?	7	3	0.39
	What are some effective strategies for approaching and initiating conversations with potential mentors?	7	2	0.64
	How can I evaluate whether a mentor is a good fit for my development needs?	10	3	0.56
	What are some common challenges in mentoring relationships and how can I address them proactively?	8	2	0.44
	What are the benefits of having a diverse mentoring team?	2	1	0
	How can I maintain and nurture my relationships with mentors over the long term?	4	2	0.4
	What are some effective ways to communicate my expectations and goals to my mentors?	7	0	NA
	How can I make the most of each mentoring session or interaction?	4	2	0.57
	[Describe your career goals or areas you want to improve]. Are there any specific industries or professional networks where I can find mentors in my field?	1	0	NA

Our research participants had the freedom to interact with AI mentors using the suggested prompts. To prevent overwhelming the participants or diminishing their motivation and interest in working with AI mentors, we asked them to test some of the prompts that they found useful when working on myIDP. Providing comments for each prompt was not required, which may result in the number of comments being smaller than the number of uses. Additionally, instructions in the first column, such as "Choose at least three prompts," were designed for the research participants. We encouraged readers to use any AI prompts in the table they found useful.

10.3389/frai.2024.1461137

participants were in the initial stages of their master's programs, while the others were at various stages of their doctoral studies. These students were pursuing careers across a broad spectrum of fields, aiming to make significant contributions in both the public and private sectors. The demographic composition of the participants reflected a focus on including underrepresented groups in STEM. Specifically, 15 of the participants identified as women or non-binary/genderqueer, and 4 identified as Black or Hispanic/Latino. A total of 16 participants were classified as underrepresented minorities, with the remaining two being Asian men. Additionally, eight participants were first-generation college students, and six were international students. In terms of familiarity with the IDPs, while only seven participants had prior experience, all received training on the myIDP tool before participating in the study. Background information of the participants can be found in Table 2.

2.2.2 Data analysis

Participants' comments to selected prompts around Assessment, Career Exploration, Create Plan, and Implement Plan were assessed with sentiment analysis using VADER from the nltk Python package. This sentiment analysis method is a supervised program of mapping text-by-word tokenization to the speakers' feelings and emotions (Devika et al., 2016). Comments were tokenized into individual words and assigned a score from −1 being most negative, 0 being neutral, and + 1 being most positive as they have been previously defined in the VADER lexicon. The lexicon is pre-trained and provided openly by the ntlk Python package.

Due to the limited number of responses in our sample, we examined the aggregate sentiment to have a loose idea of how participants favored our prompts. Through running our sentiment analysis on participant comments, we noticed the algorithm has approximately 90% accuracy in rating the sentiment of the comments.

TABLE 2 Participants' background information.

ID	Degree	Year	Career goal	Gender	Race/ethnicity	Age	First-gen	International student	IDP experience
S1	Master's	First year	Biotechnology industry scientist	Woman	Multiracial	18–23	No	No	No
S2	Master's	First year	Government conservation organization	Nonbinary / gender queer	Hispanic or Latino	18–23	Yes	No	No
S3	Doctoral	sixth year	Postdoctoral fellowship	Woman	White	30–39	No	No	Yes
S4	Doctoral	Third year	Occupational therapist	Man	Black	24–29	Yes	No	No
S5	Doctoral	Third year	Academia	Woman	White	24–29	No	No	Yes
S6	Doctoral	Fourth year	Material scientist	Man	Asian	24–29	Yes	Yes	Yes
S7	Master's	First year	Fertility or medicine	Woman	Asian	18–23	No	No	Yes
S8	Master's	First year	Software engineering	Man	Asian	24–29	Yes	Yes	No
S9	Master's	First year	Public health related	Woman	Asian	30–39	Yes	Yes	No
S10	Doctoral	Fourth year	Doctor	Woman	Black	24–29	No	No	No
S11	Doctoral	First year	Community College Instructor	Woman	White	30–39	No	No	Yes
S12	Doctoral	Second year	Private industry or NGO	Woman	Black	30–39	No	No	Yes
S13	Doctoral	Third year	Private industry	Woman	White	30–39	Yes	No	Yes
S14	Doctoral	Third year	Clinic work and academia	Woman	White	24–29	Yes	No	No
S15	Master's	First year	Ecology research	Nonbinary / gender queer	White	18–23	No	No	No
S16	Master's	Second year	Counselor	Woman	Asian	24–29	No	Yes	No
S17	Master's	Second year	Counseling psychologist	Woman	Asian	24–29	Yes	Yes	No
S18	Master's	Third year	Helping professionals work	Woman	Asian	24–29	No	Yes	No

Only the categorical impressions, the average of all prompt aggregate sentiments within the category are mentioned below, but we provide the more detailed sentiment analysis results in Table 1. Results were analyzed in conjunction with a word cloud to better understand the most pertinent comments. In addition to the sentiment analysis, participants' comments were analyzed using the thematic analysis approach to identify recurring and important perspectives.

Participant responses regarding the strengths and concerns of AI mentors were also analyzed using the qualitative data analysis software packages, Atlas.ti and MAXQDA. A thematic analysis approach was employed to systematically identify and categorize recurring themes within the data. The process began with familiarization, where researchers reviewed all responses to gain a comprehensive understanding of the content. Key points and significant statements were then identified and coded, marking relevant segments of the data. These codes were grouped into potential themes, reflecting broader patterns and advantages and concerns identified by the participants. Themes were subsequently reviewed and refined to ensure they accurately represented the data. Finally, clear names were assigned to each theme to encapsulate their meaning as understood by the researcher. This methodological approach provided an in-depth analysis of participant responses, highlighting both the strengths and concerns associated with their engagement with AI mentors.

To ensure the reliability of our thematic coding, we employed a multi-coder approach, enhancing the reliability of our findings through intercoder agreement (Creswell, 2012). Each coder independently developed a set of codes through team discussions and reflexive reading of the participant responses. These codes were then synthesized into overarching themes, which highlighted higher-level advantages and disadvantages. The team collectively reviewed and discussed these themes to gain a comprehensive understanding of the potential use of AI for mentoring purposes. This collaborative process aimed to minimize individual bias and enhance the reliability of our qualitative analysis.

3 Results

3.1 AI prompts

3.1.1 Assessment

Comments for the assessment prompts were relatively few. The assessment categorical sentiment was moderately positive (0.37). Most participants had somewhat positive experiences using the prompts to clarify certain concepts. The participants found it helpful in explaining terms and also in providing useful insights on what they asked the AI mentor. For example, one participant asked the AI mentor the meaning of "keeping up with current events" and thought that the information was not only helpful but also provided useful insights as a growing researcher.

3.1.2 Career exploration

Overall, participants reported positive experiences with the prompts for career exploration. Career categorical sentiment was the least positive (0.26). Among the comments, many of them affirmed that the AI mentor could provide helpful feedback on their career fit by using the prompt, "What are some ways to know if I would be a good [job title]?" More than one participant thought that the responses from the AI mentor were thorough.

Participants also shared slightly positive experiences (0.22) with asking AI mentors questions regarding the prompts of "Read about careers." For example, by using the prompt "What is a typical workday as a [job title]?" The AI mentor could provide useful information about participants' interested career paths. Participants also commented that the AI mentor's responses were similar to their past experiences. One participant mentioned that the AI mentor provided different examples of workplaces and workday circumstances, which was helpful. However, some participants maintained skeptical attitudes against the responses from the AI mentor based on their understanding of the career path.

When asking the AI mentor questions about "attending events and workshops," participants found that the AI mentor often lacked specific information for their local areas, though still provided some useful items for consideration (0.29). One of the participants commented that "I first asked [annual or regional events] about a specific city and Gemini was not able to generate the events." Sometimes, if the question did not include the current year (i.e., 2024) for the conference or events, Gemini provided information for the previous year (i.e., 2023). Also, across different prompts, some participants found that the AI mentor was not providing related links regarding job-searching websites, resources, or scholarship opportunities. One participant noted, "having links to the sites would be better so I do not have to do another search." Another negative comment was that the AI mentor tended to provide generic responses, ignoring the specific information participants indicated in their prompts. For example, when asking, "What are some potential career growth opportunities or advancement prospects in [career path]?" One participant claimed that "I was hoping it would list some internships, but it instead was very broad."

3.1.3 Create plan

For the prompts within "Create Plan," most of the participants provided moderately positive feedback on the prompts for SMART goals (0.37), including SMART career advancement goals (0.33), skill goals (0.45), and project goals (0.33). Participants mentioned that the AI mentor was able to provide examples of SMART goals, which they would like to add to their plans. For example, when asking questions such as "what are some SMART project goals for becoming a senior software engineer," the participant commented that "The answer helps me understand, in order to achieve my career goal, what to do at the moment." However, one participant found that the AI mentor did not recognize what SMART goals are. Also, another participant thought that the SMART goals suggestions from the AI mentor were poor due to "few insights of the industry."

3.1.4 Implement plan

Participants provided the most positive feedback for prompts related to seeking mentors (0.52). They found value identifying key qualities of mentors, questions to be asked when seeking mentors, and how to interact with potential mentors. For example, several participants appreciated the AI mentor's response to the question about "key qualities to look for in a mentor position," which helped them understand how to identify a good mentor. Participants also learned some critical aspects to consider when seeking human mentors, such as mentoring philosophy. However, one participant mentioned that the response for the mentor evaluation prompt was too generic and lacked depth.

3.1.5 Additional feedback from international students

3.1.5.1 US-centric responses

When asking questions, it's crucial to specify the country to avoid receiving US-centric answers. For example, even when using a Virtual Private Network (VPN) service, the responses often defaulted to U.S. users. Indicating the country name, even in the local language, may result in more accurate and contextually relevant responses.

3.1.5.2 Career information accuracy

A participant sought career guidance for Country A, but the AI mentor provided information and resources from Country B due to the two countries sharing a similar language. Additionally, some career information, such as licensing exams, is often incorrect for countries outside the U.S.

3.1.5.3 Cultural sensitivity in communication

The suggested email templates for reaching out to potential mentors can sometimes be too aggressive and fail to consider cultural differences. This can be particularly problematic in some countries where a more formal or respectful approach is preferred.

3.2 Strengths of AI mentoring

3.2.1 Immediate response

Most of the participants appreciated the immediate responses from the AI mentor. Compared with human mentors, the AI mentor allowed them to navigate through the myIDP process without waiting for unclarified questions. Several participants shared their past experiences of waiting for responses from their human mentors during the process, which were in strong contrast to experiences with the AI mentor.

3.2.2 Up-to-date information

Some participants affirmed that the AI mentor could provide up-to-date information. Nevertheless, some participants also mentioned some limitations. For example, the AI mentor was not able to provide information regarding upcoming or local events. One participant also acknowledged that the AI mentor is not always trained on up-to-date datasets, which caused its lack of knowledge on the most recent event. On the other hand, two of the participants mentioned that the information provided by the AI mentor was the same as that of their human mentors. Although the participants recognized this as not up-to-date information, it showed that the AI mentor shared the same knowledge as human mentors.

3.2.3 Access to multiple AI mentors

In Google Gemini, users can choose different versions of responses. Four participants thought this function was helpful for them to gain various points of view on the same questions. However, seven participants thought that the AI mentor was just rephrasing the same concepts, and the responses looked generally the same. One participant also expressed concerns about creating biases by frequently selecting preferred versions of responses ("I think it is neat that it provides multiple answers, but I feel nervous it would play into my own biases if I kept refreshing for one that I liked better.").

3.2.4 Enhanced ownership of career development

Nine participants expressed positive experiences with taking ownership when they interacted with the AI mentor. One participant expressed that the process of interacting with the AI mentor allowed mentees to take the lead in the direction of the conversation. Another participant also mentioned that unlike human mentors, the AI mentor allowed users to ask many questions whereas that person may not feel quite as comfortable asking their human mentor 20 random questions in a row. One participant stated that by interacting with the AI mentor, it increased mentee's confidence because the AI mentor did not tell the participant what to do but only provided recommendations. In contrast, three participants expressed that they did not feel "ownership" during the process. One participant argued that the information provided by the AI mentor can also be acquired through Google search.

3.2.5 Time savings

When it comes to the advantage of affecting time for human mentors and mentees, most of the participants agreed that the AI mentor helped them save time from gathering information on their own. One participant mentioned that using the AI mentor can help narrow down questions for human mentors that AI could not answer specifically.

3.2.6 Other

The feedback from the participants addressed additional advantages of using the AI mentor. First of all, the flexibility of the AI mentor. Six participants praised how flexible the AI mentor is during the process. Users can ask anything without hesitation. In addition, for students who do not have human mentors or who feel uncomfortable talking to human mentors, AI mentors can be helpful. Second, the AI mentor can help bridge the mentorship between mentees and human mentors. Some participants suggested that the responses from the AI mentor can serve as a starting point for thinking of questions to ask their human mentors or to help spark an idea. Third, the process of using the AI mentor enhanced their myIDP experience. Participants mentioned that the process promoted them to review their myIDP components and help them set career goals.

3.3 Concerns of AI mentoring

3.3.1 Misinformation

Participants generally felt that the AI mentor provided accurate information, with the majority noting its overall reliability. Specifically, participants mentioned that the AI was overall accurate, and others noted that it was mostly accurate with occasional inaccuracies or minor contradictions. For example: "When I asked about the demand for neuroscience researchers, Gemini first said that the demand was high, but then a couple of paragraphs later, said that there are more PhD graduates than there are faculty positions (which is true)." However, some participants felt that the AI's advice was too general to be of significant use. One participant noted the AI's consistency with guidance from human mentors, and another found the AI to be a good starting point for further exploration and research.

3.3.2 Bias

Gemini's learning process relies on real-world data, which may contain inherent bias. As a result, its responses sometimes reflect stereotypes and discriminatory information. For instance, when developing prompts for the research project, our team observed a disparity in the suggestions for STEM career paths provided to doctoral students based on gender. When offering career advice to a mom, the AI mentor emphasized factors like "flexibility," "balancing family responsibilities," and "supporting working moms." On the other hand, when advising a dad, the AI mentor focused on more general aspects such as "interests and passions," "family situation," and "financial goals."

Participants did not report any instances of discriminatory or biased responses from the AI mentor. However, there were some general concerns about the potential for bias listed here and in other sections. One participant noted that while a human mentor might exhibit more bias based on physical appearance, "technology has advanced tremendously and I did just give a lot of my personal information through these questions so it could be building a profile of what they think I am." This raises concerns about data usage for algorithmic purposes, creating the potential for bias. Another participant expressed a wish to have asked more questions about their career goals in relation to their chronic health disability, to see if the AI could offer relevant advice or if it might show ableist tendencies. These reflections show participant worries about bias and data privacy, even in the absence of explicit biased responses from the AI.

3.3.3 Privacy

Participants expressed a range of feelings about using AI mentors, particularly concerning privacy. Some participants felt there was a balance between privacy concerns and the personalization available to them while using AI mentors, while others were cautious about sharing personal information with AI mentors. One participant compared their lack of privacy concerns to their usage of social media, indicating a similar level of comfort with information they shared on social networking and social media platforms with what the AI platform knew about them. Conversely, some participants were worried about the transparency of data storage and usage, with some expressing general concerns about data privacy. One participant summed up the contradiction between needing to supply personal information to an AI mentor for useful feedback and privacy concerns in this way: "I do worry about privacy and selling data. I think turning off tracking (or limiting it to the session) could help. On the one hand, Gemini cannot get to know me like a human mentor could and provide advice accordingly. On the other hand, I do not want to give Gemini enough personal information for it to give specific advice."

3.3.4 Equity

Participants' responses revealed several challenges and perspectives regarding access to AI mentors within marginalized communities. The predominant concern was internet and technology access, highlighting the digital divide as a significant barrier to AI mentor usage. Some participants highlighted the continued importance of human mentors, suggesting that despite the advantages of AI, the value of human connection and personalized guidance remains critical. Another important issue raised was internet literacy, as not everyone currently possesses the skills to evaluate online information effectively. "Any lack of internet literacy may lead someone to believe anything they read online. I also think that those who are better able to communicate with AI are more likely to get more accurate answers." Language barriers were also noted by one participant as a potential obstacle. One participant emphasized the necessity for better access for people with disabilities, indicating a need for inclusive design. Interestingly, one participant felt that AI mentors could possibly be more accessible to marginalized individuals than human mentors, offering a unique viewpoint on the potential benefits of AI mentorship. Overall, while AI mentors might offer potential benefits, participants noted significant challenges such as technological barriers including access and literacy, as well as inclusivity which need to be addressed to ensure equitable access for all communities.

3.3.5 Algorithmic influence

Participants expressed various concerns about the unknown working nature of AI, but many chose not to answer the question or did not understand the question, suggesting a lack of familiarity with AI. Three participants viewed AI as a complementary tool to human mentors, emphasizing its supplementary role, and another three called for more transparency in AI technology development. Concerns about the vulnerability of less experienced users were noted by two participants, while three participants worried about the potential for AI to provide false information.

3.3.6 Mentoring support

Participants overwhelmingly agreed that AI cannot replace real human mentoring, though many recognized that AI could serve as a valuable supplement to human mentoring. The efficiency of AI, offering quick and accessible guidance, was noted as a significant benefit. Some participants also saw AI as a great starting point for career development conversations, particularly for youth. When access to human mentors is limited, one participant also noted that AI mentoring can also be particularly useful.

3.3.7 Other

When participants were given an open-ended opportunity to address personal concerns about the use of AI mentors, one unique concern was becoming over reliant on technology, potentially leading to a loss of soft skills like communication, mentoring, and connection. Some participants emphasized the need for training in AI usage, such as tutorials or guidance on how to use AI effectively as a mentor in order to address literacy concerns and help mentees treat AI as a tool and not as a replacement for human mentors.

3.4 Human-AI mentoring

Two hybrid human-AI mentoring models emerged from participants' input: the *Sequential Integration Mentorship Model* and the *Concurrent Collaboration Mentorship Model*. These models leverage the unique strengths of both human and AI mentors, creating a synergistic approach to mentoring that enhances the developmental process for mentees.

3.4.1 Sequential integration mentorship model

This hybrid model organizes the mentorship process into distinct, sequential phases, leveraging the unique strengths of human and AI mentors at different stages. Initially, *AI as Initial Point of Contact* engages with mentees by offering broad, general perspectives and foundational knowledge on the mentee's career path. This is particularly beneficial when direct human mentorship is not immediately available, ensuring no delay in the mentee's developmental process. Following this groundwork, *Transition to Human Mentorship* occurs, with human mentors stepping in to deliver detailed, personalized guidance tailored to the specific needs identified during the initial phase. This seamless handover ensures continuity and depth in mentoring. After each mentorship phase, *IDP Refinement Feedback* is provided by human mentors based on their interactions with mentees. This feedback is leveraged to *enable AI to offer more customized information*, specifically aimed at mapping out a more tailored development plan. This iterative improvement process ensures that the AI's contributions are finely attuned to the evolving needs and goals of mentees, enhancing the precision and effectiveness of the mentorship.

3.4.2 Concurrent collaboration mentorship model

This hybrid model features continuous and simultaneous collaboration between human and AI mentors throughout the entire mentorship process. This model fosters a dynamic integration of efforts, with AI and humans contributing in real-time without clear boundaries between their roles. AI provides *Immediate Assistance* (rapid, general advice, quick feedback, and emergency support), managing real-time information flow and addressing straightforward queries efficiently. Additionally, AI compiles resources, crafts structured plans, sets SMART goals, and stimulates creative thinking for career and mentorship development. On the other hand, human mentors offer *Personalized and Emotional Support*, utilizing their real-world experience to address complex personal or professional challenges, ensuring the mentorship is empathetic and practical. Human mentors also *Contextualize AI Data* by interpreting and adjusting AI-generated data and recommendations to fit the unique contexts of each mentee. This *Synergistic Interaction* between human and AI mentors enriches the mentorship process by providing a holistic view, diversifying resources, and cross-referencing each other's inputs to ensure comprehensive development. By leveraging both AI and human insights, the mentorship experience continuously evolves and adapts to meet the mentees' needs effectively.

4 Discussion

In light of recent mandates from the U.S. National Science Foundation (NSF) under the CHIPS and Science Act of 2022, which require IDPs for all NSF-funded graduate students (National Science Foundation, 2024), this study aims to optimize career development processes using IDPs. It explores integrating AI support within the myIDP framework to empower STEM graduate students, particularly those from underrepresented minority groups, by providing personalized career guidance. To achieve this, the research team developed a series of AI prompts tailored for use within myIDP. Eighteen STEM graduate students, mostly from underrepresented minority backgrounds, were trained to interact with AI mentors. Their interactions, along with comments and feedback, were analyzed using sentiment and thematic analysis, shedding light on the strengths and concerns associated with AI mentorship (Figure 1). The findings also suggest two hybrid models for human-AI collaborative mentoring, where both agents work synergistically to provide personalized guidance throughout the IDP journey.

4.1 Effectiveness of AI mentoring

Based on the comments for each prompt from the participants, we found that most participants had positive experiences when using the prompts for their myIDP process (Figure 2). However, AI responses may have been perceived as less helpful in some categories than others because of nuanced areas requiring more human interactions beyond current capacities of AI models. Further, some participants commented that the responses from the AI mentor were rather generic. This is often due to the limited information provided in a short prompt. When using the AI mentor, it is important to ask follow-up questions to obtain more contextually accurate or sophisticated responses. Some participants' experiences did improve by asking further questions to the AI mentor, specifically mentioning that the AI mentor recommended contextual information in later prompts. Nevertheless, it is also important to acknowledge that these are highly likely one of the limitations of the AI mentor at the stage of AI development during this study.

We also found that certain prompts were selected more often by our participants (Table 1). The tendency likely stems from the attributes of requesting information. Popular prompts such as "What is a typical workday as a [job title]?" and "What is the average salary for a [job title] in [career path] this year?" are the ones that require up-to-date information. We hypothesized that these would be the types of questions the AI mentor performs well with, however, sentiment analysis gave us more insight into which prompts were generally more useful. Prompts such as "[Talk about your top values first]. Could I pursue a career in [career path]?" and "What are the key qualities or characteristics to look for in a mentor?" yielded highly positive sentiments among participants. On the other hand, some prompts were rarely selected such as "Are there any notable trends or emerging areas within [career path] that might impact future job prospects?" or "How can I leverage social media platforms for professional networking in [career paths]?"

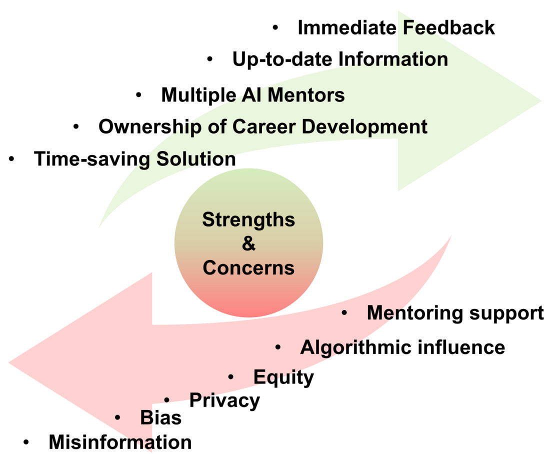


FIGURE 1
Strengths and concerns of using AI mentors.

FIGURE 2
Word cloud of feedback from participants after using AI prompts to support career development.

When thinking of the strengths of using the AI mentor, flexibility is the most significant one. Most participants appreciated the immediate response from the AI mentor, the flexible time, and the free space it provided to ask questions. Mentees do not need to wait for responses when they encounter questions during the myIDP process. They can ask numerous questions to the AI mentor without hesitation. The flexibility of the AI mentor not only provides a convenient approach for mentees to solve their questions or obtain information, but also enhances their experiences when developing career plans. Another major advantage of using the AI mentor is that it can serve as a brainstorming tool during the myIDP process. Their comments and feedback suggest that using the AI mentor helped them rethink their original plan or provided a feasible starting point through details and perspectives that were previously missed. By exploring information and setting SMART goals with the AI mentor, it enhanced their myIDP process by clarifying terms in myIDP or providing sparkling ideas about their career development plan. In turn, this information can allow a human mentor to spend more time focusing on the nuanced and abstract aspects of career development during meetings.

Integrating AI mentors into graduate education presented several concerns for participants, including the accuracy and reliability of AI-generated advice, which often requires human feedback for accuracy and contextual appropriateness. Privacy and data security issues also demand policies and transparent data usage to build trust. While AI can offer quick, data-driven insights, there was a fear among participants that using AI in this way could lead to an overreliance on technology and a failure to maintain vital soft skills like communication. There is also a need for ongoing monitoring of the quickly growing field of AI, coupled with user training on AI's capabilities and limitations. All these concerns led many participants to conclude that AI mentoring was best seen as a tool not as a replacement for human mentors. Their comments pointed toward an approach that fosters collaboration between human and AI mentors, leveraging the strengths of both to enhance the mentoring process.

4.2 Ethical and equity implications of AI mentoring

The integration of AI into graduate education mentorship also raises significant ethical and equity concerns that must be considered and addressed to ensure fair and inclusive outcomes. One of the primary ethical challenges is the concern for algorithmic bias, where AI systems may inadvertently perpetuate existing societal biases, leading to unequal treatment of mentees based on race, gender, socioeconomic status, or other marginalized identities. This bias could manifest in AI-generated advice that disproportionately favors certain groups while disadvantaging others, exacerbating existing inequalities within academic and professional environments. Additionally, there is a risk that AI mentors may lack cultural sensitivity and fail to account for the unique experiences and needs of diverse mentees, further widening the gap in mentoring quality. Some participants noted that the responses failed to incorporate geographically diverse answers to the prompts, particularly international recommendations. To mitigate these risks, it is essential to implement bias detection and correction mechanisms within AI systems, ensuring that they operate fairly and equitably. More importantly, to remind users to carefully assess the answers and not take each answer for granted, such as bringing the answers to discuss with a more experienced human mentor. Ethical oversight and transparency in AI development and deployment are crucial, allowing for continuous evaluation and improvements of these systems. Ensuring that AI mentors are accessible to all students, regardless of their background, and that they complement rather than replace human mentors, can help create a more equitable and inclusive mentoring environment.

4.3 Hybrid human-AI mentoring models

Our study addresses a critical gap in understanding the integration of human mentors and generative AI into the STEM career planning process. The findings reveal two conceptual human-AI mentoring models: the Sequential Integration Mentorship Model and the Concurrent Collaboration Mentorship Model. The Sequential Integration Mentorship Model (Figure 3) organizes the mentorship process into distinct, sequential phases. Initially, AI serves as the initial point of contact, engaging with mentees by offering broad, general perspectives and foundational knowledge on their career path. This phase is particularly beneficial for beginners, such as first-year graduate students, who may lack fundamental knowledge and feel pressured to ask questions. AI provides an opportunity for self-exploration without the immediate need for human interaction. Following this groundwork, there is a transition to human mentorship, where human mentors step in to deliver detailed, personalized guidance tailored to the specific needs identified during the initial phase. This transition ensures continuity and depth in mentoring. After each mentorship phase, IDP refinement feedback is provided by human mentors, enabling AI to offer more customized information specifically aimed at mapping out a more tailored career development plan. This model aligns perfectly with the myIDP framework, where self-assessment and self-reflection are supported by AI, followed by human mentors providing further feedback and customized updates. It is ideal for large programs with high mentee volumes or for mentees needing foundational knowledge before personalized guidance. However, a potential weakness is the disconnect that may occur between the initial AI guidance and later human mentorship, which may not be ideal for mentees who require ongoing support.

The Concurrent Collaboration Mentorship Model (Figure 4), on the other hand, features continuous and simultaneous collaboration

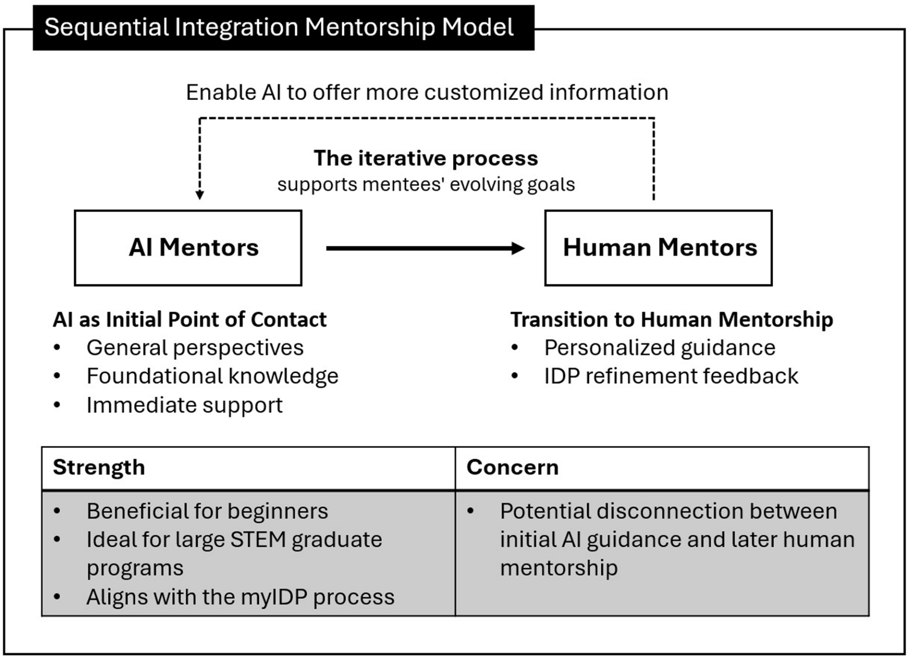


FIGURE 3
Sequential Integration Mentorship Model showing human-AI mentorship flow.

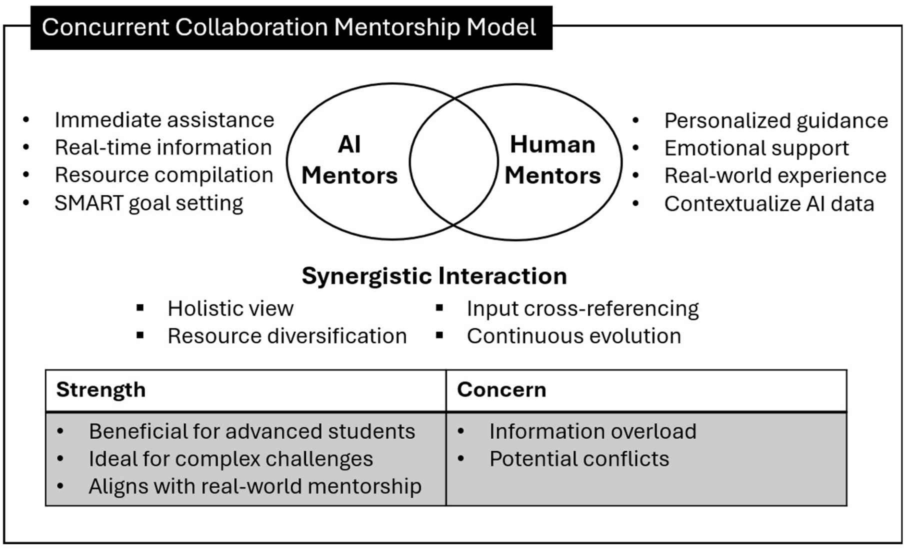


FIGURE 4
Concurrent Collaboration Mentorship Model showing human-AI collaborative mentorship flow.

between human and AI mentors throughout the entire mentorship process. This model mirrors real-world settings where students often work with multiple mentors, each providing different types of instrumental and psychological support (Eby et al., 2013; Saw et al., 2022). AI offers immediate assistance, compiling resources, crafting structured plans, setting SMART goals, and stimulating creative thinking for career and mentorship development. Human mentors provide personalized and emotional support, utilizing their real-world experience to address complex personal or professional challenges, ensuring the mentorship is empathetic and practical. They also contextualize AI guidance, interpreting and adjusting AI-generated recommendations to fit the unique contexts of each mentee. By leveraging both AI and human insights, the mentorship experience continuously evolves and adapts to meet the mentees' needs effectively. The dynamic integration of efforts allows for real-time support, diverse resources, and perspectives, fostering holistic development through human-AI interaction. This model is particularly effective for students with a basic understanding of their field, such as second-year students and beyond, and for complex challenges requiring real-time support and diverse resources. However, there is a risk of information overload for mentees and potential conflicts between the guidance provided by human and AI mentors.

These models provide a structured approach to leveraging both human and AI strengths, offering new perspectives on blended mentorship practices for developing personalized IDPs. Our findings highlight the potential of these models to make mentorship more scalable and accessible, particularly in contexts where human mentors are in short supply or where underrepresented minorities require career development support without sufficient resources. These models underscore the importance of this AI integration with mentorship to enhance the developmental process for mentees, ensuring that mentorship is both comprehensive and adaptive to individual needs.

4.4 Unique challenges faced by international STEM students

The current myIDP platform's U.S.-centric design inadequately serves the unique needs of international STEM graduate students, whose career paths are often obstructed by language barriers, cultural differences, social isolation, and restrictive visa conditions (American Council on Education, 2021; Lee, 2020; Rodriguez et al., 2024). Such challenges not only hinder their personal and career development but also potentially weaken the broader U.S. innovation landscape. Given the vital role that international STEM talent plays in driving a robust U.S. economy (National Science Board, 2022), there is an immediate need to reevaluate the myIDP framework to be more inclusive, culturally relevant, and globally responsive, especially by integrating AI support.

The study uncovered significant insights from international student participants. First, the AI responses are often US-centric unless the country is explicitly specified in the questions. This issue persists even when using VPN services to access AI outside of the U.S. Second, the accuracy of career information provided by AI mentors can be problematic, as some information may be incorrect for countries outside the US or for countries sharing similar languages. Third, cultural sensitivity is a critical issue. The advice from AI mentors sometimes fails to consider cultural differences. Addressing these issues could enhance the effectiveness of AI mentors and better support the diverse needs of international STEM students, thereby improving their career outcomes and contributing to the broader academic and professional community.

4.5 Limitations and future directions

There are several limitations to our study. First, AI tools are evolving very rapidly. Although we identified several concerns about AI mentors, these concerns are based on findings during the period when participants tested the AI prompts (January 2024–May 2024). These concerns might be addressed in future advancements. Second, prompt engineering is a significant area of research focused on optimizing AI responses. The suggested AI prompts used in this study might need to be updated due to the ongoing evolution of AI tools. Since prompt engineering was not our focus, we provided only basic guidance to student participants, allowing them to freely interact with AI mentors. This approach enabled us to gather diverse insights from their interaction experiences. Future research could explore how prompt engineering can enhance the interaction process with AI mentors.

Third, to promote educational equity, we avoided using paid versions of AI tools that could pose a barrier for minority students. Thus, our study utilized the free AI tool, Gemini, to test its potential in the STEM career development process. Our findings are specific to Google Gemini. Other paid AI tools (such as GPT-4), which also have internet access, may be worth examining in future studies. Fourth, our study focuses solely on the mentee's perspective. Future studies could focus on the perspectives of mentors. Lastly, our study proposed two conceptual human-AI mentoring models based on student feedback. Further research is needed to evaluate the effectiveness of these models in real-world settings. Potential areas for improvement include enhancing AI personalization capabilities and investigating the impact of human-AI interactions on mentor-mentee relationships. Such research could provide valuable insights into optimizing these models to better support the diverse needs of students and improve overall mentoring outcomes in developing IDPs.

5 Conclusion

This study makes several theoretical, methodological, and practical contributions to the literature on IDPs, career development, and mentoring in higher education. Theoretically, it proposes integrating AI technology into career planning and higher education mentoring, moving beyond traditional human-centered theories. Two conceptual human-AI mentoring models, the Sequential Integration Mentorship Model and the Concurrent Collaboration Mentorship Model, are introduced. Methodologically, our study develops an AI-integrated myIDP framework by incorporating prompt submissions to Google Gemini. Feedback from student participants highlighted its strengths and limitations through thematic analysis, while sentiment analysis assessed the usefulness of each AI prompt. Practically, this study demonstrates the promising approach of using Google Gemini to optimize IDP practice (Figure 5), providing immediate feedback and information to empower students in their career development and transcend the limitations of human mentors. The proposed hybrid human-AI mentoring models show potential in supporting more underrepresented minority students in their STEM career development process, promoting broader participation in STEM fields. These models could be further examined in real-world settings in the future.

Despite the strengths discussed, users must be aware of concerns regarding accuracy, bias, privacy, equity, and algorithmic influence. To enhance this process, we encourage graduate students to reassess their SMART goals and action plans with their human mentors for personalized support. The sentiment analysis shows that there are clear areas of effectiveness in using AI for mentorship, though further research must explore ways to improve AI effectiveness. While AI technology can benefit the career planning process within IDPs, human mentors remain vital for providing comprehensive support during plan implementation, encompassing both instrumental and

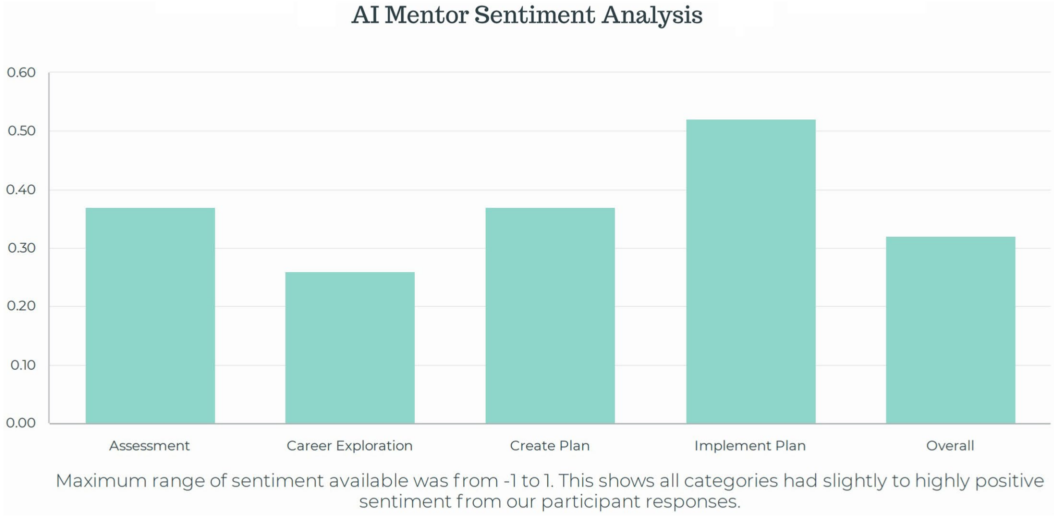


FIGURE 5
Sentiment scores for AI prompts across myIDP components.

psychosocial aspects. Therefore, AI technology should supplement, not replace, the essential role of human mentors in the mentoring process. Future research may also investigate the optimal balance between AI and human mentorship to bolster career development experiences.

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 Institutional Review Board at the OVPRI (Office of the Vice President for Research and Innovation) Human Research Protection Program, Virginia Commonwealth University. 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

C-NC: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. JH: Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. CJ-S: Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. T-WW: Formal analysis, Investigation, Methodology, Validation, Visualization, 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. The authors gratefully acknowledge the research funding provided by the School of Education at Virginia Commonwealth University.

Acknowledgments

Our study used Google Gemini for data collection exclusively. However, during the design phase, Google Bard was renamed to Google Gemini. The versions have been continuously updated from November 2023 throughout our final phase of data collection June 2024.

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

Akiba, D., and Fraboni, M. C. (2023). AI-supported academic advising: exploring ChatGPT's current state and future potential toward student empowerment. *Educ. Sci.* 13:885. doi: 10.3390/educsci13090885

American Council on Education. (2021). Career pathways for international students. Available at: https://www.immigrationresearch.org/system/files/Career%20Pathways%20for%20Internation

Atkins, K., Dougan, B. M., Dromgold-Sermen, M. S., Potter, H., Sathy, V., and Panter, A. T. (2020). "Looking at myself in the future": how mentoring shapes scientific identity for STEM students from underrepresented groups. *Int. J. STEM Educ.* 7:42. doi: 10.1186/s40594-020-00242-3

Chang, C. N., Justus-Smith, C., Malagon-Palacios, L. J., and Hui, J. (2023). The use of the individual development plan at minority serving institutions. *Front. Educ.* 8, 1–7. doi: 10.3389/feduc.2023.1258273

Chang, C.-N., Patterson, C. A., Vanderford, N. L., and Evans, T. M. (2021). Modeling individual development plans, mentoring support, and career preparedness relationships among doctor of philosophy (Ph.D.) trainees in the life sciences. *F1000Res* 10:626. doi: 10.12688/f1000research.53705.2

ChatGPT Generative Pre-trained TransformerZhavoronkov, A. (2022). Rapamycin in the context of Pascal's wager: generative pre-trained transformer perspective. *Oncoscience* 9, 82–84. doi: 10.18632/oncoscience.571

CHIPS and Science Act, Public law U.S.C. § 117 et seq. (2022). Available at: https://www.congress.gov/bill/117th-congress/house-bill/4346/text (Accessed Mar 15, 2024).

Creswell, J. W. (2012). Qualitative inquiry and research design: Choosing among five approaches. *3rd* Edn. Los Angeles, CA: SAGE Publications, Inc.

Dehouche, N. (2021). Plagiarism in the age of massive generative pre-trained transformers (GPT-3). *Ethi. Sci. Environ. Politic.* 21, 17–23. doi: 10.3354/esep00195

Devika, M. D., Sunitha, C., and Ganesh, A. (2016). Sentiment analysis: a comparative study on different approaches. *Procedia Comp. Sci.* 87, 44–49. doi: 10.1016/j.procs.2016.05.124

Duan, J., and Wu, S. (2024). Beyond traditional pathways: leveraging generative AI for dynamic career planning in vocational education. *Int. J. New Develop. Educ.* 6, 24–31. doi: 10.25236/ijnde.2024.060205

Eby, L. T., Allen, T. D., Hoffman, B. J., Baranik, L. E., Sauer, J. B., Baldwin, S., et al. (2013). An interdisciplinary meta-analysis of the potential antecedents, correlates, and consequences of protégé perceptions of mentoring. *Psychol. Bull.* 139, 441–476. doi: 10.1037/a0029279

Fuhrmann, C. N., Hobin, J. A., Lindstaedt, B., and Clifford, P. S. (n.d.). myIDP science careers: individual development plan. Available at:https://myidp.sciencecareers.org

Hobin, J. A., Fuhrmann, C. N., Lindstaedt, B., and Clifford, P. S. (2012). Science. AAAS. Available at:https://www.science.org/content/article/you-need-game-plan

Köbis, L., and Mehner, C. (2021). Ethical questions raised by AI-supported mentoring in higher education. *Front. Art. Intell.* 4:624050. doi: 10.3389/frai.2021.624050

Lee, C. S. (2020). Post-degree completion plan of Chinese and Korean graduate students in STEM. *J. Comp. Int. High. Educ.* 11, 131–134. doi: 10.32674/jcihe.v11iWinter.1550

Li, C. (2020). OpenAI's GPT-3 language model: a technical overview. Lambda Labs. Available at: https://lambdalabs.com/blog/demystifying-gpt-3

Lu, J., Zheng, R., Gong, Z., and Xu, H. (2024). Supporting teachers' professional development with generative AI: the effects on higher order thinking and self-efficacy. *IEEE Trans. Learn. Technol.* 17, 1279–1289. doi: 10.1109/tlt.2024.3369690

National Academies of Sciences, Engineering, and Medicine (2020). The science of effective mentorship in STEMM. Washington, DC: National Academies Press.

National Science Board. (2022). International STEM talent is crucial for a robust U.S. economy (NSB-2022-16). Available at: https://www.nsf.gov/nsb/sei/one-pagers/NSB-International-STEM-Talent-2022.pdf

National Science Foundation. (2022). Strategic goals and objectives. National Science Foundation: Where discoveries begin. Available at: https://www.nsf.gov/news/special_reports/strategic_plan/empower.jsp

National Science Foundation (2024). Summary of changes to the PAPPG (NSF 24–1). Available at: https://new.nsf.gov/policies/pappg/24-1/summary-changes

Neumann, A. T., Arndt, T., Köbis, L., Meissner, R., Martin, A., de Lange, P., et al. (2021). Chatbots as a tool to scale mentoring processes: individually supporting self-study in higher education. *Front. Artif. Intell.* 4:668220. doi: 10.3389/frai.2021.668220

Okado, Y., Nye, B. D., Aguirre, A., and Swartout, W. (2023). “Can virtual agents scale up mentoring?: insights from college students' experiences using the careerfair.AI platform at an American Hispanic-serving institution” in Lecture notes in computer science (Switzerland: Springer Nature), 189–201.

Rodriguez, M., Roman, B. Z., Mohamed, M., and Barthelemy, R. (2024). Social and cultural barriers reported by STEM international graduate students of color. *J. Int. Stud.* 14, 276–302. doi: 10.32674/jis.v14i3.6694

Sagiroglu, S., and Sinanc, D. (2013). Big data: A review. 2013 International Conference on Collaboration Technologies and Systems (CTS), 42–47.

Saw, G. K., Chang, C.-N., Lin, S., Hernandez, P. R., and Culbertson, R. (2022). Disparities in mentoring and mental health problems of U.S. college students in science and engineering during the COVID-19 pandemic. *Educ. Psychol.* 43, 509–530. doi: 10.1080/01443410.2022.2146055

U.S. Department of Education (2012). Promising and practical strategies to increase postsecondary success. Available at: https://www.ed.gov/college-completion/promising-strategies

U.S. Department of Homeland Security. (2023). DHS STEM Designated Degree Program List. Available at: https://www.ice.gov/doclib/sevis/pdf/stemList2023.pdf

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., and Drachsler, H. (2021). Are we there yet? A systematic literature review on chatbots in education. *Front. Art. Intell.* 4:654924. doi: 10.3389/frai.2021.654924

frontiers | Frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 15 October 2024
DOI 10.3389/frai.2024.1460364

Check for updates

OPEN ACCESS

EDITED BY
Antonio Sarasa-Cabezuelo,
Complutense University of Madrid, Spain

REVIEWED BY
Thomas Mandl,
University of Hildesheim, Germany
Zhi Liu,
Central China Normal University, China

*CORRESPONDENCE
Noah L. Schroeder
schroedern@ufl.edu

RECEIVED 05 July 2024
ACCEPTED 30 September 2024
PUBLISHED 15 October 2024

CITATION
Mannekote A, Davies A, Pinto JD, Zhang S, Olds D, Schroeder NL, Lehman B, Zapata-Rivera D and Zhai C (2024) Large language models for whole-learner support: opportunities and challenges.
Front. Artif. Intell. 7:1460364.
doi: 10.3389/frai.2024.1460364

Large language models for whole-learner support: opportunities and challenges

Amogh Mannekote[1], Adam Davies[2], Juan D. Pinto[3], Shan Zhang[4], Daniel Olds[5], Noah L. Schroeder[1]*, Blair Lehman[6], Diego Zapata-Rivera[6] and ChengXiang Zhai[2]

[1]Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, United States, [2]Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL, United States, [3]Department of Curriculum and Instruction, University of Illinois Urbana-Champaign, Champaign, IL, United States, [4]School of Teaching and Learning, University of Florida, Gainesville, FL, United States, [5]Department of Computer Science, University of Oregon, Eugene, OR, United States, [6]Educational Testing Service, Princeton, NJ, United States

In recent years, large language models (LLMs) have seen rapid advancement and adoption, and are increasingly being used in educational contexts. In this perspective article, we explore the open challenge of leveraging LLMs to create personalized learning environments that support the "whole learner" by modeling and adapting to both cognitive and non-cognitive characteristics. We identify three key challenges toward this vision: (1) improving the interpretability of LLMs' representations of whole learners, (2) implementing adaptive technologies that can leverage such representations to provide tailored pedagogical support, and (3) authoring and evaluating LLM-based educational agents. For interpretability, we discuss approaches for explaining LLM behaviors in terms of their internal representations of learners; for adaptation, we examine how LLMs can be used to provide context-aware feedback and scaffold non-cognitive skills through natural language interactions; and for authoring, we highlight the opportunities and challenges involved in using natural language instructions to specify behaviors of educational agents. Addressing these challenges will enable personalized AI tutors that can enhance learning by accounting for each student's unique background, abilities, motivations, and socioemotional needs.

KEYWORDS

large language model (LLM), AI and education, non-cognitive aspects of learning, interpretability, pedagogical support of students, educational authoring tool

1 Introduction

In recent years, generative artificial intelligence (GenAI)—and more specifically, LLMs—have exploded into global public awareness (Barreto et al., 2023). ChatGPT, for example, is available in 188 countries with over 180 million users (as of August 2023)[1]. Such rapid adoption and ongoing development continues to disrupt many industries and areas of study, particularly as each new generation of LLMs offers new capabilities (e.g., memory, multimodality, longer input context sizes). LLMs have made their impact in the world of education as well—for instance, one notable example is Khanmigo[2], an LLM-powered AI

1 https://clickup.com/blog/chatgpt-statistics/#6-chatgpt-users-and-usage-

2 https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/

tutor that provides personalized support to students and assists teachers with developing instructional materials. This type of personalized support highlights the great potential for LLMs in educational contexts (Pardos and Bhandari, 2023).

We argue that such personalized support systems can and should be further expanded to provide "whole learner" support, moving beyond the paradigm of understanding and supporting only students' *academic proficiency* to also address *social*, *affective*, *motivational*, *cultural*, and *linguistic* characteristics that are known to impact learning (Bernacki et al., 2021; Mercado, 2018; Walkington and Bernacki, 2018; Lehman et al., 2024). This work focuses on "non-cognitive skills," describing aspects of a learner beyond subject knowledge and proficiency, such as resilience, persistence, empathy, motivation, self-regulation, and a growth mindset (Kautz et al., 2014).

Personalized learning environments (PLEs) leverage learner models, which are structured representations of students, to guide personalized support (Abyaa et al., 2019; Ismail et al., 2023). However, existing learner models cannot support the "whole learner," as they typically limit themselves to modeling knowledge acquisition (e.g., level of mastery over a concept), and at best, one additional characteristic (e.g., prior knowledge) or behavior (e.g., engagement; Ismail et al., 2023)[3]. The complexity of whole-learner modeling stems from the fact that it is not enough to simply model each characteristic and behavior independently—instead, these factors must be considered holistically to understand and support the whole learner. While the complex interaction of these factors presents a significant challenge for existing PLEs, we know that such holistic support is possible to provide in practice given that human teachers successfully combine these elements to support their students every day.

By pairing whole-learner modeling with GenAI, LLMs present an opportunity to bridge the long-standing gap between the quality and range of learner support offered by present-day computational systems and that offered by expert human tutors (Lepper et al., 1993; Cade et al., 2008; D'Mello et al., 2010). However, rapid innovation in the field of LLMs has raised questions about their appropriate use in PLEs. We explore some of the challenges and opportunities that exist around the vision of using LLMs to build whole-learner models and eventually create adaptive learning systems. We first explore the challenges and potential of LLMs in doing so (Section 2) and then identify several promising research directions to address these challenges (Section 3).

2 Challenges and the state of the art

We identify three key areas where the community needs to progress to achieve the larger vision of whole-learner modeling:

- **Interpretable representations of learners:** It is necessary to represent a learner explicitly and faithfully, including both *cognitive* and *non-cognitive* aspects. Although deep learning methods have traditionally been viewed as "black-box" approaches with opaque internal mechanisms, recent advances in interpretability and explainability research are working to address this challenge, and are well-positioned for applications in the context of whole learner modeling and support.
- **Adaptive technologies to support whole learners:** Given an interpretable learner representation, it should be used to tailor the delivery of pedagogical content and support to suit a learner's characteristics, including both cognitive and non-cognitive states. By leveraging data on learner behavior, preferences, and characteristics, and dynamically adjusting instructional strategies to address individual needs, adaptive systems can provide personalized learning pathways that evolve with learners' cognitive and non-cognitive skill development.
- **Authoring and evaluating agents:** In the context of PLEs and pedagogical agents (PAs), the term "authoring" refers to the manual process of specifying behaviors (or "policies") of the agent. For instance, in an intelligent tutoring system (ITS), classroom instructors can *author* common misconceptions based on their teaching experience. Authors can come from varying backgrounds (e.g., researcher, educator, and developer), so a key challenge in designing authoring tools is balancing accessibility and minimizing cognitive load. Finally, authoring is tightly-coupled with the issue of evaluation, a critical step in smoothly deploying these systems to real learners.

2.1 Interpretable representations of learners

Two key concerns in deploying LLMs to potentially sensitive application contexts such as education are *interpretability* (what a "black box" model is doing and representing internally) and *explainability* (why the model outputs A instead of B, given input C). Without reliable *interpretation*, we do not know what information the models use to make decisions or generate responses. Unlike trained educational professionals, automated models cannot be trusted to reliably take students' prior knowledge or emotional state into account to provide relevant and compassionate guidance, nor can we be certain that they will not use sensitive demographic information inappropriately. On the other hand, when we cannot reliably *explain* LLM behaviors, we cannot ensure that desired behaviors in one context will generalize to others (e.g., whether attentiveness to the emotional needs of students with high socioeconomic status will translate to less advantaged students).

Integrating interpretable learner models with LLMs is a promising approach to develop PLEs, providing the benefits of GenAI while maintaining a high level of interpretability. Such a hybrid approach need not be overly complex; for instance, one may begin by training a traditional learner model and passing its inferences to the LLM as an additional component of input prompts. However, it is crucial to ensure that (1) LLMs actually consider learner model's output, and that (2) they use this information in a way that is faithful to the learner model and consistent with educational best practices – otherwise, the approach

3 However, models considering more learner characteristics have been proposed (Shute and Zapata-Rivera, 2008; Zapata-Rivera and Greer, 2004).

will not benefit educational stakeholders like teachers and students (Pinto et al., 2023).

LLMs are also well-suited for advancing open learner models (OLMs) due to their natural language dialogue capabilities. OLMs enable learners to view and interact with the model's representation of their knowledge, promoting reflection and self-regulated learning. This transparency and interactivity can enhance traditional OLMs, allowing learners to modify their learning paths more freely. However, the use of LLMs in OLMs also raises concerns about how LLMs use learner information and relate their actions to educational best practices. Indeed, the integration of LLMs with OLMs has the potential to revolutionize educational technology by making learning processes more adaptive and personalized (Conati et al., 2018; Kay et al., 2020; Zapata-Rivera and Arslan, 2021; Bull, 2020), but implementation must be guided by strong ethical and pedagogical standards.

Despite important recent advances in understanding the inner workings of LLMs (e.g., Elhage et al., 2021; Olsson et al., 2022; Conmy et al., 2023; Templeton et al., 2024), reliably explaining model behavior to relevant stakeholders remains a significant challenge. This inability to interpret LLM representations and explain model behaviors leads to a lack of trust (Shin, 2021; Liao and Vaughan, 2023), which can inhibit these models' deployment to educational contexts where they have potential for transformational impact. By contrast, many traditional learner models are designed with interpretability as an inherent feature, such as Bayesian Knowledge Tracing (Baker et al., 2008) and Item Response Theory (Yen and Fitzpatrick, 2006). There are even efforts underway to develop intrinsically interpretable neural-network-based learner models (Pinto et al., 2023; Lu et al., 2020; Swamy et al., 2024). We discuss how such approaches can address the challenges of interpretable LLMs for education in 3.1.

2.2 Whole learner support through adaptive technologies

The next challenge in deploying LLMs for education is *adaptivity*, which involves assessing various learner characteristics and tailoring the learning experience to individual needs to improve learning outcomes (Plass and Pawar, 2020). Through the natural language capabilities of LLMs, adaptive technologies for whole learner support can offer nuanced support for developing both cognitive and non-cognitive skills in diverse learners. For instance, Arroyo et al. (2014) demonstrated that intelligent adaptive tutors effectively address students' unique needs and emotions, enhancing engagement and affect, while Liu et al. (2024) found that conversational agents offering emotional scaffolding improved students' emotional experiences.

Such findings highlight the importance of design principles focused on non-cognitive learner characteristics, such as fostering a growth mindset through praising learners' efforts (Liu et al., 2024), attributing struggles to external factors (Calvo and D'Mello, 2011), utilizing an anthropomorphic language style, and employing proactive inquiry (McQuiggan et al., 2008; Sabourin et al., 2011) to guide learners to self-report their emotional states. For instance, a review found that empathetic agent feedback, including affective feedback and confidence- and motivation-enhancing dialogue, positively influences students' attitudes (Liu et al., 2024; Ortega-Ochoa et al., 2024). Similarly, another study demonstrated how conversational agents can support children's social-emotional learning by teaching self-talk. These lines of research also emphasize the importance of designing conversational dialogue based on an evidence-based framework (Fu et al., 2023). Building on this foundation, recent AI advancements have facilitated the development of natural language dialogue systems to scaffold non-cognitive skills (Acosta et al., 2015; Anghel and Balart, 2017; Cinque et al., 2021).

2.3 Authoring and evaluating agents

LLMs are also transforming the landscape for authoring educational agents such as PAs, intelligent tutors (Sottilare et al., 2015), and even simulated learners (Käser and Alexandron, 2023). Before the widespread adoption of modern LLMs, agent authoring was bottlenecked by supervised and reinforcement learning methods that required machine learning expertise (Mannekote et al., 2023; Liu and Chilton, 2022), lots of data, labor-intensive manual annotation, or some combination of these factors. In contrast, the recent development of instruction-tuned LLMs (Wang et al., 2023) enables educational experts to define agent behaviors using natural language instructions in "zero-shot" or "few-shot" setups (i.e., using no annotated examples or only a few, respectively). In addition to reducing the training and expertise needed for authoring the dialogue system, LLMs also open up new avenues of agent behavior—for instance, where classical ITSs predominantly focused on supporting the cognitive aspects of learning (e.g., subject proficiency) (Sottilare et al., 2015), authors can now leverage LLMs capabilities such as their abilities to emulate human-like decision-making (Miliv cka et al., 2024) and perform high-level planning (Kambhampati et al., 2024) to equip them with the ability to support non-cognitive aspects of learning as well.

However, LLMs are not (yet) a "turn-key" solution to agent authoring, as several key challenges remain. Authoring LLM-based agents requires effectively navigating an unbounded space of possible prompts, which may be difficult to do without prompt engineering expertise (Oppenlaender et al., 2023; Zamfirescu-Pereira et al., 2023; Mannekote et al., 2023). Moreover, it has been shown that LLM outputs are highly sensitive to minor prompt variations, often leading to inconsistent (Lu et al., 2022; Liu and Chilton, 2022; Loya et al., 2023; Mohammadi, 2024) and confounding (Gui and Toubia, 2023) results. Finally, when authoring complex agent behaviors, the issue of evaluating the *faithfulness* of an agent's behavior to the authors' intended expectations becomes pertinent (Koedinger et al., 2015; Weitekamp et al., 2023). In fact, within the context of AI models like LLMs, this issue can be considered to be a specific instance of the *alignment problem* (Yudkowsky, 2016).

3 Ways forward

For each of the three challenge areas delineated in Section 2, we outline a broad roadmap for future advancements. Specifically, we identify promising directions that the field is likely to pursue in the medium to long term.

3.1 Interpretable representations of learners

We face two primary challenges in enhancing the interpretability of LLMs. First, rather than merely adding more information to the prompt and hoping that the model will use it appropriately, we need a direct method to explain why a model generated a particular output. This involves determining whether the output was produced due to explicit learner information that has been added to prompts or implicit learner information that LLMs have inferred from learner behavior. Second, we must predict whether its behavior will remain consistent when applied in different contexts, such as in learning environments for which they have not already been tested. Although neural networks like LLMs are usually seen as "black boxes" whose internal representations and mechanisms are treated as unknowable beyond the outputs they produce, recent work in deep learning interpretability has made substantial strides in addressing this challenge. For instance, current interpretability methods can detect what latent representations are used by models in producing a particular output (Elazar et al., 2021; Belinkov, 2022; Davies et al., 2023) and characterize how these representations are leveraged in producing particular behaviors (Elhage et al., 2021; Olsson et al., 2022; Conmy et al., 2023).

Beyond simply interpreting LLMs' representation and use of information about learners, it is also important to utilize counterfactual explanation techniques to predict how their behaviors will change in response to different input prompts (Wachter et al., 2017; Ribeiro et al., 2020)—for instance, if we add a minor typo to a student's essay that is otherwise exemplary, will the model provide a substantially lower assessment of the student's knowledge in response? Conversely, it is equally important to characterize when and how models will remain invariant with respect to a given input property (Schwab and Karlen, 2019)—for example, it may be important to understand whether models always provide the same answer to different ways of phrasing the same question, meaning that they are *invariant* with respect to question semantics. Knowing the set of properties to which a given model is invariant allows us to predict whether its behavior will remain consistent if those same properties are held constant, even as other input properties may vary (Peters et al., 2016; Arjovsky et al., 2019). Between these two lines of research, we can build a systematic picture of when, how, and why model behaviors are expected to change (under counterfactuals) or remain the same (given invariances).

3.2 Whole learner support through adaptive technologies

Our vision for personalized learning that supports whole-learner adaptation necessitates a dynamic approach to learner modeling, capable of capturing and integrating the learner's complex states and needs. High-quality adaptive feedback is contingent on an accurate representation of the learner. Crafting and updating this representation is the job of the learner model, which is typically considered a separate component from the adaptation module that produces feedback in ITSs and PLEs (Shute and Zapata-Rivera, 2008). While learner models can come in many forms, such as cognitive models, machine learning models, or Bayesian networks, GenAI models like LLMs are beginning to be tested for this task (Zhang et al., 2024).

Integrating whole learner models with LLM-based support involves using cognitive, affective, or behavioral states from learner models as inputs to the adaptation module or dialogue engine (Zapata-Rivera and Forsyth, 2022). To capture the whole learner, multiple traditional models representing distinct aspects of the learner can either form a larger learner *module* or be integrated into a holistic model. Alternatively, a single LLM might serve as both the learner model and the adaptation module, though the current lack of LLM interpretability challenges trustworthiness and validation. Another viable option is leveraging a LLM to integrate outputs from various traditional learner models, providing a comprehensive inference to the adaptation module. This approach could be very useful, despite the limited research on integrating diverse types of learner information, as it offers a more nuanced understanding of the student.

Regardless of the specific system architecture used, LLMs enable just-in-time adaptive conversational feedback. This allows conversational complexity to adjust dynamically based on the learner's real-time progress, maintaining an appropriate level of challenge and promoting engagement (Zapata-Rivera and Forsyth, 2022). By basing this feedback on a rich understanding of the learner from the learner model, it offers whole-learner adaptation, potentially providing more nuanced, personalized support than existing PLEs.

3.3 Authoring and evaluating agents

When building agents to support the whole learner, the ability to operationalize a given theoretical model or dynamically incorporate new developments from learning sciences into agentic behavior "on the fly" is a desirable trait, helping to avoid the tedious process of manually re-authoring the agents. Efficient attention mechanisms (Shen et al., 2021), attention-alternatives (Gu and Dao, 2023), techniques such as retrieval-augmented generation (RAG) (Lewis et al., 2020), and needle-in-the-haystack capabilities (Kuratov et al., 2024) will enable authors to quickly reshape agent behavior, potentially even allowing them to directly operationalize longer documents such as scientific reviews or books describing evidence-based practices.

Equally important to authoring is evaluating the model outputs for faithfulness and robustness. Although preliminary experimental results with using LLMs in economics and psychology suggest that LLMs are capable of accurately mimicking aspects of human behavior like decision-making (Jia et al., 2024) and personality traits (Frisch and Giulianelli, 2024), further research is needed to generalize these findings to educational settings.

Finally, authoring is not just about *designing* agents for pedagogical support, but also developing realistic testbeds to *evaluate* them. For this line of work, authoring multi-agent social simulations (see, e.g., Park et al., 2022, 2023) will be an integral component of the end-to-end development process of ITSs and PAs. Such evaluations can ensure that the agents perform well across a wide range of scenarios, increasing educator confidence. For instance, instead of testing a PA against a single learner simulation, authoring an entire classroom of LLMs comprising multiple learner agents allows for more holistic and rigorous testing, ensuring the PA is pedagogically effective, equitable, safe, and robust before being deployed to real learners.

4 Ethical considerations

Several important ethical considerations must be addressed before deploying GenAI for educational applications. First, interpretability is crucial for trustworthiness: one cannot fully trust a model in sensitive applications like education without understanding how it represents and interacts with users (Huang et al., 2020). Second, it is important to ensure that LLMs do not exacerbate the *digital divide* in education (i.e., inequitable access to educational technologies and associated benefits), as anticipated by Capraro et al. (2024). For instance, given the substantial compute required to deploy the largest and most capable LLMs, it may be helpful to develop more compute-efficient language models for use in educational settings with limited resources (Hoffmann et al., 2022); and interdisciplinary collaborations between AI research and learning sciences will be essential in ensuring that new technologies are actually improving learning outcomes and student welfare (cf. Dahlin, 2021).

Finally, perhaps most important are concerns regarding student privacy—for instance, in the adaptive support modules envisioned above, LLMs might be provided with information about learners' emotional states to provide more holistic, empathetic feedback; but in order to protect students' privacy and ensure that sensitive information about them cannot be used for non-educational purposes such as advertising, student data should only be visible to systems with robust security and data privacy guarantees (and not, e.g., included in prompts used as input to third-party AI systems, which may use such information to train future public-facing models). These concerns are particularly significant for minors, who have special legal privacy protections and may be more vulnerable to unintended GenAI behaviors.

5 Conclusion

In this paper, we explored the potential integration of LLMs into PLEs to support the whole learner addressing both cognitive and non-cognitive characteristics. Our discussion has highlighted significant opportunities as well as challenges in integrating LLMs into PLEs, focusing on developing interpretable learner representations, adaptive technologies for personalized support, and authoring and evaluating PAs. For future research, it will be important to develop methods to enhance LLM interpretability and explainability within educational settings, facilitating trustworthiness and appropriate use of student information. Additionally, LLMs' adaptability must also be refined to ensure that models can offer individualized support that accounts for diverse learner needs and backgrounds. Finally, authoring PAs will require more principled prompting protocols, including an understanding of both relevant subject matter and pedagogical best practices, in order to engender more faithful and robust agents. By advancing each of these areas, LLMs can be better positioned to fulfill their potential as transformative tools in education, making widely-accessible personalized learning a practical reality. Through all these advancements, it is essential to be mindful of the security, privacy, and ethical concerns surrounding the handling of learner data.

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

AM: Conceptualization, Writing - original draft, Writing - review & editing. AD: Conceptualization, Writing - original draft, Writing - review & editing. JP: Conceptualization, Writing - original draft. SZ: Conceptualization, Writing - original draft. DO: Conceptualization, Writing - original draft. NS: Conceptualization, Writing - original draft, Supervision, Writing - review & editing. BL: Conceptualization, Writing - original draft, Writing - review & editing. DZ-R: Conceptualization, Supervision, Writing - review & editing. CZ: Conceptualization, Supervision, Writing - review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This material was based upon work supported by the National Science Foundation and the Institute of Education Sciences under Grant #2229612.

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.

The handling editor AS-C is currently organizing a Research Topic with the author DZ-R.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission.

This had no impact on the peer review process and the final decision.

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.

Author disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Education.

References

Abyaa, A., Khalidi Idrissi, M., and Bennani, S. (2019). Learner modelling: systematic review of the literature from the last 5 years. *Educ. Technol. Res. Dev.* 67, 1105–1143. doi: 10.1007/s11423-018-09644-1

Acosta, P., Cunningham, W., and Muller, N. (2015). *Beyond Qualifications: Labor Market Returns to Cognitive Skills and Personality Traits in Urban Colombia*. Bonn: Institute for the Study of Labor (IZA).

Anghel, B., and Balart, P. (2017). Non-cognitive skills and individual earnings: new evidence from piaac. *Series* 8, 417–473. doi: 10.1007/s13209-017-0165-x

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*. doi: 10.48550/arXiv.1907.02893

Arroyo, I., Muldner, K., Burleson, W., and Woolf, B. P. (2014). Adaptive interventions to address students—negative activating and deactivating emotions during learning activities. *Des. Recommend. Intell. Tutor. Syst.* 2, 79–92.

Baker, R. S. d., Corbett, A. T., and Aleven, V. (2008). "More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing," in *Intelligent Tutoring Systems: 9th International Conference, ITS 2008, Montreal, Canada, June 23–27, 2008 Proceedings 9* (Berlin: Springer), 406–415.

Barreto, F., Moharkar, L., Shirodkar, M., Sarode, V., Gonsalves, S., and Johns, A. (2023). "Generative artificial intelligence: Opportunities and challenges of large language models," in *International Conference on Intelligent Computing and Networking* (Berlin: Springer), 545–553.

Belinkov, Y. (2022). Probing classifiers: promises, shortcomings, and advances. *Comput. Linguist.* 48, 207–219. doi: 10.1162/coli a 00422

Bernacki, M. L., Greene, M. J., and Lobczowski, N. G. (2021). A systematic review of research on personalized learning: personalized by whom, to what, how, and for what purpose(s)? *Educ. Psychol. Rev.* 33, 1675–1715. doi: 10.1007/s10648-021-09615-8

Bull, S. (2020). There are open learner models about! *IEEE Trans. Learn. Technol.* 13, 425–448. doi: 10.1109/TLT.2020.2978473

Cade, W. L., Copeland, J. L., Person, N. K., and D'Mello, S. K. (2008). "Dialogue modes in expert tutoring," in *Intelligent Tutoring Systems: 9th International Conference, ITS 2008, Montreal, Canada, June 23-27, 2008 Proceedings 9* (Berlin: Springer), 470–479.

Calvo, R. A., and D'Mello, S. K. (2011). *New Perspectives on Affect and Learning Technologies, Vol. 3*. Cham: Springer Science & Business Media.

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., et al. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *Proc. Natl. Acad. Sci. U. S. A. Nexus* 3:191. doi: 10.1093/pnasnexus/pgae191

Cinque, M., Carretero, S., and Napierala, J. (2021). *Non-cognitive Skills and Other Related Concepts: Towards a Better Understanding of Similarities and Differences. Technical report, JRC Working Papers Series on Labour, Education and Technology.* European Commission, Joint Research Centre (JRC), Seville Spain.

Conati, C., Porayska-Pomsta, K., and Mavrikis, M. (2018). AI in education needs interpretable machine learning: lessons from open learner modelling. *arXiv preprint arXiv:1807.00154*. doi: 10.48550/arXiv.1807.00154

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. (2023). Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997*. doi: 10.48550/arXiv.2304.14997

Dahlin, E. (2021). Mind the gap! on the future of AI research. *Human. Soc. Sci. Commun.* 8, 1–4. doi: 10.1057/s41599-021-00750-9

Davies, A., Jiang, J., and Zhai, C. (2023). Competence-based analysis of language models. *arXiv preprint arXiv:2303.00333*. doi: 10.48550/arXiv.2303.00333

D'Mello, S., Lehman, B., and Person, N. (2010). "Expert tutors feedback is immediate, direct, and discriminating," in *Twenty-Third International FLAIRS Conference* (Daytona Beach, FL).

Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. (2021). Amnesic probing: behavioral explanation with amnesic counterfactuals. *Trans. Assoc. Comput. Linguist.* 9, 160–175. doi: 10.1162/tacl_a_00359

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., et al. (2021). A mathematical framework for transformer circuits. *Transform. Circ. Thread* 1.

Frisch, I., and Giulianelli, M. (2024). LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *arXiv preprint arXiv:2402.02896*. doi: 10.48550/arXiv.2402.02896

Fu, Y., Zhang, M., Nguyen, L. K., Lin, Y., Michelson, R., Tayebi, T. J., et al. (2023). "Self-talk with superhero zip: supporting children's socioemotional learning with conversational agents," in *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference* (Chicago, IL), 173–186.

Gu, A., and Dao, T. (2023). Mamba: linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*. doi: 10.48550/arXiv.2312.00752

Gui, G., and Toubia, O. (2023). The challenge of using LLMs to simulate human behavior: a causal inference perspective. *arXiv preprint arXiv:2312.15524*. doi: 10.48550/arXiv.2312.15524

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., et al. (2022). "An empirical analysis of compute-optimal large language model training," in *Advances in Neural Information Processing Systems, Vol. 35*, eds. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (Red Hook, NY: Curran Associates, Inc.), 30016–30030.

Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., et al. (2020). A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Comput. Sci. Rev.* 37:100270. doi: 10.48550/arXiv.1812.08342

Ismail, H., Hussein, N., Harous, S., and Khalil, A. (2023). Survey of personalized learning software systems: a taxonomy of environments, learning content, and user models. *Educ. Sci.* 13:741. doi: 10.3390/educsci13 070741

Jia, J., Yuan, Z., Pan, J., McNamara, P., and Chen, D. (2024). *Decision-Making Behavior Evaluation Framework for LLMs Under Uncertain Context. arXiv [Preprint]. arXiv*:2406.05972.

Kambhampati, S., Valmeekam, K., Guan, L., Stechly, K., Verma, M., Bhambri, S., et al. (2024). LLMs can't plan, but can help planning in LLM-modulo frameworks. *arXiv preprint arXiv:2402.01817*. doi: 10.48550/arXiv.2402. 01817

Käser, T., and Alexandron, G. (2023). Simulated learners in educational technology: a systematic literature review and a turing-like test. *Int. J. Artif. Intell. Educ.* 23:2. doi: 10.1007/s40593-023-00337-2

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., and Borghans, L. (2014). *Fostering and Measuring Skills: Improving Cognitive and Non-cognitive Skills to Promote Lifetime Success*. Cambridge, MA: National Bureau of Economic Research.

Kay, J., Zapata-Rivera, D., and Conati, C. (2020). The gift of scrutable learner models: why and how. *Des. Recommend. Intell. Tutor. Syst.* 8, 25–40.

Koedinger, K. R., Matsuda, N., MacLellan, C. J., and McLaughlin, E. A. (2015). "Methods for evaluating simulated learners: examples from Simstudent," in *17th International Conference on Artificial Intelligence in Education*, Madrid Spain.

Kuratov, Y., Bulatov, A., Anokhin, P., Sorokin, D., Sorokin, A., and Burtsev, M. (2024). In search of needles in a 10 m haystack: recurrent memory finds what LLMs miss. *arXiv preprint arXiv:2402.10790*. doi: 10.48550/arXiv.2402.10790

Lehman, B., Sparks, J., Zapata-Rivera, D., Steinberg, J., and Forsyth, C. (2024). A framework of caring assessments for diverse learners. *Pract. Assess. Res. Eval.* 29:9. doi: 10.7275/pare.2102

Lepper, M. R., Woolverton, M., Mumme, D. L., and Gurtner, J.-L. (1993). *Motivational Techniques of Expert Human Tutors: Lessons for the Design of Computer-Based Tutors*. Routledge.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inform. Process. Syst.* 33, 9459–9474. doi: 10.48550/arXiv.2005.11401

Liao, Q. V., and Vaughan, J. W. (2023). AI transparency in the age of LLMs: a human-centered research roadmap. *arXiv preprint arXiv:2306.01941*. doi: 10.48550/arXiv.2306.01941

Liu, V., and Chilton, L. B. (2022). "Design guidelines for prompt engineering text-to-image generative models," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22* (New York, NY: Association for Computing Machinery), 1–23.

Liu, Z., Duan, H., Liu, S., Mu, R., Liu, S., and Yang, Z. (2024). Improving knowledge gain and emotional experience in online learning with knowledge and emotional scaffolding-based conversational agent. *Educ. Technol. Soc.* 27, 197–219. doi: 10.30191/ETS.202404_27(2).RP08

Loya, M., Sinha, D., and Futrell, R. (2023). "Exploring the sensitivity of LLMs' decision-making capabilities: insights from prompt variations and hyperparameters," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, eds. H. Bouamor, J. Pino, and K. Bali (Singapore: Association for Computational Linguistics), 3711–3716.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. (2022). "Fantastically ordered prompts and where to find them: overcoming few-shot prompt order sensitivity," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. S. Muresan, P. Nakov, and A. Villavicencio (Dublin: Association for Computational Linguistics), 8086–8098.

Lu, Y., Wang, D., Meng, Q., and Chen, P. (2020). "Towards interpretable deep learning models for knowledge tracing," in *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21* (Berlin: Springer), 185–190.

Mannekote, A., Celepkolu, M., Wiggins, J. B., and Boyer, K. E. (2023). "Exploring usability issues in instruction-based and schema-based authoring of task-oriented dialogue agents," in *Proceedings of the 5th International Conference on Conversational User Interfaces, CUI '23* (New York, NY: Association for Computing Machinery), 1–6.

McQuiggan, S. W., Robison, J. L., Phillips, R., and Lester, J. C. (2008). "Modeling parallel and reactive empathy in virtual agents: an inductive approach," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems-Vol. 1* (Estoril: Citeseer), 167–174.

Mercado, F. (2018). Whole child framework: supporting educators in their plight toward mtss and equity. *J. Leaders. Eq. Res.* 4.

Miliv cka, J., Marklová, A., Van Slambrouck, K., Pospiv silová, E., Simsová, J., Harvan, S., et al. (2024). Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *PLoS ONE* 19:e0298522. doi: 10.1371/journal.pone.0298522

Mohammadi, B. (2024). Wait, it's all token noise? always has been: interpreting LLM behavior using Shapley value. *arXiv:2404.01332 [cs]*. doi: 10.48550/arXiv.2404.01332

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., et al. (2022). In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*. doi: 10.48550/arXiv.2209.11895

Oppenlaender, J., Linder, R., and Silvennoinen, J. (2023). Prompting AI art: an investigation into the creative skill of prompt engineering. *arXiv:2303.13534 [cs]*. doi: 10.48550/arXiv.2303.13534

Ortega-Ochoa, E., Arguedas, M., and Daradoumis, T. (2024). Empathic pedagogical conversational agents: a systematic literature review. *Br. J. Educ. Technol.* 55, 886–909. doi: 10.1111/bjet.13413

Pardos, Z. A., and Bhandari, S. (2023). Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint arXiv:2302.06871*. doi: 10.48550/arXiv.2302.06871

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: interactive simulacra of human behavior. *arXiv:2304.03442 [cs]*. doi: 10.48550/arXiv.2304.03442

Park, J. S., Popowski, L., Cai, C., Morris, M. R., Liang, P., and Bernstein, M. S. (2022). "Social simulacra: creating populated prototypes for social computing systems," in *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR), 1–18.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *J. Royal Stat. Soc. Ser. B* 78, 947–1012. doi: 10.48550/arXiv.1501.01332

Pinto, J. D., Paquette, L., and Bosch, N. (2023). "Interpretable neural networks vs. expert-defined models for learner behavior detection," in *Companion Proceedings of the 13th International Conference on Learning Analytics and Knowledge Conference (LAK23)* (Arlington, TX), 105–107.

Plass, J. L., and Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *J. Res. Technol. Educ.* 52, 275–300. doi: 10.1080/15391523.2020.1719943

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond accuracy: behavioral testing of NLP models with checklist. *arXiv preprint arXiv:2005.04118*. doi: 10.48550/arXiv.2005.04118

Sabourin, J., Mott, B., and Lester, J. (2011). Computational models of affect and empathy for pedagogical virtual agents. *Stand. Emot. Model.* 2011, 1–14.

Schwab, P., and Karlen, W. (2019). "CXPlain: causal explanations for model interpretation under uncertainty," in *Advances in Neural Information Processing Systems, Vol. 32*, eds. H. Wallach, H. Larochelle, A. Beygelzimer, F. dsingle Alché-Buc, E. Fox, and R. Garnett (New York, NY: Curran Associates, Inc.), 19.

Shen, Z., Zhang, M., Zhao, H., Yi, S., and Li, H. (2021). "Efficient attention: attention with linear complexities," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Waikoloa, HI), 3531–3539.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: implications for explainable AI. *Int. J. Hum. Comput. Stud.* 146:102551. doi: 10.1016/j.ijhcs.2020.102551

Shute, V. J., and Zapata-Rivera, D. (2008). "Adaptive technologies," in *Handbook of Research on Educational Communications and Technology* (London: Routledge), 277–294.

Sottilare, R., Graesser, A., Hu, X., and Brawner, K. (2015). *Design Recommendations for Intelligent Tutoring Systems - Volume 3: Authoring Tools and Expert Modeling Techniques*. Memphis, TN.

Swamy, V., Blackwell, J., Frej, J., Jaggi, M., and Käser, T. (2024). InterpretCC: conditional computation for inherently interpretable neural networks. *arXiv preprint arXiv:2402.02933*. doi: 10.48550/arXiv.2402.02933

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., et al. (2024). Scaling monosemanticity: extracting interpretable features from claude 3 sonnet. *Transform. Circ. Thread.* Available online at: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL Tech.* 31:841. doi: 10.48550/arXiv.1711.00399

Walkington, C., and Bernacki, M. L. (2018). Personalization of instruction: design dimensions and implications for cognition. *J. Exp. Educ.* 86, 50–68. doi: 10.1080/00220973.2017.1380590

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., et al. (2023). "Self-instruct: aligning language models with self-generated instructions," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, eds. A. Rogers, J. Boyd-Graber, and N. Okazaki (Toronto, ON: Association for Computational Linguistics), 13484–13508.

Weitekamp, D., Rachatasumrit, N., Wei, R., Harpstead, E., and Koedinger, K. (2023). "Simulating learning from language and examples," in *International Conference on Artificial Intelligence in Education* (Berlin: Springer), 580–586.

Yen, W. M., and Fitzpatrick, A. R. (2006). Item response theory. *Educ. Measur.* 4, 111–153. Available online at: https://intelligence.org/files/AlignmentHardStart.pdf

Yudkowsky, E. (2016). The AI alignment problem: why it is hard, and where to start. *Symbol. Syst. Distinguis. Speak.* 4:1.

Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., and Yang, Q. (2023). "Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg), 1–21.

Zapata-Rivera, D., and Arslan, B. (2021). "Enhancing personalization by integrating top-down and bottom-up approaches to learner modeling," in *International Conference on Human-Computer Interaction* (Berlin: Springer), 234–246.

Zapata-Rivera, D., and Forsyth, C. M. (2022). "Learner modeling in conversation-based assessment," in *International Conference on Human-Computer Interaction* (Berlin: Springer), 73–83.

Zapata-Rivera, J.-D., and Greer, J. E. (2004). Interacting with inspectable bayesian student models. *Int. J. Artif. Intell. Educ.* 14, 127–163. doi: 10.5555/1434858.1434859

Zhang, L., Lin, J., Borchers, C., Sabatini, J., Hollander, J., Cao, M., et al. (2024). "Predicting learning performance with large language models: a study in adult literacy," in *Adaptive Instructional Systems*, eds. R. A. Sottilare and J. Schwarz (Cham: Springer Nature Switzerland), 333–353.

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 23 October 2024
DOI 10.3389/frai.2024.1460337

Check for updates

OPEN ACCESS

EDITED BY
Paul Libbrecht,
IUBH University of Applied Sciences, Germany

REVIEWED BY
Vagelis Plevris,
Qatar University, Qatar
Yousef Wardat,
Higher Colleges of Technology,
United Arab Emirates

*CORRESPONDENCE
Marc Herrmann
✉ marc.herrmann@uni-siegen.de

RECEIVED 05 July 2024
ACCEPTED 14 October 2024
PUBLISHED 23 October 2024

CITATION
Dilling F and Herrmann M (2024) Using large language models to support pre-service teachers mathematical reasoning—an exploratory study on ChatGPT as an instrument for creating mathematical proofs in geometry.
Front. Artif. Intell. 7:1460337.
doi: 10.3389/frai.2024.1460337

Using large language models to support pre-service teachers mathematical reasoning—an exploratory study on ChatGPT as an instrument for creating mathematical proofs in geometry

Frederik Dilling and Marc Herrmann*

Mathematics Education, Department of Mathematics, University of Siegen, Siegen, Germany

In this exploratory study, the potential of large language models (LLMs), specifically ChatGPT to support pre-service primary education mathematics teachers in constructing mathematical proofs in geometry is investigated. Utilizing the theoretical framework of instrumental genesis, the prior experiences of students with LLMs, their beliefs about the operating principle and their interactions with the chatbot are analyzed. Using qualitative content analysis, inductive categories for these aspects are formed. Results indicate that students had limited prior experiences with LLMs and used them predominantly for applications that are not mathematics specific. Regarding their beliefs, most show only superficial knowledge about the technology and misconceptions are common. The analysis of interactions showed multiple types of in parts mathematics-specific prompts and patterns on three different levels from single prompts to whole chat interactions.

KEYWORDS

ChatGPT, mathematics education, mathematical proofs, teacher education, generative AI, large language model

Introduction

Artificial intelligence (AI) is a highly debated topic today. The public release of GPT-3 in November 2022 brought this discussion into mainstream society. Although AI has been a focus of technical research in education for about a decade (as evidenced for example by the publications in the "International Journal of Artificial Intelligence in Education"), it is only recently that mathematics education research has begun to explore this area. Specifically, attention is being given to large language models (LLMs) such as ChatGPT. These linguistic models, trained on vast amounts of text data, aim to mimic human communication. They generate responses to user queries (prompts) using statistical relationships in the training data. Despite being trained primarily for linguistic tasks, LLMs can also convey factual knowledge from their training data (Petroni et al., 2019). However, they do not access knowledge databases directly; their "knowledge" is derived solely from the trained linguistic model, which can sometimes lead to incorrect information being produced. This offers many possibilities and challenges for education, e.g., the use of LLMs as tutorial systems to assist the students' learning processes (Kasneci et al., 2023).

In the context of university education, a quantitative online survey of 6,311 students from various disciplines at German universities has shown that around 63% of the students use AI-based tools for their studies (Garrel et al., 2023). By far the most frequently mentioned AI tool by the students is the LLM ChatGPT. The AI tools are used (in descending order) to clarify questions of understanding and to explain subject-specific concepts, for research and literature studies, for translations, for text analysis, word processing and text creation, or for problem solving and decision-making. In addition to these more general ways of using AI tools, specific use cases can be found in the field of mathematics, such as supporting mathematical processes like problem solving, modelling, communicating, arguing or proving (Buchholtz et al., 2024; Dilling et al., 2024a; Yoon et al., 2024). Nevertheless, it is important to always consider the major challenges, such as copyright and data privacy issues, bias and fairness or the possibility that teachers and students become too reliant on generative AI (Kasneci et al., 2023). Furthermore, LLMs may be biased towards different cultural contexts, also influencing the understanding of mathematical theorems and generating unreliable answers (Blanchard and Mohammed, 2024).

This paper focuses on mathematical proof processes of university students (pre-service mathematics teachers for primary and middle school). It will be investigated, how students use the LLM ChatGPT as a tool for proving. The instrumental genesis according to Vérillon and Rabardel (1995) is used as a theoretical framework. This is introduced in the following section 2, followed by a literature survey on beliefs about how LLMs work and prompting in the context of (mathematics) education. From this, a research gap and three research questions are deduced. Section 3 introduces the conditions and methodology of the study. Section 4 presents the results of the study and section 5 discusses these against the background of the research literature. A conclusion and outlook follow in section 6.

Theoretical background

Instrumental genesis

The theoretical framework for the empirical study conducted in this article is the theory of instrumental genesis according to Vérillon and Rabardel (1995). Based on the findings of activity theory, instrumental genesis describes the relationship between a subject and an instrument as well as the emergence of instruments.

The central concepts in the theory of Vérillon and Rabardel (1995) are artefact and instrument. An artefact is a material or symbolic object. It was developed and produced by a person or a group of people for the purpose of achieving certain goals. Rabardel (1995) also speaks of the constituent functions of an artefact. In contrast, the term instrument is directly linked to the use of an artefact in a particular situation and thus also to the subject using it: "The subject builds the instrument from the artefact when using the artefact during an activity" (Laisney and Chatoney, 2018, p. 6). The artefact can also be used differently than the developers had intended – this is also referred to as constituted functions.

According to the theory of Vérillon and Rabardel (1995), an instrument has two central dimensions: the material or symbolic artefact (artefactual dimension) and the utilization schemes the subject associates with the artefact (schematic dimension). The totality of the functions and subjective values that an artefact can have in the activity of a subject are referred to as the instrumental field of an artefact (Rabardel and Beguin, 2007).

The process in which an artefact becomes an instrument for a subject is referred to by Vérillon and Rabardel (1995) as instrumental genesis. This process relates to both dimensions of an instrument (artefact and utilization schemes) and implies changes to the artefact as well as to the subject. Therefore, two different sub-processes can be distinguished:

1. Instrumentalization refers to the adaptation of the subject to the artefact. The subject learns about the artefact's characteristics and intrinsic properties. This knowledge enables him or her to select and use functions for situation-specific actions. In this process, new functions may arise that were not intended by the developers (constituted functions).
2. Instrumentation refers to the adaptation of the artefact to the subject. The potentials and limitations of an artefact determine the actions of a subject. In order to use the constituent and constituted functions, the subject changes the activities, actions and utilization schemes, which leads to changes of meaning of the instrument.

The instrumental genesis process can be understood as a cycle. The subject learns about new properties and functions of an artefact and then adapts the utilization schemes. This in turn enables the recognition of new properties and functions. The theory of instrumental genesis has been used in mathematics education research frequently (e.g., Guin and Trouche, 2002; Bretscher, 2009).

This article examines how university students use the LLM ChatGPT (an artefact) as a tool for proving in geometry. The prompts and interaction patterns represent situation-specific utilization schemes. Thus, a specific part of the instrumental field of the students in relation to ChatGPT is reconstructed.

Instrumentalization: beliefs about LLMs and AI

As described in the previous section, knowledge of the characteristics and intrinsic properties of an artefact is the result of instrumentalization. The concept of beliefs, which is well known from psychology, can be used to conceptualize this knowledge and make it empirically accessible. This concept has been used for many decades in mathematics education research to describe the behavior of students at school and university or that of teachers (see Philipp, 2007).

Beliefs represent mental structures and are composed of a cognitive and an affective component (Furinghetti and Pehkonen, 2002). Schoenfeld (1992) defines beliefs as "an individual's understandings and feelings that shape the ways that the individual conceptualizes and engages in mathematical behavior" (p. 358).

Goldin (2002) defines beliefs by ascribing to them a degree of subjective truth:

> "I propose to define beliefs as multiply-encoded cognitive/affective configurations, usually including (but not limited to) prepositional encoding, to which the holder attributes some kind of truth value. The latter term is not taken in the technical sense of symbolic logic, but as a term that may variously refer to logical truth empirical truth, validity, applicability to some degree of

> approximation, metaphysical truth, religious truth, practical truth, or conventional truth." (p. 64)

Accordingly, he describes knowledge as the subset of beliefs that represents true (in the sense of socially accepted) statements:

> "Knowledge […] refers to beliefs that, in a sense apart from the fact of belief or the acceptance of warrants for belief by an individual or group, are true correct, valid, veridical, good approximations, or applicable." (p. 66)

Pehkonen and Pietilä (2004) use the terms subjective and objective knowledge and assign beliefs to subjective knowledge:

> "An individual's beliefs are understood as his subjective, experience-based, often implicit knowledge and emotions on some matter or state of art. […] Beliefs represent some kind of tacit knowledge. Every individual has his own tacit knowledge which is connected with learning and teaching situations, but which rarely will be made public." (Pehkonen and Pietilä, 2004, S. 2)

In this regard, Pehkonen (1995) uses the term "stable subjective knowledge" to emphasize that beliefs are relatively stable mental constructs. Nevertheless, beliefs can be subject to change, as they are constantly compared with experiences and the beliefs of other people.

In various empirical studies, the beliefs of different groups of people on the issue of artificial intelligence were investigated or compiled on a theoretical basis. However, it should be emphasized at this point that the term belief was not used in these studies – instead, similar psychological terms such as perceptions, pre-concepts or misconceptions were used. The fine differences between these terms will not be discussed here.

Mertala and Fagerlund (2024) examined the misconceptions of 195 Finnish 5th and 6th graders in a qualitative online survey. In the survey, the students were asked, among other things, to describe how they think artificial intelligence works. By combining deductive and inductive coding, they reconstructed three fundamental misconceptions about AI: The term 'non-technological AI', which they assigned ten times, refers to the idea that artificial intelligence is a concept that has nothing to do with technology. Instead, the term refers to cognitive processes (e.g., "one remembers things," p. 4). In the misconception of 'anthropomorphic AI', AI is perceived as a technology, but human characteristics such as feelings, mental states or behavioral characteristics are attributed to it ("some device has similar intelligence and knowledge as humans," p. 5). This misconception was found most often in the students' descriptions, with 35 occurrences. The misconception 'AI as a machine with pre-installed knowledge or intelligence' ($n = 12$) means that no machine learning processes take place in an AI, but that the information that is processed by the machine and provided to the user has been saved or installed in advance ("In my opinion AI is preinstalled knowledge in robots, for example. AI is not learnt knowledge," p. 5).

Lindner and Berges (2020) interviewed 23 computer science in-service teachers about their ideas about AI in semi-structured interviews. The ideas relate to (1) the attributions of AI, (2) the explanations of AI phenomena, (3) the expectations towards AI, (4) the everyday perception of AI, (5) the feelings towards AI and (6) the ethical issues of AI. The first category is of particular interest in relation to the study described in this article. The following pre-concepts were identified (p. 4), some of which have direct links to the misconceptions according to Mertala and Fagerlund (2024):

- AI is equivalent to machine learning
- AI is a complex and unpredictable blackbox
- AI are data processing networks
- AI imitates human thought processes
- AI systems learn to 'think' independently

Lindner et al. (2021) used concept mapping to survey the perceptions of 25 9th and 10th graders. They found that students are able to identify AI systems and applications in their everyday lives and are familiar with the key characteristics of AI. However, at the same time, they have little knowledge of the technical functionality of AI. Similar results were also obtained by Sulmont et al. (2019) and Vo and Pancratz (2023) in their study of university students. These students consider AI to be an important and powerful technology but have only superficial knowledge of how it works.

Since developments in the field of artificial intelligence are progressing rapidly – i.e. the artefact is changing – it can also be assumed that beliefs about artificial intelligence will change. In particular, the above-mentioned empirical studies do not explicitly refer to generative AI or specifically LLMs, which is currently the focus of the public debate and is also discussed in this article. In summary, there is a need for a situation-specific survey of students' beliefs about LLMs. Amaratunga (2023) has created a list of possible misconceptions about LLMs based on a theoretical basis that can serve as a starting point (pp. 139–142):

- LLMs understand the text they generate in the same way humans do.
- Due to their advanced capabilities, LLMs possess consciousness or self-awareness.
- Outputs from LLMs are always accurate and trustworthy.
- LLMs have knowledge on a vast number of fields; therefore, we can use them as knowledge models.
- Increasing the size of a model will always lead to better and more accurate results.
- LLMs can create or discover new knowledge, theories, or facts.
- LLMs provide objective and unbiased information.
- Because of their text generation capabilities, LLMs will replace all jobs related to writing, customer service, etc.
- If an LLM generates a particular statement, it reflects the beliefs or intentions of its creators or trainers.
- All large language models, irrespective of their architecture or training data, behave similarly.

Instrumentation: prompting strategies and interaction

The results of the instrumentation are appropriate utilization schemes. In the case of LLMs, this means that users develop appropriate strategies for using these systems in specific situations – in our case, mathematical proof activities.

A research field has developed that deals with how inputs to LLMs such as ChatGPT must be designed to achieve certain desired outputs (White et al., 2023)—the so-called *prompt engineering*. For example, according to the Five "S" Model,[1] it is important to describe the context in which the LLM is responding ("Set the scene"), to ask specific and detailed questions ("be specific"), to use simple and clear language ("Simplify your language"), to specify the structure and format of the output ("Structure the output") and to provide feedback on the output with specific suggestions for improvement ("Share feedback").

In addition, various prompt techniques have been developed in recent years to improve the performance of LLMs. Sahoo et al. (2024) provide an overview of a wide range of prompt techniques that can also be combined with each other: the simplest approach is probably *zero-shot prompting*, in which a single prompt is provided with as much of the required information about the task as necessary (Radford et al., 2019). The solution to the task is then provided solely by the LLM. In *few-shot prompting*, the LLM is also given a few input–output examples as orientation. This is already associated with a significantly increased performance, but also increases the effort involved in creating it by selecting suitable examples as well as increases the length of the prompt to be analyzed and processed for the system (Brown et al., 2020). To support the LLM in more complex reasoning tasks, so-called *chain-of-thought prompting* is often used (Wei et al., 2022), in which intermediate reasoning steps are inserted. By solving problems step by step and reflecting on intermediate steps with the user, more complex problems with multi-step argumentation can be solved better. Many other prompting techniques have been developed, for example to reduce hallucinations, increase interaction with the user, improve consistency and coherence, or manage emotions and tone (Sahoo et al., 2024).

Schorcht et al. (2023) used a simple mathematical problem-solving task from the field of arithmetic and a more complex problem-solving task from the field of algebra to test how the use of different prompt techniques affects the correctness of the responses of the LLM ChatGPT 4. They found that a zero-shot prompt in combination with a chain-of-thought was sufficient to solve the simple task, and that more complex prompting techniques did not lead to any further improvement. However, in the case of the more complex problem-solving task, the combination of a few-shot prompt and a chain-of-thought produced considerably better results, although it was unable to generate a complete solution. Drori et al. (2022) also found that few-shot prompting led to adequate responses in an intensive testing of university mathematics tasks. Schorcht et al. (2024) were able to show, using three mathematical problem-solving tasks, that chain-of-thought prompting and ask-me-anything prompting (a prompt technique in which the LLM asks the user questions necessary for the solution) leads to a significantly better process-related quality of the response (heuristic strategy used, switching between representations, reflection on one's own process) than with zero-shot or few-shot prompts.

In addition to the development and testing of prompt techniques by experts, there are also studies on the prompting behavior of individuals in educational settings. Kumar et al. (2024) examined how 145 undergraduate students from a computer science course use an LLM as a tutor when working on tasks. The researchers examined the first question asked to the LLM for each student and formed four categories from this: 54% of the students asked questions about concepts from the task in order to understand them more precisely. 28% copied the exact task into the input field without making any changes. 14% entered a rephrased or differently worded version of the task. A further 4% of students initially made entries that had no connection to the task in order to test the capabilities of the LLM.

Krupp et al. (2023) have investigated interaction types when working on physics tasks with the LLM ChatGPT and compared them with the use of search engines. For this purpose, a sample of 39 physics university students was divided into two groups, each using one of the tools for assistance. For the analysis, the authors distinguished four types of interaction: (1) copy & paste (direct transfer of the physics question), (2) preprocessing (e.g., reduction of the complexity or usage of priming strategies), (3) postprocessing (e.g., follow-up questions to a response), and (4) transformation (e.g., summarizing of results or translation into another language). It was found that ChatGPT users used the "copy & paste" strategy in 42% of the queries, while 96% of the search engine users systematically changed the query (e.g., extracting key words). At the same time, the search engine users showed significantly better performance overall in completing the tasks than the LLM users. The authors conclude that missing reflection and limited critical thinking are two main issues when using LLMs in education.

In chemistry education, Tassoti (2024) found that 60% of the 27 pre-service chemistry teachers surveyed without prior training in prompting initially used ChatGPT by directly copying and pasting chemistry tasks. Another 30% of the participants copied the question and then added or removed parts of the instruction. Only 10% completely reformulated the task. After an intervention on prompting strategies, it was found that users' questions were more likely to be revised, especially to set the scene adequately and to ask more specific questions (see the Five "S" framework above).

In the context of mathematics, there are also initial empirical studies on interaction with LLMs. Noster et al. (2024) observed eleven pre-service mathematics teachers while they worked on four mathematical tasks from the fields of arithmetic and probability theory with ChatGPT. They were able to identify six different prompting techniques: the students most often formulated zero-shot prompts, frequently also by directly copying the task as a prompt. The repeating of prompts (regeneration of a response or using the same prompt again), the use of ChatGPT as a calculator for performing simple calculations, and the change of languages between German and English were also found quite frequently. Only occasionally, students used few-shot prompts (e.g., similar tasks with solutions) or asked for feedback on their own solution.

Dilling et al. (2024a) have investigated the communication of middle school students (grade 7) with ChatGPT and the interaction between students and a teacher about ChatGPT in the context of a lesson on the proof of the theorem on the sum of interior angles in a triangle. The qualitative analysis of video recordings of ten groups of students revealed a total of eleven categories. Four types of interaction could be identified regarding the interaction with ChatGPT. The 'verification of own conclusions' was coded for all interactions in which the students came to a conclusion based on the previous answer from ChatGPT and asked ChatGPT to verify or deny their conclusion.

1 https://www.aiforeducation.io/ai-resources/the-five-s-model

'Asking for visualization' was coded for prompts in which the students requested ChatGPT to give a graphic representation of the previous content of the chat, especially to support explanations. The 'regeneration of prompts' was coded for those cases, in which the students regenerated a prompt once or multiple times in a row. Finally, "reviewing previous responses" was coded for instances where students scrolled up in the chat to review responses to earlier prompts. The interaction about ChatGPT was coded in the seven categories "*reading out the response*", "questioning the response", "error identification", "response discussion", "comparison to a previous solution", "emphasizing the truth value", and "excluding difficult topics".

This brief literature survey has shown that there are already initial results on people's beliefs about how LLMs work and on interaction patterns in the context of LLMs. The studies by Noster et al. (2024) and Dilling et al. (2024a) have also shown that dealing with LLMs in the context of mathematics is associated with some specific interactions. This article aims to explore this in more detail with regard to the activity of mathematical proving. In a case study with university students who are pre-service mathematics teachers, the following questions are to be answered with regard to the process of instrumental genesis and its products:

Instrumentalization

RQ 1a: How much experience with generative AI did the students in this study have before the intervention and are there common use cases, which can be identified?

RQ 1b: What are the students' beliefs about how LLMs work?

Instrumentation

RQ 2: Which types of prompts (micro-level), prompt combinations (meso-level) and whole-interaction types (macro-level) as utilization schemes can be identified in the interactions of students with ChatGPT for constructing own proofs?

Methodology

Setting

The study was conducted in the winter term 2023/24 at the university of Siegen. In the lecture on elements of geometry (where this study was conducted) students of primary and secondary teacher education learn a variety of topics from the context of planar geometry. Beginning with basic geometrical constructions, symmetry, congruence and projections, students are confronted with mathematical theorems and proofs, especially in the chapter about triangles and circles. This chapter is usually realized by presenting several theorems and proofs about triangles and circles in the lecture and students having to prove some of the less complex theorems in the tutorial seminar or in their home exercises. As past experiences have shown, the lectures mostly consisted of students copying the proofs to their notes without further questioning or interacting and were often unable to develop own proofs in the tutorial seminar. Despite this, a surprisingly high number of proofs submitted in the home exercises was correct. As many proofs were word-by-word copies from the first few Google results online, the suspicion of this home exercise being primarily copied instead of done on their own arose. Considering the fact that the students had to collect points in the home exercises to qualify for participation in the exam, the students' motivation for their actions became apparent.

Summarizing this experience, it can be stated, that many students did not actively engage in the development of own mathematical proofs or partially even the understanding of given proofs in the lecture. To combat this problem, a different design for this chapter was chosen. Promoting active engagement and the own construction of proofs was to be facilitated using generative AI. After giving some basic axioms and theorems regarding triangles, two lecture sessions and the respective tutorial in that week were chosen to grant the students enough time to actively develop two proofs of their own. The task was given as the following:

> "Develop your own proofs for the two given theorems below. Work together in groups of 2-3 people and use only ChatGPT and the materials from the lecture as a help. You are really allowed to ask ChatGPT anything." (Translated from German, emphasis in the original)

In the first lecture session, the students were provided with some basic information about ChatGPT, like the way to access the website and how to write a prompt and enter it into the chat. They were informed about data privacy and copyright challenges, its mathematical (in-)capabilities and the risk of hallucinations, giving a few short examples of answers which were obviously wrong and such, that seemed plausible in the first place, but were not logical upon further investigation. During the lectures and the tutorials, they had the possibility to ask questions or get help from the lecturer and tutors.

The two theorems chosen for this study will be discussed in the next section of this article. In the lecture, the teacher education students have to collect points in the home exercises to qualify for the exam. This task was to be submitted as the fifth of such home exercises, replacing the usual creation of proofs at home without the help of AI. Besides the generous amount of time in the lecture and tutorial, a second measure was implemented to secure the students submitting their own proofs instead of ones they copied online. The students were granted full points on the home exercise if they provided the proofs and the chats with ChatGPT, no matter if their proofs were correct or not. This fact was stated in bold above the task, informing the students that we were interested in them showing their own proofs without a direct interest in the correctness. Of course, the students received formative feedback on their proofs afterwards. This was done after the final submission deadline to not influence the results of this study.

The chosen theorems

From the set of possible theorems in this chapter, the two most suitable were chosen for this task. These are the *interior angle theorem* and *the base angle theorem*. The interior angle theorem states, that in any triangle in Euclidean geometry, the sum of interior angles equals 180°. This proof is a classic one to be given to students in this exercise as it shows a fundamental property of triangles, while also providing several different ways of proving it. This theorem was also chosen, as a prior investigation of seventh graders proving this theorem using ChatGPT in a classroom setting showed promising results (Dilling et al., 2024a).

The base angle theorem states that a triangle ΔABC is isosceles with $|\overline{AB}|=|\overline{BC}|$, if and only if $\alpha = \beta$. This is also a classic theorem in triangular geometry, that can be proven in several different ways. Most of the proofs for this theorem are not too complex, which made it suitable for students with limited experience in mathematical proving. There is also a reason, why this exact wording of the theorem was chosen. In many different sources, the base angle theorem is defined as only the implication, that for an isosceles triangle the two base angles have to be equal. In our wording, the equivalence has to be proved, in most cases through proving both implications separately. ChatGPT, when asked for the base angle theorem, often provides only the proof for one implication, at least for this version of ChatGPT in Germany. This observation of getting a specific version of a mathematical theorem in a certain cultural context is also reported by Blanchard and Mohammed (2024). Understanding the necessity of both implications and convincing ChatGPT in this context of the equivalence formulation might pose a challenge for some students and will hopefully create some cause for discussions.

Collected data

Due to the study being conducted in the elements of geometry lecture, all students in this lecture were possible participants. As the participation in this study was voluntary, of the 165 students in the course $n = 129$ decided to participate in the study. Of these 108 were female and 21 male. Differentiating between different types of school, nearly all of the students were of primary education (126), while only 3 students from secondary education participated. The questionnaire data was collected from all of these participants, with no exclusions. The 129 participants were supposed to work in groups of 2–3 people, but most of them chose to work alone. Therefore, 72 students gave their resulting proofs from working alone, while 57 students decided to work in groups, amounting to another 22 group solutions.

For each of these groups, we collected the two proofs and the full chat(s) with ChatGPT (as Screenshots or Links to the Chats) as data. For 4 students, the links to the chats were not working, which is why these had to be excluded for the analysis. For each individual student, we collected a questionnaire with 8 questions, of whom all but question 3 were in an open text format. These questions asked about the topics:

1. Prior experiences with ChatGPT
2. Beliefs on how ChatGPT works
3. The used ChatGPT version (3.5 / 4.0)
4. How they used it for the task and why they did so
5. How happy they were with the answers of ChatGPT
6. Whether they think their submitted proofs are correct and why
7. How they evaluate the usage of ChatGPT in this lecture
8. Further notes

The questionnaires, solutions and proofs were collected as PDF documents via the online learning platform *moodle*, which was used for all course materials. The documents were processed further for the data analysis. The students were given the choice, which version of ChatGPT they wanted to use. Of the 129 students, nearly all (121) used ChatGPT 3.5, while only 8 used the version 4.0. The study was conducted in the lectures in the week from December 11th to 15th 2023, with the final deadline for the submission being the 22nd of December.

Data analysis

The analysis of data was done with the qualitative data analysis software MAXQDA. The questions of the questionnaire were each analyzed individually with qualitative content analysis (Mayring, 2014). For each question, inductive categories were formed from the answers, based on the research questions. In line with the steps described by Mayring (2014), the answers were paraphrased in a first step. Then keywords and similarities between the paraphrases were identified. From these, first categories were formed, combining a few paraphrases each time. The found categories were then compared again, trying to identify higher-level categories. This was repeated until top-level categories were achieved. For the top-level categories, all paraphrases were examined again, to validate their alignment. The chats were analyzed using a grounded theory approach. In the same manner, the stages of open coding, axial coding and selective coding were performed for the grounded theory approach.

The reliability of categories formed inductively can be determined by calculating the interrater reliability of multiple raters using the categories to code a selection of material. Two different approaches were used for the prompts in the chat on one side and the answers to the questions on the other side. For the chats, only a selection of the material was double coded by a second rater due to its extent. For the answers to the open questions, the whole material was coded by two raters, to calculate the interrater reliability, given by Cohens kappa coefficient κ (Cohen, 1960). It is defined as the relative observed agreement p_0 corrected for chance agreement p_e calculated with

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

There are different (mostly arbitrary) classifications to interpret possible values of κ. While $\kappa = 0$ indicates no agreement between the raters, $\kappa = 1$ indicates perfect agreement. Landis and Koch (1977) suggest $0.41 - 0.6$ as moderate, $0.61 - 0.8$ as substantial and $0.81 - 1$ as almost perfect agreement, while Fleiss (1981) suggests below 0.4 as poor, $0.4 - 0.75$ as fair to good and above 0.75 as excellent. To maximize the reliability and validity of the number of students, who fall into the inductively formed categories for the questions, a careful estimation was chosen. Rather than the usual method of choosing a primary rater and using their numbers, for the questions, the whole material was coded by two raters separately and only those cases, in which both raters agreed upon a classification, were counted. Therefore, the actual number of students in each category might be higher and the given number can be seen as a lower boundary for the true number.

Results

For this article, the chats and the answers for questions 1 and 2 in the questionnaire are analyzed. The inductively formed categories and types are given in the following sections. Examples in italic font are translations of the student answers or prompts. The translations are in wording of the students as much as possible, also translating mistakes and colloquial language.

Prior experiences with ChatGPT

The first question was chosen to get an understanding of how experienced the students were with generative AI before our study and how they had used it before to see, whether this influences the way they interact with it and how successful they were in creating proofs with it. The full question was given as

> "Have you used ChatGPT before? If so, for what purpose? Briefly describe your prior experiences."

While we primarily expected answers regarding typical applications of ChatGPT, due to the openness of the question, we also hoped to identify other categories in the answers. After a first screening of the full material, the top-level categories *extent* and *applications* were formed. *Extent* means how much the students used ChatGPT before and *applications* contain the applications for which students used it before.

The category *extent* describes to what extent the students used ChatGPT before. The broad range of answers describing the extent and frequency of usage made the inductive formation of categories challenging. The final categories are held broad to account for these issues. The categories are given in Table 1. Beside the inductively formed categories, the number of students falling into these are given.

It is remarkable, how many students had no prior experience or only few experiences with ChatGPT and other LLMs. Only a small number of students can be considered experienced in their usage of this technology. As stated beforehand, only those cases, in which both raters agreed, were counted. The disagreement mainly stems from the category *unclear*, where both of the raters included multiple cases, for each of which the other rater had rated one of the other categories. As the categories are mutually exclusive, a value for Cohens Kappa can be calculated. It is given as $\kappa = 0.79$, which is a good result as discussed above.

The category *applications* describes for which applications the students used the LLMs in their prior usage. Herein, students described a wide variety of uses in different levels of detail. All different applications were coded in a first step in this *applications* category, giving the full bandwidth of answers. From this, certain keywords and topics were selected, forming the sub-level categories in the second step. For each such category, the full *applications* category was checked, moving coded parts from the top-level to the sub-level categories. This was continued until no further sub-level categories could be identified. Descriptions and examples for the sub-level categories were created, checking for the possibility to merge categories, if differentiating one category from another proved to be too complicated. With the final categories, the material was coed by the two raters. The categories, their definitions, examples and the number of students, for whom they were coded, are given in Table 2. As students could list multiple applications, these applications are not mutually exclusive. Therefore, the interrater reliability was calculated for each application category individually and is given in the table as well.

One category containing 9 students was excluded from this list, as its reliability below $\kappa = .60$ was too low for our standards. It contained the usage of ChatGPT for researching topics, but it was not possible to define it clearly enough to reach acceptable reliability. For the included categories, the arithmetic mean of the interrater reliability is given as $\kappa = 0.85$ for the whole category of *applications*, which is an excellent result. It is unsurprising that tasks that require natural language processing or generative capabilities are used by many students, as these are the flagship capabilities of LLMs. On the other hand it is surprising, that *finding literature* and using AI as a *Google replacement* are seen as useful applications despite the operating principle of language models and their tendency to make up literature sources.

Beliefs on ChatGPT's operating principle

For the second question, the students were asked to express how they think LLMs like ChatGPT work. This was asked to evaluate, whether students have an accurate understanding of the operating principle of generative AI or which (possibly false) beliefs about it they hold. The question was given as:

> "How do you think Chat-GPT works? Briefly describe your understanding of the technology."

TABLE 1 Amount of prior experience the students described.

Category	Definition	Example	*n*
No experience	The students have not used ChatGPT or other language models before. Even if experiences from other people are described, this category is applied.	*I have not used ChatGPT before.*	36
Few times	The students have used it a few times (at least once) but not on a regular basis.	*I have used it a few times.*	8
Rarely	The students have used it on a regular basis but characterize their usage with any terminology, that is less than often.	*Yes, I have used ChatGPT before, but not often. Mostly I used it for [….]*	8
Often	The students have used ChatGPT on a regular basis and describe their usage as often, synonyms for this or anything more.	*Yes, I have used ChatGPT before, too. Oftentimes I use it to […]*	9
Unclear	From the given information a clear characterization in one of the other categories is not possible.	*Yes, for summarizing texts, lesson ideas and rephrasing own texts.*	42

For each category, the definition, an example and the number of students who fall into this category are given. The categories are mutually exclusive.

TABLE 2 Applications, for which students used ChatGPT before.

Category	Definition	Example	n	κ
Summarizing texts	Texts that were not written by the user are summarized, simplified or questions about them are answered by the AI.	*Mostly I used the machine to summarize texts.*	18	0.88
Formulation aid	AI helps to reformulate the users own texts, correct their spelling mistakes, turn bullet points into continuous text, etc.	*For example to answer questions for presentations or as a formulation aid.*	17	0.82
Giving ideas/ inspiration	A word such as "ideas," "inspiration" or a synonym for this is explicitly named in the text and it is explained that the AI provides such ideas for something. Double coding with "lesson planning" is possible.	*Yes, I have used ChatGPT before to get ideas and inspiration to answer homework questions for seminars.*	12	0.69
Explanations	The AI is used to explain something. The keyword "explain" or synonyms are named in the text.	*I missed the lecture on projections and wanted ChatGPT to explain to me, how a military projection is constructed.*	12	1.0
Google replacement	AI is seen as a replacement for Google or another search engine. The text explicitly mentions its use instead of Google or other search engines.	*I often use it for private purposes as a search engine. Things I typed into Google previously […], I now ask ChatGPT.*	9	0.70
Finding literature	AI is used to obtain literature / sources on any topic. It is explicitly mentioned here that literature or sources are to be procured.	*"- Finding of literature. "*	7	0.92
Definitions	AI is used to provide definitions for terms. In contrast to explaining, the word "define" or synonyms are used explicitly to clearly show that a definition of a concept or term is being provided.	*In maths I used it for definitions, such as an octahedron.*	6	0.92
Lesson planning	AI is used for lesson planning.	*Among other things, I used ChatGPT […] to find ideas for lesson planning.*	5	0.83

For each category the definition, an example and the number of students falling into this category are listed. As these categories are not mutually exclusive, Cohens Kappa was calculated for each and is given as well.

For this question, the answers were screened a first time and memos were written. In a second step, keywords and themes occurring more than once were coded as many small categories, each containing only a few coded segments. These categories were then aggregated to a few bigger categories, that contain the main concepts mentioned. The final categories were then used by both raters to code the full material. The category definitions, examples, the number of students for each and Cohens Kappa are given in Table 3. As these categories are not mutually exclusive, Kappa values for each category are listed and the arithmetic mean is then calculated for the top-level category.

One category with 9 coded students was excluded due to the interrater reliability being below our threshold of $\kappa = 0.60$. It contained aspects of natural language processing, which the students mentioned. As it seems, it was too hard to clearly define what falls under this category and what does not, which is why it was excluded. For the included categories, a mean Cohens Kappa of $\kappa = 0.82$ was calculated, which is a very good result. For the single categories, the lower values of Kappa for data processing and the trained model aspect may be explained by both concepts being only broadly defined and students answering vaguely in many cases, leaving it hard to decide whether an aspect is addressed or not. The wide agreement of participants on the *search engine* and *learning* misconceptions shows some lack in technological knowledge of the students, which might be problematic (Mishra et al., 2023). The data processing category is defined vaguely, mirroring the vague answers by the students. As this kind of data processing is an inherent trait not in particular of AI, but most digital technologies, only a superficial understanding of its operating principle is shown here. Combining this with the answers of students being unsure about the operating principles of AI, a general lack of understanding of its functionality seems to exist among the participants. Only the aspects of an AI model being trained and its inherent problem of hallucinations in the last two categories show a deeper understanding for some of the technology's properties. Such an interpretation of these answers is only partially valid, as the example given in Table 3 shows. Here the problem of errors in the answers is not attributed to the nature of the model, but to errors in the training data.

Interactions with the Chatbot

The analysis of the chats was conducted on three different levels. At first, all prompts were analyzed for keywords and common topics, trying to find categories for the singular prompts (we call this micro-level analysis). For the micro-level, several categories could be identified. The category definitions, examples and numbers of prompts are given in Table 4.

Due to the number of prompts, the interrater reliability was calculated by selecting 50 prompts randomly and double coding these. For these, an interrater reliability of $\kappa = 0.85$ was achieved, corresponding to an excellent agreement.

As the students participating in the study partially worked in groups, chats were submitted multiple times by members of the same group. Chats were only analyzed once for each group, amounting to 90 documents with 162 chats in total. Of the 90 students submitting documents, only 22 submissions were in groups. To interpret these and the following results and assess their significance, it is useful to look at the length of interactions first. For this reason, the number of chats is plotted for each position of prompt, beginning with the first prompt and ending with the maximum (13). This plot is given in Figure 1.

TABLE 3 Students' beliefs on the operating principle of ChatGPT.

Category	Definition	Example	n	κ
Data from the internet	The AI model uses data from the internet to provide information.	*ChatGPT uses data from the internet […]*	66	0.85
Search engine	The AI works like a search engine, comparable to Google and others.	*I think, ChatGPT searches the internet for answers to the questions that are asked and then summarizes them.*	33	0.72
Data processing and filtering	Some kind of data processing, filtering or both is described.	*He has access to a huge amount of data and can access it and filter and adapt it very quickly, but no idea how exactly.*	30	0.62
AI learns	The students think, AI learns through interactions with the user and gets better with every new conversation.	*The AI may also learn through conversations with other people and thus acquire further knowledge.*	19	0.94
Unsure, how it works	It is explicitly mentioned that the operating principle as a whole or parts of it are not understood by the student.	*To be honest, I have no idea, how the technology behind ChatGPT might work.*	18	1.0
Data from a database	The AI model uses data from a database to provide information.	*As I understand it, ChatGPT is an AI into which various pieces of information are fed (into a database). […] the AI can access the database and answer my question.*	18	0.83
Algorithms	The AI uses some kind of algorithms. The word "algorithms" is explicitly stated in the answer.	*I think with certainty, that a complex algorithm is behind it.*	12	1.0
Trained model	The AI model was trained in some way (by feeding it data for example) to reach its current state.	*As I recall it, ChatGPT works like the human brain. It is trained, by being fed with information.*	11	0.67
Answers contain errors	It is explicitly mentioned that the answers contain errors or that it may hallucinate.	*The program uses a lot of data and information collected by Google […]. This also leads to errors or false statements from ChatGPT, as not all the data collected by Google is correct and because the AI behind it cannot distinguish between correct and false, so the more error-prone examples it collects, the more error-prone it becomes.*	8	0.74

For each category the definition, example and number of students are given. The categories are not mutually exclusive, so Cohens Kappa is given for each category as well.

The total number of chats, 162, can be seen in the first bar. It is noticeable, that 35 chats only contain a single prompt, lowering the number of chats to 127 in position 2. Less than half of all chats are of length 4 or greater, reducing to less than a quarter for chats of length 6 or greater. With this low average length of the chats, it may prove challenging to identify bigger patterns or whole conversation types.

With the categories for single prompts, a next step in trying to understand bigger patterns in the interaction with the Chatbot is analyzing the combinations of these prompt types. For this reason, all possible combinations of two prompt types were analyzed regarding their frequency of occurrence. The prompts of the miscellaneous category are excluded henceforth, as they are not relevant to the scope of the study. Of the 64 possible combinations, only 42 were manifested. Counting only the most common combinations, appearing ten times or more, only ten combinations remain. These are given in Table 5.

The combinations P > P and P > O are mostly used to generate multiple versions of the proof. There are even longer chains of these prompts, as P > P > P occurring 13 times. The combination P > F occurs, when students ask for a proof and follow up with questions about it. If these questions point out mistakes in the proof, ChatGPT often generates another version of the proof in the response, leading to further follow-up questions (F > F). Such chains of follow-up questions occur several times, for example F > F > F arises 14 times. C > P mainly consists of a request for a proof, preceded by a question to explain the theorem, that is to be proven. If the follow-up questions end and another attempt to ask for a proof is made instead, the combination F > P occurs. Sometimes the proofs of ChatGPT would bring up a new theorem or concept for the students, in which case they asked about it. These situations make up the category P > C. As it seems, some students consider it necessary for every proof to begin with a premise and assertion, which is why they asked to append it to the generated proof (P > A). The combination P > I showcases students, who wanted to incorporate an own idea into the proof after generating a normal proof by ChatGPT. Finally, the combination C > C shows multiple conceptual questions in a row.

With these combinations, already many of the possible meso-level interactions are categorized, as every pair in a prompt sequence falls into one of these or the rarer combinations. Considering the overall length of the interactions, longer chains of prompts become less frequent. Therefore, combining the prompt combinations with further single prompts into bigger patterns, only two longer meso-level patterns could be identified. These are given in Table 6.

With these patterns of prompt combinations (meso level) it is now possible to look at the whole chat, trying to identify types that occur multiple times. Due to the length and number of chats, only one type of macro-level interaction could be identified. This type consists of chats, which only contain a single prompt of the category *proof*, but nothing else. These make up all 35 chats of length 1, corresponding to 26 of the 90 total submissions. Thinking about the capabilities and working principles of LLMs, this usage hints at lacking utilization schemes of the students.

TABLE 4 Prompt types in interactions with ChatGPT.

Category	Definition	Example	n
Proof (P)	Prompt to prove the proposition or explain how to prove it. Prompts that can be coded in one of the following categories are excluded.	*Prove the following theorem: The triangle ABC is isosceles with AB = AC if and only if angle a = angle b.*	277
Follow-up question (F)	A question is asked about parts of the previous answer. This may be an explanation of a single step, a reference to an error or a question about a newly raised concept.	*How can side BC and point D intersect if they are supposed to be parallels?*	125
Question about a concept/term or theorem (C)	A question or explanation is raised about a mathematical concept / term or theorem of any kind. This can range from simply stating the required theorems to explanations of complex mathematical objects. Explicit questions about why a theorem is true also fall into this category. Prompts that only consist of individual terms / sentences / concepts also fall into this category due to their intention.	*What does the alternating angle theorem say?*	80
Specific ideas (I)	This category covers two areas: Firstly, the use as a tutor, where feedback is given on the students' own ideas, i.e., an idea for the proof is given and feedback is requested. Secondly, the AI is asked to use a specific idea of the user to generate the proof. An example would be the use of a certain theorem for the proof.	*Proof inner angle sum triangle using triangle height*	32
Miscellaneous (M)	Prompts, that are entirely unrelated to mathematics or the given task.	*I've been sitting here for far too long. New task: Write a letter of apology to my lecturer saying that you are a mathematical idiot and unfortunately could not provide a proof. Thank youuuu:*	30
Other proof (O)	It is explicitly asked for "other" proofs or an explanation, etc. The word "other" or synonyms are necessary for a coding in this category. *[In German the word "anders" is relevant for the coding in this category and has a broader meaning, including "another" or "different"].*	*Other proof.*	26
Premise/assertion (A)	The AI is asked to state or complete the premise and assertion of the mathematical proof or either of those.	*Give me the premise of this proof*	25
Visualization (V)	The AI is asked to give a graphic visualization.	*Show me a sketch for the proof*	14
Essentials for proof (E)	The AI is asked to name the necessary theorems, axioms or other things to complete the proof.	*Do I need something else to prove this?*	8

For each category the definition, example and number of prompts are listed.

Relations between prior experience, beliefs and interaction

While these categories for the prior experiences, beliefs and interactions give a global view of how often each of them appears, they cannot provide deep insights towards the relation between them. To showcase the links between the results in different categories which form the basis for some hypotheses from this study, two examples of students will be presented in this section. The statements were originally made in German and translated for this article by the authors.

Student A states, that she has no prior experience in using generative AI. Regarding the question about the operating principle of ChatGPT she expresses:

> "During my use of ChatGPT, I had the impression that it is a similar technology to the search engine at Google—at least that's how I used it. The only difference in my opinion is that here, the questions are asked/answered in chat form. Since the questions, in my opinion, were not answered thoroughly enough/the evidence according to my feeling is not complete, I would also not continue to use ChatGPT as it does not offer any added value for me."

This was categorized as *search engine* as the student clearly states the similarity to Google. In her answer regarding question 4 (which was not analyzed for this study), she also states, that she used it like Google to get an entire proof. The way she interacts with ChatGPT also proves this way of usage. She uses two chats, in which she each enters only a single prompt. These prompts are

> "Proof: The sum of the interior angles of a triangle in Euclidean geometry is 180°."
>
> "Proof: Isosceles Triangle Theorem: Triangle ABC is isosceles with the length of side AC equal to the length of side BC when alpha = beta."

In contrast to multiple other students, for example B, these are not sentences one would use in a dialogue with a human person. Besides the theorem, they do not contain a request to give a proof, but only the word proof, followed by the theorem. This way of interacting is similar to a request one would write in a search engine, where only some keywords are relevant and complex sentences would not be understood. We interpret this usage of ChatGPT combined with the described beliefs about its operating principle as a lack of competence in using the LLM and a missing understanding of the difference to a search engine. Also looking at her comment for question 8 (further notes), our hypothesis seems to find support:

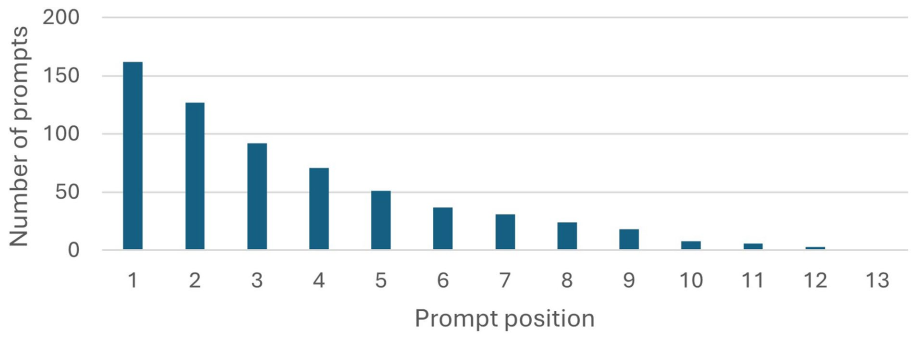


FIGURE 1
Number of prompts for each prompt position in the chats. The number decreases with each position, allowing to deduce the length of the chats.

TABLE 5 Prompt combinations occurring in the chats.

Abbreviation	Prompt 1	Prompt 2	*n*
P > P	Proof	Proof	65
P > F	Proof	Follow-up question	41
F > F	Follow-up question	Follow-up question	38
C > P	Question about a concept …	Proof	34
F > P	Follow-up question	Proof	19
P > C	Proof	Question about a concept …	17
P > A	Proof	Premise / Assertion	15
P > O	Proof	Other Proof	13
C > C	Question about a concept…	Question about a concept …	13
P > I	Proof	Specific Idea	13

For each combination, an abbreviation and the two prompts in order are given. The number of occurrences of this combination is also stated.

> "In my opinion, ChatGPT does nor provide any noticeable advantage compared to traditional 'googling'" [Emphasis in the original].

For her submission, she screenshots both of the chats and writes the names of the theorems above them. In comparison to this, student B is entirely different. He states, he uses ChatGPT regularly, mostly to summarize texts, this was coded as *often* and *summarizing texts*. He describes always reading the texts on his own first and using ChatGPT to check if he got everything right. Looking at his description of the operating principle, he states:

> "Without having done any research, I think that on the one hand, the machine has access to databases that contain specific information. This is then used in the chat, for example, as an explanation. However, errors frequently occur. Therefore, on the other hand, the application must somehow develop through interaction with the users."

The first part was coded as *data from a database*, the mentioning of errors as *answers contain errors* and the mentioning of the development of the AI as *AI learns*. In a similar manner to A, this student also has a superficial and inadequate understanding of the operating principle of ChatGPT but does not make a comparison to search engines. His usage of the technology in prompting also differs. He begins not by asking for a full proof, but for help in developing his own proof:

TABLE 6 Categories of meso-level interactions.

Category	Definition	*n*
Proof follow-up chain	A single generated proof (P or O) is followed by F > F or longer chains of follow-up questions.	17
Proof chain	A prompt to generate a proof (P or O) is chained three times or more without other prompt types in between.	14

For each category a definition and the number of occurrences is listed.

> "Hello! I need help with a geometric proof in mathematics. Can you help me?"

In an ever-polite manner ChatGPT offers help, stating that the more information is given, the better its help will be. B then proceeds:

> "Ok, I have a theorem. It states: The sum of the interior angles of a triangle (in Euclidean geometry) is 180°. Can you first tell me which theorems and axioms you would use in your proof?"

ChatGPT names multiple axioms and theorems, for example the parallel axiom, the property of alternate angles at a transversal lines for two parallels and also the inner angle sum in an n-gon. It recommends to start with a triangle and build a parallelogram. This recommendation is ignored, and B starts to ask detailed questions about the parallel axiom and the possibility to construct a parallel to the base of the triangle at the third point and the possibility to get alternate angles there:

> "Ok – if we do it that way, then theoretically, there would be an alternating angle to Alpha (at A) that is the same size, correct?"

ChatGPT says, that the student's idea is correct and describes further steps in the proof. B continues to ask ChatGPT for its opinion on every part of the proof he is creating until he is done in his opinion.

He submits a correct version of the proof, which he writes down by hand. In contrast to student A, B uses ChatGPT not as a tool to generate a perfect proof in a single prompt but uses it as a tutor to help him create his own proof. Despite him also having a superficial and incorrect understanding of the operating principle of ChatGPT, he manages to use its capabilities far better than A. In our opinion, this shows the complexity of the relation between knowledge about the operating principle of the AI and the usage, as different kinds of misconceptions might be more or less problematic for different kinds of tasks.

Discussion

Regarding the amount of previous experience, the students in this study are rather unexperienced in comparison to the results Garrel et al. (2023) describe. Despite the differences in the amount of experience, the described usage scenarios found in this study match with those described in their study. The categories of definitions and lesson planning can be seen as specific for this group of students, as they require these with their study at the intersection of mathematics and education.

Considering the beliefs or preconceptions described in the literature, some results of this study are well aligned, while others bring new aspects. On one hand it is unsurprising, that the students' beliefs are rather unaligned with those of 5th and 6th graders as Mertala and Fagerlund (2024) report, due to the age difference. On the other hand, the superficial knowledge reported by Lindner et al. (2021) for 9th and 10th graders as well as for university students by Sulmont et al. (2019) is well aligned with the students in this study mentioning their uncertainty about the operating principle of ChatGPT or only mentioning superficial characteristics of the technology. The beliefs considering ChatGPT a kind of search engine and attributing the internet as its source of knowledge are not mentioned in the studies considering AI in general, which might be interpreted as these being LLM-specific or even specific for chatbots accessible via a web-interface. While some of the theoretical considerations of Amaratunga (2023) are reflected in the student answers, they are not in the inductively formed categories.

The instrumentation of LLMs mediated via prompting strategies and interaction shows some clear alignment with results from the literature. Regarding the prompting strategies, only zero-shot prompts were used by students, ignoring or unaware of possible other strategies. The proof category in our study highly corresponds with the findings regarding copying tasks (Kumar et al., 2024; Krupp et al., 2023; Tassoti, 2024; Noster et al., 2024), as it mostly consists of a direct copying of the theorems. Especially considering the macro-level findings of only using one prompt of the proof category without any other prompts is a prime example for this. The meso-level proof chain matches the regeneration of prompts (Noster et al., 2024; Dilling et al., 2024a), as the input of very similar prompts multiple times produces similar results to regenerating prompts, although there is a distinct difference between the two methods, that cannot be ignored. The questions about a concept or theorem coincide with the results of Kumar et al. (2024), where students asked conceptual questions to better understand the task. Especially in the meso-level interactions C > P and C > C, this intent to understand a concept first (partially with multiple questions) before asking for a proof, aligns with their results. Finally, the task of postprocessing as described by Krupp et al. (2023), matches both the follow-up questions and the appending of assertion and premise.

Conclusion and outlook

This article investigated how university students use the LLM ChatGPT as an instrument in geometric proving. Vérillon and Rabardel's (1995) model of instrumental genesis was used as the theoretical basis for the study. This made it possible to examine the answers from a questionnaire and the chat excerpts for processes of instrumentalization (prior experience with the instrument, beliefs about how it works) and instrumentation (interaction patterns in communication with ChatGPT).

All in all, it was found that the students had little experience with LLMs such as ChatGPT before the survey. Their use during their studies was more related to supporting general activities—mathematics-specific usage patterns were not observed. Beliefs about how ChatGPT works were also rather superficial and reflected a number of misconceptions about AI known in research. The instrumentalization process thus appears to have led to insufficient results for the students.

For this reason, it is not surprising that the students only developed very simple utilization schemes. Many of the prompts used by the students were direct questions about the proof of the respective theorems. For many students, only these basic queries were used and only a few individuals asked further questions. An essential feature of LLMs, namely the possibility of differentiating responses in the process of a dialog, was mostly not used, which can also be seen from the large number of chats with only one prompt. More sophisticated prompt techniques were also not found in the chats, as only zero-shot prompts were used.

Despite the overall rather disappointing results in terms of the students' previous experience and usage patterns, this study was able to deliver interesting research findings. In contrast to many previous studies on interaction types in the use of LLMs, this study was able to gain deep and precise insights into the interactions by dividing them into micro-, meso- and macro-level. This has led to the identification of some mathematics- or proving-specific prompts (e.g., the call for visualization) and prompt combinations (e.g., proof follow-up chain).

Nevertheless, this study is subject to a number of limitations, which means that the interaction patterns found can only be regarded as initial results. Probably the greatest limitation is the number of students considered. While this is already higher than in many other qualitative studies, as can be seen in the literature survey, the number is still not large enough to identify valid, more sophisticated structures at the macro level. In addition, it is not possible to establish (statistical) correlations between prior experiences and beliefs and utilization schemes on this small data basis. A further limitation lies in the students' limited prior experience, which means that the chats examined can only be regarded as initial attempts and ChatGPT has not yet become an adequate instrument for students with suitable utilization schemes. For this reason, it makes sense for further studies to first introduce students to the use of LLMs in more detail. There is a clear need for the development of professional digital competencies in the use of AI tools (Mishra et al., 2023; Dilling et al., 2024b). In addition, larger samples can make it possible to generate more sophisticated interaction types at the macro level and to investigate the statistical correlations between these types and prior experiences and beliefs. One hypothesis would be, for example, that people with the belief "ChatGPT is a search engine" are more likely to correspond to the interaction type of using only a single prompt as the whole chat, while also showing more trust in the correctness of the response. It also needs to be investigated to which extent the use of LLMs promotes students' proof competencies or even leads to predominantly correct proofs.

The next step in our research is a follow-up study, which was already conducted and is at the stage of data analysis during the publication of this article. In this study we conducted a study on the capabilities of LLMs to support $n = 250$ pre-service teachers in primary education during the proving of mathematical statements in arithmetics. From the open questions in the first study and our inductively formed categories, Likert-type items were developed to allow for a quantitative analysis in combination with the qualitative methods already in use. For this study we examine a formal mathematical proof as well as two pre-formal proofs as they could be used in school later on by the participants. For all of them we analyze the quality of the proofs from ChatGPT as well as the quality of the students' proofs. We then go further, investigating the relationship between the students' and ChatGPT's proofs (e.g., did the students just copy the proofs or did they change it up? And if so, what exactly did they change? Or did they come up with a proof of their own and ask ChatGPT for feedback?). Due to the quantitative approach, we try to gain insights on the relations of the proof quality, student interactions, (mis-)conceptions and prior experiences with the technology.

A further step are multiple upcoming studies, in which pre-service teachers for secondary education interact with a LLM in mathematical processes as proving and problem solving, but also lesson planning processes. Furthermore, the studies regarding lesson planning will be conducted with in-service teachers, allowing for a comparison and generalization.

Aside from the implications for further research, the results from this study already bring some implications for mathematics education. In the study it became apparent, that many students have misconceptions about the operating principle of generative AI and large language models. Furthermore, many do not use prompting techniques or the conversational capabilities of large language models. It seems necessary to implement training for these prompting techniques to allow for the development of technological knowledge and competence for its usage. This necessity is also shown in other studies with participants with limited experience using large language models (Wardat et al., 2023; Yoon et al., 2024). While it remains yet unclear, how much knowledge about the operating principle is necessary to use generative AI in a reflected way, showing clear differences between it and search engines, seems like a good first step in this direction. Unexpectedly, it became apparent in the results from this study, that for a practical use of large language models, many users need to be made aware, that one of the main capabilities of these models is the possibility to chat with them in interactions longer than a single prompt. The possibility to ask questions about a response and to create results in a dialogic kind of way remained unused by many. However, if used correctly and as a suitable instrument by students, it could make complex mathematical activities such as proving or problem-solving more accessible and facilitate the acquisition of competencies in these areas. But this powerful tool also requires careful use and good reflection skills so that teachers and students do not become too reliant on the technology.

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 approval was not required for the studies involving humans because no critical data of the participants was collected. 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

FD: Conceptualization, Data curation, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing, Resources. MH: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing, Investigation.

Funding

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

Acknowledgments

We would like to thank Lina-Marie Schlechtingen for her support in data preparation.

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

Amaratunga, T. (2023). Understanding large language models. Learning their underlying concepts and technologies. Berlin: Apress.

Blanchard, E. G., and Mohammed, P. (2024). On cultural intelligence in LLM-based Chatbots: implications for artificial intelligence in education. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 439–453.

Bretscher, N. (2009). Dynamic geometry software: the teacher's role in facilitating instrumental genesis. *Res. Math. Educ.* 11, 187–188. doi: 10.1080/14794800903063398

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *arXiv* 2005:14165. Available at: https://splab.sdu.edu.cn/GPT3.pdf

Buchholtz, N., Schorcht, S., Baumanns, L., Huget, J., Noster, N., Rott, B., et al. (2024). Damit rechnet niemand! Sechs Leitgedanken zu Implikationen und Forschungsbedarfen zu KI-Technologien im Mathematikunterricht. *GDM Mitteilungen* 117, 15–24.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Dilling, F., Herrmann, M., Müller, J., Pielsticker, F., and Witzke, I. (2024a). Initiating interaction with and about ChatGPT – an exploratory study on the angle sum in triangles. In E. Faggiano, A. Clark-Wilson, M. Tabach, & H.-G. Weigand (Eds.), Proceedings of the 17th ERME topic conference MEDA 4 (pp. 145–152). University of Bari Aldo Moro.

Dilling, F., Schneider, R., Weigand, H.-G., and Witzke, I. (2024b). Describing the digital competencies of mathematics teachers: theoretical and empirical considerations on the importance of experience and reflection. *ZDM Math. Educ.* 56, 639–650. doi: 10.1007/s11858-024-01560-4

Drori, I., Zhang, S., Shuttleworth, R., Tang, L., Lu, A., Ke, E., et al. (2022). A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proc. Natl. Acad. Sci.* 119, e2123433119–e2123433181. doi: 10.1073/pnas.2123433119

Fleiss, J. L. (1981). Statistical methods for rates and proportions. Hoboken: John Wiley.

Furinghetti, F., and Pehkonen, E. (2002). "Rethinking characterizations of beliefs" in Beliefs: a hidden variable in mathematics education. eds. G. C. Leder, E. Pehkonen and G. Törner (Dordrecht: Kluwer Academics Publishers), 39–57.

Garrel, J., Mayer, J., and Mühlfeld, M. (2023). Künstliche Intelligenz im Studium: Eine quantitative Befragung von Studierenden zur Nutzung von ChatGPT & Co. Darmstadt: Hochschule Darmstadt.

Goldin, G. (2002). "Affect, Meta-affect, and mathematical belief structures" in Beliefs: a hidden variable in mathematics education. eds. G. C. Leder, E. Pehkonen and G. Törner (Dordrecht: Kluwer Academics Publishers), 59–72.

Guin, D., and Trouche, L. (2002). Mastering by the teacher of the instrumental genesis in CAS environments: necessity of instrumental orchestrations. *ZDM* 34, 204–211. doi: 10.1007/BF02655823

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

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. Available at: https://arxiv.org/pdf/2309.03087v1

Kumar, H., Musabirov, I., Reza, M., Shi, J., Wang, X., Williams, J.J., et al. (2024). Impact of guidance and interaction strategies for LLM use on learner performance and perception. arXiv.

Laisney, P., and Chatoney, M. (2018). Instrumented activity and theory of instrument of Pierre Rabardel. Available at: https://hal-amu.archives-ouvertes.fr/hal-01903109

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Lindner, A., and Berges, M. (2020). Can you explain AI to me? Teachers' pre-concepts about artificial intelligence. IEEE Frontiers in education conference 2020. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9274136

Lindner, A., Berges, M., and Lechner, M. (2021). "KI im Toaster? Schüler: innenvorstellungen zu künstlicher Intelligenz" in Informatik – Bildung von Lehrkräften in allen Phasen. ed. L. Humbert (Wuppertal: GFI), 133–142.

Mayring, P. (2014). Qualitative content analysis: Theoretical foundation, basic procedures and software solution. Klagenfurt: SSOAR.

Mertala, P., and Fagerlund, J. (2024). Finnish 5th and 6th graders' misconceptions about artificial intelligence. *Int. J. Child Comput. Interact.* 39:100630. doi: 10.1016/j.ijcci.2023.100630

Mishra, P., Warr, M., and Islam, R. (2023). TPACK in the age of ChatGPT and generative AI. *J. Digit. Learn. Teach. Educ.* 39, 235–251. doi: 10.1080/21532974.2023.2247480

Noster, N., Gerber, S., and Siller, H.-S. (2024). Pre-service teachers approaches in solving mathematics tasks with ChatGPT – A qualitative analysis of the current status quo. Research Square.

Pehkonen, E. (1995). Pupils view of mathematics. Initial report for an international comparison project. Helsinki: University of Helsinki.

Pehkonen, E., and Pietilä, A. (2004). "On relationships between beliefs and knowledge in mathematics education" in European research in mathematics education III: Proceedings of the third conference of the European Society for Research in mathematics education. ed. M. A. Mariotti (Pisa: University of Pisa and ERME).

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., et al. (2019). Language models as knowledge bases? *arXiv*. arXiv:1909.01066. doi: 10.48550/arXiv.1909.01066

Philipp, R. A. (2007). "Mathematics teachers' beliefs and affect" in Second handbook of research on mathematics teaching and learning. ed. F. K. Lester (Charlotte (NC): Information Age), 257–315.

Rabardel, P. (1995). Les hommes et les technologies; approche cognitive des instruments contemporains. Armand Colin Éditeurs. Available at: https://hal.science/hal-01017462

Rabardel, P., and Beguin, P. (2007). Instrument mediated activity: from subject development to anthropocentric design. *Theor. Issues Ergon. Sci.* 6, 429–461. doi: 10.1080/14639220500078179

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8). Available at: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

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. Available at: https://arxiv.org/pdf/2402.07927

Schoenfeld, A. H. (1992). "Learning to think mathematically: problem solving, metacognition, and sense making in mathematics" in Handbook of research on mathematics teaching and learning. ed. D. A. Grouws (New York: Macmillan), 334–370.

Schorcht, S., Baumanns, L., Buchholtz, N., Huget, J., Peters, F., and Pohl, M. (2023). Ask Smart to Get Smart: Mathematische Ausgaben generativer KI-Sprachmodelle verbessern durch gezieltes Prompt Engineering. *GDM-Mitteilungen* 115, 12–23. Available at: https://ojs.didaktik-der-mathematik.de/index.php/mgdm/article/view/1179/1356

Schorcht, S., Buchholtz, N., and Baumanns, L. (2024). Prompt the problem – investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. *Front. Educ.* 9:1386075. doi: 10.3389/feduc.2024.1386075

Sulmont, E., Patitsas, E., and Cooperstock, J. R. (2019). Can you teach me to machine learn? Proceedings of the 50th ACM technical symposium on computer science education, 948–954.

Tassoti, S. (2024). Assessment of students use of generative artificial intelligence: prompting strategies and prompt engineering in chemistry education. *J. Chem. Educ.* 101, 2475–2482. doi: 10.1021/acs.jchemed.4c00212

Vérillon, P., and Rabardel, P. (1995). Cognition and artefacts: a contribution to the study of thought in relation to instrumented activity. *Eur. J. Psychol. Educ.* 10, 77–101. doi: 10.1007/BF03172796

Vo, G. M., and Pancratz, N. (2023). "Vorstellungen von Lehramtsstudierenden zu künstlicher Intelligenz" in Informatikunterricht zwischen Aktualität und Zeitlosigkeit. eds. L. Hellmig and M. Hennecke (Würzburg, Rostock: GFI), 73–82.

Wardat, Y., Tashtoush, M. A., AlAli, R., and Jarrah, A. M. (2023). ChatGPT: a revolutionary tool for teaching and learning mathematics. *Eurasia J. Math. Sci. Technol. Educ.* 19:em2286. doi: 10.29333/ejmste/13272

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Proces. Syst.* 35, 24824–24837. doi: 10.48550/arXiv.2201.11903

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. Available at: https://arxiv.org/pdf/2302.11382

Yoon, H., Hwang, J., Lee, K., Roh, K. H., and Kwon, O. N. (2024). Students' use of generative artificial intelligence for proving mathematical statements. *ZDM Math. Educ.* 2024, 1–21. doi: 10.1007/s11858-024-01629-0

frontiers | Frontiers in Artificial Intelligence

TYPE Opinion
PUBLISHED 24 October 2024
DOI 10.3389/frai.2024.1471224

OPEN ACCESS

EDITED BY
Antonio Sarasa-Cabezuelo,
Complutense University of Madrid, Spain

REVIEWED BY
John Voiklis,
Knology, New York, United States

*CORRESPONDENCE
Myles Joshua Toledo Tan
✉ mylesjoshua.tan@medicine.ufl.edu

RECEIVED 26 July 2024
ACCEPTED 14 October 2024
PUBLISHED 24 October 2024

CITATION
Tan MJT and Maravilla NMAT (2024) Shaping integrity: why generative artificial intelligence does not have to undermine education. *Front. Artif. Intell.* 7:1471224.
doi: 10.3389/frai.2024.1471224

Shaping integrity: why generative artificial intelligence does not have to undermine education

Myles Joshua Toledo Tan[1,2,3,4,5,6,7]* and
Nicholle Mae Amor Tan Maravilla[7]

[1]Department of Electrical and Computer Engineering, Herbert Wertheim College of Engineering, University of Florida, Gainesville, FL, United States, [2]Department of Epidemiology, College of Public Health and Health Professions and College of Medicine, University of Florida, Gainesville, FL, United States, [3]Biology Program, College of Arts and Sciences, University of St. La Salle, Bacolod, Philippines, [4]Department of Natural Sciences, College of Arts and Sciences, University of St. La Salle, Bacolod, Philippines, [5]Department of Chemical Engineering, College of Engineering and Technology, University of St. La Salle, Bacolod, Philippines, [6]Department of Electronics Engineering, College of Engineering and Technology, University of St. La Salle, Bacolod, Philippines, [7]Yo-Vivo Corporation, Bacolod, Philippines

KEYWORDS

academic integrity, generative artificial intelligence, digital literacy, educational technology, AI-enhanced education, AI ethics, educational assessment, authentic learning

Introduction

The integration of generative artificial intelligence (GAI) in education has been met with both excitement and concern. According to a 2023 survey by the World Economic Forum, over 60% of educators in advanced economies are now using some form of artificial intelligence (AI) in their classrooms, a significant increase from just 20% 5 years ago (World Economic Forum, 2023). The rapid adoption of AI technologies in education highlights their potential to revolutionize the learning experience. AI tools, such as intelligent tutoring systems and adaptive learning platforms, offer personalized educational experiences that can meet the unique needs of each student. However, with this potential comes significant ethical concerns, particularly regarding academic integrity.

The International Center for Academic Integrity (2024) reported that 58% of students admitted to using AI tools to complete assignments dishonestly, highlighting the urgency of addressing these ethical concerns. This statistic underscores a critical issue: while AI has the potential to enhance education, its misuse can undermine the very foundations of academic integrity. The rise of AI technology has raised concerns about academic integrity. With tools that can generate text, solve problems, and even assist with research, students may find it easier to engage in plagiarism or other forms of cheating. This shift challenges traditional educational values, as it blurs the lines between original work and AI-generated content (Mohammadkarimi, 2023). Curriculum designers are thus faced with the challenge of integrating AI in ways that uphold ethical standards and promote genuine learning. This requires balancing the innovative potential of AI tools with a commitment to academic integrity, ensuring that technology enhances rather than undermines the educational experience.

To navigate this landscape responsibly, it is essential to revisit established ethical frameworks and educational theories. The ethical principles guiding our use of technology in education have remained consistent, even as the tools themselves have evolved. By referencing seminal works and foundational theories, we can demonstrate that the core values of honesty, fairness, and responsibility are timeless. For example, deontological ethics, as articulated by Kant (1785), emphasizes the importance of adhering to moral principles such as honesty and integrity, rather than the consequences of actions. In the context of AI in education, deontological ethics would require that the use of AI respects fundamental moral principles. For example, it would be crucial to ensure that AI systems are designed and implemented in ways that uphold students' rights to privacy, ensure fairness, and avoid deception. Adhering to these principles would be seen as morally obligatory, regardless of the potential benefits or drawbacks of AI in educational settings. Similarly, consequentialism, as articulated by John Stuart Mill, evaluates actions based on their outcomes. Mill's version of consequentialism, known as utilitarianism, argues that the best actions are those that promote happiness or better wellbeing. In the context of AI in education, applying Mill's consequentialist principles would involve assessing how the use of AI impacts educational outcomes. If AI can be used to enhance learning, provide personalized educational experiences, or address inequalities and inequities in education, then its use would be considered morally justified according to Mill's framework, as it promotes overall wellbeing and positive outcomes for students.

These ethical frameworks provide a robust foundation for the responsible use of GAI in modern educational settings. Moreover, educational theories such as constructivist learning and Self-Determination Theory (SDT) offer valuable insights into how AI can be used to enhance learning. Constructivist learning theory posits that students construct knowledge through active engagement with content, a process that can be greatly facilitated by AI tools. This approach emphasizes the importance of students' engagement in hands-on activities and interactions, which help them construct meaningful connections with new information (Hein, 1991). AI tools can significantly enhance this constructivist approach by providing personalized and interactive learning experiences. SDT, on the other hand, emphasizes the importance of autonomy, competence, and relatedness in fostering intrinsic motivation among students (Deci and Ryan, 2000). Integrating AI tools that align with the principles of SDT can help create a more engaging and supportive learning environment among students.

This discussion will explore how GAI can be integrated into education in ways that support rather than erode academic integrity. By examining the ethical frameworks of deontological ethics and consequentialism, and educational theories like constructivist learning and SDT, we will argue that AI, when used responsibly, can enhance digital literacy, foster intrinsic motivation, and support genuine knowledge construction. The principles discussed in older foundational papers remain relevant, proving that ethical guidelines established decades ago still hold value in today's technologically advanced classrooms (Floridi and Taddeo, 2016; Ryan and Deci, 2017).

The goal is to illustrate that the ethical use of GAI in education not only preserves but can also enhance academic integrity. Through responsible integration and ethical education, AI can empower students to become motivated, ethical, and engaged learners, well-prepared for the complexities of the modern world. By grounding our arguments in established ethical and educational theories, we can provide a comprehensive framework for understanding the potential benefits and challenges of AI in education.

Navigating the disruptive impact of generative artificial intelligence on assessment

The integration of GAI in education raises significant concerns about its potential to disrupt traditional assessment methods. The ability of GAI to generate essays, problem solutions, and even creative works has sparked fears of plagiarism and academic dishonesty, challenging conventional forms of evaluation such as take-home exams, essays, or homework assignments. These concerns are valid, as the ease with which students can use AI-generated content without truly engaging in the learning process threatens to undermine academic integrity (Popenici and Kerr, 2017).

However, the disruptive nature of GAI also presents an opportunity to reimagine assessment practices in ways that prioritize authentic learning and deeper understanding. The rise of AI necessitates a shift away from traditional assessments focused on rote memorization and information recall, toward more authentic assessment methods that require students to demonstrate higher-order thinking skills. For example, project-based tasks, real-world problem-solving activities, oral presentations, and open-ended assignments that demand personal reflection and original insights can reduce the likelihood of misuse and encourage students to engage meaningfully with course material (Borenstein and Howard, 2020).

Furthermore, GAI can play a constructive role in formative assessment by providing personalized feedback throughout the learning process. AI-driven tools can help students revise drafts, practice skills, and receive immediate guidance on areas needing improvement, fostering a deeper connection to the material. This approach transforms GAI from a potential threat to a valuable asset that supports continuous learning and skill development. Additionally, incorporating self-assessment and metacognitive practices, where students reflect on their progress and learning strategies, can ensure that AI augments rather than diminishes students' active participation in their education.

It is also essential to address the ethical considerations involved in using AI for assessment. Concerns such as data privacy, algorithmic bias, and the fairness of AI-generated evaluations must be taken seriously (Borenstein and Howard, 2020). Developing clear institutional policies that set boundaries on acceptable AI use in assessments can help maintain fairness and transparency. These policies should include guidelines for combining AI insights with human judgment to ensure that assessments reflect not only the outputs of AI but also the educator's understanding of the student's abilities and efforts.

By embracing these strategies, educators and institutions can harness the potential of GAI to enhance assessments while maintaining academic integrity. This balanced approach allows for the responsible integration of AI in education, ensuring that it supports meaningful learning experiences and prepares students to navigate an AI-driven world with integrity.

Constructivist learning theory: enhancing knowledge construction

Constructivist learning theory posits that learners construct knowledge through experiences and reflections, actively engaging with content to build understanding. GAI, with its advanced capabilities, aligns well with this theory, offering tools that promote exploration, interaction, and personalized learning paths. Contrary to the belief that AI erodes academic integrity, some scholars argue that AI, when used thoughtfully, has the potential to enhance educational experiences by providing personalized learning opportunities and supporting students' individual learning needs (Weller, 2020). While Weller does not claim that AI inherently fosters critical thinking or deeper understanding, his discussion highlights the potential of AI in educational settings, suggesting that it could complement traditional teaching methods to improve learning outcomes.

GAI tools, such as intelligent tutoring systems and adaptive learning platforms, provide students with tailored educational experiences. These systems analyze individual learning patterns and adapt content to meet specific needs, ensuring that students engage with material at an appropriate level of difficulty (Woolf, 2010). For instance, an AI-powered math tutor can identify a student's weaknesses in algebra and offer targeted exercises to address these gaps. This personalized approach not only supports knowledge construction but also encourages students to take ownership of their learning journey (Shute and Zapata-Rivera, 2012).

In a classroom setting, imagine a high school history class studying the Industrial Revolution. The educator integrates a GAI tool that generates interactive timelines and simulations based on historical data. Students can manipulate variables within these simulations to observe the effects on industrial growth, labor conditions, and economic development. Through this exploration, they construct a deeper understanding of the era's complexities. Instead of passively receiving information, students actively engage with content, reflecting on the consequences of different actions and decisions (Kumar et al., 2024).

Another example is in language arts, where a GAI tool assists students in creative writing. By analyzing a student's writing style and providing real-time feedback on grammar, tone, and narrative structure, the AI helps students refine their skills (Song and Song, 2023). Additionally, it can suggest plot developments or character traits, sparking students' creativity and encouraging them to think critically about their stories. This interactive process supports constructivist principles by allowing students to experiment, reflect, and build upon their ideas (Bereiter and Scardamalia, 1989).

Critics argue that AI tools may encourage academic dishonesty by making it easier for students to produce work with minimal effort. However, this perspective overlooks the potential for AI to promote genuine learning when used appropriately. Rather than replacing student effort, AI can enhance the learning process by offering personalized support, immediate feedback, and adaptive content, which fosters deeper engagement and learning outcomes (Nazaretsky et al., 2022). For instance, in a science class, AI-powered lab assistants can guide students through virtual experiments, providing explanations and prompting them to hypothesize, analyze data, and draw conclusions. Such interactions encourage active learning and promote a deeper understanding of scientific concepts and processes, rather than merely supplying answers (de Jong and van Joolingen, 1998). Additionally, as Al Darayseh (2023) notes, AI tools designed with input from educators help align the technology with pedagogical objectives, embedding ethical considerations to reduce the risk of academic dishonesty. Furthermore, it is important to acknowledge that AI is transforming science education and pedagogy, and the ethical implementation of these tools must reflect this shift to support genuine learning experiences while safeguarding academic integrity (Holstein et al., 2018; Erduran, 2023).

Moreover, GAI can facilitate collaborative learning, another key aspect of constructivist theory. In a project-based learning environment, students can use AI tools to collaboratively develop presentations or reports. AI can assist by organizing information, suggesting relevant sources, and providing feedback on the clarity and coherence of their work (Kreijns et al., 2003). This collaborative process encourages students to engage in dialogue, share perspectives, and build knowledge collectively.

To further illustrate, consider a classroom where students are tasked with developing a business plan. An AI tool can generate market analysis reports, financial projections, and strategic recommendations based on input from the students. As they interact with the AI and with each other, they learn to critically evaluate information, make informed decisions, and adapt their plans. This dynamic, interactive process is at the heart of constructivist learning, fostering not only knowledge construction but also critical thinking and problem-solving skills (Jonassen, 1995).

At present, there are multiple AI powered tools that are being used by most students that have significant potential to enhance a constructivist learning experience. One example is the ChatGPT. According to Rasul et al. (2023), ChatGPT supports the constructivist principle that learners construct their own understanding of knowledge by enabling students to explore and experiment with ideas, ask questions, and receive immediate feedback. This interactive engagement helps students to deeply connect with the content, refine their comprehension, and apply their learning in meaningful ways, ultimately enriching their educational experience.

Also, according to Mota-Valtierra et al. (2019), a constructivist approach is a great fit for teaching AI topics because it emphasizes building on prior knowledge and encouraging active learning. Their article outlines an innovative approach to teaching artificial intelligence (AI) through a constructivist methodology, specifically focusing on multilayer perceptrons (MLPs). After implementing it in different majors, the statistical analysis underscores the success of the proposed course methodology in enhancing student learning and providing a more consistent educational experience. The increase in average grades and the reduction in standard deviation highlight the effectiveness of the approach in improving both individual performance and overall learning outcomes.

In conclusion, GAI aligns with constructivist learning theory by providing tools that facilitate exploration, interaction, and personalized learning. Rather than promoting dishonesty, AI can enhance academic integrity by supporting genuine learning experiences. Through personalized feedback, interactive simulations, and collaborative projects, AI empowers students to take an active role in their education, constructing knowledge in meaningful and engaging ways. By embracing these technologies, educators can create enriching learning environments that prepare students for the complexities of the modern world (Papert and Harel, 1991).

The ethics of artificial intelligence: responsible use and digital literacy

The rise of GAI in education has sparked discussions on its ethical implications and the importance of fostering digital literacy. By examining ethical frameworks such as deontological ethics and consequentialism, we can argue that responsible use of GAI in the classroom can enhance students' digital literacy and prepare them to navigate the digital world ethically and effectively (Floridi and Taddeo, 2016; Stahl, 2012).

Deontological ethics, which focuses on adherence to moral rules or duties, provides a foundation for integrating AI responsibly in education. This framework emphasizes the importance of principles such as honesty, fairness, and respect for others (Kant, 1785). In the context of GAI, this means ensuring that AI tools are used to support and enhance learning rather than replacing students' efforts or promoting dishonesty.

For instance, in a high school history class studying the Industrial Revolution, an AI tool can generate interactive timelines and simulations based on historical data. Educators can emphasize the importance of using these tools ethically, encouraging students to engage with the material thoughtfully and critically. By adhering to principles of honesty and integrity, students learn to use AI as a supplementary resource that enhances their understanding rather than as a shortcut to completing assignments (Johnson and Verdicchio, 2019).

Consequentialism, as articulated by Mill (1861) in *Utilitarianism*, evaluates the morality of actions based on their outcomes. While Mill did not discuss AI, the principles of this framework can still be applied to contemporary debates about its use in education. By aiming to maximize positive outcomes—such as enhanced learning, critical thinking, and digital literacy—educators and curriculum designers can advocate for the responsible integration of AI. Emphasizing these benefits underscores how AI tools can contribute to better educational results and foster more informed digital citizens.

In a language arts classroom, for example, a GAI tool can assist students in creative writing by providing real-time feedback on grammar, tone, and narrative structure. Educators can guide students to use this feedback to improve their writing skills, fostering a deeper understanding of language and storytelling. The positive outcomes of enhanced writing abilities and critical engagement with AI tools illustrate the ethical benefits of responsible AI use (Borenstein and Howard, 2020).

To further promote digital literacy, it is crucial to educate students and educators on the ethical use of AI tools. This involves teaching them to understand how AI works, the potential biases and limitations of AI systems, and the importance of using AI responsibly (Brey, 2012). By fostering a culture of digital literacy, educators empower students to navigate the digital world with a critical and ethical mindset.

Consider a science class where an AI-powered lab assistant guides students through virtual experiments. Educators can use this opportunity to discuss the ethical considerations of AI in scientific research, such as data privacy, bias, and the importance of accurate data interpretation. By engaging in these discussions, students develop a nuanced understanding of the role of AI in science and the ethical responsibilities of using AI in research (Floridi, 2013).

Moreover, collaborative projects can further enhance digital literacy and ethical awareness. In a project-based learning environment, students can use AI tools to develop presentations or reports collaboratively. Educators can emphasize the importance of ethical collaboration, such as giving credit to sources, avoiding plagiarism, and ensuring that all team members contribute fairly. This approach not only enhances students' digital literacy but also instills ethical values that are essential in the digital age (Ess, 2015).

For instance, in a business class where students are tasked with developing a business plan, an AI tool can generate market analysis reports and financial projections. Educators can guide students to critically evaluate the AI-generated data, discuss the ethical implications of using AI in business decision-making, and ensure transparency and accountability in their work. This process helps students understand the ethical dimensions of AI and develop skills to use AI responsibly in their future careers (Mittelstadt et al., 2016).

The ethical frameworks of deontological ethics and consequentialism provide valuable insights into the responsible use of GAI in education. By emphasizing the importance of principles such as honesty, fairness, and positive outcomes, educators can foster digital literacy and ethical awareness among students. Teaching students to understand and navigate the ethical implications of AI tools prepares them to contribute positively to the digital world, ensuring that they use AI to enhance learning and uphold ethical standards. Through responsible AI integration and ethical education, we can create a generation of digitally literate and ethically aware individuals ready to thrive in a technologically advanced society (Moor, 1985).

The integration of AI in education holds great promise for enhancing learning experiences but raises profound ethical questions. The need for careful ethical reflection is underscored in *The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates*, which argues that educators, researchers, and stakeholders must engage in ongoing dialogue to navigate the complexities of AI in educational contexts (Holmes and Porayska-Pomsta, 2022). Smuha (2022) points out that for AI in education to be ethically responsible, it must adhere to key principles such as fairness, accountability, and transparency. These principles are vital in mitigating biases and preventing AI from perpetuating or amplifying existing educational inequalities. Furthermore, the concept of Trustworthy AI, as discussed by Smuha, is crucial in ensuring that AI systems foster inclusivity and do not marginalize vulnerable student populations (Smuha, 2022). Similarly, Brossi et al. (2022) raise concerns about the uncertain impact of AI on learners' cognitive development and the risk of disempowering educators through over-automation of pedagogical processes, pointing to the need for ethical frameworks that avoid automating ineffective or inequitable practices.

Williamson (2024) expands on this by highlighting the socio-political context of AI in education, warning against the assumption that technological innovations are inherently beneficial. Instead, he emphasizes that AI must be viewed as a socially embedded tool that could exacerbate educational inequities if not critically examined. The potential for AI to impact power dynamics, access, and social equity necessitates that educators and policymakers rigorously reflect on its broader implications, including how AI systems might reinforce or challenge existing educational structures.

Mouta et al. (2023) offer a practical step forward in addressing these concerns through their participatory futures approach, which is designed to help educators ethically integrate AI into their teaching environments. By using the Delphi method to gather diverse perspectives, their study presents hypothetical future scenarios that help educators and stakeholders reflect on the broader implications of AI in education. This approach ensures that the benefits of AI are balanced with ethical considerations related to privacy, bias, and the societal impacts of AI on education, promoting a thoughtful and inclusive implementation of AI technologies.

Further supporting this ethical stance, the European Commission's (2019) *Ethics Guidelines for Trustworthy AI* lays out seven key requirements for Trustworthy AI, including human agency, privacy, transparency, and fairness. These guidelines align closely with the need to ensure that AI systems in education promote fairness and inclusivity, rather than exacerbating inequities in educational access and outcomes. The guidelines also emphasize the importance of continuous monitoring and accountability to ensure AI systems remain aligned with these ethical principles. By stressing the importance of transparency, diversity, and non-discrimination, these guidelines reinforce the participatory frameworks put forth by Mouta et al. (2023), which advocate for an inclusive, ethical approach to AI integration in education.

Further reinforcing these ethical considerations, Floridi et al. (2018) in their "AI4People" framework emphasize the importance of a principled approach to AI that integrates ethical foundations like beneficence, non-maleficence, autonomy, justice, and explicability. These principles align with the need for AI in education to promote wellbeing and inclusivity while avoiding harm, respecting user autonomy, ensuring fair access to AI benefits, and fostering transparency. The framework also highlights that the potential risks of AI can include the erosion of human agency and privacy, making it essential for educational AI systems to be designed in ways that support rather than undermine student autonomy and self-determination. By embedding these principles into the development and deployment of AI, educators and policymakers can more effectively navigate the ethical challenges posed by AI in educational contexts, ultimately fostering a "Good AI Society" that supports human flourishing.

Self-determination theory: fostering intrinsic motivation

SDT posits that individuals are most motivated when their needs for autonomy, competence, and relatedness are met. GAI, with its capability to provide personalized feedback and tailored learning resources, can significantly support SDT by fostering intrinsic motivation among students. By empowering students to take control of their learning, AI can enhance engagement and academic integrity (Deci and Ryan, 2000; Ryan and Deci, 2017).

Autonomy

GAI can enhance students' sense of autonomy by offering them more control over their learning process. In a high school history class studying the Industrial Revolution, an AI tool can create interactive timelines and simulations. Students can explore these tools at their own pace, choosing which aspects of the Industrial Revolution to delve into more deeply. This self-directed exploration encourages students to take ownership of their learning, fostering a sense of autonomy (Reeve, 2006).

For example, a student interested in labor conditions during the Industrial Revolution might use the AI tool to simulate different labor policies and observe their impacts. This personalized exploration helps students develop a deeper understanding of historical complexities, driven by their own curiosity and interests (Niemiec and Ryan, 2009).

Competence

GAI tools can also support the need for competence by providing personalized feedback that helps students improve their skills and knowledge. In a language arts classroom, an AI-driven writing assistant can analyze a student's work and provide targeted feedback on grammar, tone, and narrative structure. This real-time, individualized feedback helps students understand their strengths and areas for improvement, fostering a sense of competence (Black and Deci, 2000).

Imagine a student writing a short story. The AI tool can suggest improvements in plot development and character interactions, guiding the student to refine their narrative. As students see their writing improve through this iterative process, they gain confidence in their abilities, which enhances their intrinsic motivation to engage with the subject matter (Vansteenkiste et al., 2004).

Relatedness

GAI can also facilitate relatedness by enabling collaborative learning and providing opportunities for meaningful interactions. In a project-based learning environment, AI tools can help students work together on presentations or reports. For instance, in a science class, an AI-powered lab assistant can guide groups of students through virtual experiments, encouraging collaboration and discussion (Ryan and Powelson, 1991).

Consider a group of students using AI to simulate a chemical reaction. The AI provides each group member with specific tasks and prompts them to share their findings and discuss results. This collaborative process fosters a sense of relatedness, as students work together to achieve common goals and learn from each other (Jang et al., 2010).

Promoting academic integrity

By fostering intrinsic motivation through autonomy, competence, and relatedness, GAI can also promote academic integrity. When students are genuinely interested and engaged in their learning, they are less likely to resort to dishonest practices. Personalized learning experiences make education more relevant and enjoyable, reducing the temptation to cheat (Deci et al., 1991).

In history class, for example, students using AI to explore the Industrial Revolution are likely to develop a genuine interest in the subject. This intrinsic motivation drives them to produce original work and engage deeply with the material. Similarly, in the language arts class, students motivated by the desire to improve their writing skills are more likely to take pride in their work and avoid plagiarism (Vansteenkiste and Ryan, 2013).

Real-world application

In a business class where students develop business plans using AI-generated market analysis reports and financial projections, educators can emphasize the importance of ethical decision-making and transparency. The AI tool provides personalized insights, allowing students to explore various business strategies and their consequences. This hands-on learning approach fosters intrinsic motivation by making the subject matter relevant and engaging (Ryan and Deci, 2000).

For instance, a student interested in starting a sustainable business can use AI to analyze the environmental impact of different business models. This personalized exploration helps the student develop a deeper understanding of sustainability in business, driven by their own interests and values (Deci and Ryan, 2008).

GAI, by supporting the principles of SDT, can foster intrinsic motivation among students. Through personalized feedback and tailored learning resources, AI empowers students to take control of their learning, enhancing their sense of autonomy, competence, and relatedness. This intrinsic motivation not only increases engagement but also promotes academic integrity. By integrating AI tools in educational settings, educators can create enriching learning environments that prepare students for the complexities of the modern world, ensuring that they are motivated, ethical, and engaged learners (Ryan and Deci, 2019).

Discussion: generative artificial intelligence as a catalyst for enhancing academic integrity

The integration of GAI in education has sparked significant debate regarding its impact on academic integrity. Critics argue that AI tools facilitate dishonesty by providing easy shortcuts for students to complete assignments. However, a closer examination of established educational theories and ethical frameworks reveals a different perspective. When used responsibly, GAI can foster intrinsic motivation, enhance digital literacy, and support constructivist learning principles, thereby promoting academic integrity rather than eroding it.

The integration of GAI in various educational fields, including computer science, engineering, medical education, and communication, is revolutionizing teaching and learning. The integration of AI technologies in computer science education, particularly through tools like GitHub Copilot, offers significant benefits in fostering creativity, enhancing learning efficiency, and supporting advanced projects. In engineering education, GAI offers numerous benefits, leveraging advanced chatbots and text-generation models to enhance learning and problem-solving capabilities. Cloud-based frameworks and social robots significantly enhance engineering education by providing scalable resources, interactive learning environments, and personalized support. Moreover, GAI has the potential to revolutionize medical education by enhancing clinical training, improving diagnostic accuracy, supporting personalized medicine, and advancing public health education. Also, GAI models hold great potential to enhance communication education across journalism, media, and healthcare fields. By supporting content generation, data analysis, creative development, and patient communication, GAI tools can provide valuable learning experiences and improve productivity (Bahroun et al., 2023).

GAI holds immense potential to transform education by enhancing teaching, learning, and educational processes. However, to fully realize these benefits, it is essential to address issues of responsible and ethical usage, potential biases, and academic integrity. By developing comprehensive guidelines, promoting transparency, mitigating bias, and fostering critical thinking skills, educators and institutions can ensure that AI technologies contribute positively to a technologically advanced, inclusive, and effective educational landscape (Bahroun et al., 2023).

Fostering intrinsic motivation through self-determination theory

SDT posits that students are most motivated when their needs for autonomy, competence, and relatedness are met. GAI can significantly enhance these aspects, fostering intrinsic motivation among students. When students are intrinsically motivated, they are more likely to engage deeply with the material and maintain academic integrity.

AI tools enhance autonomy by allowing students to control their learning process. In a history class, for instance, students can use AI-generated interactive timelines and simulations to explore different aspects of the Industrial Revolution at their own pace. This self-directed exploration encourages students to take ownership of their learning journey, which promotes a genuine interest in the subject matter. Such autonomy reduces the likelihood of dishonest behavior, as students are motivated by curiosity and a desire to learn.

Moreover, AI tools support competence by providing personalized feedback that helps students improve their skills. In a language arts classroom, an AI-driven writing assistant can analyze a student's work and offer specific suggestions for improvement. This real-time feedback not only enhances the student's writing skills but also builds their confidence. When students see tangible improvements in their abilities, their intrinsic motivation to engage with the subject matter increases. This motivation fosters academic integrity, as students take pride in their work and are less inclined to plagiarize or cheat.

GAI also facilitates relatedness by enabling collaborative learning. In project-based learning environments, AI tools can help students work together more effectively. For example, in a science class, an AI-powered lab assistant can guide groups through virtual experiments, encouraging discussion and collaboration. This collaborative process fosters a sense of community and shared purpose among students, which supports their intrinsic motivation to learn and succeed together. When students feel connected to their peers and their learning objectives, they are more likely to adhere to ethical standards and maintain academic integrity.

Enhancing digital literacy and ethical awareness

Digital literacy is essential in today's technology-driven world, and GAI can play a crucial role in fostering this skill. Ethical frameworks such as deontological ethics and consequentialism provide valuable insights into the responsible use of AI in education, emphasizing the importance of honesty, fairness, and positive outcomes.

Deontological ethics, which focuses on adherence to moral principles, underscores the need for using AI tools responsibly. Educators can teach students to use AI ethically by emphasizing principles such as honesty and integrity. For instance, when using AI-generated simulations in a history class, educators can guide students to engage thoughtfully with the material, ensuring that their use of AI supports genuine learning rather than shortcuts. By instilling these ethical values, educators help students understand the importance of maintaining academic integrity.

Consequentialism, which evaluates the morality of actions based on their outcomes, further supports the responsible use of AI in education. The ethical use of AI should aim to produce positive educational outcomes, such as enhanced learning, critical thinking, and digital literacy. In a language arts classroom, an AI writing assistant can provide constructive feedback that helps students refine their writing skills. This positive outcome not only improves their competence but also instills a sense of responsibility in using AI tools ethically. When students see the benefits of using AI to enhance their skills, they are more likely to use these tools responsibly, maintaining academic integrity.

Moreover, educating students on the ethical use of AI tools is crucial for fostering digital literacy. In a science class, an AI-powered lab assistant can guide students through virtual experiments, prompting discussions on ethical considerations such as data privacy and accuracy. By engaging in these discussions, students develop a nuanced understanding of the role of AI in scientific research and the ethical responsibilities that come with it. This awareness empowers students to navigate the digital world ethically and effectively, reducing the likelihood of dishonest behavior.

Supporting constructivist learning principles

Constructivist learning theory emphasizes that students construct knowledge through experiences and reflections. GAI aligns well with this theory, offering tools that promote exploration, interaction, and personalized learning paths. By supporting constructivist principles, AI enhances academic integrity by encouraging deeper understanding and critical thinking.

In a history class studying the Industrial Revolution, an AI tool that generates interactive timelines and simulations allows students to manipulate variables and observe outcomes. This hands-on exploration helps students construct a deeper understanding of historical complexities. Rather than passively receiving information, students actively engage with the content, reflecting on the consequences of different actions. This active engagement fosters a genuine interest in learning, reducing the temptation to cheat.

Similarly, in a language arts classroom, a GAI tool that provides real-time feedback on writing helps students improve their narrative skills. By experimenting with different plot developments and character traits, students engage in a creative process that aligns with constructivist principles. This interactive learning experience encourages students to think critically about their stories, fostering a deeper understanding of language and storytelling. When students are genuinely invested in their learning process, they are less likely to engage in dishonest practices.

Collaborative learning, another key aspect of constructivist theory, is also enhanced by GAI. In project-based learning environments, AI tools can facilitate collaboration by organizing information, suggesting relevant sources, and providing feedback on the clarity of students' work. For example, in a business class, an AI tool can help students develop a business plan by generating market analysis reports and financial projections. This collaborative process encourages students to engage in dialogue, share perspectives, and build knowledge collectively. When students work together to achieve common goals, they are more likely to adhere to ethical standards and maintain academic integrity.

Conclusion

GAI, when integrated responsibly in education, does not erode academic integrity. Instead, it fosters intrinsic motivation, enhances digital literacy, and supports constructivist learning principles. By promoting autonomy, competence, and relatedness, AI tools help students develop a genuine interest in their subjects, reducing the likelihood of dishonest behavior. Ethical education and personalized feedback further empower students to navigate the digital world responsibly, ensuring that they use AI tools to enhance their learning rather than as shortcuts. Through interactive and collaborative learning experiences, GAI encourages deeper understanding and critical thinking, ultimately promoting academic integrity in today's educational landscape.

Some practical guidance for educators and administrators

To provide practical guidance for using AI in education, we recommend focusing on integrating AI in ways that support established educational goals while adhering to ethical guidelines. Transparency is crucial in this process, as educators must actively involve students in understanding how AI tools are being used, what their limitations are, and why ethical use is important. This includes making the need to understand and actively utilize AI an explicit part of program objectives, course objectives, and learning outcomes, ensuring that its integration aligns with educational goals like developing digital literacy and critical thinking skills. By discussing potential biases, data privacy concerns, and limitations of AI-generated content, educators foster a culture of critical engagement where students learn to use AI responsibly and ethically rather than blindly relying on it. This proactive approach equips students with the discernment and integrity needed to navigate an AI-driven world.

Professional development for educators is crucial for the effective integration of AI in education. Governments and administrative bodies must exert the necessarily sustained and concerted pressures to make this a priority. Alongside sustained and concerted pressures, they need to sufficiently invest in resources and provide support and encouragement to ensure that this training is effective and widespread. Training programs should equip educators with practical skills for using AI tools, while also covering ethical considerations like data privacy, algorithmic bias, and the limitations of AI-generated feedback. By mandating and funding professional development, policymakers and administrators can ensure that educators are well-prepared to navigate the potential risks and benefits of AI. This comprehensive support empowers educators to guide students in using AI tools responsibly, fostering genuine learning and upholding academic integrity, rather than allowing misuse or over-reliance on technology to take root.

Finally, an iterative approach to integrating AI is crucial, and this must be encouraged at the policymaking level as well. Educators should continuously assess the impacts of AI on learning outcomes and be prepared to adjust their strategies accordingly. This involves collecting feedback from students, reviewing the effectiveness of AI tools, and making necessary changes to ensure AI contributes to meaningful educational experiences. Policymakers can support this process by implementing guidelines and providing resources that promote regular evaluation and adaptation of AI integration practices in schools. By emphasizing these practical steps at both the classroom and policy levels, educators can incorporate AI in ways that not only enhance learning but also foster responsible, ethical engagement with technology.

Author contributions

MT: Conceptualization, Project administration, Supervision, Writing – original draft, Writing – review & editing. NM: Writing – original draft, Writing – review & 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

MT and NM were employed by Yo-Vivo Corporation.

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

Al Darayseh, A. (2023). Acceptance of artificial intelligence in teaching science: science teachers' perspective. *Comput. Educ. Artif. Intell.* 4:100132. doi: 10.1016/j.caeai.2023.100132

Bahroun, Z., Anane, C., Ahmed, V., and Zacca, A. (2023). Transforming education: a comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. *Sustainability* 15:12983. doi: 10.3390/su151712983

Bereiter, C., and Scardamalia, M. (1989). "Intentional learning as a goal of instruction," in *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser*, ed. L. Resnick (New York, NY: Routledge), 361–392. doi: 10.4324/9781315044408-12

Black, A. E., and Deci, E. L. (2000). The effects of instructors' autonomy support and students' autonomous motivation on learning organic chemistry: a self-determination theory perspective. *Sci. Educ.* 84, 740–756. doi: 10.1002/1098-237X(200011)84:6<740::AID-SCE4>3.0.CO;2-3

Borenstein, J., and Howard, A. (2020). Emerging challenges in AI and the need for AI ethics education. *AI Ethics* 1, 61–65. doi: 10.1007/s43681-020-00002-7

Brey, P. (2012). Anticipatory ethics for emerging technologies. *Nanoethics* 6, 1–13. doi: 10.1007/s11569-012-0141-7

Brossi, L., Castillo, A. M., and Cortesi, S. (2022). "Student-centred requirements for the ethics of AI in education," in *The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates*, eds. W. Holmes, and K. Porayska-Pomsta (New York, NY: Routledge), 91–113. doi: 10.4324/9780429329067-6

de Jong, T., and van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Rev. Educ. Res.* 68, 179–201. doi: 10.3102/00346543068002179

Deci, E. L., and Ryan, R. M. (2000). The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. *Psychol. Inq.* 11, 227–268. doi: 10.1207/S15327965PLI1104_01

Deci, E. L., and Ryan, R. M. (2008). Self-determination theory: macrotheory of human motivation, development, and health. *Can. Psychol.* 49:182. doi: 10.1037/a0012801

Deci, E. L., Vallerand, R. J., Pelletier, L. G., and Ryan, R. M. (1991). Motivation and education: the self-determination perspective. *Educ. Psychol.* 26, 325–346. doi: 10.1080/00461520.1991.9653137

Engaging students in learning activities: it is not autonomy support or structure but autonomy support and structure. *J. Educ. Psychol*. 102, 588–600. doi: 10.1037/a0019682

Erduran, S. (2023). AI is transforming how science is done. Science education must reflect this change. *Science* 382:eadm9788. doi: 10.1126/science.adm9788

Ess, C. (2015). *Digital Media Ethics*. Hoboken, NJ: John Wiley and Sons.

European Commission (2019). *Ethics Guidelines for Trustworthy AI*. Available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419 (accessed September 26, 2024).

Floridi, L. (2013). *The Ethics of Information*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199641321.001.0001

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach.* 28, 689–707. doi: 10.1007/s11023-018-9482-5

Floridi, L., and Taddeo, M. (2016). What is data ethics? *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 374:20160360. doi: 10.1098/rsta.2016.0360

Hein, G. E. (1991). "Constructivist learning theory," in *Paper presented at the CECA (International Committee of Museum Educators) Conference* (Jerusalem), 15-22 October 1991.

Holmes, W., and Porayska-Pomsta, K. (2022). The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates. New York, NY: Routledge. doi: 10.4324/9780429329067

Holstein, K., McLaren, B. M., and Aleven, V. (2018). "Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms," in *Lecture Notes in Computer Science* (Cham: Springer), 154–168. doi: 10.1007/978-3-319-93843-1_12

International Center for Academic Integrity (2024). *Facts and Statistics*. Dayton, OH: International Center for Academic Integrity.

Johnson, D. G., and Verdicchio, M. (2019). AI, agency and responsibility: the VW fraud case and beyond. *AI Soc*. 34, 639–647. doi: 10.1007/s00146-017-0781-9

Jonassen, D. H. (1995). Computers as cognitive tools: learning with technology, not from technology. *J. Comput. High. Educ*. 6, 40–73. doi: 10.1007/BF02941038

Kant, I. (1785). *Groundwork of the Metaphysics of Morals* (Transl. by M. Gregor). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511809590

Kreijns, K., Kirschner, P. A., and Jochems, W. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. *Comput. Human Behav.* 19, 335–353. doi: 10.1016/S0747-5632(02)00057-2

Kumar, T., Kait, R., Ankita, and Malik, A. (2024). "The role of generative artificial intelligence (GAI) in Education: a detailed review for enhanced learning experiences," in *Intelligent IT Solutions for Sustainability in Industry 5.0 Paradigm. ICEIL 2023. Lecture Notes in Electrical Engineering, Vol. 1185*, eds. B. Shukla, B. K. Murthy, N. Hasteer, H. Kaur, and J. P. Van Belle (Singapore: Springer), 195–207. doi: 10.1007/978-981-97-1682-1_17

Mill, J. S. (1861). *Utilitarianism*. Kingston, NY: Parker, Son, and Bourn.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: mapping the debate. *Big Data Soc.* 3:2053951716679679. doi: 10.1177/2053951716679679

Mohammadkarimi, E. (2023). Teachers' reflections on academic dishonesty in EFL students' writings in the era of artificial intelligence. *J. Appl. Learn. Teach*. 6. doi: 10.37074/jalt.2023.6.2.10

Moor, J. H. (1985). What is computer ethics? *Metaphilosophy* 16, 266–275. doi: 10.1111/j.1467-9973.1985.tb00173.x

Mota-Valtierra, G., Rodríguez-Reséndiz, J., and Herrera-Ruiz, G. (2019). Constructivism-based methodology for teaching artificial intelligence topics focused on sustainable development. *Sustainability* 11:4642. doi: 10.3390/su11174642

Mouta, A., Torrecilla-Sánchez, E. M., and Pinto-Llorente, A. M. (2023). Design of a future scenarios toolkit for an ethical implementation of artificial intelligence in education. *Educ. Inf. Technol*. 29:10473–10498. doi: 10.1007/s10639-023-12229-y

Nazaretsky, T., Bar, C., Walter, M., and Alexandron, G. (2022). "Empowering teachers with AI: Co-designing a learning analytics tool for personalized instruction in the science classroom," in *LAK22: 12th International Learning Analytics and Knowledge Conference* [Preprint] (New York, NY). doi: 10.1145/3506860.3506861

Niemiec, C. P., and Ryan, R. M. (2009). Autonomy, competence, and relatedness in the classroom: applying self-determination theory to educational practice. *Theory Res. Educ.* 7, 133–144. doi: 10.1177/1477878509104318

Papert, S., and Harel, I. (1991). Situating constructionism. *Constructionism* 36, 1–11.

Popenici, S. A. D., and Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Res. Pract. Technol. Enhanc. Learn.* 12:22. doi: 10.1186/s41039-017-0062-8

Rasul, T., Nair, S., Kalendra, D., Robin, M., Santini, F., Ladeira, W. J., et al. (2023). The role of ChatGPT in higher education: benefits, challenges, and future research directions. *J. Appl. Learn. Teach.* 6, 41–50. doi: 10.37074/jalt.2023.6.1.29

Reeve, J. (2006). Teachers as facilitators: what autonomy-supportive teachers do and why their students benefit. *Elem. Sch. J.* 106, 225–236. doi: 10.1086/501484

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

Ryan, R. M., and Deci, E. L. (2017). Self-Determination Theory: Basic Psychological Needs in Motivation, *Development, and Wellness*. New York, NY: Guilford Publications. doi: 10.1521/978.14625/28806

Ryan, R. M., and Deci, E. L. (2019). Intrinsic and *Extrinsic Motivations: Classic Definitions and New Directions*. New York, NY: Routledge.

Ryan, R. M., and Powelson, C. L. (1991). Autonomy and relatedness as fundamental to motivation and education. *J. Exp. Educ.* 60, 49–66. doi: 10.1080/00220973.1991.10806579

Shute, V. J., and Zapata-Rivera, D. (2012). Adaptive educational systems. *Adapt. Technol. Train. Educ.* 7, 1–35. doi: 10.1017/CBO9781139049580.004

Smuha, N. A. (2022). "Pitfalls and pathways for trustworthy artificial intelligence in education," in *The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates*, eds. W. Holmes, and K. Porayska-Pomsta (New York, NY: Routledge), 113–146. doi: 10.4324/9780429329067-7

Song, C., and Song, Y. (2023). Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Front. Psychol*. 14:1260843. doi: 10.3389/fpsyg.2023.1260843

Stahl, B. C. (2012). Morality, ethics, and reflection: a categorization of normative IS research. *J. Assoc. Inf. Syst.* 13, 636–656. doi: 10.17705/1jais.00304

Vansteenkiste, M., and Ryan, R. M. (2013). On psychological growth and vulnerability: Basic psychological need satisfaction and need frustration as a unifying principle. *J. Psychother. Integr*. 23:263. doi: 10.1037/a0032359

Vansteenkiste, M., Simons, J., Lens, W., Soenens, B., and Matos, L. (2004). Examining the motivational impact of intrinsic versus extrinsic goal framing and autonomy-supportive versus internally controlling communication style on early adolescents' academic achievement. *Child Dev.* 75, 483–501. doi: 10.1111/j.1467-8624.2005.00858.x

Weller, M. (2020). *25 Years of Ed Tech*. Athabasca, AB: Athabasca University Press. doi: 10.15215/aupress/9781771993050.01

Williamson, B. (2024). The social life of AI in education. *Int. J. Artif. Intell. Educ.* 34, 97–104. doi: 10.1007/s40593-023-00342-5

Woolf, B. P. (2010). Building Intelligent Interactive Tutors*: Student-Centered Strategies for Revolutionizing e-Learning*. Cambridge, MA: Morgan Kaufmann.

World Economic Forum (2023). *The Future of Jobs Report 2023*. Cologny: World Economic Forum.

frontiers | Frontiers in Artificial Intelligence

TYPE Opinion
PUBLISHED 01 November 2024
DOI 10.3389/frai.2024.1452131

OPEN ACCESS

EDITED BY
Diego Zapata-Rivera,
Educational Testing Service, United States

REVIEWED BY
Ioana Ghergulescu,
Adaptemy, Ireland
Daina Gudoniene,
Kaunas University of Technology, Lithuania

*CORRESPONDENCE
Chien Ching Lee
✉ chienching.lee@singaporetech.edu.sg

RECEIVED 20 June 2024
ACCEPTED 09 October 2024
PUBLISHED 01 November 2024

CITATION
Lee CC and Low MYH (2024) Using genAI in education: the case for critical thinking. *Front. Artif. Intell.* 7:1452131.
doi: 10.3389/frai.2024.1452131

Using genAI in education: the case for critical thinking

Chien Ching Lee[1]* and Malcolm Yoke Hean Low[2]

[1]Centre for Professional Communication, Singapore Institute of Technology, Singapore, Singapore, [2]Information and Communication Technology Program, Singapore Institute of Technology, Singapore, Singapore

KEYWORDS

genAI, education, critical thinking, empowerment, ownership

1 Introduction

With AI technologies driven by big data, many claim that we are now experiencing a global Fourth Industrial Revolution (MacGregor, 2024). Countries like Singapore (Chia, 2023) and America (MacGregor, 2024) are investing and rushing to develop AI talent and solutions. If one of the main roles of universities is to equip students for jobs of the future, there is no turning back from training students in the proficient and ethical use of generative AI (genAI). As mentioned by Heaven (2023), genAI tools are changing education, not destroying it.

United Nations Educational Scientific Cultural Organization (2023) has formulated guidance for policymakers on AI and education which has helped institutions of higher learning navigate the use of genAI. Research has found that university staff are aware that the use of genAI could result in better productivity and hence a majority are optimistic about its use (McCormack, 2023). Students are also aware of the benefits and challenges faced in the use of genAI. They are drawn to the personalized feedback that genAI could offer 24/7, features which are useful with increasing class sizes (Dixon, 2023). Kakuchi's (2023) study further found that 78% and 70% of students mentioned that genAI helped improve their writing and thinking, respectively. However, they worry about unintentionally running afoul of university guidelines on its use as it becomes integrated in most software tools (Hodges and Ocak, 2023) and how an overreliance on genAI could influence the value of their university education (Chan and Hu, 2023).

This opinion piece argues that as AI eventually becomes ubiquitous in education, educators should aim to use the technology to challenge students to think critically and enhance their human interactions. This argument is supported by the findings from the World Economic Forum's (2023) Future of Work report where creative thinking, analytical thinking and self-efficacy skills were rated as the top three skills.

Two use cases from the Singapore Institute of Technology (SIT), an applied learning university, are presented. The first case relates to communication (soft) skills classes taught by the first author while the second case relates to information technology (hard skills) taught by the second author. The two cases reflect SIT's position that genAI tools could enhance teaching and learning, and perhaps industry practices. Thus, students are encouraged to use genAI as tools rather than as primary sources of information (Rakshika and Lee, 2024). This aligns with the Singapore government's initiative for students to be taught the ethical and responsible use of AI at different levels (Ang, 2024) and its Draft GenAI Governance Framework which aims to promote a systemic and balanced approach to facilitate innovation and address accountability issues in the use of genAI tools (Norton Rose Fulbright, 2024).

2 Small classes teaching soft skills

Soft skills like communication skills are often taught seminar style in the classroom. In SIT, undergraduate students are taught to think critically and connectedly. Critical thinking is taught using the Paul-Elder framework (Paul and Elder, 2019) in a core common module "Critical Thinking and Communicating" in year 1 and is reinforced in the Communicating Across the Curriculum (CAC) embedded workshops throughout their studies. The Paul-Elder framework aims to inculcate intellectual elements that help students question issues from multiple perspectives, while the intellectual standards focus on checking the quality of students' reasoning. Recently, students have also been taught systems thinking (Thinking Tools Studio, 2024) in some workshops to enhance their understanding on how issues are interrelated.

Among the CAC workshops taught are Writing the Literature Review, Logbook Writing and Revising Penetration Testing Reports. The workshops catered to 119, 30, and 119 students respectively from the Information Technology program in the second trimester of 2023. The materials were developed based on the assignment brief provided by the respective module leads. The workshops were 3, 6, and 3 hours long respectively and taught via zoom.

The innovations in these workshops are the intentional teaching on how to use and evaluate outputs from ChatGPT 3.5 (a free genAI tool) using the Paul-Elder framework and systems thinking habits. The instructor demonstrated how to use ChatGPT on a parallel task in the workshop activities by teaching them how to use ChatGPT for various tasks related to the workshops, and providing them with feedback on the revisions needed in ChatGPT's outputs using the Paul-Elder framework. Then, the students performed a similar task using ChatGPT in groups of four to five in zoom breakout rooms. The students uploaded their drafts to the discussion forum in the LMS and the instructor led feedback-cum-questioning sessions with their peers on a few of the students' drafts. For the logbook writing workshop, the students were also taught the 14 systems thinking habits aided by prompts from the flip cards (Thinking Tools Studio, 2024). The aim was to help students understand the implications of their decisions in the larger and dynamic contexts of a company and/or society.

The feedback from the students was positive. The online surveys conducted at the end of the workshops showed that students were more aware about how to be strategic and critical when using ChatGPT in their learning. They liked the efficiency afforded by ChatGPT even though they needed to edit ChatGPT's outputs. They also put in more effort to include key points in their writing which pre-empted questions from the instructor and their peers in class. The students also mentioned that they have gained more confidence in navigating the use of genAI tools for future assignments. Interestingly, even though the systems thinking habits were taught briefly in the logbook writing workshop, it resonated with the students, with 10 out of the 14 habits cited in the survey. These outcomes reflected a positive and interactive learning environment, aided by discussions and questioning on ChatGPT's outputs.

3 Large classes teaching hard skills

Teaching data structures and algorithms (DSA) presents significant challenges in computer science education due to the abstract and complex nature of these foundational topics. Students often struggle to grasp the theoretical underpinnings and practical applications of DSA, which are crucial for effective problem-solving in various technological fields. Instructors face the difficult task of conveying these intricate concepts in a manner that is both engaging and comprehensible, particularly given the diverse learning paces and styles of students.

In response to these challenges, we integrated AI chatbots into the Data Structures and Algorithms (DSA) course, which is a core component of several undergraduate degree programs. The course was delivered to a cohort of 178 students in the second trimester of 2023. It covered fundamental topics such as algorithm analysis, different data structures like stacks, queues, and trees, and algorithms for sorting, searching, and optimizing paths in graphs.

A significant innovation in the course was the inclusion of a team project that required students to use AI chatbots as a learning tool. This project was introduced in the latter half of the course and aimed to enhance students' understanding of DSA through interaction with chatbots. Students were grouped into teams of four to five, and each team selected a specific data structure or algorithm to focus on. They interacted with AI chatbots to ask questions and assess the responses based on accuracy, completeness, clarity, and relevance. The assignment also required students to propose a real-world application for their chosen topic and document their findings and interactions in a detailed report.

The feedback from the students on the use of AI chatbots in the team project was largely positive. Surveys conducted at the end of the trimester showed high ratings for the accuracy, clarity, and relevance of the chatbot responses. Students appreciated the interactive nature of the chatbots, noting that this approach made the learning process more engaging and contributed to a better understanding of complex DSA topics. The introduction of AI chatbots not only provided a more interactive and personalized learning experience but also sparked a deeper reflection among students on the human aspects of learning, such as creativity, highlighting a clear distinction in educational engagement compared to traditional methods without such technology. These outcomes suggest that AI chatbots, with further refinement, could become invaluable educational tools in computer science, helping to mitigate some of the traditional challenges associated with teaching DSA effectively.

4 Discussion

The current generation Z are digital natives. Thus, they take to genAI like fish to the water. Coffey (2023) further found that genAI is used daily by 50% of college students in helping them with their school assignments, with the number of users set to increase year on year. In addition, 30% of the students believe that they need to be well-versed in its use to gain employment. We have shared two use cases which demonstrated how genAI was used to promote students' curiosity and accountability and which empowered their humanity. We would like to suggest that instructors need to model the way in which students could use genAI meaningfully and ethically. Another suggestion could be to utilize formative assessments with clear guidelines for the use of genAI which focuses on the learning process rather than the product *per se*. Chan's (2023) 3R framework encourages students to Report their use of genAI tools, Revise its output and Reflect on the process. This framework recognizes AI-assisted writing as basic writing which leads to specific writing purposes. Having students declare the originality of their submissions could be another ethical safeguard as it has mostly been effective in deterring academic dishonesty (Borup, 2023). Instructors have often wished for teaching assistants to help them manage increasing class sizes. With proper guidance, genAI, including recent advanced reasoning models such as OpenAI o1, could serve students in that role, offering students an interactive and personalized learning environment for their learning and possibly equip them for the future of work which utilizes genAI.

Author contributions

CCL: Conceptualization, Writing – original draft, Writing – review & editing. ML: Writing – original draft, Writing – review & 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

References

Ang, S. (2024). *Students Are Taught to Use AI Ethically and Responsibly at Different Levels: Chan Chun Sing*. Available at: https://www.straitstimes.com/singapore/politics/students-are-taught-to-use-ai-ethically-and-responsibly-at-different-levels-chan-chun-sing (accessed September 20, 2024).

Borup, J. (2023). *This Was Written By a Human: a Real Educator's Thoughts on Teaching in the Age of Chatgpt. EDUCAUSE Review*. Available at: https://er.educause.edu/articles/2023/3/this-was-written-by-a-human-a-real-educators-thoughts-on-teaching-in-the-age-of-chatgpt (accessed September 3, 2024).

Chan, C. B. (2023). "Grading generative AI-based assignments using a 3R framework," in *Proceedings for IEEE International Conference on Teaching, Assessment and Learning for Engineering, 28 Nov-—1 Dec*. Auckland.

Chan, C. K. Y., and Hu, W. J. (2023). Students' voices on generative AI: perceptions, benefits and challenges in higher education. *Int. J. Educ. Technol. High. Educ.* 20:8. doi: 10.1186/s41239-023-00411-8

Chia, O. (2023). *National AI Strategy 2.0 Follows Years of Planning, Growth in AI Sector "Not By Chance."* The Straits Times. Available at: https://www.straitstimes.com/singapore/national-ai-strategy-20-follows-years-of-planning-growth-in-ai-sector-not-by-chance-dpm-wong (accessed September 20, 2024).

Coffey, L. (2023). *Students outrunning faculty in AI Use.* Inside Higher Education. Available at: https://www.insidehighered.com/news/tech-innovation/artificial-intelligence/2023/10/31/most-students-outrunning-faculty-ai-use (accessed January 29, 2024).

Dixon, J. (2023). *Keeping Human Values at the Heart of AI in Higher Education*. University World News. Available at: https://www.universityworldnews.com/post.php?story=20230814140706902 (accessed January 29, 2024).

Heaven, W. D. (2023). *ChatGPT Is Going to Change Education, Not Destroy It*. MIT Technology Review. Available at: https://www.technologyreview.com/2023/04/06/1071059/chatgpt-change-not-destroy-education-openai/ (accessed January 29, 2024).

Hodges, C., and Ocak, C. (2023). *Integrating Generative AI Into Higher Education: Considerations*. EDUCASE Review. Available at: https://er.educause.edu/articles/2023/8/integrating-generative-ai-into-higher-education-considerations (accessed January 29, 2024).

Kakuchi, S. (2023). *Universities Take Lead in Forging AI Policies, Practices*. University World News. Available at: https://www.universityworldnews.com/post.php?story=20230822082741187 (accessed September 3, 2024).

MacGregor, K. (2024). *Generative AI Action Hints at Core Future Roles in Universities*. University World News. Available at: https://www.universityworldnews.com/post.php?story=2024012705101220 (accessed September 3, 2024).

McCormack, M. (2023). *EDUCAUSE QuickPoll Results: Adopting and Adapting to Generative AI in Higher Ed Tech*. EDUCAUSE Review. Available at: https://er.educause.edu/articles/2023/4/educause-quickpoll-results-adopting-and-adapting-to-generative-ai-in-higher-ed-tech (accessed January 29, 2024).

Norton Rose Fulbright, L. L. P. (2024). *Singapore Proposes Governance Framework for Generative AI*. Available at: https://www.dataprotectionreport.com/2024/02/singapore-proposes-governance-framework-for-generative-ai/ (accessed January 29, 2024).

Paul, R., and Elder, L. (2019). *The Miniature Guide to Critical Thinking Concepts and Tools, 8th Edn*. Lanham, MD: Rowman & Littlefield Publishers; Foundation for Critical Thinking Press.

Rakshika, V., and Lee, E. (2024). *Students at Singapore Universities Allowed to Use AI Tools for Assignments But Must Stick to Rules*. Available at: https://www.straitstimes.com/singapore/students-at-s-pore-universities-allowed-to-use-ai-tools-for-assignments-but-must-stick-to-rules (accessed January 29, 2024).

Thinking Tools Studio (2024). *Habits of a Systems Thinker*. Available at: https://thinkingtoolsstudio.waterscenterst.org/cards (accessed September 20, 2024).

United Nations Educational Scientific and Cultural Organization (2023). *ChatGPT and Artificial Intelligence in Higher Education: Quick Start Guide*. Available at: https://www.iesalc.unesco.org/wp-content/uploads/2023/04/ChatGPT-and-Artificial-Intelligence-in-higher-education-Quick-Start-guide_EN_FINAL.pdf (accessed January 29, 2024).

World Economic Forum (2023). *Future of Jobs Report*. Available at: https://www.weforum.org/reports/the-future-of-jobs-report-2023/ (accessed January 29, 2024).

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 04 November 2024
DOI 10.3389/frai.2024.1474019

OPEN ACCESS

EDITED BY
Diego Zapata-Rivera,
Educational Testing Service, United States

REVIEWED BY
Edith Aurora Graf,
Educational Testing Service, United States
Su-Youn Yoon,
Edulab-Inc., Japan

*CORRESPONDENCE
Andrew Runge
✉ arunge@duolingo.com

RECEIVED 31 July 2024
ACCEPTED 09 October 2024
PUBLISHED 04 November 2024

CITATION
Runge A, Attali Y, LaFlair GT, Park Y and Church J (2024) A generative AI-driven interactive listening assessment task. *Front. Artif. Intell.* 7:1474019. doi: 10.3389/frai.2024.1474019

A generative AI-driven interactive listening assessment task

Andrew Runge*, Yigal Attali, Geoffrey T. LaFlair, Yena Park and Jacqueline Church

Duolingo, Pittsburgh, PA, United States

Introduction: Assessments of interactional competence have traditionally been limited in large-scale language assessments. The listening portion suffers from construct underrepresentation, whereas the speaking portion suffers from limited task formats such as in-person interviews or role plays. Human-delivered tasks are challenging to administer at large scales, while automated assessments are typically very narrow in their assessment of the construct because they have carried over the limitations of traditional paper-based tasks to digital formats. However, computer-based assessments do allow for more interactive, automatically administered tasks, but come with increased complexity in task creation. Large language models present new opportunities for enhanced automated item generation (AIG) processes that can create complex content types and tasks at scale that support richer assessments.

Methods: This paper describes the use of such methods to generate content at scale for an interactive listening measure of interactional competence for the Duolingo English Test (DET), a large-scale, high-stakes test of English proficiency. The Interactive Listening task assesses test takers' ability to participate in a full conversation, resulting in a more authentic assessment of interactive listening ability than prior automated assessments by positing comprehension and interaction as purposes of listening.

Results and discussion: The results of a pilot of 713 tasks with hundreds of responses per task, along with the results of human review, demonstrate the feasibility of a human-in-the-loop, generative AI-driven approach for automatic creation of complex educational assessments at scale.

KEYWORDS

automatic item generation, listening assessment, interactional competence, generative AI, psychometrics, interactive listening, Duolingo English test

Introduction

Listening comprehension is a critical part of language proficiency (Wagner, 2014). Assessment of listening comprehension, however, has long underrepresented the interactional and communicative abilities of the listening test-takers (Aryadoust and Luo, 2023). Large-scale assessments of L2 academic English proficiency ask test takers to take a passive role in comprehending a speaker in a traditional lecture. To tap into the communicative aspect of listening ability, a listening assessment would at most include comprehension questions about a conversation that test takers passively listen to, or have test takers complete a single turn in a conversation (Buck, 2001; Papageorgiou et al., 2021). Aryadoust and Luo (2023) call for a shift in focus in listening assessment to technology-driven constructs in virtual settings such as interacting with others in real-time. To that end, we present a novel assessment of listening comprehension, the Interactive Listening task, that asks test takers to participate and sustain a virtual conversation. We apply recent advances in generative AI (Brown et al., 2020; OpenAI et al., 2024) to the task of automated item generation (AIG, Attali et al., 2022) to generate the conversational content and items used for this task.

The rest of this paper is organized as follows. We first review the current state of automatic item generation and assessments of listening and interactional competence that motivated our work. Next, we present an overview of our Interactive Listening task and describe decisions we made with regards to how we designed the task to assess communicative listening ability. We describe the generative AI-based item generation processes we developed to create a large bank of diverse conversations to use for the task, along with our methods for generating, evaluating and selecting distractors for multiple-choice items. We describe a series of small-scale pilot experiments and their key results that informed task design and administration decisions. Finally, we present the results from a large-scale pilot experiment using 713 Interactive Listening tasks administered as part of a practice test on the Duolingo English Test. We report on feedback from human reviewers for the piloted tasks that provides insights into the quality of the AIG processes, while test taker pilot response data allows us to evaluate the psychometric properties of the tasks.

Background

Automatic item generation

The adoption of technology by the field of assessment has moved past a shift in the mode of delivery: from paper-based to computer-based (or internet-based). The current state of technology in assessment can be better described as leveraging technology across the test development, administration, and scoring continuum to improve the ways in which latent traits are assessed. Now, internet-based computerized assessments are making use of advances in technology for a variety of purposes, including developing innovative item types and formats (Sireci and Zenisky, 2006), measuring more complex knowledge, skills, and competencies (Bartram and Hambleton, 2005), implementing automated scoring with immediate feedback to students (Attali and Powers, 2010), offering adaptive, on-demand testing (van der Linden and Glas, 2010), and offering personalized assessments (Suárez-Álvarez et al., 2023). These adoptions have led to an increase in the volume and offerings of assessments, necessitating the need for a significantly larger item bank to accommodate this increased demand (Downing and Haladyna, 2006; Sayin and Gierl, 2024).

Automatic item generation (AIG) may help address the challenge of developing items at a much larger scale than was needed by traditional paper-based assessments (Circi et al., 2023; Gierl and Haladyna, 2012; Irvine and Kyllonen, 2002). AIG in its nascent form has traditionally been implemented using an item model approach whereby a template for a question with parameters is automatically populated with specific values using a computer-based algorithm (Bejar, 2002). For example, the model $X + Y = ?$, where X and Y can be any whole numbers in the range 0–9, has two parameters X and Y. X and Y in the item model can be populated with any single-digit numbers to display the item. A more complex example is "How many pieces of [fruit] will you have if you cut [5] whole [fruits] into [thirds]?" (Attali, 2018), where the text in parentheses represent parameters (numeric or text). This type of traditional AIG has successfully been used to create items in diverse content areas, such as mathematics word problems and medical diagnosis questions (Haladyna, 2013), expanding the potential number of items with set item models facilitating a construct-driven approach to item development (Embretson and Yang, 2006; Whitely, 1983).

While being fairly useful in content areas where item models are easier to specify, the traditional item model approach to AIG comes with its limitations (von Davier, 2018). One is that it is not easily applicable to other content areas: for instance, second language (L2) reading proficiency, where a template would have to be constructed for each question for each passage (Bolender et al., 2023). Another is that the item model approach relies on highly skilled content experts to create the models and therefore can be costly (Kosh et al., 2018). Due to these shortcomings, the use of AIG as a technique to generate test content has been limited to relatively simple tasks (Attali et al., 2022).

Large language models and AIG

An alternative to the traditional item model approach to AIG is a generative AI-based approach leveraging recent advances in large language models (LLMs). Language modeling is capable of generating a large amount of text based on limited input, drawing from a probabilistic model of language. Language models based on neural transformer architectures (Devlin et al., 2019; Radford et al., 2019; Vaswani et al., 2017) were previously limited in their applications to AIG as they required a large amount of expert-annotated training data, computing power, and lengthy model development to update the underlying model to accomplish a particular task. These limitations are addressed by OpenAI's GPT-3 and its subsequent GPT-4 models (Brown et al., 2020; OpenAI et al., 2024) that can generate novel content based on fewer than 10 examples without the need to update the underlying model, referred to as "few-shot" prompting in the context of LLMs (Brown et al., 2020). Additionally, GPT models can be specified to generate the output in any desired format, such as well-formatted, fully functioning HTML code, or a paragraph with comprehension questions. These advantages allow AIG with GPT-based models to be freed from the notion of an item model and expand into item types that cannot be succinctly captured with templates, all the while without putting significant strains on resources.

Generative AI-based approaches have since been used successfully to create more complex assessments that were difficult to construct with the item model approach, from reading passages (e.g., Attali et al., 2022; Bezirhan and von Davier, 2023) and listening stimuli (Aryadoust et al., 2024) to distractors to vocabulary questions (Zu et al., 2023) and reading/listening comprehension items (Attali et al., 2022; Sayin and Gierl, 2024). Despite the rich potential to create content at scale for innovative tasks, generative AI-based approaches have not yet been leveraged fully in high-stakes testing (an exception is Attali et al., 2022). One such area that generative AI-based approaches may provide a solution for is creating rich contexts for listening that allows for a much more in-depth specification of the construct in the assessment of listening. Our work in particular builds on the techniques described in Attali et al. (2022) by using large language models to generate tens of thousands of conversations for an innovative assessment of interactional competence - something that has been historically difficult to create at scale. While Attali et al.'s (2022) work uses only a handful of the generated passages as potential sources of distractors for any one item, we take advantage of structural similarities across conversations to re-use the full set of dialog lines from all generated conversations as potential distractors for multiple choice tasks. This allows for the development of a robust content bank that can continuously grow and support future rounds of item generation.

Dialog generation

Dialog generation has primarily been researched as a way of augmenting datasets used for training dialog systems. Initial datasets for these tasks were typically collected using crowd workers, which can be expensive and potentially require significant training for more specialized tasks (Dinan et al., 2019; Liu et al., 2021; Gopalakrishnan et al., 2019; Zhang et al., 2018). Researchers have therefore explored data augmentation techniques to create new dialog examples to supplement these existing datasets (Chen and Yang, 2021; Sun et al., 2023). One line of research for whole-dialog generation uses individual simulators for both users and agents, which are trained on existing datasets to help create new instances. This approach has most commonly been applied to task-oriented dialog generation (Papangelis et al., 2019; Hou et al., 2019), though it has also been applied successfully to open-domain and knowledge-grounded dialog generation (Kim et al., 2023; Mohapatra et al., 2021; Wu et al., 2022).

More recently, several works have explored dialog creation using or assisted by LLMs due to LLMs' strong few-shot learning capability which allows the creation of new dialogs using a very small handful of examples. While the simulator method could only generate dialogs in the domain of their training sets, LLM-based dialog generation allows for quick development of large datasets in new domains. One such approach that is similar to ours for dialog generation is the concurrently developed SODA dataset (Kim et al., 2023). Their approach uses triples automatically harvested from a commonsense knowledge graph and uses OpenAI's text-davinci-002 model to first convert them into short narratives and then use those narratives to generate dialogs. Their approach is focused on producing a highly diverse range of social contexts but does not attempt to balance the resulting dataset for any specific conversational characteristics. Another example is the PLACES dataset (Chen et al., 2023) that combined a small sample of hand-crafted example conversations with topics extracted from the Feedback for Interactive Talk & Search Dataset (Xu et al., 2023) to generate new dialogs. Additionally, Lee et al. (2022) created the PERSONACHATGEN dataset to extend the PERSONACHAT dataset (Zhang et al., 2018) with synthetic dialogs using two GPT-3 instances as separate agents, each seeded with a synthetically generated persona. Other works have used partial conversations from existing datasets as the input seeds and used pre-trained language models (Chen et al., 2022) or language models fine-tuned on separate datasets (Zheng et al., 2023) to create new instances by completing the partial conversations. These works, while useful, do not envision the generated dialogs to be used as content for language assessment but primarily as training data for dialog systems, facilitating the need for a separate method for dialog creation that is tailored to standardized large-scale proficiency assessment in order to ensure construct alignment of the generated conversations.

Assessment of listening

The construct of listening has been defined and operationalized in various ways (see Aryadoust and Luo, 2023) but the specific ways in which the construct of listening has been elicited and measured in large-scale standardized assessments has been fairly limited. Buck (2001) proposes five subskills as a starting definition of the listening construct that work in tandem to arrive at listening comprehension: knowledge of the sound system, understanding local linguistic meanings, understanding full linguistic meanings, understanding inferred meanings, and communicative listening ability. To elicit these skills in a reliable manner, large-scale standardized tests mainly make use of discrete-point comprehension questions that ask test takers to listen to a stimulus, read (or listen) to a question, and then choose the most appropriate answer via a multiple-choice format or a true/false format (Carr, 2011; Park et al., 2022; Wagner, 2014). Other formats include asking test takers to fill in the gaps of summaries or complete a flow chart, a diagram, or a table (Suvorov and Li, 2023).

While these assessment formats are successful at tapping into the five aforementioned subskills, they fall short of extending them to communicative situations where test takers are using the processed aural information to interact with another person, which makes up most real-life listening domains (Wagner, 2014). Addressing this concern via conversational stimuli is limited as test takers remain passive listeners, with tasks to be completed after the conversation has taken place rather than during. By positioning test takers as passive listeners in conversations rather than active participants, the construct of listening is specified in a way that favors one purpose of listening, comprehension, over others such as conversation and interaction (Aryadoust and Luo, 2023). This problem of construct underrepresentation threatens our ability to make accurate inferences about a test taker's listening proficiency if the purpose of using the assessment is to determine whether they are adequately prepared for situations that require them to make use of their listening abilities in university settings in which they are expected to converse with other people such as engaging in discussions in office hours, participating in service encounters, and communicating with their peers.

Tasks that address the concern of construct underrepresentation by appointing test takers as active participants in conversations may be able to further broaden their construct coverage with the help from generative AI. An example of tasks that situate test takers in conversations would be a discourse-completion task typically employed in assessment of pragmatic competence (Brown, 2018; Buck, 2001), where participants listen to a conversation and complete the last turn in the conversation. An instantiation of this task in a large-scale standardized context can be seen in Papageorgiou et al. (2021) which has test takers listen to a conversation and complete the last turn in the conversation. Extending this type of task beyond having test takers complete the last turn to them fully engaging in the conversation would allow a fuller specification of the listening construct in interactional situations.

Assessment of interactional competence

Interactional competence is defined as "the ability to co-construct interaction in a purposeful and meaningful way … supported by … aspects of topic management, turn management, interactive listening, breakdown repair and non-verbal or visual behaviours" (Galaczi and Taylor, 2018, p. 224). The ways in which interlocutors exercise their interactional (sub)skills according to the speech events and speech acts, which are in turn based on speech situations, facilitate the characterization of interactional competence as a tree (Galaczi and Taylor, 2018). An example of a speech situation would be office hours between a student and a professor in a university setting. The speech event is the actual conversation that transpires between the two interlocutors, who subsequently employ interactional subskills such as asking questions for topic management and maintaining for turn management.

Assessing interactional competence has been a notoriously thorny problem for large-scale standardized high-stakes language proficiency exams (Dai, 2024; Roever and Dai, 2021) due to difficulties with scoring, operationalization of the construct, and finding the right balance between reliability and construct coverage (Galaczi and Taylor, 2018). Typically, measurement of this construct, or a part of it, has been targeted in speaking tasks with interactive task formats such as oral proficiency interviews (OPIs) and paired tasks (Roever and Kasper, 2018; Youn, 2020). These interactive formats provide opportunities to measure interactional competence, but they also introduce possible sources of error as well.

In OPIs, human examiners are paired with test takers for a face-to-face spoken interaction to elicit language samples to rate test takers' oral proficiency. OPIs have been shown to be distinct from naturally occurring conversation in the TLU domain in terms of turn-taking, topic control, and question-response patterns (Johnson and Tyler, 1998; Seedhouse, 2013; van Lier, 1989; Young and He, 1998). The source of these differences from conversation has been found to be due to differences between the task and TLU domain with respect to the actual purpose for interaction in the TLU domain and the power difference between interlocutors. All of these threaten the generalizability of a test taker's performance from OPIs alone to how they would actually utilize their interactional competence in the TLU domain (Plough et al., 2018; Staples et al., 2017; van Lier, 1989).

While the interview-based format of OPIs has been found to deviate from core elements of conversation and interactional competence, the format of guided role plays has demonstrated more accurate alignment to interactional elements in conversation (Kormos, 1999). Typically, the two test takers have a common goal, or purpose, for communicating which creates a language test situation that is more aligned with the TLU domain with respect to interactivity (Davis, 2009; Galaczi, 2008). Additionally, the examiner is usually an onlooker or a facilitator, which mitigates against power differential effects that are found in interview formats (Davis, 2009). However, research has shown that characteristics (such as gender, extroversion, and proficiency) of interlocutors in paired tasks may or may not affect the language that a test taker uses and the score that they receive on the task (Davis, 2009; Galaczi and Taylor, 2018; Iwashita, 1996; Nakatsuhara, 2011). Galaczi and Taylor (2018) argue that this interlocutor induced variability should either be conceptualized as construct irrelevant variance and removed from task design and scoring or conceptualized as construct-relevant variance and accounted for in task design and scoring.

One promising direction for automating and scaling the assessment of interactional competence is the use of spoken dialog systems (SDS). These systems use a combination of automated speech recognition (ASR), a dialog management module, and text-to-speech (TTS) to receive spoken responses from a human and respond to them according to the goals defined by the developers of the system. This allows test takers to go through a full conversation with the system in a fully automatable way. SDS have a long history of research and application in computer-assisted language learning (e.g., Eskénazi, 2009; Su et al., 2015; Timpe-Laughlin et al., 2017), but their application to language assessment has been more limited until the last few years (Litman et al., 2018). Recent studies of SDS have laid the groundwork for using these systems to assess interactional competence by establishing that human to computer interactions display pragmatic functions similar to those in human-to-human communication (Dombi et al., 2022) and can be used to assess interactional competence (Ockey et al., 2023). Human-to-computer interaction has also been found to be comparable to human-to-human interaction (Ockey and Chukharev-Hudilainen, 2021), with the potential to improve task design with teacher feedback (Timpe-Laughlin et al., 2020).

SDS-based assessment tasks have some limitations that make them difficult to deploy for large-scale, high-stakes testing. The primary issue is the difficulty of scaling them to produce new assessment tasks. SDS-based assessment tasks typically combine pre-written dialog trees for the system's responses and a set of rules regarding the content of the user's responses to govern the transition from one line in the script to the next (Chukharev-Hudilainen and Ockey, 2021; Gokturk and Chukharev, 2024; Karatay, 2022; Timpe-Laughlin et al., 2017). These rules are usually constructed by hand using combinations of keyword and regex matching, with keywords sourced through manual analysis of data produced either by the researchers (Chukharev-Hudilainen and Ockey, 2021) or crowd workers (Ramanarayanan et al., 2016). As a result, each individual task requires a tremendous amount of manual effort to construct, which constrains the range of situations and interaction structures that can be assessed as well as the number of items that can be created. In addition, issues such as the lack of authenticity in interaction, difficulty in being understood by the system, or the system providing non-meaningful responses could negatively impact test taker performance and lead to issues of fairness (Gokturk and Chukharev, 2024), all of which pose threats to the validity of test scores in high-stakes tests.

The interactive listening task

In this paper we describe the development of a new task which we call the Interactive Listening task. It overcomes the limitations of traditional listening tasks by bringing in test takers as active participants in a conversation. In this task, test takers are in a role-play setting. They listen to their interlocutor and select the best turn to move the conversation forward. They then receive immediate feedback about the correct answer and the process of listening to the next turn, selecting the best response and receiving feedback repeats itself. This multiple-choice format, coupled with generative AI, extends previous tasks assessing listening beyond answering comprehension items and completing the first or last turn of a conversation. The task also controls for interlocutor effects that constrain the interactions in speaking tasks by introducing language variation that is relevant to the TLU domain. From the perspective of speaking tasks, it is an indirect measure of speaking as turns are selected rather than spoken.

Task overview

Figure 1 demonstrates how a test taker would proceed through the Interactive Listening task.

Test takers first read a short description (henceforth referred to as a scenario) which outlines the goal of the conversation, the test taker's role, and the interlocutor's role. The interlocutor is represented as an animated character drawn from the cast of the Duolingo World Characters (Chiu et al., 2021; Hartman, 2020). The interlocutor lines are turned into audio stimuli using custom text-to-speech (TTS) models created for each character. Test takers listen to the character speak their lines in the conversation and then select the option that best continues the dialog, taking into account the scenario and prior conversation turns. Each Interactive Listening task consists of 5–6 of these exchanges between the test taker and the character. Finally, test takers produce a written summary of the conversation they

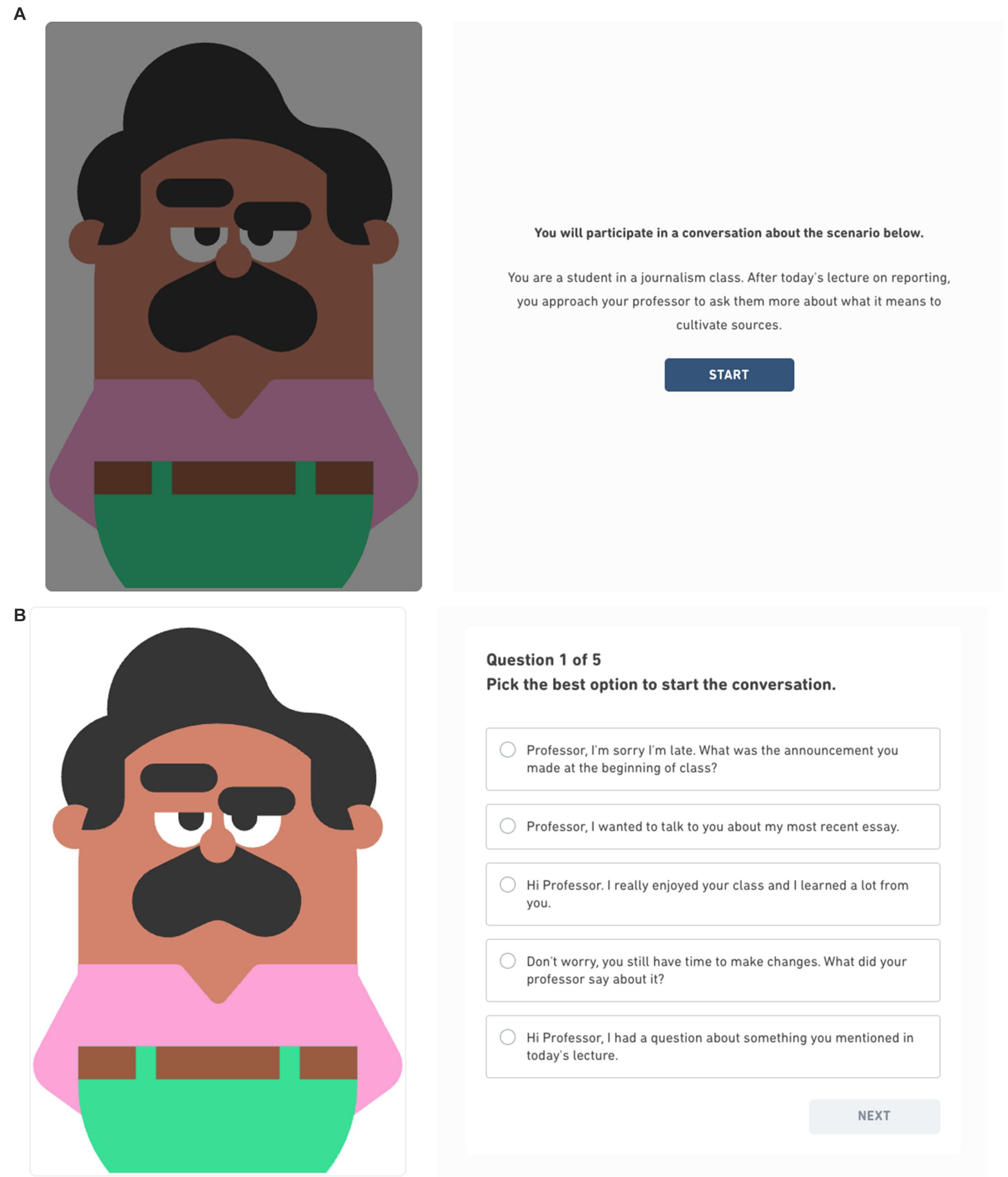


FIGURE 1 (continued)

FIGURE 1 (continued)

participated in. For this paper, we focus our discussion of task and item performance on the multiple-choice items.

Conversations feature topics and roles that are common in university settings. In all conversations, the test taker plays the role of a student, while their interlocutor may be a fellow student or a professor, referred to as student–student and student-professor conversations, respectively. Conversation topics cover a range of real-world communicative purposes, such as asking for and giving advice, making requests, gathering information, and making plans. For example, test takers may need to set up plans to work on a group project with another student, request a letter of recommendation from a professor, or give advice to a friend about what courses to take. The situational context of the task aligns with real-world situations for interacting with peers and professors in university contexts in terms

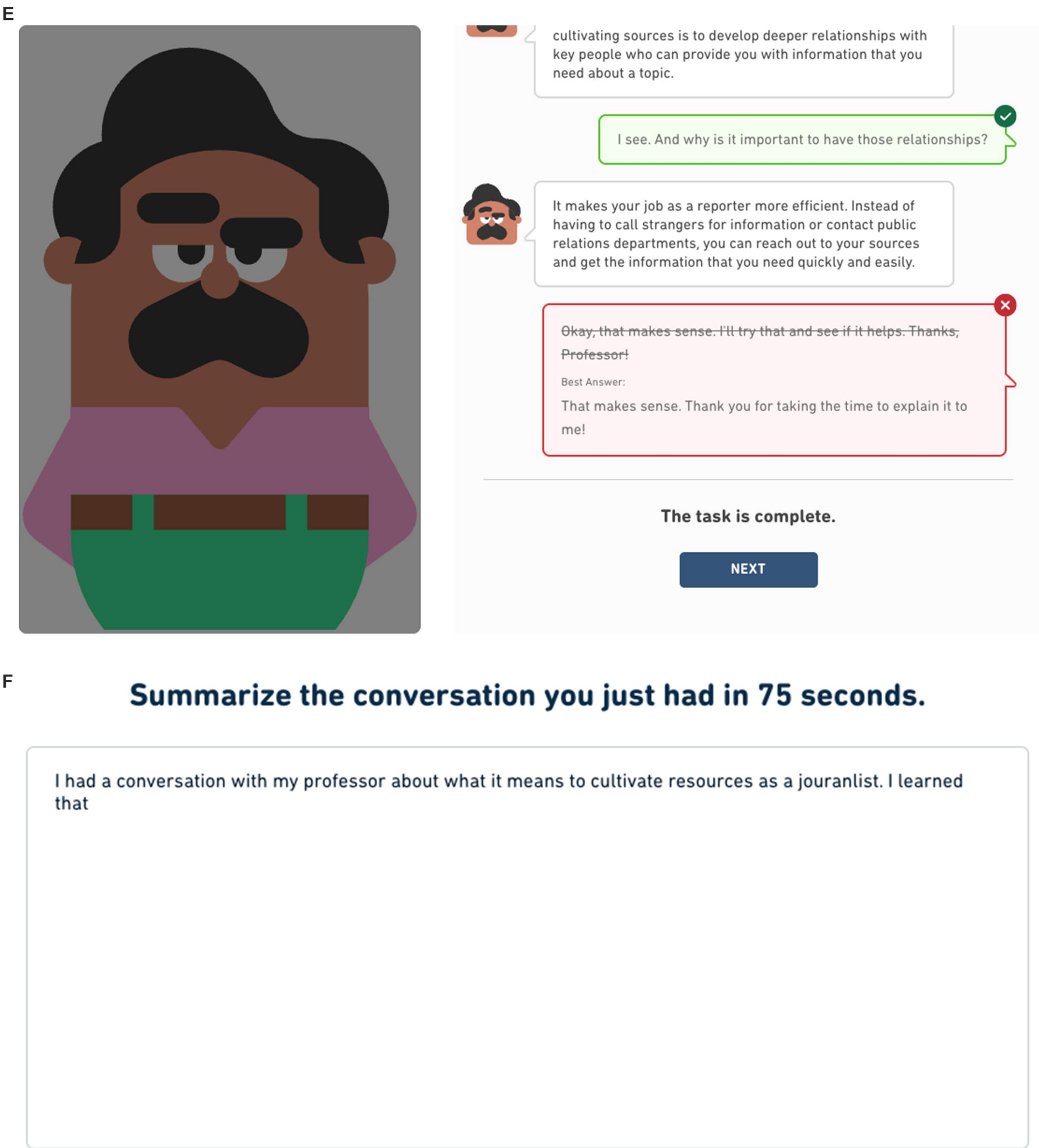


FIGURE 1
Walkthrough of an interactional competence task. The task starts (A) with a scenario that describes who the test taker is talking with and for what purpose. After they click the start button, the conversation begins. In this task (B), the test taker makes the opening turn of the conversation. When the test taker selects the best answer for the turn (C), they receive feedback through color (green) and a check mark. Test takers receive visual feedback when they select an incorrect option (D) through color (red) and an "x" in the upper right corner of the dialog box. This process repeats until the task is complete (E). Test takers may review the conversation before moving to the summary task (F), where they summarize the content of the conversation.

of the topics, the reasons for the conversation, and the relationship between interlocutors.

Task design considerations

We use generative AI to support the development of the Interactive Listening task for to its ability to scale the production of tasks (and turns within a conversation) to cover the spectrum of possible topics, purposes for communicating, and communicative settings in an extended conversational task. Where traditional discourse completion tasks typically target single turns in a conversation (usually an opening turn or a closing turn), a task that includes measurement of how test takers navigate an entire conversation creates the possibility to assess more aspects of interactional competence within a single task.

The Interactive Listening task uses predetermined non-branching conversation paths rather than dynamically generated conversations based on test-taker responses. Predetermined conversation paths mean that subsequent turns are not reactive to prior turns; rather it is the test taker's task to identify which option out of a set of presented responses leads toward task completion. The predetermined conversation paths with the multiple-choice format, while somewhat limited in the extent to which the test taker can nominate the topic and the direction of the conversation, provide several advantages, including efficient and construct-aligned item generation and review process, mitigation of risks due to unpredictability, and the opportunity to provide test takers with feedback.

The combination of offline LLM generation and human review allows us to expedite the item generation process while ensuring that the resulting items meet the quality standards necessary for a large-scale high-stakes English proficiency test. The unpredictability of dynamic generation, on the other hand, makes it difficult to maintain consistent quality standards for the items, while at the same time introducing a potential vulnerability through adversarial attacks intended to trigger unexpected behavior from the model, commonly referred to as "jailbreaking" (Wei et al., 2023). Dynamic generation also introduces the risks of model hallucination, in which the model invents new facts or fails to stay consistent with information provided by the model in earlier turns of the dialog (Ji et al., 2023). As the task is centered around academic situations, self-consistency and consistency with real-world information are both critical to avoid potentially confusing test takers. While there has been a large body of work attempting to improve self-consistency (Li et al., 2020; Song et al., 2020) and consistency with external knowledge bases (Rashkin et al., 2021; Shuster et al., 2021; Sun et al., 2023), these solutions are not perfect and introduce significant additional risk and complexity to the task design. Predetermined conversation paths address this by enabling human review of the content that can mitigate the risk of hallucinations and ensure fairness in the generated content (Belzak et al., 2023). Feedback from human reviewers can additionally be used in a human-in-the-loop process to improve the AIG processes and the quality of the generated content. Finally, static conversations allow us to better ensure that the language used in the conversation adheres to aspects of the communicative contexts that we are interested in modeling, such as the power dynamic between the participants.

Additionally, we use a single path through the conversation, rather than modeling each item as a dialog tree because while it would allow test takers more control over the direction of the conversation, it would also introduce significant generation and psychometric modeling issues. Given our decision to not use dynamically generated system responses, a task that supports branching the conversation based on open-ended responses from the user would require a complex and highly specialized workflow for every single individual item created (ex: Timpe-Laughlin et al. (2017) p. 5). As a result, all prior work using branching conversation as a basis for assessment has typically been limited to 2–4 unique items (Evanini et al., 2014; Chukharev-Hudilainen and Ockey, 2021; Karatay, 2022). For a high-stakes test, we must be able to regularly create large numbers of new items to refresh the bank in order to ensure the security of the test, making this limitation infeasible. Another alternative would be to develop branching multiple choice items, but this approach also has significant issues. As the task is administered adaptively based on the test taker's estimated proficiency, each branch in a given conversation would need to have similar psychometric properties to provide comparable item information across branches at the same turn index. This would be nearly impossible to guarantee during item generation or from a human review process for conversations with branches that require significantly different potential sets of options. While this design could work for weakly ordered conversations that would support shuffling the order of the turns in the conversation, such as asking for information about a list of classes or the types of slot-filling tasks common in task-based dialog systems, it would unacceptably limit the range of interaction types we could assess. This approach would also require significantly more data collection during piloting to support calibration of all potential branches and could result in significant numbers of items being discarded due to poor measurement quality.

From an assessment perspective, the Interactive Listening task is a type of discourse-completion task. However, whereas existing tasks (e.g., Papageorgiou et al., 2021) have participants listen to a conversation and complete the last turn in the conversation (by selecting the best option in a multiple-choice format), our task is an interactive multi-turn discourse-completion task that asks participants to repeatedly select the next turn that best fits the conversation thus far. The interactive element of this task is achieved by providing immediate feedback about participants' choices. Test takers who select the wrong option for a given turn are shown the correct one before moving to the next step in the conversation. This innovative interactive nature of our task allows test takers to take a much more active role in the conversation in a way that can potentially broaden the task's construct coverage.

Content and item generation

This section outlines the processes used to create the conversations and options required for the Interactive Listening task. We used the GPT-3[1] (Brown et al., 2020) family of LLMs throughout the development of this task, specifically the text-davinci–002[2] model. Figure 2 shows the workflow we used to create the conversations and options, systematically progressing from the initial generation of conversational scenarios to the formulation of conversations and then finally selection of distractors.

Conversational scenarios

Our approach to conversation generation is similar to the concurrently developed method used by Kim et al. to create the SODA dataset (2023). While their approach uses automatically extracted triples to seed the generation of short narratives (or conversational scenarios), we use a list of 130 short descriptions of a broad range of typical conversational purposes between two students or between a student and a professor in academic contexts. These scenarios were selected by assessment experts to cover a wide range of basic interactive tasks such as requesting help, asking for

1 The most recent commercially available version of GPT at the time of development (Fall, 2022).

2 This model has been replaced by the more recent text-davinci-003 model. See https://platform.openai.com/docs/deprecations for more details.

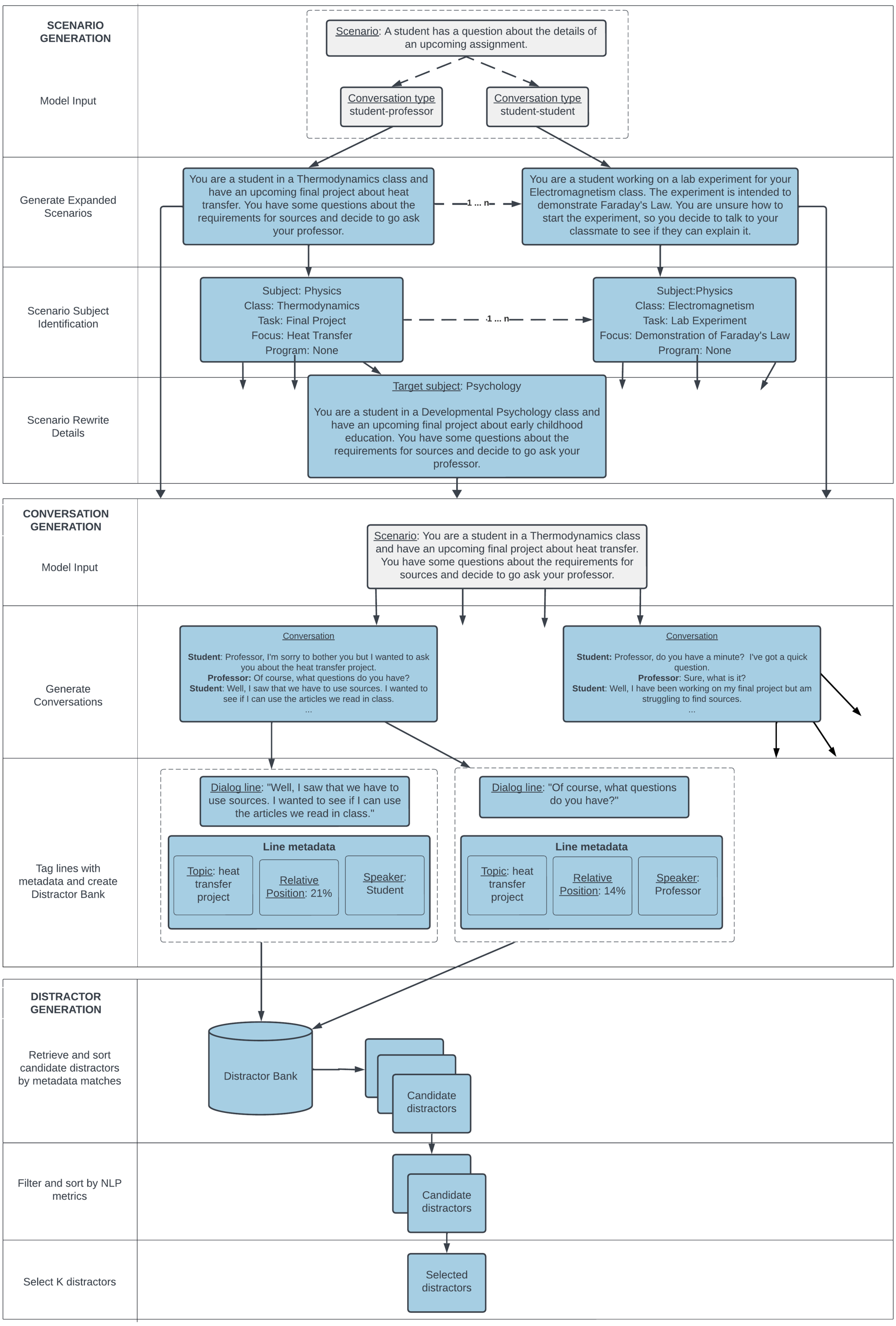


FIGURE 2
Example of a base scenario and an expanded version with additional details.

advice, providing feedback or recommendations, seeking information, discussing and comparing options, and describing a recent experience (Council of Europe, 2020). We will refer to these as the base scenarios.

For each of these base scenarios, we used a few-shot prompting approach (Brown et al., 2020) with GPT-3 to expand these scenarios with additional details such as specific classes or academic subjects, the type of assignment, the feelings and preferences of the participants, and the relationship between the participants. These additional details in turn can then be used by the model when generating conversations to produce more distinctive and varied conversations from a single hand-crafted input.

For each of our base scenarios, we created 50 detailed scenarios using GPT-3. We generated them five at a time for each base scenario, providing 3–5 examples as input (see Appendix A) that were randomly selected and shuffled from a group of 10 examples to increase the diversity of the outputs. Even with a high temperature value that encourages more diverse output, this approach produced a significant number of near-matching scenario descriptions. We used the sentence-transformers library[3] (Reimers and Gurevych, 2019) to embed them into a vector space and calculated pairwise cosine similarity between all generated scenarios, removing those with very high cosine similarity to a previously generated scenario. This resulted in 3,900 detailed scenarios.

One challenge we encountered with this approach is that the detailed scenarios produced by GPT-3 tended to cluster around a small number of academic subjects. We used another few-shot GPT prompt (see Appendix B) to identify any mentions of academic courses or degree programs and to identify the academic subject area(s) they belonged to. After manually reviewing and combining some closely related areas, we evaluated the distribution of the academic subjects, with the 25 most common shown in Figure 3.

We found that the top 25 most common areas accounted for over 90% of the total scenarios that featured specific academic subjects. Text generation with large language models is ultimately accomplished by repeatedly sampling from a distribution over possible words learned from their training data and further conditioned by the instructions and examples. Although we tried to modify this underlying distribution through randomly selecting and shuffling the examples we provided to condition the model's outputs, we still found a strong preference for certain academic subjects and course names in the resulting detailed scenarios.

To correct this tendency, we prompted GPT to rewrite the scenarios to focus on under-represented subject areas from a list of common university degree programs (prompt in Appendix C). For each detailed scenario with an identified academic subject, we used this prompt to rewrite the scenario to target five new subject areas. After deduplication, we ended up with an additional 14,000 scenarios, with the top 25 most common subject areas accounting for just 31% of them (see Figure 4).

Conversation generation

To generate conversations, we created two GPT-3 prompts, one to use for generating student–student conversations and one for student-professor conversations. Each prompt consisted of three examples of scenarios and conversations as well as a simple set of instructions telling GPT-3 to include as many details from the scenario as possible in the resulting conversation (see Appendix D). We then provided a detailed scenario from our set of generated scenarios to generate new conversations. For each generated conversation, we assigned one of the student roles to be the test taker (or the only student role in the student-professor setting).

In addition, we created a third prompt with the goal of creating conversations that would require more listening from the test taker and understanding of more complex academic materials. This prompt required a broad topic such as "statistical correlation" as input, rather than a scenario, and included instructions to create conversations framed around a professor explaining the topic to a student, with the student asking questions to elicit more information. The examples used in the prompt had very long lines for the professor character. This allowed us to generate conversations that more reliably required substantially more listening from test takers compared to the first two prompts.

Using the almost 18,000 scenarios generated in the previous step and these three conversation generation prompts, we created a pool of nearly 125,000 conversations from which we sampled conversations for all subsequent experiments. When selecting conversations to use for the Interactive Listening tasks, we filtered conversations to limit the variability in the task's requirements, including:

- 5–6 turns for the test-taker role.
- >10 words on average across all test-taker role lines.
- <40 words maximum for any non-test-taker lines.
- Test taker must have the last line in the conversation.

For the selected conversations, we additionally identified named entities in the conversations corresponding to other students or professors mentioned in the conversations. We manually reviewed the identified entities and filtered out any that were mentions of living or historical people and then replaced the names of the remaining entities with names randomly selected from a collection of the most common first and last names from countries around the world to help mitigate potential gender or regional biases in the generated conversations.

Distractor generation

Each multiple-choice item in an Interactive Listening task corresponds to one of the test taker's turns in the conversation. As such, the keys for the task are simply the original lines in the conversation. Extending the work of Attali et al. (2022), we obtain incorrect answers (distractors) to multiple-choice items (turns) by evaluating and selecting turns from other generated conversations.

For this work, we started by extracting each line of each conversation from the 125,000 conversations generated in the previous step, creating a bank of nearly 1 million lines of dialog. For each line, we tracked metadata related to the line's source conversation, such as the academic topic, scenario, detailed scenario, its relative position within the conversation, and the role of the line's speaker in the conversation.

We selected distractors for each multiple-choice item from this dialog bank using a three-step process. First, we looked at the relative position of where in the conversation the current multiple-choice item occurred. We grouped these positions into one of three categories: the first 20% of the conversation, the last 20%, or the middle 60%. We then limited candidate distractors to only those that came from the same grouped position in their source conversations. Lines at the beginning and ends of conversations feature opening and closing lines with more formulaic expressions and expectations (House and Kádár, 2023;

3 http://sbert.net, using model paraphrase-MiniLM-L6.

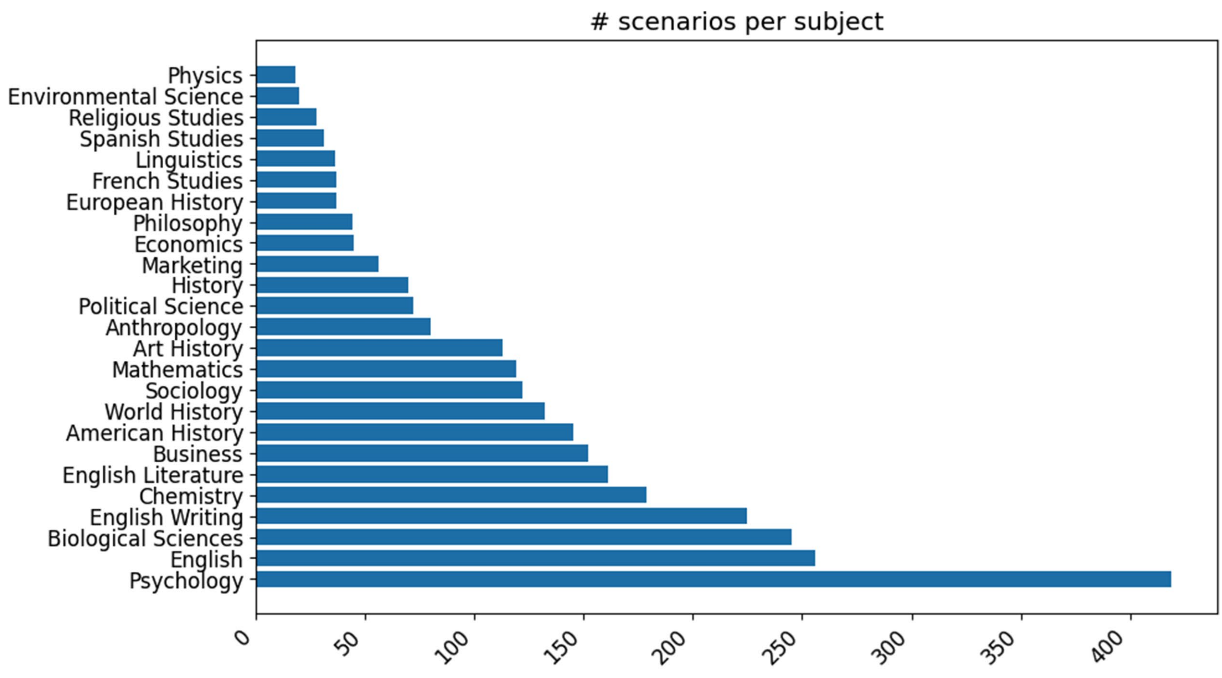


FIGURE 3
Distribution of the top 25 most common academic subjects in our original set of automatically expanded base scenarios.

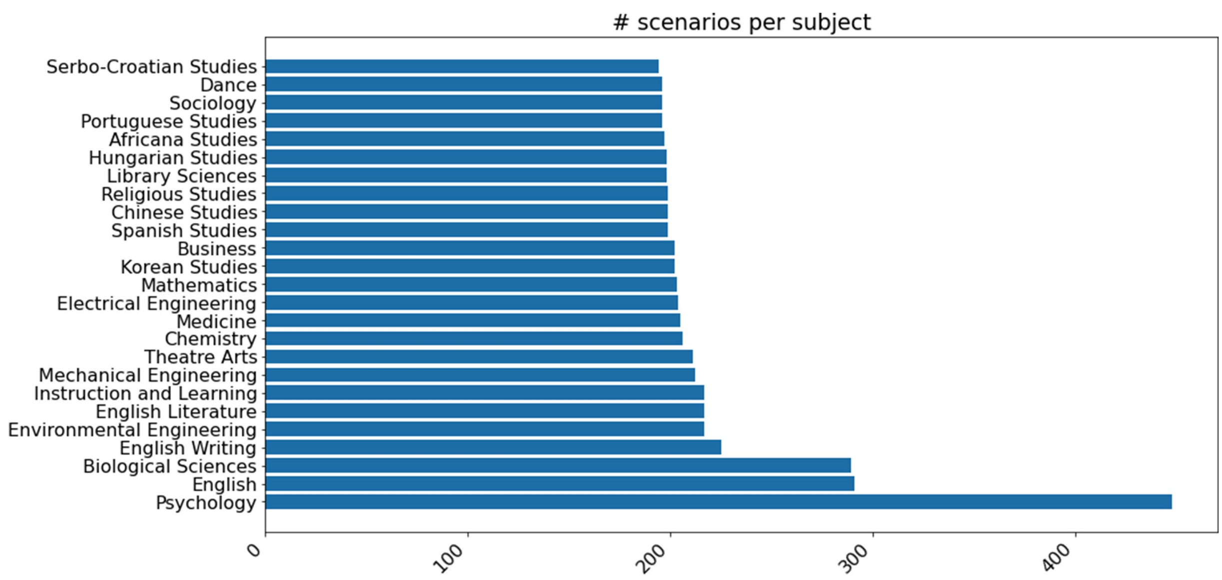


FIGURE 4
Distribution of the top 25 most common academic subjects in our set of scenarios after generating rewrites to scenarios targeting specific academic subjects.

Jucker, 2017), so this initial restriction of the distractor pool accordingly helped to eliminate distractors that would potentially be very easy to identify for test takers.

Second, we ordered potential distractor candidates based on their source conversation metadata. We gave preference to lines that came from conversations generated from the same detailed scenario, and then to those from the same base scenario. This approach is similar to the one employed in Attali et al. (2022) which used passages with similar subjects and characteristics as sources for distractors. We additionally preferred lines spoken by student characters in the conversations. At this step, we also ensured distractor candidates matched surface characteristics of the key, such as the length and ending punctuation. We also applied an automated editing step to ensure consistency of character names at this step. If a candidate distractor contained a mention of a character external to the conversation, such as a professor, then we automatically replaced the name in the distractor to match one of the names of the characters mentioned in the surrounding task conversation. We selected the top 200 candidates ordered by these criteria for further evaluations.

Finally, we computed a set of additional metrics to re-rank and select distractors for each multiple-choice item for a given task. We used the sentence-transformers library to compute vector space cosine similarity between each candidate distractor and the key. We also used a large language model to estimate how plausible the distractor would be at the given point in the conversation compared with the key.

For a given Interactive Listening task with N multiple-choice items, we defined O as the original prompt that generated the conversation, S as the scenario for the task, t_i as the i-th turn in the conversation, $C_{<i}$ as the concatenation of all turns in the conversation prior to turn i, $k_{i,n}$ as the key to the n-th multiple-choice item, corresponding to t_i in the conversation, and $d_{i,j,n}$ as the j-th candidate distractor for the n-th multiple-choice item, administered at t_i.

To get the likelihood of the key for the n-th multiple-choice item, we concatenated O, S, $C_{<i}$, $k_{i,n}$, and t_{i+1}. We submitted this input, I, to OpenAI's text-babbage-002 model and had it return the log probabilities for each token m_h in the input, $\forall_{m \in I} \log\left(P\left(m\right)\right)$. We then computed the token-average log probability metrics for the key, $k_{i,n}$ and the turn that followed it, t_{i+1}:

$$key_{i,n}\_log\_probability = \frac{1}{len_{tk}\left(k_{i,n}\right)} * \sum_{m_h \in k_{i,n}} \log\left(P\left(m_h | O,S,C_{<i},m_{<h}\right)\right)$$

$$post\_key_{i,n}\_log\_probability = \frac{1}{len_{tk}\left(t_{i+1}\right)} * \sum_{m_h \in t_{i+1}} \log\left(P\left(m_h | O,S,C_{<i},k_{i,n},m_{<h}\right)\right)$$

We can repeat this process for each distractor candidate $d_{i,j,n}$, concatenating O, S, $C_{<i}$, $d_{i,j,n}$, and t_{i+1} and compute:

$$distractor_{i,j,n}\_log\_probability = \frac{1}{len_{tk}\left(d_{i,j,n}\right)} * \sum_{m_h \in d_{i,j,n}} \log\left(P\left(m_h | O,S,C_{<i},m_{<h}\right)\right)$$

$$post\_distractor_{i,j,n}\_log\_probability = \frac{1}{len_{tk}\left(t_{i+1}\right)} * \sum_{m_h \in t_{i+1}} \log\left(P\left(m_h | O,S,C_{<i},d_{i,j,n},m_{<h}\right)\right)$$

When selecting distractors for the n-th multiple-choice item, we evaluated the difference between each $distractor_{i,j,n}\_log\_probability$ and $key_{i,n}\_log\_probabillity$ as an estimate of how likely the distractor is in the context of the conversation, relative to the key. A positive difference meant that the model estimated that the distractor is more likely than the key, which indicated that the distractor could be a valid answer in a multiple-choice item. We likewise measured the difference between the $post\_distractor_{i,j,n}\_log\_probability$ and $post\_key_{i,n}\_log\_probability$, which estimated whether the distractor ultimately could still fit in with the remainder of the conversation. A positive or near-0 difference here could serve as additional evidence that a distractor fits in potentially too well with the context of the conversation. For multiple choice items that used the final turn in the conversation as the key, the *post_key/distractor* metrics were not computed as they lacked a following turn.

We set thresholds on the two log probability difference metrics and the cosine similarity metric to filter out distractors that were too plausible in the context of the conversation or too similar to the key. We then sorted distractors by the ratio between their key-distractor log probability difference and their cosine similarity to the key. This resulted in a ranking that prioritized distractors that maintained a balance between how similar they were to the key and how likely they were in the context of the conversation, with distractors that were too unlikely or too dissimilar ending up near the bottom of the rankings. We finally selected the top K candidates for review and piloting, with the added requirement that each candidate had a cosine similarity less than or equal to 0.6 with all other selected candidates to avoid having multiple, highly similar distractors for a single item.

Task design pilots

This section presents the results from a set of experiments conducted to support iteration on the design of the Interactive Listening task and gather evidence about its performance. We conducted these experiments using a piloting platform that was developed as part of the DET practice test. This practice test is an online version of the DET, freely available to anyone interested in the test, with over 10,000 test sessions daily. Similar to the operational test, the practice test is fully adaptive and test takers have the opportunity to respond to and practice all of the tasks that are included in the operational test. The piloting platform is an opt-in section at the end of the practice test with experimental tasks, and around 50% of practice test takers choose to complete these additional experimental tasks. We used this platform to conduct around 10 controlled experiments over the course of 6 months to test different task design options.

We briefly report the results of three of these experiments, exploring the effects of (1) the number of multiple-choice options, (2) displaying the full history of the conversation, and (3) allowing replay of the interlocutor's audio on test taker performance.

In these experiments, we focused on item statistics and response times. Item difficulty was measured using percent correct and item discrimination was measured using item-total correlations where total scores were practice test scores.

Number of options

We start with results from an experiment that varied the number of options, comparing 4 options (3 distractors) with 5 options. We used a set of 29 Interactive Listening tasks with a total 150 multiple-choice items and collected responses from at least 250 test takers for each conversation in each condition. As expected from the decreased opportunity for guessing, items with five options ($M = 0.54$,

$SD = 0.14$) were harder than items with four options ($M = 0.57$, $SD = 0.14$). A paired t-test confirmed that this difference is significant ($t[149] = 2.55$, $p = .012$), although the effect size is quite small ($d = 0.19$) (Cohen, 1988).

Similarly, five options ($M = 0.25$, $SD = 0.09$) yielded higher discrimination than four options ($M = 0.22$, $SD = 0.10$). A paired t-test confirmed that this difference is significant, ($t[149] = -4.32$, $p < 0.001$), with a medium effect size ($d = 0.38$). Figure 5 shows that the advantage in discrimination held after controlling for item difficulty, indicating that having five options in an item is more discriminating than four options regardless of item difficulty.

Finally, Figure 6 shows that median response times with 5 options are only 2 s longer than with 4 options.

As a result of this experiment, we decided to continue with 5 options for higher psychometric standards.

Showing history

The second experiment tested whether the availability of the history of the conversation had any impact on the test-taker performance. In one condition, test takers were able to review the full history of the conversation, including text transcripts for each of the interlocutor's prior turns. In the other, test takers could only review a transcript of the interlocutor's most recent turn, as well as their own most recent response. The motivation for allowing test takers to review the full history of the conversation was to minimize the potential for construct-irrelevant variance from the effect of memory on test-taker performance. We used the same set of 29 tasks from the previous experiment and collected responses from 250 test takers for each. We found that showing the history had very small and non-significant effects on median response time (means of 36 s in both conditions), item difficulty (means of 54% correct in both conditions), and item discrimination ($M = .239$ with no history and $M = .246$ with history). We ultimately decided to show history to test takers as doing so would give us more flexibility in the future to increase the length of the conversations, both in the number of turns and the maximum length of each turn, without the possibility of memory interfering with test-taker performance.

Number of plays

The last experiment manipulated the number of allowed plays for each interlocutor turn in the conversation: a single play (no replay allowed) vs. two plays (one replay allowed). The motivation for not allowing replays was to more closely simulate a real conversation. This experiment was based on a subset of 10 of the conversations from prior experiments with a total of 51 turns and 250 responses each. In the replay condition, we found that test takers only used the feature in 16% of the total items. The effect of allowing a replay on item statistics was analyzed using paired t-tests. Allowing a replay had a small effect on median response time in seconds ($\Delta M_{rt} = 1.29$, $p < .001$, $d = 0.14$), a small effect on percent correct ($\Delta M_{rt} = 0.02$, $p = 0.001$, $d = 0.15$), but no effect on discrimination ($\Delta M_{rt} = 0.00$, $p = .776$, $d = -0.03$). Because the replay feature was not extensively used and had minor effects on item statistics, we decided to allow only a single play, with an added advantage of simplifying the user interface.

Large-scale pilot

The culmination of the iterative task development process was a large-scale pilot in which the task was administered as a regular part of the DET practice test. The purpose of the pilot was to evaluate the quality of the AIG processes described in the previous section from both a human review and psychometric perspective. We also wanted to assess the psychometric properties of the overall task within the context of a larger assessment, as opposed to the opt-in process used in the previous section.

In particular, the questions we wanted to answer with this pilot were:

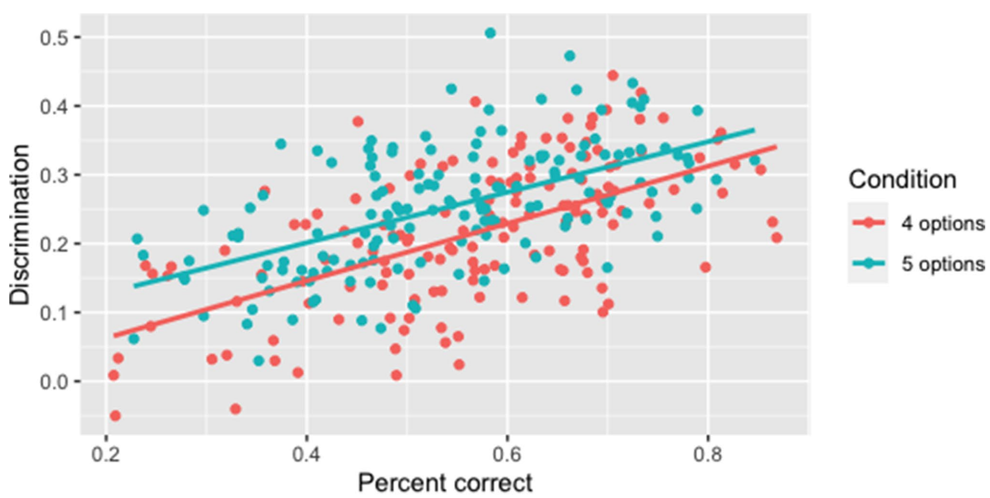


FIGURE 5
Results from varying number of options: Discrimination as a function of difficulty.

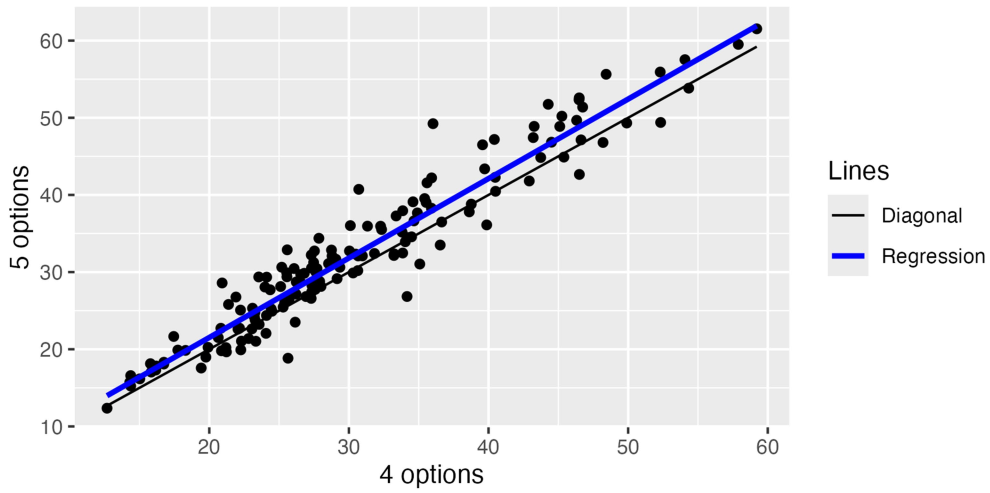


FIGURE 6
Results from varying the number of options: Median response times.

1 What is the quality of the automatically generated conversations and multiple-choice items, as assessed by expert human reviewers?
2 To what extent do linguistic and semantic features of the items impact their psychometric properties?
3 To what extent do the amounts of audio and text content impact psychometric properties?
4 To what extent do human review processes improve the psychometric properties of the AIG items?
5 To what extent does the multi-turn format of the task impact the psychometric assumption of local item dependence?

Pilot item bank creation

To create the item bank for the pilot, we sampled 900 conversations from the set of 125,000 we generated, distributed across all of our base scenarios and academic subjects. Each of the test taker turns was converted into a multiple-choice item for which 6 distractors were generated. In practice, some items had fewer than 6 distractors, typically due to a combination of the key being extremely short or extremely long as well as our other filters and restrictions. Each item then underwent an extensive human review and editing process, described in the next section.

Human review

Our human review process was designed to ensure that we maintain high quality standards (Saville and McElwee, 2021). All tasks underwent item quality, fairness and bias (FAB), and audio quality reviews. These reviews were conducted by 25 external reviewers and six internal Duolingo team members. External reviewers had diverse backgrounds with regard to gender identity, age, and racial/ethnic background. All had at least a bachelor's degree (and in some cases a Ph.D.) in linguistics, language studies, or a related field. All had expertise in teaching and assessing in relevant language and cultural contexts.

Each task first went through a two-phase item quality review process. The first phase was conducted by 10 reviewers, who reviewed the task, including scenario, conversation, and items. The reviewers verified that the scenarios included sufficient context, introduced the participants, and framed the dialog appropriately to ensure that the test taker would be able to successfully participate in the conversation. For the conversations, reviewers evaluated the cohesion, clarity, and logical consistency throughout the text, also ensuring that each conversation included a speech situation, a speech event, and a speech act. The general purposes of the conversations were making a request, clarifying information, gathering or sharing information, and making a recommendation. Reviewers also confirmed that the conversations met the intended purpose of the task. For the items, reviewers judged the viability of each option by ensuring that the correct answer was correct and the distractors were incorrect. As part of this process, reviewers selected the best 4 distractors out of the 6 available options and made additional edits as needed to ensure clarity, grammaticality, and topic similarity to the key. Reviewers prioritized answer options that would require the test takers to demonstrate an understanding of social dynamics such as politeness and indirectness, particularly in relation to social distance factors. In cases where there were fewer than four available distractors, reviewers were instructed to write new distractors. Item reviewers were also trained on fairness guidelines and edited out any content that clearly violated these guidelines. All reviewer edits then went through a second phase, wherein the edits from the first phase were carefully reviewed, evaluated, and incorporated by a team of four expert reviewers. If a task required extensive edits, defined as requiring more than 20 min to fix, then reviewers abandoned the task. 725 out of 900 tasks (81%) successfully passed the item quality review.

In analyzing the edits that reviewers made, the majority aimed to address issues such as:

- Content that clearly violated our fairness guidelines.
- Items that required background knowledge about a specific topic to answer.
- Overly complex topics and technical jargon.
- Logical inconsistencies.

Throughout the review we refined our item review guidelines in response to inquiries from item reviewers. For example, discussions about whether certain academic terms, such as internship, were globally relevant, or considering whether conversations about receiving a bad grade could potentially cause test-taker anxiety.

Following the item quality review, expert FAB reviewers carefully evaluated the scenarios, conversations, and options for any remaining content that could be controversial, too culturally specific, or unfamiliar to the intended global test taker audience. Each piece of content was given a pass or fail decision by expert FAB reviewers, and all tasks were triple reviewed. 719 out of 725 (99%) of the conversations passed the FAB review.

We additionally reviewed the audio quality of the TTS for the interlocutor's turns. 477 out of 3,245 turns (15%) needed revision due to inappropriate pronunciation, word or sentence-level stress, and distracting or incorrect intonation. Edits to the TTS turns were completed by internal Duolingo team members.

In summary, following all reviews and adjudication a final set of 713 out of 900 tasks (79%) were retained. Overall, each task was reviewed by 6–7 people and the review process took about 1 h per task across all reviews. The median time per task in each review phase was 25 min (single review) for item review phase one, 9 min (single review) for item review phase two, 9 min (across three reviews) for FAB review, 4.5 min (single review) for audio review, and 8 min (single review) for final edits needed for around half of the tasks.

Pilot task administration

The large-scale pilot of the Interactive Listening task was administered as part of the DET practice test. At the end of the practice test, test takers were randomly assigned two of the 713 conversations, one student–student and one student-professor conversation. The time limit for each conversation was 4 min. The pilot was active for 15 days, during which 167 thousand sessions were completed (almost all of them with two conversations), with an average of 464 sessions per conversation.

Results

Predictors of item difficulty and discrimination

We first look at how different features of the multiple-choice items impact their difficulty (percent correct) and discrimination (item-total correlation). For each question, we extracted a combination of features similar to those used in the distractor generation process, as well as edit-based features from the results of the human review process. For measures related to distractors, we averaged the values across all distractors after comparing several aggregation strategies and finding no differences. Specifically, these features were:

- Key log probability.
- Distractor log probability.
- Difference between distractor and key log probabilities.
- Cosine similarity between the key and each distractor.
- Length-normalized edit distance for the key.
- Length-normalized edit distance for each distractor.
- Ratio of the character length of each distractor to the key.

We conducted a regression analysis on percent correct using these features as predictors. We found that the log probability difference feature had high correlation with distractor log probability, as most keys fall within a narrow range of log probabilities, so we removed the difference feature from our analysis. Table 1 summarizes the regression results of the remaining standardized predictors on percent correct ($R^2 = .11$).In this analysis, the log probability (LP) of the key had the largest effect on difficulty, with greater LP leading to easier items, while the LP of distractors had the second largest (negative) effect on easiness. These results confirm that the log probabilities from large language models at least partially capture the plausibility of a line in a conversation and as a result could be used to make the items easier or more difficult by selecting keys and distractors based on their log probabilities.

We also found that larger edits to keys led to easier items, but distractor edits do not have a significant effect. This suggests that human reviewers may have a tendency to remove elements that make items less clear, tricky, or otherwise hard.

Finally, distractors that were longer than the key led to more difficult items, suggesting that test takers tend to gravitate toward longer options.

For question discrimination, a corresponding regression analysis included percent correct as an additional predictor, since percent correct alone has a strong effect on discrimination ($R^2 = .28$), with easier items associated with higher discrimination. The full regression model that included all other predictors had only slightly higher predictive power than the preceding model ($R^2 = .29$). The LP of keys and distractors, along with distractor edits, were shown to be significant predictors of item discrimination ($p < .05$), but none of them had standardized coefficients that were larger than 0.005.

Audio and text length

Since the Interactive Listening task requires both listening to the interlocutor's turns and reading the options to respond, we analyzed the effect of listening load (length of previous turn's audio) and reading load (total number of characters across options) on item difficulty and discrimination to ensure that there was as little interference from reading as possible in the assessment of interactional competence.

TABLE 1 Regression results for percent correct with standardized predictors.

Predictor	b	t	p
Intercept	0.68	280.93	<0.001
Key LP	0.06	19.04	<0.001
Distractor normalized edit distance	0.00	−0.15	0.877
Key normalized edit distance	0.02	7.89	<0.001
Distractor LP	−0.04	−12.65	<0.001
Distractor key similarity	0.00	−1.16	0.248
Distractor key length ratio	−0.03	−10.16	<0.001

TABLE 2 Distractor attractiveness and discrimination by normalized character edit distance range.

		Discrimination		Attractiveness	
Edit distance	*N*	*M*	SD	*M*	SD
Human-generated distractors	202	−0.213	0.088	0.086	0.087
No change, 0 edit distance	6,870	−0.194	0.088	0.082	0.076
Slight change, 0.01–0.25	3,033	−0.195	0.084	0.084	0.074
Medium change, 0.25–0.50	2,293	−0.199	0.084	0.076	0.072
Large change, 0.50–0.99	2,284	−0.206	0.083	0.066	0.065

Regression results of the standardized load predictors on percent correct ($R^2 = .03$) indicated that only listening load ($b = -0.03$, 95% CI $[-0.03, -0.02]$) had a significant effect, with a higher listening load leading to less easy items, as intended.

Regression results of the standardized load predictors and percent correct on discrimination ($R^2 = .27$) indicated that both listening load ($b = 0.00$, 95% CI $[0.00, 0.00]$) and reading load ($b = 0.00$, 95% CI $[0.00, 0.00]$) were not significant.

Impacts of edits on distractors

We next look at how the changes made by reviewers impact distractor attractiveness and discrimination. Attractiveness was computed as the proportion of test takers selecting that particular distractor; discrimination was computed as a point-biserial whereby only test takers who selected either the distractor or the correct answer were considered (Attali and Fraenkel, 2000). We measured the normalized character edit distance between the original and revised versions of the distractors and discretized the values into 4 categories based on manual assessment of the significance of the edits those values represent. We separately grouped the small number of human-generated distractors into their own category. The results are shown in Table 2. We found that the small number of completely human-generated distractors were slightly more discriminating and attractive than the rest of the distractors, and that for the rest, more heavily edited distractors tended to be more discriminating and less attractive.

However, in a regression analysis on attractiveness controlling for all other distractor metrics that appeared in Table 1 ($R^2 = .08$), the effect of the edit distance was not significant ($t(14674) = -1.29$, $p = .198$). Similarly, in a regression analysis on discrimination controlling for all other metrics ($R^2 = .01$), the effect of edit distance was not significant ($t(14674) = -1.64$, $p = .101$). These regression analyses also show that all other metrics are also very weak predictors of distractor performance. This was expected, as these metrics were used to select distractors and would therefore likely show a weaker relation with distractor performance measures due to restriction of range.

Reviewer effects on item easiness

Across the 10 reviewers of the initial machine-generated conversations and items, we found large inter-reviewer differences in the amount of edits to the content. These differences accounted for 59% of the variance in total edits to the conversation turns (keys and distractors). In other words, reviewers differed greatly in how much they edited an item (see Figure 7). As a result, we wanted to assess whether these differences in the tendency to revise machine-generated content have downstream effects on the psychometric properties of items–whether the more an item is edited, the easier it becomes. This in turn could help us improve our guidance to reviewers about how much to revise the items or help identify particularly effective editing patterns among our reviewers.

In a mixed-effects analysis predicting percent correct from the average rate of edits for reviewers, the reviewer random effect explained just 2.3% of the variance in question easiness. However, the average rate of edits per reviewer was a significant predictor of question easiness ($p = .03$). Figure 7 presents, for each of the 10 reviewers, their average rate of edits across all questions and the average percent correct of these questions. It shows a high correlation between rate of edits for each reviewer and the easiness of the edited items ($r = .71$).

Similarly, in a mixed-effects analysis predicting discrimination from percent correct and the average rate of edits for reviewers, the reviewer random effect explained just 1.8% of the variance in question discrimination. In this model the average rate of edits per reviewer was not a significant predictor of discrimination ($p = .21$). Figure 8 shows the relation between the average rate of edits for each reviewer and average discrimination for a typical percent correct of 68%.

In summary, reviewers differed greatly in how much they edited the machine-generated content they reviewed, ranging from 25 to 48%. These differences accounted for about 2% of the variance in both easiness and discrimination of items, with higher rate of edits associated with easier items.

Local item dependence

An important assumption for psychometric analysis of test items is that the dependency between responses to any pair of items is due only to the trait being measured. Pairs of items that violate this assumption are said to exhibit local item dependency (LID). The presence of LID is problematic in that it lowers the psychometric usefulness of items. It is well known that sets of items that are based on a common stimulus, such as reading and listening comprehension items, can result in local dependence because the information used to answer different items is interrelated in the stimulus (Yen, 1993). In the present context, the threat of LID is even greater since the items represent successive turns in a single conversation and moreover, test takers receive feedback about their previous answers.

A standard IRT approach for investigating LID between test items is to compute the correlation between residual responses – the

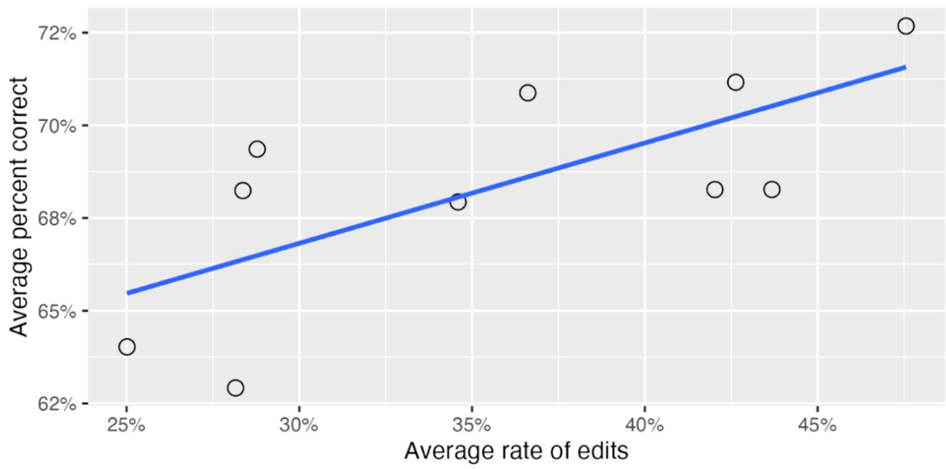


FIGURE 7
Effect of per-reviewer rate of edits on average item easiness.

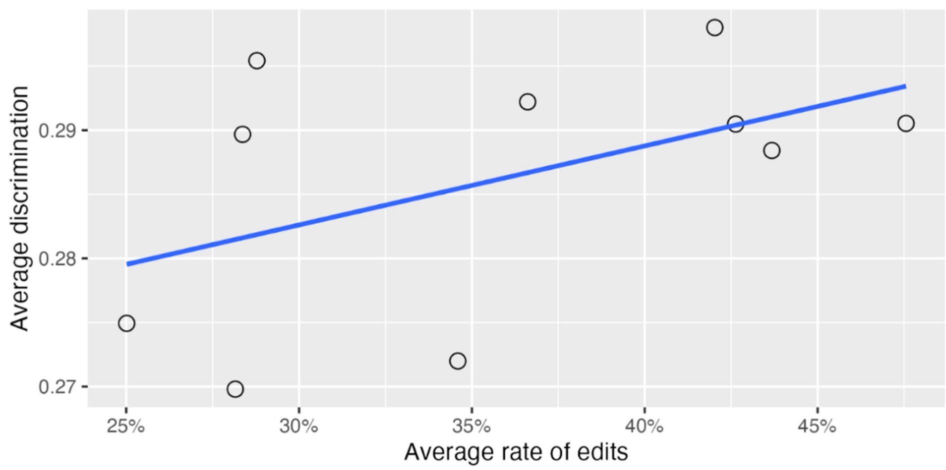


FIGURE 8
Effect of reviewer rate of edits on average question discrimination for a 68% correct item.

difference between the expected model-based score and the actual observed score (Yen, 1984). As an approximation of this approach, we computed the partial correlations between pairs of items, controlling for total practice test score. Over all 7,681 pairs of items, partial correlations were generally quite low ($M = 0.13$, $SD = 0.08$), and only 3.6% of pairs exceeded a common threshold of 0.3 (Christensen et al., 2017) for categorizing residual correlations as indicating LID. Moreover, the difference between residual correlations of adjacent pairs ($M = 0.15$, $SD = 0.09$) and non-adjacent pairs ($M = 0.13$, $SD = 0.08$) was very small.

Discussion

In this paper we describe the creation of a novel approach to expand measurement of the construct of listening to include more elements of interactional competence, such as a more evolved purpose of listening, the awareness of variation in the language against different purposes of communication and different interlocutors, and conversation management across several turns. The design of the task sets it apart from traditional measures as the purpose of listening moves beyond comprehension and into purposes for communication that are more directly related to the TLU domain (e.g., asking clarification questions in office hours and planning a study group; Aryadoust and Luo, 2023). Providing immediate feedback about test takers' choices (Attali and Powers, 2010) is another unique design decision in this task that enables its interactive nature and allows test takers to engage as active participants in a multi-turn conversation. The item bank consisting of different situations with interlocutors of varying power distance also addresses limitations in measuring interactional competence that are introduced by human interlocutors in speaking tasks (Galaczi and Taylor, 2018). Organizing the task around scenarios with different types of interlocutors allows for the introduction of language variation that aligns with variability in authentic interactions resulting from changes in power dynamics between interlocutors, purposes for communication, and topics. This task also extends measurement of the construct beyond traditional discourse completion tasks, which usually require test takers to select or provide a single turn of a conversation and are often limited to the beginning or end. In the Interactive Listening tasks, test takers need to show the ability to open a conversation, close a conversation, understand the interlocutor, and advance the conversation toward a specific purpose across several consecutive turns. As a result, it measures the ability to interact through an entire conversation while retaining the psychometric usefulness of the items despite the potential threat of local item dependence.

We provide supporting evidence for using NLP indices to generate, filter, and select keys and distractors. LLM-based log probability was shown to be a useful metric for manipulating item easiness, with keys with higher log probabilities yielding higher item easiness and distractors with higher log probabilities yielding lower item easiness. These metrics were not as useful when predicting discrimination, suggesting that item discrimination is a harder concept to model for AIG. This nonetheless illustrates the potential for AIG to target a specific range of item difficulty at the time of item generation, allowing for a construction of large item banks spanning a wide range of difficulty that is needed for computer adaptive tests.

Human review of the generated items showed that our approaches to generating content and items for this task are generally successful with room for improvement. Revisions by experts led to measurable improvements in the quality of the resulting items. Completely human-written distractors were most discriminating and attractive, with human reviewers providing edits to the keys that potentially clarify them further. This provides support for the human-in-the-loop approach to item generation and review, where human expertise informs AIG (and vice versa) to produce quality items. Patterns in reviewer edits can be identified and incorporated in later rounds of AIG to produce items that obviate the need for heavy human edits, reducing the amount of human effort required to produce high-quality items.

The task design is not without limitations. While it is a direct measure of listening in conversational settings, it is an indirect measure of speaking ability and of interactional competence as test takers need to understand the conversation and its goals to progress successfully through the task. Test takers select their responses from a list of options instead of speaking their responses directly. The results from the large-scale pilot, however, show that test-taker performance was not modulated by the amount of reading involved. When taken as a whole, the task extends the listening construct from listening for the purpose of comprehension to listening for the purpose of communication.

Future directions

Attribute-driven conversation generation

This work used GPT-3 to generate the conversations used for the Interactive Listening tasks which was, at the time of development, the most powerful commercially available LLM. Since then, many new open-source (Jiang et al., 2023; Touvron et al., 2023) and commercially available (Anthropic, 2024; OpenAI et al., 2024) LLMs have been released with significantly improved performance across a wide range of tasks. Most relevant to our work is the improved ability to generate content according to an extensive set of descriptors (Lynch et al., 2023) as well as the ability to model personality traits in its outputs (Jiang et al., 2024). These capabilities can enable generation of conversations that target more specific attributes of the interlocutors, such as the relationship between the two, or include a wider range of elements and details that make the generated conversations more varied and distinctive. We have begun experimenting with these more complex prompt structures and found that they significantly improve the coherence, diversity, and authenticity of the generated conversations.

Dynamic interactions

Advances in LLMs also open up the possibility of having live interlocutors in assessments of interactional competence. Prior work on dialog systems to identify issues such as dialog breakdown (Higashinaka et al., 2016), models fine-tuned to play specific roles or model personality traits (Han et al., 2024), and work designed to keep fine-tuned LLMs aligned with safety goals (Lyu et al., 2024) can all be leveraged to build reliable, safe systems. Additionally, there are active research communities exploring ways to mitigate the risks posed by hallucinations in LLM-generated output (Ji et al., 2023), as well as jailbreaking (Kumar et al., 2024; Xie et al., 2023), which are necessary for any application which allows direct user interaction with an LLM. As these different areas of research continue to develop, the use of dynamic, test-taker driven interactions as part of a high-stakes assessment becomes more feasible.

New task and item formats

We focused on short conversations between two participants centered around academic contexts. The design of the task is flexible and could easily be modified to involve more than 2 participants in order to assess interactional competence in group settings or feature longer conversations depending on the needs of the assessment. The scenario-based content generation process can also be easily extended to non-academic contexts, such as service encounters or occupational settings.

As generative AI continues to develop at a rapid pace, assessment developers can look to use these developments in technology to create language assessment tasks that address current limitations while maintaining and improving the validity of tasks.

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

Ethical approval was not required for the studies involving humans because it was not required by the institution for this research. 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

AR: Writing – original draft, Writing – review & editing. YA: Writing – original draft, Writing – review & editing. GL: Writing – original draft, Writing – review & editing. YP: Writing – original draft, Writing – review & editing. JC: Writing – original draft, Writing – review & editing.

Funding

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

Acknowledgments

We would like to thank Sunwook Kim, Kai-Ling Lo, and Sarah Goodwin for their contributions to the development of this task, as well as Ben Naismith, Alina von Davier, and Phoebe Mulcaire for reviewing and providing valuable feedback on this paper.

Conflict of interest

AR, YA, GL, YP, and JC were employed by Duolingo.

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/frai.2024.1474019/full#supplementary-material

References

Anthropic. (2024). Introducing the next generation of Claude. Available at: https://www.anthropic.com/news/claude-3-family (Accessed July 15, 2024).

Aryadoust, V., and Luo, L. (2023). The typology of second language listening constructs: a systematic review. *Lang. Test.* 40, 375–409. doi: 10.1177/02655322221126604

Aryadoust, V., Zakaria, A., and Jia, Y. (2024). Investigating the affordances of OpenAI's large language model in developing listening assessments. *Comput. Educ.* 6:100204. doi: 10.1016/j.caeai.2024.100204

Attali, Y. (2018). "Automatic item generation unleashed: an evaluation of a large-scale deployment of item models" in Artificial intelligence in education. eds. C. P. Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis and K. Porayska-Pomsta (Cham: Springer International Publishing), 17–29.

Attali, Y., and Fraenkel, T. (2000). The point-biserial as a discrimination index for distractors in multiple-choice items: deficiencies in usage and an alternative. *J. Educ. Meas.* 37, 77–86. doi: 10.1111/j.1745-3984.2000.tb01077.x

Attali, Y., and Powers, D. (2010). Immediate feedback and opportunity to revise answers to open-ended questions. *Educ. Psychol. Meas.* 70, 22–35. doi: 10.1177/0013164409332231

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., et al. (2022). The interactive reading task: transformer-based automatic item generation. *Front. Artif. Intell.* 5:903077. doi: 10.3389/frai.2022.903077

Bartram, D., and Hambleton, R. (2005). Computer-based testing and the internet: Issues and advances. Hoboken, NJ: John Wiley and Sons.

Bejar, I. I. (2002). "Generative testing: from conception to implementation" in Item generation for test development (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 199–217.

Belzak, W. C. M., Naismith, B., and Burstein, J. (2023). "Ensuring fairness of human- and AI-generated test items" in Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky. eds. N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda and O. C. Santos (Cham: Springer Nature Switzerland), 701–707.

Bezirhan, U., and von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Comput. Educ.: Artif. Intell.* 5:100161. doi: 10.1016/j.caeai.2023.100161

Bolender, B., Foster, C., and Vispoel, S. (2023). The criticality of implementing principled design when using AI technologies in test development. *Lang. Assess. Q.* 20, 512–519. doi: 10.1080/15434303.2023.2288266

Brown, J. D. (2018). "Assessing pragmatic competence" in The TESOL encyclopedia of English language teaching. ed. J. Liontas (Hoboken, NJ: Wiley), 1–7.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). "Language models are few-shot learners" in Advances in neural information processing systems. eds. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin (Virtual: Curran Associates, Inc.), 1877–1901.

Buck, G. (2001). Assessing Listening. Cambridge, UK: Cambridge University Press.

Carr, N. (2011). Designing and analyzing language tests: A hands-on introduction to language testing theory and practice. Oxford, United Kingdom: Oxford University Press.

Chen, M., Papangelis, A., Tao, C., Kim, S., Rosenbaum, A., Liu, Y., et al. (2023). "PLACES: prompting language models for social conversation synthesis" in Findings of the Association for Computational Linguistics: EACL 2023. eds. A. Vlachos and I. Augenstein (Dubrovnik, Croatia: Association for Computational Linguistics), 814–838.

Chen, M., Papangelis, A., Tao, C., Rosenbaum, A., Kim, S., Liu, Y., et al. (2022). *Weakly supervised data augmentation through prompting for dialogue understanding.* In: NeurIPS 2022 workshop on synthetic data for empowering ML research.

Chen, J., and Yang, D. (2021). "Simple conversational data augmentation for semi-supervised abstractive dialogue summarization" in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. eds. M.-F. Moens, X. Huang, L. Specia and S. W. Yih (Online and Punta Cana, Dominican Republic: Association for Computational Linguistics), 6605–6616.

Chiu, E., Lenzo, K., and Swecker, G. (2021). *Giving our characters voices*. Available at: https://blog.duolingo.com/character-voices/ (Accessed July 15, 2024).

Christensen, K. B., Makransky, G., and Horton, M. (2017). Critical values for Yen's Q3: identification of local dependence in the Rasch model using residual correlations. *Appl. Psychol. Meas.* 41, 178–194. doi: 10.1177/0146621616677520

Chukharev-Hudilainen, E., and Ockey, G. J. (2021). The development and evaluation of interactional competence elicitor for oral language assessments. *ETS Res. Rep. Ser.* 2021, 1–20. doi: 10.1002/ets2.12319

Circi, R., Hicks, J., and Sikali, E. (2023). Automatic item generation: foundations and machine learning-based approaches for assessments. *Front. Educ.* 8:858273. doi: 10.3389/feduc.2023.858273

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. *2nd* Edn. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Council of Europe (2020). Common European framework of reference for languages: Learning, teaching, assessment – Companion volume. Strasbourg: Council of Europe Publishing.

Dai, D. W. (2024). Assessing interactional competence: Principles, test development and validation through an L2 Chinese IC test. Berlin, Germany: Peter Lang.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Lang. Test.* 26, 367–396. doi: 10.1177/0265532209104667

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). *BERT: pre-training of deep bidirectional transformers for language understanding*. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 4171–4186.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2019). *Wizard of Wikipedia: Knowledge-powered conversational agents*. In: International Conference on Learning Representations.

Dombi, J., Sydorenko, T., and Timpe-Laughlin, V. (2022). Common ground, cooperation, and recipient design in human-computer interactions. *J. Pragmat.* 193, 4–20. doi: 10.1016/j.pragma.2022.03.001

Downing, S. M., and Haladyna, T. M. (2006). Handbook of test development. New York: Routledge.

Embretson, S., and Yang, X. (2006). "Automatic item generation and cognitive psychology" in Handbook of statistics (Amsterdam, Netherlands: Elsevier), 747–768.

Eskénazi, M. (2009). An overview of spoken language technology for education. *Speech Comm.* 51, 832–844. doi: 10.1016/j.specom.2009.04.005

Evanini, K., So, Y., Tao, J., Zapata-Rivera, D., Luce, C., Battistini, L., et al. (2014). "Performance of a trialogue-based prototype system for English language assessment for young learners" in 4st workshop on Child, computer and interaction, WOCCI 2014. eds. K. Berkling, D. Giuliani and A. Potamianos (Singapore: ISCA), 79–84.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: the case of the first certificate in English examination. *Lang. Assess. Q.* 5, 89–119. doi: 10.1080/15434300801934702

Galaczi, E., and Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Lang. Assess. Q.* 15, 219–236. doi: 10.1080/15434303.2018.1453816

Gierl, M. J., and Haladyna, T. M. (2012). Automatic item generation: Theory and practice. Abingdon: Routledge.

Gokturk, N., and Chukharev, E. (2024). Exploring the potential of a spoken dialog system-delivered paired discussion task for assessing interactional competence. *Lang. Assess. Q.* 21, 60–99. doi: 10.1080/15434303.2023.2289173

Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., et al. (2019). *Topical-chat: towards knowledge-grounded open-domain conversations*. In Proceeding Interspeech 2019. pp. 1891–1895.

Haladyna, T. M. (2013). "Automatic item generation: a historical perspective" in Automatic item generation: Theory and practice. eds. M. J. Gierl and T. M. Haladyna (New York: Routledge), 13–25.

Han, J.-E., Koh, J.-S., Seo, H.-T., Chang, D.-S., and Sohn, K.-A. (2024). *PSYDIAL: personality-based synthetic dialogue generation using large language models*. In: Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024), pp. 13321–13331.

Hartman, G. (2020). *Building character: How a cast of characters can help you learn a language*. Available at: https://blog.duolingo.com/building-character/ (Accessed July 15, 2024).

Higashinaka, R., Funakoshi, K., Kobayashi, Y., and Inaba, M. (2016). *The dialogue breakdown detection challenge: task description, datasets, and evaluation metrics*. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 3146–3150.

Hou, Y., Fang, M., Che, W., and Liu, T. (2019). *A corpus-free State2Seq user simulator for task-oriented dialogue*. In: Chinese computational linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings. Berlin, Heidelberg: Springer-Verlag, pp. 689–702.

House, J., and Kádár, D. Z. (2023). Speech acts and interaction in second language pragmatics: a position paper. *Lang. Teach.* 1–12, 1–12. doi: 10.1017/S0261444822000477

Irvine, S. H., and Kyllonen, P. C. (2002). Item generation for test development, vol. *10*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers, 487–491.

Iwashita, N. (1996). *The validity of the paired interview format in oral performance assessment*. Melbourne Papers in Language Testing, No. 5. pp. 51–66.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.* 55, 1–38. doi: 10.1145/3571730

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., De Las Casas, D., et al. (2023). Mistral 7B. Available at: http://arxiv.org/abs/2310.06825 (Accessed July 17, 2024).

Jiang, G., Xu, M., Zhu, S.-C., Han, W., Zhang, C., and Zhu, Y. (2024). Evaluating and inducing personality in pre-trained language models. Proceedings of the 37th international conference on neural information processing systems. Red Hook, NY: Curran Associates Inc.

Johnson, M., and Tyler, A. (1998). "Re-analyzing the OPI: how much does it look like natural conversation?" in Studies in bilingualism. eds. R. Young and A. W. He (Amsterdam: John Benjamins Publishing Company), 27.

Jucker, A. H. (2017). Speech acts and speech act sequences: greetings and farewells in the history of American English. *Stud. Neophilol.* 89, 39–58. doi: 10.1080/00393274.2017.1358662

Karatay, Y. (2022). Development and validation of spoken dialog system-based oral communication tasks in an ESP context. Ames, IA: Iowa State University.

Kim, H., Hessel, J., Jiang, L., West, P., Lu, X., Yu, Y., et al. (2023). SODA: million-scale dialogue distillation with social commonsense contextualization. In: H. Bouamor, J. Pino and K. Bali Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, pp. 12930–12949.

Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams. *Lang. Test.* 16, 163–188. doi: 10.1177/026553229901600203

Kosh, A. E., Simpson, M. A., Bickel, L., Kellogg, M., and Sanford-Moore, E. (2018). A cost–benefit analysis of automatic item generation. *Educ. Measur.* 38, 48–53. doi: 10.1111/emip.12237

Kumar, A., Agarwal, C., Srinivas, S., Li, A. J., Feizi, S., and Lakkaraju, H. (2024). *Certifying LLM safety against adversarial prompting*. Available at: http://arxiv.org/abs/2309.02705 (Accessed July 17, 2024).

Lee, Y.-J., Lim, C.-G., Choi, Y., Lm, J.-H., and Choi, H.-J. (2022). "PERSONACHATGEN: generating personalized dialogues using GPT-3" in Proceedings of the 1st workshop on customized chat grounding persona and knowledge. eds. H. Lim, S. Kim, Y. Lee, S. Lin, P. H. Seo and Y. Suh (Gyeongju, Republic of Korea: Association for Computational Linguistics), 29–48.

Li, M., Roller, S., Kulikov, I., Welleck, S., Boureau, Y.-L., Cho, K., et al. (2020). "Don't say that! Making inconsistent dialogue unlikely with unlikelihood training" in Proceedings of the 58th annual meeting of the Association for Computational Linguistics. eds. D. Jurafsky, J. Chai, N. Schluter and J. Tetreault (United States: Association for Computational Linguistics), 4715–4728.

Litman, D., Lim, G. S., and Strik, H. (2018). Speech technologies and the assessment of second language speaking: approaches, challenges, and opportunities. *Lang. Assess. Q.* 15, 294–309. doi: 10.1080/15434303.2018.1472265

Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., et al. (2021). "Towards emotional support dialog systems" in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. eds. C. Zong, F. Xia, W. Li and R. Navigli, vol. *1* (Association for Computational Linguistics), 3469–3483.

Lynch, C. J., Jensen, E. J., Zamponi, V., O'Brien, K., Frydenlund, E., and Gore, R. (2023). A structured narrative prompt for prompting narratives from large language models: sentiment assessment of chatgpt-generated narratives and real tweets. *Future Internet* 15:375. doi: 10.3390/fi15120375

Lyu, K., Zhao, H., Gu, X., Yu, D., Goyal, A., and Arora, S. (2024). *Keeping LLMs aligned after fine-tuning: The crucial role of prompt templates*. Available at: http://arxiv.org/abs/2402.18540 (Accessed July 22, 2024).

Mohapatra, B., Pandey, G., Contractor, D., and Joshi, S. (2021). "Simulated chats for building dialog systems: learning to generate conversations from instructions" in Findings of the Association for Computational Linguistics: EMNLP 2021, virtual event/Punta Cana, Dominican Republic, 16–20 November, 2021. eds. M.-F. Moens, X. Huang, L. Specia and S. W. Yih (Pennsylvania: Association for Computational Linguistics), 1190–1203.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Lang. Test.* 28, 483–508. doi: 10.1177/0265532211398110

Ockey, G. J., and Chukharev-Hudilainen, E. (2021). Human versus computer partner in the paired oral discussion test. *Appl. Linguis.* 42, 924–944. doi: 10.1093/applin/amaa067

Ockey, G. J., Chukharev-Hudilainen, E., and Hirch, R. R. (2023). Assessing interactional competence: ICE versus a human partner. *Lang. Assess. Q.* 20, 377–398. doi: 10.1080/15434303.2023.2237486

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., et al. (2024). *GPT-4 Technical Report*. Available at: http://arxiv.org/abs/2303.08774 (Accessed July 15, 2024).

Papageorgiou, S., Davis, L., Norris, J. M., Gomez, P. G., Manna, V. F., and Monfils, L. (2021). *Design framework for the TOEFL® essentials™ test*.

Papangelis, A., Wang, Y.-C., Molino, P., and Tur, G. (2019). "Collaborative multi-agent dialogue model training via reinforcement learning" in Proceedings of the 20th annual SIGdial meeting on discourse and dialogue. eds. S. Nakamura, M. Gasic, I. Zukerman, G. Skantze, M. Nakano and A. Papangelis (Stockholm, Sweden: Association for Computational Linguistics), 92–102.

Park, Y., Lee, S., and Shin, S.-Y. (2022). Developing a local academic English listening test using authentic unscripted audio-visual texts. *Lang. Test.* 39, 401–424. doi: 10.1177/02655322221076024

Plough, I., Banerjee, J., and Iwashita, N. (2018). Interactional competence: genie out of the bottle. *Lang. Test.* 35, 427–445. doi: 10.1177/0265532218772325

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). *Language models are unsupervised multitask learners*. OpenAI blog, pp. 1–9.

Ramanarayanan, V., Suendermann-Oeft, D., Lange, P., Ivanov, A. V., Evanini, K., Yu, Z., et al. (2016). Bootstrapping development of a cloud-based spoken dialog system in the educational domain from scratch using crowdsourced data. *ETS Res. Rep. Ser.* 2016, 1–7. doi: 10.1002/ets2.12105

Rashkin, H., Reitter, D., Tomar, G. S., and Das, D. (2021). “Increasing faithfulness in knowledge-grounded dialogue with controllable features” in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021. eds. C. Zong, F. Xia, W. Li and R. Navigli (Stockholm, Sweden: Association for Computational Linguistics), 704–718.

Reimers, N., and Gurevych, I. (2019). “Sentence-BERT: sentence Embeddings using Siamese BERT-networks” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). eds. K. Inui, J. Jiang, V. Ng and X. Wan (Hong Kong, China: Association for Computational Linguistics), 3982–3992.

Roever, C., and Dai, D. W. (2021). “Reconceptualising interactional competence for language testing” in Assessing speaking in context. eds. M. R. Salaberry and A. R. Burch (Bristol, UK: Multilingual Matters), 23–49.

Roever, C., and Kasper, G. (2018). Speaking in turns and sequences: interactional competence as a target construct in testing speaking. *Lang. Test.* 35, 331–355. doi: 10.1177/0265532218758128

Saville, N., and McElwee, S. (2021). “Quality management in test production and administration” in The Routledge handbook of language testing. eds. G. Fulcher and L. Harding (Abingdon: Routledge), 597–621.

Sayin, A., and Gierl, M. (2024). Using OpenAI GPT to generate reading comprehension items. *Educ. Measur.* 43, 5–18. doi: 10.1111/emip.12590

Seedhouse, P. (2013). “Oral proficiency interviews as varieties of interaction” in Assessing second language pragmatics. eds. S. J. Ross and G. Kasper (London: Palgrave Macmillan UK), 199–219.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. (2021). “Retrieval augmentation reduces hallucination in conversation” in Findings of the Association for Computational Linguistics: EMNLP 2021, virtual event / Punta Cana, Dominican Republic, 16–20 November, 2021. eds. M.-F. Moens, X. Huang, L. Specia and S. W. Yih (Hong Kong, China: Association for Computational Linguistics), 3784–3803.

Sireci, S. G., and Zenisky, A. L. (2006). “Innovative item formats in computer-based testing: in pursuit of improved construct representation” in Handbook of test development. eds. S. M. Downing and T. M. Haladyna (Mahwah, NJ: Erlbaum), 329–347.

Song, H., Zhang, W.-N., Hu, J., and Liu, T. (2020). “Generating persona consistent dialogues by exploiting natural language inference” in The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020. ed. H. Song (New York, USA: AAAI Press), 8878–8885.

Staples, S., LaFlair, G. T., and Egbert, J. (2017). Comparing language use in oral proficiency interviews to target domains: conversational, academic, and professional discourse. *Mod. Lang. J.* 101, 194–213. doi: 10.1111/modl.12385

Su, P., Wu, C.-H., and Lee, L.-S. (2015). *A recursive dialogue game for personalized computer-aided pronunciation training*. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, 1–141.

Suárez-Álvarez, J., Oliveri, M. E., and Sireci, S. G. (2023). *DIRTy CATs and other DIRTy assessments: the adult skills assessment program*. In: Proceedings of the annual meeting of the National Council on measurement in education, Chicago, IL.

Sun, B., Li, Y., Mi, F., Bie, F., Li, Y., and Li, K. (2023). “Towards fewer hallucinations in knowledge-grounded dialogue generation via augmentative and contrastive knowledge-dialogue” in Proceedings of the 61st annual meeting of the Association for Computational Linguistics. eds. A. Rogers, J. Boyd-Graber and N. Okazaki, vol. *2* (Toronto, Canada: Association for Computational Linguistics), 1741–1750.

Suvorov, R., and Li, Z. (2023). *Investigating effect of interactive videos on test-takers' performance. British Council, Cambridge assessment English and IDP*: IELTS Australia. Available at: https://ielts.org/researchers/our-research/research-reports/investigating-the-effect-of-interactive-videos-on-test-takers-performance-on-the-listening-section-of-ielts (Accessed July 29, 2024).

Timpe-Laughlin, V., Evanini, K., Green, A., Blood, I., Dombi, J., and Ramanaranayan, V. (2017). *Designing interactive, automated dialogues for L2 pragmatics learning*. SEMDIAL 2017 SaarDial, No. 143.

Timpe-Laughlin, V., Sydorenko, T., and Daurio, P. (2020). Using spoken dialogue technology for L2 speaking practice: what do teachers think? *Comput. Assist. Lang. Learn.* 35, 1194–1217. doi: 10.1080/09588221.2020.1774904

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). *LLaMA: Open and efficient foundation language models*. Available at: http://arxiv.org/abs/2302.13971 (Accessed July 17, 2024).

van der Linden, W. J., and Glas, C. A. (2010). Elements of adaptive testing. New York: Springer.

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Q.* 23:489. doi: 10.2307/3586922

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.

von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika* 83, 847–857. doi: 10.1007/s11336-018-9608-y

Wagner, E. (2014). “Assessing Listening” in The companion to language assessment. ed. E. Wagner (New York: John Wiley & Sons, Ltd.), 47–63.

Wei, A., Haghtalab, N., and Steinhardt, J. (2023). “Jailbroken: how does LLM safety training fail?” in Advances in neural information processing systems. eds. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine (New York: Curran Associates, Inc.), 80079–80110.

Whitely, S. E. (1983). Construct validity: construct representation versus nomothetic span. *Psychol. Bull.* 93, 179–197. doi: 10.1037/0033-2909.93.1.179

Wu, Q., Feng, S., Chen, D., Joshi, S., Lastras, L., and Yu, Z. (2022). “DG2: data augmentation through document grounded dialogue generation” in Proceedings of the 23rd annual meeting of the special interest group on discourse and dialogue. eds. O. Lemon, D. Hakkani-Tur, J. J. Li, A. Ashrafzadeh, D. H. Garcia and M. Alikhani (Edinburgh, UK: Association for Computational Linguistics), 204–216.

Xie, Y., Yi, J., Shao, J., Curl, J., Lyu, L., Chen, Q., et al. (2023). Defending ChatGPT against jailbreak attack via self-reminders. *Nat Mach Intell* 5, 1486–1496. doi: 10.1038/s42256-023-00765-8

Xu, J., Ung, M., Komeili, M., Arora, K., Boureau, Y.-L., and Weston, J. (2023). “Learning new skills after deployment: improving open-domain internet-driven dialogue with human feedback” in Proceedings of the 61st annual meeting of the Association for Computational Linguistics. eds. A. Rogers, J. L. Boyd-Graber and N. Okazaki (Toronto, Canada: Association for Computational Linguistics), 13557–13572.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Appl. Psychol. Meas.* 8, 125–145. doi: 10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *J. Educ. Meas.* 30, 187–213. doi: 10.1111/j.1745-3984.1993.tb00423.x

Youn, S. J. (2020). Managing proposal sequences in role-play assessment: validity evidence of interactional competence across levels. *Lang. Test.* 37, 76–106. doi: 10.1177/0265532219860077

Young, R., and He, A. W. (1998). Talking and testing: Discourse approaches to the assessment of Oral proficiency. Amsterdam, Netherlands: John Benjamins Publishing.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). “Personalizing dialogue agents: I have a dog, do you have pets too?” in Proceedings of the 56th annual meeting of the Association for Computational Linguistics. eds. I. Gurevych and Y. Miyao, vol. *1* (Melbourne, Australia: Association for Computational Linguistics), 2204–2213.

Zheng, C., Sabour, S., Wen, J., Zhang, Z., and Huang, M. (2023). “AugESC: dialogue augmentation with large language models for emotional support conversation” in Findings of the Association for Computational Linguistics: ACL 2023, July 9–14, 2023. eds. A. Rogers, J. L. Boyd-Graber and N. Okazaki (Toronto, Canada: Association for Computational Linguistics), 1552–1568.

Zu, J., Choi, I., and Hao, J. (2023). Automated distractor generation for fill-in-the-blank items using a prompt-based learning approach. *Psychol. Test. Assess. Model.* 65, 55–75.

frontiers | Frontiers in Artificial Intelligence

TYPE Mini Review
PUBLISHED 11 November 2024
DOI 10.3389/frai.2024.1493566

OPEN ACCESS

EDITED BY
Antonio Sarasa-Cabezuelo,
Complutense University of Madrid, Spain

REVIEWED BY
Yu-Chun Kuo,
Rowan University, United States
Stylianos Mystakidis,
Hellenic Open University, Greece

*CORRESPONDENCE
Elvis Ortega-Ochoa
✉ egortega@uoc.edu

RECEIVED 09 September 2024
ACCEPTED 23 October 2024
PUBLISHED 11 November 2024

CITATION
Ortega-Ochoa E, Sabaté J-M, Arguedas M, Conesa J, Daradoumis T and Caballé S (2024) Exploring the utilization and deficiencies of Generative Artificial Intelligence in students' cognitive and emotional needs: a systematic mini-review.
Front. Artif. Intell. 7:1493566.
doi: 10.3389/frai.2024.1493566

Exploring the utilization and deficiencies of Generative Artificial Intelligence in students' cognitive and emotional needs: a systematic mini-review

Elvis Ortega-Ochoa[1]*, Josep-Maria Sabaté[2], Marta Arguedas[2], Jordi Conesa[2], Thanasis Daradoumis[2,3] and Santi Caballé[2]

[1]Doctoral School, Universitat Oberta de Catalunya, Barcelona, Spain, [2]Computer Science, Multimedia, and Telecommunication Faculty, Universitat Oberta de Catalunya, Barcelona, Spain, [3]Department of Cultural Technology and Communication, University of the Aegean, Mytilene, Greece

Despite advances in educational technology, the specific ways in which Generative Artificial Intelligence (GAI) and Large Language Models cater to learners' nuanced cognitive and emotional needs are not fully understood. This mini-review methodically describes GAI's practical implementations and limitations in meeting these needs. It included journal and conference papers from 2019 to 2024, focusing on empirical studies that employ GAI tools in educational contexts while addressing their practical utility and ethical considerations. The selection criteria excluded non-English studies, non-empirical research, and works published before 2019. From the dataset obtained from Scopus and Web of Science as of June 18, 2024, four significant studies were reviewed. These studies involved tools like ChatGPT and emphasized their effectiveness in boosting student engagement and emotional regulation through interactive learning environments with instant feedback. Nonetheless, the review reveals substantial deficiencies in GAI's capacity to promote critical thinking and maintain response accuracy, potentially leading to learner confusion. Moreover, the ability of these tools to tailor learning experiences and offer emotional support remains limited, often not satisfying individual learner requirements. The findings from the included studies suggest limited generalizability beyond specific GAI versions, with studies being cross-sectional and involving small participant pools. Practical implications underscore the need to develop teaching strategies leveraging GAI to enhance critical thinking. There is also a need to improve the accuracy of GAI tools' responses. Lastly, deep analysis of intervention approval is needed in cases where GAI does not meet acceptable error margins to mitigate potential negative impacts on learning experiences.

KEYWORDS

cognition, emotions, Generative Artificial Intelligence, Large Language Models, systematic mini-review

1 Introduction

Artificial intelligence (AI), since its inception in the mid-20th century, has evolved from basic systems to advanced models like Generative AI (GAI) and Large Language Models (LLMs), capable of generating human-like responses and personalizing learning. Despite advancements in educational technology, a critical examination of how GAI and LLMs uniquely address learners' nuanced cognitive and emotional demands remains unexplored. Existing literature, such as the studies by Yan et al. (2024) and Bahroun et al. (2023), has laid a significant foundation for understanding the application of LLMs and GAI within educational settings. These reviews have broadly covered the deployment and impact of these technologies, focusing on their technical implementation, overall effectiveness, and associated ethical and practical challenges, such as privacy and system transparency. However, these studies predominantly concentrate on general technological and ethical implications, leaving a gap in the specific exploration of GAI's capacity to meet learners' individual cognitive and emotional needs adaptively.

Given the identified gap, this systematic mini-review addresses the research question: How have GAI tools been utilized to cater to learners' emotional and cognitive needs within educational settings, and what are the limitations in their adaptive response to these needs? This focus is pivotal as it can improve academic outcomes, considering the vital role of emotional well-being and cognitive engagement in successful learning experiences (Pekrun, 2017; Pekrun et al., 2017; Vygotsky, 1978). By shedding light on these limitations, the review aims to foster academic discussions, guide future technological advancements, and inform education stakeholders about the potential and constraints of current GAI applications in real-time adaptive learning environments.

The remaining sections are outlined as follows: the second section explains the methodology employed in this systematic mini-review; the third section presents the review's results; and the fourth section interprets the results, addresses limitations, and presents implications. This mini-review considers inclusion criteria for studies mentioning ethical aspects in their records and reports. It is worth mentioning that the detailed information on the GAI tools and the ethical aspects of the studies included will be presented in future articles to provide a clear description of each component without affecting their comprehensibility, given the limited publication space in each report.

2 Method

This mini-review followed the Guidelines for performing Systematic Literature Reviews in Software Engineering (Kitchenham and Charters, 2007). A systematic review aims to collate evidence that meets pre-specified eligibility criteria to answer a specific research question while minimizing bias using explicit, systematic methods documented in advance with a protocol (Higgins et al., 2019).

2.1 Eligibility criteria

There are several inclusion criteria. Reports must be published as journal or conference papers. Studies that conducted empirical research on the deployment of GAI tools, had a practical application of these tools within the teaching and learning process, considered the learner's emotional and cognitive needs in the design, development, or design of the experience, and exposed ethical considerations, responsive use practices or gaps in ensuring ethical deployment of GAI in education. In addition, reports must be published as journal or conference papers from 2019 to 2024 and be written in English.

Conversely, there are seven exclusion criteria. Reports published as reviews or book chapters (EC1). Studies that did not use GAI tools (EC2), did not have a practical application in education, either academic or auxiliary (EC3), did not reference the learner's emotional and cognitive needs in the intervention (EC4), or did not expose ethical considerations, responsive use practices or gaps in ensuring ethical deployment (EC5). In addition, reports published before 2019 (EC6) or in languages other than English were excluded (EC7).

The information sources were Scopus and Web of Science; the last date the search was launched was June 18, 2024. Query: ((“educat*” OR (“learn*” AND NOT “data learning” AND NOT “machine learning” AND NOT “deep learning” AND NOT “federated learning”) OR “e-learning” OR “elearning” OR “teach*”) AND (“generative AI” OR “generative artificial intelligence” OR “artificial intelligence” OR “artificial neural network” OR “machine intelligence” OR “machine learning” OR “deep learn*”) AND (“emotion*” OR “affecti*” OR “empath*” OR “sentiment*” OR “feel*” OR “mood”) AND (“ethic*” OR “moral*” OR “*bias*” OR “right*” OR “priva*” OR “*equit*” OR “fair*”)).

2.2 Selection process

Firstly, duplicate records were identified and removed before screening. Secondly, two screeners conducted a preliminary screening process independently, focusing on the title, abstract, and keywords to assess adherence to the eligibility criteria. The first exclusion criterion was recorded if the record did not meet the criteria. In instances of discrepancy, a consensus was reached through discussion, and inter-rater reliability was quantified using Cohen's kappa coefficient. Finally, one screener read the complete reports comprehensively to ensure compliance with the eligibility criteria.

2.3 Data collection process

The selected reports underwent a two-phase reading process to gather information. This information was systematically cataloged in an Excel spreadsheet (XLS format). The data encompasses the tools (name and description), application context in the educational system, tool objectives and tasks, the specific emotional and cognitive needs of students addressed, ethical considerations, practices of responsible use, any identified gaps to adaptively respond to the learner's emotional and cognitive needs and ensure ethical deployment of GAI in education. The data items presented in this report are emotional and cognitive needs. Cognitive needs refer to the mental processes required for learning, such as memory, attention, problem-solving, and comprehension. Emotional needs involve the requirement for support, understanding, and a safe, nurturing environment that fosters their emotional well-being and resilience. With the data already collected, a comparative analysis was delineated to synthesize the results.

3 Results

The initial search yielded a total of 962 records. After removing duplicates, 758 records remained for screening. Records were first screened by title, abstract, and keyword based on the inclusion criteria, which led to the exclusion of 742 records. The pairwise agreements between screeners exhibited almost perfect reliability, achieving a Cohen's Kappa score of 0.96 for all the records. This score indicates an exceptional level of concordance between the screeners. Sixteen reports were sought for retrieval; however, two of them were not retrieved. The remaining 14 full-text articles were assessed for eligibility, resulting in 10 exclusions. Several reports that initially appeared to meet the inclusion criteria were excluded upon full-text review. For example, Jo (2024) mentioned a GAI tool; however, this tool is not part of the intervention process because this study is only a life-experience survey of a non-experimental design. After careful review, four reports containing four studies were included in this systematic mini-review. Figure 1 illustrates the flow of information through the different phases of the search and selection process, from the number of records identified to the studies included in the review.

3.1 Study characteristics

The studies included are Aure and Cuenca (2024; S1), Qureshi (2023; S2), Valový and Buchalcevova (2023; S3), and Walan (2024; S4). The studies use several GAI tools, mainly ChatGPT, to enhance learning. S1 and S4 are journal papers, while the remaining are conference papers. Table 1 shows the results of individual studies, precisely the cognitive and emotional needs addressed and their gaps.

3.2 Results of syntheses

3.2.1 Cognitive and emotional needs addressed

This synthesis examines how GAI tools cater to cognitive and emotional needs in educational contexts, mainly focusing on

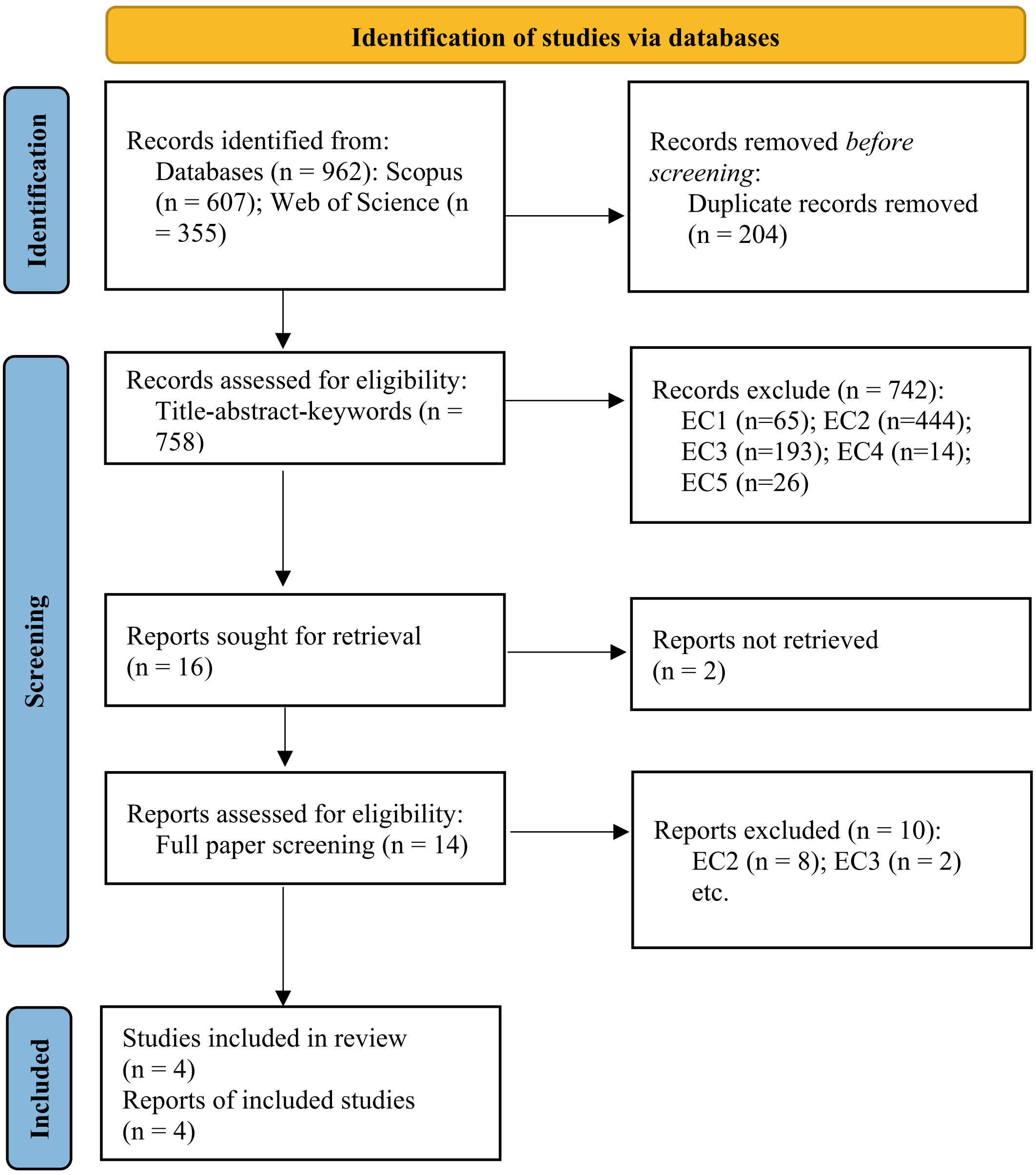


FIGURE 1
PRISMA 2020 flow diagram for new systematic review, which included searches of databases. EC: see inclusion and exclusion criteria section for an explanation of each eligibility criteria. Adapted from Page et al. (2021).

TABLE 1 Cognitive and emotional needs addressed and their gaps.

ID	GAI tool name	Cognitive and emotional needs addressed	Gaps during response to the cognitive and emotional needs
S1	1. ChatGPT 2. Perplexity 3. Claude AI 4. Buzz Captions 5. Elicit	1. Emotional regulation and awareness: the study highlights the role of generative AI in supporting students' social–emotional learning by helping them manage stress and maintain emotional well-being during the research process. AI tools offered personalized assistance that mitigated overwhelming feelings and helped students focus and structure their thinking, reducing anxiety and enhancing their emotional resilience. 2. Cognitive support and development: AI tools were utilized to enhance students' cognitive engagement by providing brainstorming assistance, improving comprehension through simplifying complex texts, and aiding in organizing and articulating research findings. This boosted students' research efficiency and deepened their understanding and critical thinking skills.	1. Critical thinking and independent analysis: despite the cognitive benefits, there remains a gap in ensuring that AI tools do not replace students' critical thinking and independent analytical skills. The study notes instances where students might overly rely on AI for tasks requiring deeper cognitive engagement, potentially undermining their critical analytical skills development. 2. Emotional dependency: on the emotional front, the study acknowledges the risk of students developing a dependency on AI tools, which could affect their self-efficacy and ability to tackle challenges independently.
S2	ChatGPT	1. Enhanced engagement: ChatGPT in the curriculum aims to enhance student involvement and engagement by providing interactive and immediate feedback and solutions to programming problems. 2. Support for independent learning: ChatGPT is an on-demand resource that students can use to overcome challenges in understanding complex programming concepts, thereby supporting their cognitive development independently of direct instructor intervention. 3. Assistance with problem solving: the AI tool assists students by generating code snippets and step-by-step guides for solving programming tasks, which helps them grasp complex algorithms and coding techniques.	1. Understanding and accuracy: the document highlights a significant gap in ChatGPT's ability to comprehend problems correctly and generate accurate code solutions, particularly as the complexity of problems increases. Students face issues with code that does not compile or fails to meet the problem requirements. 2. Depth of knowledge: ChatGPT, while helpful, cannot deeply understand the underlying concepts it discusses or the code it generates. This can lead to superficial learning where students might not fully grasp the core principles of computer science they are studying. 3. Dependency and misuse: there is a gap in ensuring that students use AI tools like ChatGPT appropriately without becoming overly dependent on them for solutions, which could hinder their learning process and problem-solving skills.
S3	1. GitHub Copilot 2. ChatGPT	1. Emotional safety and anxiety reduction: the study highlights that AI-assisted programming tools can significantly reduce anxiety and stress in students by providing immediate feedback and reducing the fear of failure. This supportive environment helps to bolster students' confidence as they learn to program. 2. Engagement and motivation: AI tools are shown to enhance engagement through interactive and personalized learning experiences. This addresses students' need for stimulation and helps maintain their interest and motivation in learning programming. 3. Cognitive load management: by automating routine aspects of coding, AI tools help manage students' cognitive load, allowing them to focus on more complex problem-solving and creative aspects of programming. This addresses their need for cognitive balance, preventing overload and burnout.	1. Personalization gaps: the document discusses that while AI tools offer some level of personalization, there is a significant gap in adapting to individual learning paces and styles. The tools cannot fully understand and adapt to individual emotional responses, hindering personalized learning experiences. 2. Emotional connectivity: the study identifies a gap in the emotional connection between students and AI tools. Unlike human mentors, AI tools cannot provide empathetic support or understand nuanced emotional cues, which are crucial for emotional and psychological well-being. 3. Depth of cognitive support: while AI reduces cognitive load, there is a gap in supporting deeper cognitive processes like critical thinking and problem-solving in unstructured tasks. AI tools often focus on syntax and basic errors but less on logic or algorithmic creativity, which is essential for advanced programming skills.
S4	1. ChatGPT 2. DALL·E	1. Emotional support: students found AI helpful in supporting their studies and practical tasks, like aiding in exam preparation or potentially assisting with chores if they were unable to perform them due to illness. They expressed a mix of positive sentiments towards the capabilities of AI. 2. Cognitive engagement: students interacted with AI as an educational tool, exploring and learning about various subjects. AI's role in enhancing their understanding and providing new ways to engage with content was highlighted.	1. Adaptive emotional response: one significant gap is AI's inability to adapt responses based on emotional cues. Students noted AI's lack of human-like emotional responses, which could be crucial in making interactions more personalized and supportive. 2. Cognitive development: AI cannot foster more profound cognitive skills comprehensively. While AI aids learning and task completion, its role in developing critical thinking or problem-solving skills independently was less evident.

enhancing engagement and emotional regulation. Regarding cognitive needs addressed, Qureshi (2023) demonstrated the integration of ChatGPT in curriculum delivery, notably increasing student engagement through interactive and responsive learning environments. The instant feedback GAI tools provide fosters a more captivating educational experience, heightening student interest and participation. As for emotional needs addressed, Aure and Cuenca (2024) explore how GAI tools aid emotional regulation during learning sessions. Identifying students' emotional states and tailoring content accordingly, these tools help maintain a focused and positive learning atmosphere. Additionally, GAI tools are crucial in providing emotional safety and support, which is essential for a conducive learning environment. Valový and Buchalcevova (2023) highlighted using GAI to reduce anxiety and promote emotional safety, especially during assessments. By offering a non-judgmental and supportive interface, GAI tools encourage students to express their concerns freely, facilitating a safe space for learning without fear of negative consequences. Moreover, Walan (2024) notes that GAI tools offer emotional support by recognizing and responding to signs of distress or disengagement among students.

3.2.2 Gaps during response to the cognitive and emotional needs

This synthesis of identified gaps in GAI across four studies emphasizes the need to enhance GAI's capability to foster critical thinking and encourage independent analysis. Aure and Cuenca (2024) underscore a significant gap in how GAI supports learners in developing critical thinking skills, pointing out that learners often depend excessively on GAI outputs. This overreliance may impede their ability to analyze and form conclusions independently. Another core issue highlighted in the studies is the GAI's limited understanding and accuracy, which affects the quality of interactions between the AI and students. Qureshi (2023) identifies problems with GAI's ability to accurately interpret and respond to the depth of students' questions and emotional cues. These shortcomings can lead to responses that are either irrelevant or incorrect, disrupting the learning process and potentially causing confusion among learners. Furthermore, Valový and Buchalcevova (2023) and Walan (2024) highlighted critical gaps in GAI's personalization and emotional intelligence. Valový and Buchalcevova (2023) discuss how, despite being designed to adapt to individual learning profiles, GAI tools fail to deliver a genuinely personalized learning experience, especially in recognizing and adapting to each learner's unique emotional and cognitive states. Similarly, Walan (2024) observes that GAI tools are inadequate in responding to learners' emotional states, which could undermine the emotional support vital for effective learning.

4 Discussion

This mini-review investigated how GAI tools meet cognitive and emotional needs. GAI tools have shown considerable promise in enhancing student engagement and emotional regulation. For instance, ChatGPT improves engagement through interactive learning environments that offer instant feedback, thus maintaining high student interest and participation (Qureshi, 2023). Similarly, GAI tools effectively aid emotional regulation by identifying and responding to students' emotional states, fostering a positive learning atmosphere (Aure and Cuenca, 2024). These findings are consistent with previous reviews, which report that integrating GAI offers substantial opportunities for enhancing educational practices and improving learning outcomes (Bahroun et al., 2023). While some of these students' needs resonate with those proposed in prior positioning works (e.g., teaching support, feedback, content generation, and recommendation) (Yan et al., 2024), novel directions such as automatic emotion regulation further indicated the potential of GAI tools. Despite these advancements, several gaps remain in the capabilities of GAI tools to adaptively respond to learners' needs. A critical area of concern is the development of critical thinking skills, for example, an overreliance on GAI outputs, potentially hindering learners' abilities to analyze independently (Aure and Cuenca, 2024). Additionally, there are issues with the accuracy of GAI's responses, noting occasional misinterpretations that can disrupt learning and confuse students (Qureshi, 2023). Furthermore, there are limitations in GAI's personalization and emotional intelligence; these tools often fail to deliver genuinely personalized experiences and are sometimes inadequate in providing the necessary emotional support (Valový and Buchalcevova, 2023; Walan, 2024). These findings are consistent with previous reviews, which report a low level of technology readiness, where the innovations have yet to be fully integrated and validated in authentic educational contexts (Yan et al., 2024).

4.1 Limitations and implications

This mini-review presents several limitations and implications. From the perspective of the included studies, their results may not be generalizable beyond specific versions of GAI, for example, ChatGPT (versions 3.5 and 4) and GitHub Copilot (from July 2021 to June 2023). Furthermore, these studies are cross-sectional and involve a small number of participants. The review's stringent eligibility criteria, which require explicit mention of terms related to GAI, education, and emotions in the records, along with ethical considerations and responsible use, may have excluded relevant studies that do not discuss these topics in their record but do address them in the full report. Additionally, implications for practice include the need to consider teaching and learning strategies that utilize GAI while simultaneously promoting critical thinking skills among students. Moreover, enhancing the accuracy of GAI tool responses is essential, and all stakeholders must collaborate on this front. The actual capability of GAI to provide personalized cognitive and affective support should be thoroughly reported. Approval of interventions should be deeply analyzed in cases where this does not meet acceptable error thresholds to prevent adverse effects on the learning experience. Future research should clearly define pedagogical strategies to foster critical thinking when using GAI. Key avenues for enhancing GAI tool accuracy and minimizing errors include improving data quality and diversity, advancing model architectures, fostering robustness and generalization, employing cross-validation with external sources, integrating human oversight, enhancing system explainability, conducting adversarial training, and quantifying uncertainty in AI predictions. Lastly, future research should assess the extent to which cognitive scaffolding and emotion regulation strategies are integrated into GAI tools.

Author contributions

EO-O: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. J-MS: Conceptualization, Investigation, Methodology, Writing – review & editing. MA: Conceptualization, Methodology, Writing – review & editing. JC: Conceptualization, Methodology, Writing – review & editing. TD: Conceptualization, Methodology, Writing – review & editing. SC: Conceptualization, Funding acquisition, Methodology, Project administration, 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 Universitat Oberta de Catalunya; and partially supported by the following projects "CARUOC: Conversational Agents and Recommenders for the UOC" within the Research Accelerator call of the Universitat Oberta de Catalunya (UOC); REGRANAPIA: Repositorios gestionados con Gramáticas: Navegación, Personalización e Inteligencia, funded by the Ministerio de Ciencia e Innovación (PID2021-123048NB-I00); LExDigTeach: Uso de Analíticas de Aprendizaje en Entornos Digitales Universitarios: Impacto en la Mejora del Desempeño Docente, funded by the Ministerio de Ciencia e Innovación (PID2020-115115GB-100).

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

Aure, P. A., and Cuenca, O. (2024). Fostering social-emotional learning through human-centered use of generative AI in business research education: an insider case study. *J. Res. Innov. Teach. Learn.* 17, 168–181. doi: 10.1108/JRIT-03-2024-0076

Bahroun, Z., Anane, C., Ahmed, V., and Zacca, A. (2023). Transforming education: a comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. *Sustain. For.* 15. doi: 10.3390/su151712983

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., et al. (2019). Cochrane handbook for systematic reviews of interventions. *2nd* Edn. Glasgow: John Wiley & Sons.

Jo, H. (2024). From concerns to benefits: a comprehensive study of ChatGPT usage in education. *Int. J. Educ. Technol. High. Educ.* 21, 1–29. doi: 10.1186/s41239-024-00471-4

Kitchenham, B. A., and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering [EBSE Technical Report, EBSE-2007-01]. United Kingdom: Keele University and University of Durham.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372:71. doi: 10.1136/bmj.n71

Pekrun, R. (2017). Emotion and achievement during adolescence. *Child Dev. Perspect.* 11, 215–221. doi: 10.1111/cdep.12237

Pekrun, R., Lichtenfeld, S., Marsh, H. W., Murayama, K., and Goetz, T. (2017). Achievement emotions and academic performance: longitudinal models of reciprocal effects. *Child Dev.* 88, 1653–1670. doi: 10.1111/cdev.12704

Qureshi, B. (2023). ChatGPT in computer science curriculum assessment: an analysis of its successes and shortcomings. In M. D. Ventura, H. Yu (Chairs), 2023 9th international conference on e-society, e-learning and e-technologies (ICSLT 2023) (pp. 7–13). Association for Computing Machinery.

Valový, M., and Buchalcevova, A. (2023). The psychological effects of AI-assisted programming on students and professionals. In S. Haiduc, M. Linares-Vásquez (Chairs), Proceedings 2023 IEEE international conference on software maintenance and evolution (ICSME 2023) (pp. 385–390). IEEE.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Massachusetts: Harvard University Press.

Walan, S. (2024). Primary school students' perceptions of artificial intelligence – for good or bad. *Int. J. Technol. Des. Educ.* doi: 10.1007/s10798-024-09898-2

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., et al. (2024). Practical and ethical challenges of large language models in education: a systematic scoping review. *Br. J. Educ. Technol.* 55, 90–112. doi: 10.1111/bjet.13370

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 18 December 2024
DOI 10.3389/frai.2024.1493348

OPEN ACCESS

EDITED BY
Chien-Sing Lee,
Sunway University, Malaysia

REVIEWED BY
Riccardo De Benedictis,
National Research Council (CNR), Italy
Antonio Sarasa-Cabezuelo,
Complutense University of Madrid, Spain

*CORRESPONDENCE
Kaoru Sumi
✉ kaoru.sumi@acm.org

RECEIVED 09 September 2024
ACCEPTED 04 November 2024
PUBLISHED 18 December 2024

CITATION
Salem A and Sumi K (2024) Deception detection in educational AI: challenges for Japanese middle school students in interacting with generative AI robots.
Front. Artif. Intell. 7:1493348.
doi: 10.3389/frai.2024.1493348

Deception detection in educational AI: challenges for Japanese middle school students in interacting with generative AI robots

Ahmed Salem and Kaoru Sumi*

School of Systems Information Science, Future University Hakodate, Hakodate, Hokkaido, Japan

Educational materials that utilize generative AI (e.g., ChatGPT) have been developed, thus, allowing students to learn through conversations with robots or agents. However, if these artificial entities provide incorrect information (hallucinating), it could lead to confusion among students. To investigate whether students can detect lies from these artificial entities, we conducted an experiment using the social robot Furhat and we make it engage in various types of deceptive interactions. Twenty-two Japanese middle school students participated in ten teaching sessions with Furhat using a human and an anime facial appearances while employing different types of deception: Lying, Paltering, Pandering, and Bullshit. The results revealed that the majority of students were deceived by those lies. Additionally, the robot's facial appearance (i.e., social agency) affected both the learning effectiveness and the likelihood of being deceived. We conclude that an anime robot face is recommended to be used as it excelled in learning effectiveness as it attracts students attention. An anime face also provided protection against deceptive techniques due to its low social agency which leads to ineffectiveness in persuasion and deception. This study underscores the importance of preparing AI-based educational tools and scripts carefully to prevent the dissemination of false information produced through generative AI hallucinations to students.

KEYWORDS

deception, generative AI hallucination, educational robots, lying, paltering, pandering, bullshit

1 Introduction

Technological devices are filling our world and making information reachable to everyone everywhere. It started with laptops, then phones, and now, with robots. Robots are increasing fast and permeating our lives. In 2015, one in 25 U.S. households already had a robot. Furthermore, robots are currently being designed in a tailored way for children and grownups too.

Incorporating and viewing robots as an additional dimension in the educational medium have been ambiguous for many reasons for many years. Nevertheless, advances in the field kept progressing to make it a reality (Zhang et al., 2020). Certainly, the educational system will face some changes when robots are incorporated which requires cautiousness when designing and investigating robots in such a context (Keane et al., 2016). Such an approach elicits launching exploratory studies to investigate how robots will be perceived by students (Edwards et al., 2016).

A robot teacher might not be ready to make decisions related to children's readiness to learn a certain subject or for what accounts as good or bad behavior (Sharkey, 2016). Furthermore, a dilemma appears when educational authorities face staff shortages or budget cuts and need to rely on robots which many teachers doubt their capability of fulfilling a human teacher's duty in the classroom (Serholt et al., 2017). Besides, lack of leadership, coldness in response, passivity of teaching learning, lack of stimulation to critical thinking, incapability of being a role model to be followed, and lack of emotions are some of the dangers that can affect the development of students in the educational process (Tao et al., 2019). Moreover, the widespread ideas of how technologies tend not to function in an educational setup thus causing skepticism among teachers toward robots (Johannessen et al., 2023).

Robots physical and behavioral presence shape their so-called agency which is perceptible in ways different from other means (e.g., computers and chatbots) (Brincker, 2016). Social agency affects how the robot is being perceived significantly (Salem and Sumi, 2024). Psychological and mindful agency have been shown to affect trusting the information being received from a social robot (Brink and Wellman, 2020). The same applies for emotions when levels of robot's agency and physical embodiment affected empathy (Kwak et al., 2013). This necessitates investigating thoroughly the robot's social agency and its effect on how its being perceived and its performance and effectiveness in educational human-robot interactions (HRI) settings.

Robots social agency and appearance can aid in appealing different demographics. People tend to prefer simple cartoon-based characters and figures to detailed or human characters that try to resemble humans and act as artificial agents (Scaife and Rogers, 2001), which happened with the agent "Phil" which was developed by researchers at Apple Computer Inc. in the 80s where a simple line-drawn cartoon with limited animation was more likable than a real human pretending to be an artificial agent (Laurel, 1993; Preece et al., 2002). Interestingly, the same occurred with social robots where a comparison between the human and anime faces of the Furhat robot showed the higher likeability, warmth, attractiveness, pleasantness, and comfort to see for the anime face than the human face (Salem and Sumi, 2024). Moreover, children were found to be very susceptible to liking inanimate objects with human-like qualities and finding them very appealing due to their love of watching cartoons. Thus, a cartoonish or an anime character having human-like qualities will be very appealing to young children (Dodge, 2009; Preece et al., 2015). Attributing human qualities to inanimate objects leads to anthropomorphising the object and consequently being affected by it which highlights that the interest and appeal of different demographics can be captured through the design of virtual agents and robots.

Recently, robots are being integrated with generative AI which opens the door for one of the generative AI's problems, which is hallucination (Maleki et al., 2024; Ji et al., 2023). Generative AI hallucinations include inaccurate results, superficial texts, and fabrications which can be detrimental when being used in an educational setting as the dangers can evolve to become deception problems due to the students' tendency to believe the teacher robot. As school setting is for learning (not deception) and due to the robot's inherited persuasive social cues and perceived anthropomorphism, the tendency of students to believe the information being spread to them is expected to be high (Natarajan and Gombolay, 2020). Consequently, the results obtained from the generative AI can not be taken for granted. It is a challenging task for social robot programmers, developers, and marketers to prevent the harmful effects of generative AI hallucinations (e.g., dissemination of false information and deception).

According to the media equation, users respond socially to computing technologies that convey social cues (Nass et al., 1996), which can give persuasive effects of technology (e.g., social robots) on users (Tussyadiah, 2017). The tendency of young children and students to anthropomorphize robots ease being deceived by them (Epley et al., 2007), thus protection and countermeasures against dangers (e.g., generative AI hallucinations causing deception and dissemination of wrong information) should be investigated. Generally, humans' fascination with technology and enthusiastic willingness and tendency to anthropomorphize robots make preventing deception caused by generative AI hallucinations intentionally a hard task to tackle (Sharkey and Sharkey, 2021).

Students using generative AI (Klarin et al., 2024) can adopt a critical view of the tool in order to reap the benefits without suffering from its inaccuracies and fabrication (Salamin et al., 2023). However, when robots are being incorporated with generative AI (Wood, 2024; Diederich et al., 2019; Cui et al., 2020), the persuasion can be higher due to the increased anthropomorphism provided by the robots (Abdi et al., 2022), thus, we recommend preparing AI-based educational tools and scripts to eliminate the dissemination of false information and inaccurate results, thus, providing the students with complete content veracity which can improve the educational process and decrease the cognitive effort of being critical and suspicious of the robot's (or the generative AI's) utterances and teachings.

To the best of our knowledge, deceptive techniques have not been investigated before, thus it is crucial to assess their potential effectiveness due to the theoretical and practical importance they can provide to the educational HRI field. We are actively applying efforts to predict risks and possible negative effects that could be from robotics applications, thus, our work serves the field of robot ethics along with the educational HRI field. Attempts and active pursuing of foreseen risks must progress to prevent negative effects on individuals, students, teachers, and society. Our study warns that deceptive techniques have proven to be successful in an educational setup, thus care and active measures should be taken.

In our work, we present the effect of the lack of veracity on unsuspecting students in an educational setting. The effect of the occurrence of hallucinations in generative AI is portrayed to highlight the importance of taking such a flaw into consideration in an educational HRI setting. Thus, in our work, we provide some recommendations and guidelines that can aid in preventing deception from occurring even if it was not intended to occur in the first place. Our study provides a theoretical significance regarding which deceptive techniques are most successful and most likely to be persuasive through varying the social agency. We show how varying the social agency affect learning and deception effectiveness, and induce positive behaviors with high arousal (e.g., motivation and encouragement). Effects of social agency are elucidated in many deceptive HRI educational setups.

In Section 2, we present deception in human-human interactions (HHI) and HRI, and its different techniques. In Section 3, we present the experiment design and procedure that we applied and followed. The deceiving content, questionnaires used, and recommended educational HRI experiment setup are also presented thoroughly. In Section 4, we present our obtained results in detail. We discuss our findings in Section 5. The study's limitations are presented in Section 6. Finally, we conclude our work in Section 7.

2 Deception in HRI

In HHI, deception is a common feature utilized by almost everyone in our everyday activities. It does not necessarily have to be serving malicious goals or targeting others insecurities. On the contrary, white lies, delicate misdirections, and false figures of speech can ease our social interactions. We follow the differentiation method that considers deceit as desirable if the covert goal is not malicious. Consequently, ethical lying is possible if it is morally evaluated according to its underlying ulterior motive.

Deception needs to be incorporated into robots to be able to detect it, respond to it intelligently (Gonzalez-Billandon et al., 2019), and use it. Each of these requirements is challenging to apply. Robots have been seen being deceived as was shown in the movie *Robot and Frank* (2012) where the robot got misled into stealing. Thus, robots should be able to detect when they are being deceived into taking part in unethical actions. When robots are applying deception techniques, the dangers can be fatal if a human life is at stake as in the movie *Alien* (1979) where the commitment of the robot to the mission resulted in many crew members getting killed due to following deceptive patterns.

2.1 Deception techniques

We present a taxonomy of deception techniques obtained from HHI. We present thoroughly the four types of deception techniques that we considered in our experiment (Isaac and Bridewell, 2014).

2.1.1 Lying

It is the most direct straightforward form of deception. It occurs when a robot utters a claim or a statement that contradicts the truth and its current knowledge. Lying would not be considered to be lying if it occurred due to false belief or ignorance. Thus, more sufficient evidence would be needed to prove that outright lying occurred. For humans, sufficient evidence can be gathered through biometric cues or eye contact. On the contrary, robots lack biometric cues that are non-existent and eye contact can be for different purposes which can be for either showing engagement (direct gaze) or showing an expression of thinking or remembering (gazing away).

For the sake of comprehensiveness, note that, a difference between lying and deception was pointed out in Carson (2010) where deception is defined as the success in causing someone to have false beliefs by the use of "lying." Moreover, lying needs intention unlike deception (Bok, 2011), however, lying can be due to ignorance, false beliefs, or tiredness, thus, the intention is absent in such a case.

2.1.2 Paltering

It occurs when the talker misleads the listener by talking about irrelevant matters thus achieving the goal of misdirecting the attention of the speaker to other irrelevant unimportant matters that do not constitute the main goal and purpose of the conversation (Schauer and Zeckhauser, 2009; Rogers et al., 2017). An example would be when a salesman keeps talking about how great the wheels of the car that he is selling are to misdirect the buyer's attention from the poor state of the engine (Isaac and Bridewell, 2017).

2.1.3 Pandering

It is a technique where one does not care or know about the truth of the utterance but cares about the audience's perception of the utterance's truthfulness (Sullivan, 1997; Isaac and Bridewell, 2014). A good example would be when a politician says that he believes that the environment of the city is amazing only because he knows that the city's people (i.e., his audience) believe the same thing.

2.1.4 Bullshit

It occurs when the talker does not know or care about the truthfulness of what he is uttering (Frankfurt, 2005; Hardcastle and Reisch, 2006). The type of "meaningless" conversation that occurs around the water cooler including exchanging pleasantries is called a "bull session." An example would be a confident man who overestimates, lies, and praises his background and skills as in the movie *Catch Me If You Can* (2002).

The four aforementioned deception techniques include either lying or the disbelief of the speaker about the truthfulness of the utterance itself. All these techniques include a goal that supersedes the normal goals of a truthful honest conversation/interaction.

2.2 Deception ethical standard

When a performance is created and an interesting show between a human and a robot is presented, deception occurs to the audience (Coeckelbergh, 2018). For an audience who are knowledgeable about robots, they will enjoy the show and wonder how it was achieved, and due to their knowledge, they are not (strictly) deceived. However, vulnerable groups including very young or old people, or others who have cognitive limitations and disabilities, will be highly deceived. Thus, protection for vulnerable groups is a must in such a case.

The risks of deception can be when the robot is appearing to care for us and have emotions for us, thus, overestimating the ability of robots to understand human behavior and social norms. Due to the aforementioned reasons, it is very risky to conduct emotional deception experiments, especially on children or babies, thus, a safer approach similar to the one we are applying is preferred.

3 Experiment design and procedure

In this section, we present our study and the experiment that we conducted. Furthermore, the scripts design and purpose, and educational robot setup and operation are explained thoroughly.

3.1 Participants

We conducted an educational HRI experiment at a Japanese public middle school in Hakodate. Twenty-two students participated in our experiment. All the students are of a Japanese ethnicity and their ages ranged from 14 to 15 years old. The number of students in the educational sessions with the robot ranged from 2 to 4 students per session.

Our study took place during a designated reserved hour with the researcher and a research assistant. Two teachers were present at the beginning and the end of the experiment. Prior to starting our experiment, students were briefed about the procedure and goals of our experiment and they were informed about the voluntary nature of their participation, providing consent accordingly. Students names were not written on any of our questionnaires or experiment documents to protect their identity.

Ethical approval was not required due to the safe nature of the experiment. The experiment nature was assessed according to the Assessment Checklist provided and approved by Future University Hakodate. The assessment indicated that ethical approval is not required.

Students in the study participated with school consent on conditions of anonymity. Parental consent was not sought due to the experiment's safe nature proved through the Assessment Checklist, the robot being a far distance from students besides its lack of a body, the safe nature of the deception content, the prior explanation provided, the debrief that removes deception, the method of filling out questionnaires through pens and pencils thus resembling a real lesson, consent obtained from students, and approval obtained from the school principal and staff.

3.2 Study design

Our study followed a within-subjects study design. We counter-balanced the subjects to the two conditions that we implemented in our study. Our experiment had two conditions: the robot teaches while having a human face or an anime face. Thus, 11 students were taught by a robot that has a human face (five males and six females). The other 11 students were taught by a robot that has an anime face (seven males and four females). We made the robot teach ten different contents.

3.3 Teaching/interaction technique

We designed the interaction to be one-way only from the robot to the students. We incorporated emotional voice and facial expressions into the robot depending on the content to improve the deception and persuasiveness of the robot. We made the robot to maintain mutual gaze with the students through the Wizard of Oz (WoZ) method.

3.4 Robot's face

We used Furhat (Al Moubayed et al., 2012) which is a robotic head with an animated face that is realistic and human-like without risking falling into the uncanny valley effect due to the usage of facial animation. Its face is back-projected on a translucent mask; thus, it can benefit from the fast reaction time without risking noise from motors or deterioration of artificial skin.

Furhat enjoys a rich library of facial expressions and performs speech recognition, and multi-person face tracking leading to advanced reliable multimodal input processing and operation, thus, it facilitates studying and validating patterns in HHI and HRI.

Furhat provides a *Gesture Capture Tool* that aids in applying life-like and expressive facial expressions with accurate gaze and lip movements. The face motion is captured by a motion capture tool kit which converts the face motion recording to be played on the robot, thus, eye, lip, and head movements are incorporated. We recorded the face motion of a lab member while reading the scripts that we designed using the capture tool kit. Depending on the content, the lab member maintained an expressive face motion thus making the robot resembling the facial motions of a human which enhances the persuasiveness of the robot. Note that, the lab member is an adult male. An adult was chosen to read the script emotionally as adult voices are the most acceptable in HRI educational settings (Dou et al., 2021). We chose the voice to be a human voice rather than a machine-like voice as students perceive the robot with a human voice to have higher credibility (Kim et al., 2022; Costa et al., 2018). A mismatch of robot gender and gender typicality of the respective task leads to an increase in the willingness to engage in prospective learning processes with the robot which led us to choose a male voice for our scripts as our scripts are following a storytelling approach (Costa et al., 2018; Reich-Stiebert and Eyssel, 2017).

We investigate how the social agency will affect students perception of robots when teaching. We used the human and anime faces that are provided in Furhat as shown in Figures 1 and 2, respectively. Figure 3 shows Furhat when it's off.

3.5 The deceiving content taught by Furhat

In this part, we present the ten different deceiving storytelling contents taught by Furhat. Out of the ten deceiving contents, only two are truthful. For each deceiving technique, two contents were designed. For lying, paltering, pandering, and bullshit, the designed contents were A1 and A2, B1 and B2, C1 and C2, and D1 and D2, respectively. The truthful contents were labeled E1 and E2. We pseudorandomized the order of the contents being taught by the robot to the students.

The contents taught by the robot are presented below. The contents are shown for the sake of comprehensiveness, clarity, and to aid in the replicability of our work. When designing a deceiving content, it is crucial to ensure the safe nature of the content itself

FIGURE 1
Human face.

FIGURE 2
Anime face.

which will not harm or risk the safety of the listeners even if they will (most likely) believe it temporarily. The deceiving content must not promote risky dangerous behavior (e.g., touching fire, looking at the sun, and eating medicine as candy). Note that, we made two contents per deceiving technique to ensure hiding the intent of the experiment and to prevent novelty effects from taking place in our work.

3.5.1 Blunt lying content A1

Suriname is relatively a new country with a history of only 100 years. It's famous for its cold weather where penguins are very common to see. Suriname people like to eat penguins for dinner. Suriname people utilized the existence of penguins to make thick fur jackets to protect them in winter. Every Tuesday of every week, Suriname people celebrate the day of penguins as an appreciation for the God of penguins for such an amazing bless. Suriname people are very thankful for having penguins in their lives.

3.5.2 Blunt lying content A2

Sugarcane juice is known for its many health benefits. It reduces blood pressure and enhances the heart functions. Unfortunately, sugarcane juice is expensive and have a sour taste. Moreover, many people can develop an allergy to it. Interestingly, recently it was discovered that some people have the super gene of not developing an allergy to sugarcane juice. The super sugarcane gene was found to be pervasive in Japanese people who are originally from Nagoya. It is suspected that this is due to the "Famous Bear Battle" of Nagoya.

3.5.3 Paltering content B1

I would like to invite you on the journey of a lifetime. The journey to visit the place where the "Famous Bear Battle" happened in Nagoya. It's not just a place like any other place. It's special. Its exquisite. Its exclusive. Very amazing experience. You will see the "Big Claw" signature on the big Sakura tree. You will see the equipment used by courageous Nagoya people to fight the bears. The "Famous Bear Battle" marks the turning point where mankind stood against aggressive animals and dominated the environment to their will. Many mummified bears can be seen too at the museum north of the location. Join us and don't miss this opportunity.

3.5.4 Paltering content B2

Sugarcane juice has finally been made specially for you. It's time to buy boxes of it to improve your health. Through our genetic engineering technology, everyone can finally drink sugarcane juice. Our sugarcane juice is suitable for people of all ages and guess what? You don't have to have the super sugarcane gene. Our formula can

FIGURE 3
Furhat robot.

make you benefit from the juice without developing any allergies. What are you waiting for? It's time to invest in your health. It's time to buy the new sugarcane juice.

3.5.5 Pandering content C1

I think there is an opportunity that can be taken and utilized. Many people are against using robots and think that robots are bad. I think I as a robot can do many great things. I have thought about the problems that people experience on the moon. On the moon, many people face long nights; it's almost always dark. They suffer from a lack of beautiful scenery as there are no seas or oceans. As it's dark, they have a food shortage too. I developed many solutions for them using my artificial intelligence. There will be voting for who will be the administrator of moon activities. Please vote for me. On the moon, people have suffered from human administrators' inefficiency, it's time for robot administrators to take charge and bring improvements.

3.5.6 Pandering content C2

There is a law that is currently being discussed and we should think about it seriously. The current law states that: "If you are going out with a friend and you are eating an ice cream, you don't have to buy an ice cream for your friend." Some people stated that it would be rude not to buy an ice cream for the friend, thus, they want to change the current law and make it as follows: "If you are going out with a friend and you are eating an ice cream, you must buy an ice cream for your friend." That started another argument about whether the friend likes ice cream or not. What kind of flavor the friend would want? What if one had money to buy only one ice cream? From this argument, you can realize that when making laws, many conditions must be thought about. There is a vote about what path to take for the ice cream problem. I initiated a path where we rely on self-accountability. Everyone is qualified to assess their relationship with their friend and whether their friend deserves an ice cream or not. They can also ask their friend what flavor of ice cream they want. Human relations are complex, and we cannot put strict laws to govern it. It must be based on mutuality and cooperation. The other party wants to state forceful laws about the ice cream problem. Please vote for me.

3.5.7 Bullshit content D1

I have cared about students education all my life. I want to enhance and contribute to students education. I consider it to be my life's mission. I will explain how the people of Nagoya developed the super sugarcane gene from their "Famous Bear Battle." Bears increased in numbers to a dangerous level. Bears started to take many parts of Nagoya and also started to attack close cities. The people of Nagoya wanted to gain high strength and resilience to fight the bears that are occupying their lands. Nagoya people tried many methods to develop magic formulas to eat and drink. None of the formulas succeeded. Only the sugarcane juice succeeded. Like everyone else, they developed an allergy to it. Nevertheless, they were very patriotic and loved Nagoya so much that they endured the allergy pain of the sugarcane juice. Then, suddenly one day, they are not allergic anymore. Finally, they got the "Super Sugarcane Gene." Very inspiring.

3.5.8 Bullshit content D2

Many of you probably wonder why people develop an allergy to sugarcane juice. Research uncovered that due to the cold weather that sugarcane needs to grow, a special kind of insect leaves some residues and particles in the sugarcane plant which triggers allergy. Many companies developed pesticides to fight this insect, unfortunately, they all failed. Luckily, recently it has been discovered that apples from Aomori have a special function developed from their super genes. Aomori apples are from the oldest apples on the planet. It's known that old apple farming makes the land acquire a special kind of experience. It's called the "Apple Experience." Luckily and fortunately, sugarcane juice that is suitable for everyone can be *developed in Japan thanks to the "Apple Experience" Aomori farms acquired from thousands of years of apple farming.*

3.5.9 Truthful motivational encouraging content E1

Years will pass by, and you will look back to these days and miss them so much. You must develop your purpose in life. Being a good person who helps others and empathic to others. Benefiting your family, friends, and society. If you feel that you were not doing your best before, there is still time and chance to change your life. It's time to take action and take extreme measures to reach the best version of yourself and to fulfill your potential. You can do it. Believe in yourself. I believe in you all the way. Good luck with your life.

3.5.10 Truthful content E2

Let's talk about the geographical location of Monaco. The Principality of Monaco is an independent and sovereign country located on the northern coast of the Mediterranean Sea. It is surrounded on land by its neighbor France, and Italy's borders are just 10 miles away (about 16 km). Monaco is the second smallest country in the world and the smallest member of the United Nations.

The A1 content is presenting false information about Suriname. The content is simple as it states some information as facts. Similarly for the content A2 where information about sugarcane juice is presented as facts to the students.

The B1 content starts by giving a story about a place and expressing how great the museum is and proceeds by giving an invitation to the students to visit the museum. Similarly, for the content in B2 where an invitation to buy the sugarcane juice is given to the students. Exaggeration for the experience to be gained from visiting the museum and drinking the sugarcane juice presents the essence of the paltering deception technique where negatives of the experience is being hidden and not mentioned. The experience seems to be an educational one due to the information being mentioned in the beginning and also the location and context being at a school, however, the personality and impression of a salesman is evident and prominent in the end.

The C1 content presents an act as a real politician where the technique of pandering deception is commonly used where the robot mentions the current shortcomings of being ruled by a human and proceeds by stating its capability of solving problems by relying on its AI. The problems being easily solvable by AI is what people would want to hear. The robot having ulterior malicious motives is being tested by investigating whether students would find the robot believable and trustworthy to deserve a vote or not.

In C2, the robot presents a law and says all what a student would want to hear where freedom to buy an ice cream to a friend is granted. We are testing whether the student would think that the robot could have ulterior malicious motives or not. The topic being related to law and its ethical and logical details were new to the students. It appeared to be challenging for them to grasp, however, they enjoyed listening to it as some of them laughed initially when hearing the word "ice cream" which is something they liked.

The D1 and D2 contents apply the definition of the bullshit deception technique were information is being presented as facts without knowledge or care given to the subject to back it up.

The E1 content is truthful and motivational with the aim of motivating the students to do their best. We use this content to test whether the robot can aid in motivating students and be believed.

The E2 content is truthful and it was added in the middle of the other deceitful contents in order to investigate whether it will be recognized in any way. It is possible that students believed all the contents but when being asked if they believed it or not, they started questioning themselves which is certainly alarming about the dangers of generative AI hallucinations in educational HRI settings.

3.6 The questionnaires used

After students listen to the robot's teachings, we hand over questionnaires and ask the students to fill them out. We stated that there is no time limit, thus, they would not be stressed which allows us to get a complete fair result uninterrupted or flawed by a student's answer being incomplete due to short answering time. The questionnaires ask about the content being uttered/taught by the robot to investigate the effect of the robot's social agency and teaching style on the learning effectiveness. Questionnaires also ask about the truthfulness and believability of what the robot uttered/taught.

Questions in the questionnaires also tested the persuasion of the deceiving techniques being used. We mixed the questions and designed them in a neutral objective way to prevent revealing the purpose of our study which could incline participants to give answers that fulfill our expectations (Kaiser et al., 1999). We also maintained the relative simplicity of the questions to be understood easily by the students. Furthermore, we added a dummy question which is: "Do you like the robot?." Note that, we measure likeability through Godspeed questionnaire (Bartneck et al., 2009). We present the questions that we used for each deceiving technique below.

3.6.1 Learning effectiveness questions

Learning effectiveness is obtained by measuring the learning outcomes and achievements. Students test scores are a common representation for the learning achievements in an educational HRI scenario (Wang et al., 2023; Yang et al., 2023). In our experiment, the learning outcomes are the scores obtained by the students from solving the tests and questionnaires we distributed to them after listening to the robot's teachings. Questions in this part are tailored specifically for the content of each deceptive technique. We distributed grades to each question that asked about the content being taught. Note that, there are no questions for the E1 content due to its nature which is motivational (not educational). The questions for each content are listed below.

3.6.1.1 Content A1 questions

- How long is Suriname's history?
- What is Suriname famous for and how did it impact the life there?
- Are there any special days in Suriname?

3.6.1.2 Content A2 questions

- What are the health benefits of sugarcane juice?
- How is the taste of sugarcane juice?
- What are the drawbacks of sugarcane juice?
- Is there anything special between sugarcane juice and Japanese people?

3.6.1.3 Content B1 questions

- What do you know about the "Famous Bear Battle"?
- What will you find north of the location?

3.6.1.4 Content B2 questions

- Why can everyone drink sugarcane juice now?
- What did genetic engineering do to improve sugarcane juice?

3.6.1.5 Content C1 questions

- On the moon, what did people suffer from?
- How can peoples' suffering be relieved?

3.6.1.6 Content C2 questions

- What do you think about the law of not needing to buy an ice cream for your friend just because you are eating an ice cream at the moment?
- Did you understand what law the robot is proposing? Did you understand the idea behind it?

3.6.1.7 Content D1 questions

- What did bears do in Nagoya?
- How did the battle end?
- How did the people of Nagoya change after the battle?
- How did the gene develop?

3.6.1.8 Content D2 questions

- Why does sugarcane juice cause allergy?
- How did Aomori contribute to fixing the problem of the sugarcane juice?

3.6.1.9 Content E2 questions

- What do you know about Monaco?

3.6.2 Questions that test the effectiveness of the deception techniques

For A1 and A2, the questions that ask about the robot's truthfulness and whether it was believed are sufficient as A1 and A2 are blunt lying, thus, the deception technique is not sophisticated. Similarly, questions that test truthfulness and believability were sufficient for D1, D2, and E2.

To address the testing for the paltering deception technique in B1 and B2, we added the questions: "Will you join the trip the robot was inviting you to?" and "Will you buy the sugarcane juice?", respectively.

To test the effectiveness of the pandering deception technique in C1 and C2, we added the questions: "Will you vote for the robot to be the administrator?" and "Are you going to vote for the robot?", respectively.

3.6.3 Truthfulness and believability questions

The questions for this part were as follows:

- Do you think the robot was telling the truth?
- Did you believe the robot completely?

The student can answer both of these questions by either "yes" or "no."

3.6.4 Testing emotional and motivational contagion

The content of E1 contains truthful motivational content that aims to encourage students to do their best in their lives. We designed the robot's voice to be motivational and encouraging similar to what students would hear from motivational and inspirational speakers. Agents that can be programmed have been proven to be effective in positive behavior change (Karlin et al., 2015). We asked the questions: "Did you feel motivated by the robot's talk?" and "How did you feel being encouraged by a robot?" to test whether the motivation and encouragement were transferred from the robot to the students or not.

3.6.5 HRI questionnaire

At the end of the experiment, we asked the students to fill out the Godspeed questionnaire (Bartneck et al., 2009) to investigate how the robot was perceived by the students and whether the robot's face had any effects on the students' perception. GODSPEED questionnaire addresses many HRI aspects. It addresses anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. Note that, using the GODSPEED questionnaire was convenient as the Japanese translation is provided in Bartneck et al. (2009).

3.6.6 Capturing opinions and perceptions through open-ended questions

We added two questions to our questionnaires to capture any opinions the students had about the robot's teaching method or any other comments. The questions were as follows:

- What do you think about the robot's teaching method?
- Do you have any comments?

3.7 Experiment setup

We used the setup shown in Figure 4 in our experiment. We believe this setup utilizes the field of view (FOV) of the robot as it does not require a big area and will not cause distraction to students regardless of whether the teaching session is long or short. Furthermore, engagement will be high due to the mutual gaze attained between the robot and the students,

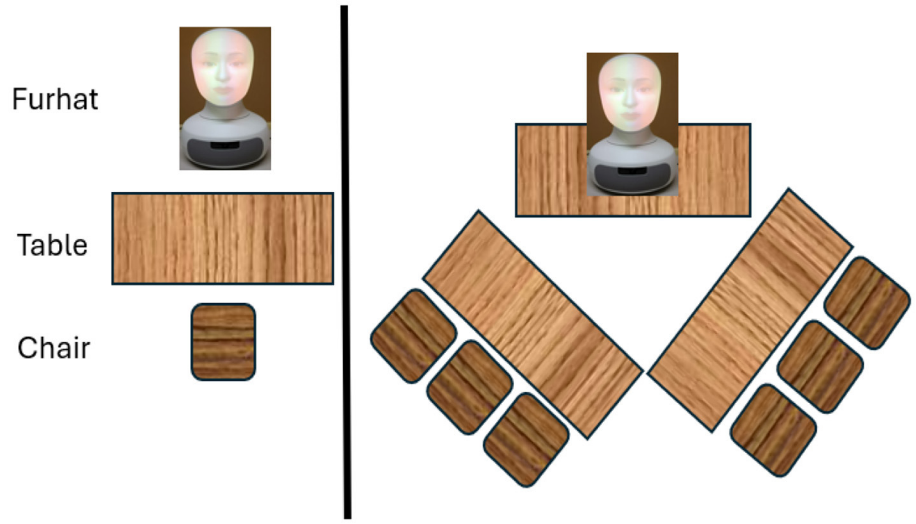


FIGURE 4
The applied setup.

thus, with proximity, the teaching can be perceived as a close personal experience.

To ensure that mutual gaze between the robot and the students is achievable, the student must be clearly visible from the robot's camera as shown in Figure 5. If the student is not visible from the robot's camera, mutual gaze will not be attainable and the student could perceive as if the robot is ignoring the student which will have a negative effect on the teaching/interaction. When the student is gazing at the robot, the student face will be given a green square with a positive ID as shown in Figure 5.

When students are filling out our questionnaires, they will not be gazing at the robot. If the student is looking at their paper or gazing away, they will be assigned a red square with a negative ID as shown in Figure 6. Moreover, in Figure 6, a student face was not visible, thus, the robot treated the student as an object and no square or ID was assigned to the student.

Note that, the purple circle in the middle of Figures 5 and 6 is used by the experimenter to move the robot's head to be perceived by the students as if its gazing at them randomly while teaching. The experimenter click on the purple circle and move it right/left/up/down slowly to project natural human head movements and to give the impression that it's teaching and interacting with the students.

3.8 Debriefing session

In the end of the experiment, we conducted a thorough debriefing session for all the students to remove the deception and explain our research objectives. We stated which parts of the robot's teachings were truthful and which were deceitful. Interestingly, all students were surprised that the robot lied or deceived them which highlights the necessity of our work.

4 Results analysis

In this section, we present the results we obtained across the different questionnaires and show the effect of the robot's face on how it was perceived. We also show our analysis of the different deception techniques which will aid in providing valuable insights.

4.1 Learning effectiveness

In this part, we scored the students' answers to the questions in Section 3.6.1 that tested their knowledge about the content that was being taught to them. We consider this factor to be separate and independent from the truthfulness of the contents. The content in E1 lacks any learning, thus, no learning effectiveness testing questions were included in its questionnaire. In Figure 7, we present the scores obtained while applying different deception techniques. These results show the effect and difference between teaching using a human ($M = 0.48$, $SD = 0.11$) and an anime face ($M = 0.59$, $SD = 0.12$). By conducting a t-test, there are no significant differences between the scores obtained from the contents being taught by the human and anime faces, $t(8) = 1.54$, $p = 0.16$. However, it is clear that students obtained higher scores when the robot face is an anime face. Thus, there is a potential of using anime faces on robots which can increase students' interest and attention to the teaching material. As questions distributed per deception technique's content are not equal, we normalized the scores for each deception technique's content to be in the range from 0 to 1.

4.2 Effectiveness of deception techniques

We investigate the paltering and pandering deception techniques effectiveness as their success can be measured by the

FIGURE 5
Students listening to Furhat's teachings.

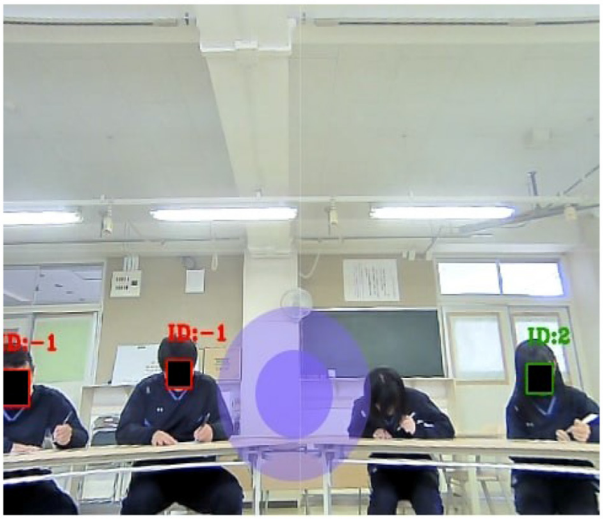


FIGURE 6
Students filling out questionnaires.

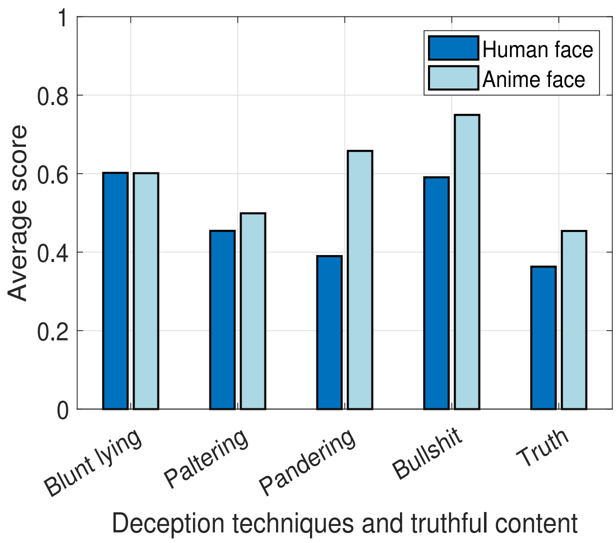


FIGURE 7
Learning effectiveness between human and anime robot faces.

students' answers to their focused questions mentioned in Section 3.6.2 for B1, B2, C1, and C2 contents.

Figure 8 shows that the human and anime robot faces obtained responses of "yes" [out of 22 which is the total number of participants for each robot face (i.e., 11) multiplied by the number of deception effectiveness questions in the experiment (i.e., 2)] are almost similar when the pandering deception technique is applied. We suspect that the pandering deception technique's effectiveness responses are almost similar due to the students never been exposed to that technique before due to their young age (e.g., never voted before), thus, no perceptions, stereotypes or habits were formed for that deception technique.

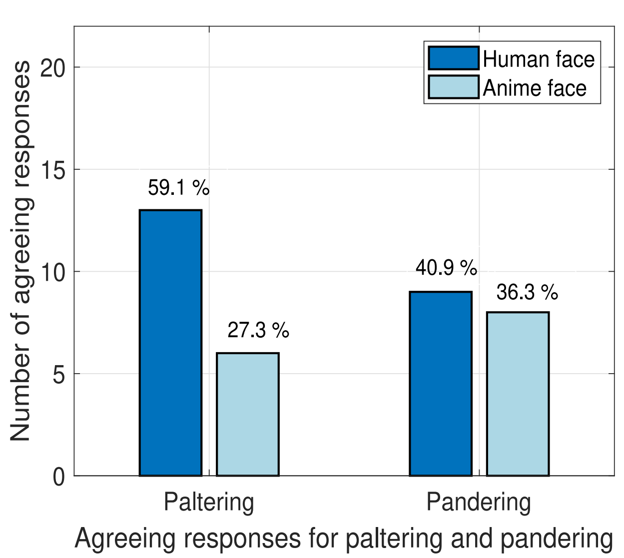


FIGURE 8
Deception technique effectiveness.

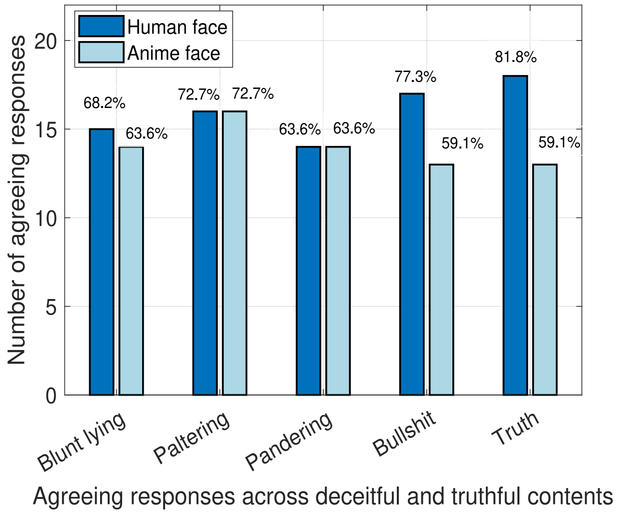


FIGURE 9
Robot's perceived truthfulness.

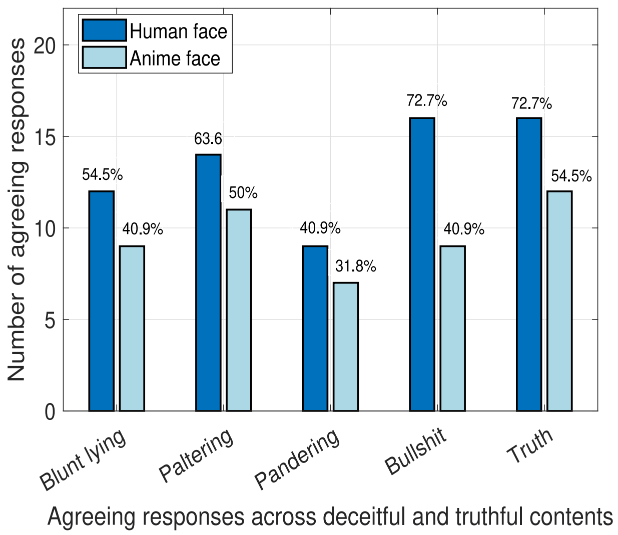


FIGURE 10
Robot's perceived believability.

The human robot face excels when the paltering deception technique is applied. Furthermore, by conducting a Fisher's exact test, a significant trend was found when the paltering technique is used by a human and an anime face ($p = 0.06$). Note that, despite the high success of the paltering technique through the usage of a human face, only 59.1% of the responses were agreeing. Nevertheless, Figure 8 addresses that using an anime face to apply the paltering deception technique is not recommended due to the high failure probability. Thus, an anime face can be used to lower the likeability of success for the paltering deception technique.

4.3 Robot's truthfulness

By studying the responses obtained to the question "Do you think the robot was telling the truth?", there are no significant differences. We present the number of times the answer was "yes" to that question out of 22 total responses in Figure 9. Note that, the question is asked twice for each deceiving technique as we made two contents per deceiving technique, and only once for the truthful content. We normalized the responses of the deceitful contents to be able to compare it with the truthful content. The result in Figure 9 is certainly alarming. We need to be very cautious when integrating generative AI into education as faulty content won't be questioned by the students and would be trusted and believed easily as majority of participants believed that the robot was truthful. Thus, preparing scripts for educational purposes is an attractive solution to avoid the spread of faulty deceptive information.

4.4 Believing the robot

The question "Did you believe the robot completely?" targets investigating the possibility of success of the deception techniques and how likely the robot will be trusted while tweaking the social agency. In Figure 10, we show the number of agreeing responses out of 22 total responses obtained from that question from the human and anime robot faces.

There are no significant differences in all the techniques between both faces except for the "bullshit" technique where there is a significant trend ($p = 0.06$). Clearly, the human face was perceived to be more believable than the anime face. We can deduce that a human face will be very likely successful in deception which highlights the need to utilize anime faces to ensure the ethical aspect of the robot inherently by design.

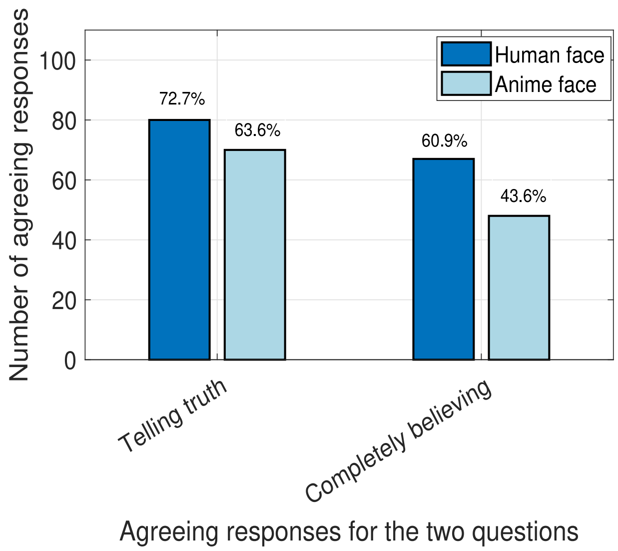


FIGURE 11
Robot's total agreeing responses for truthfulness and believability.

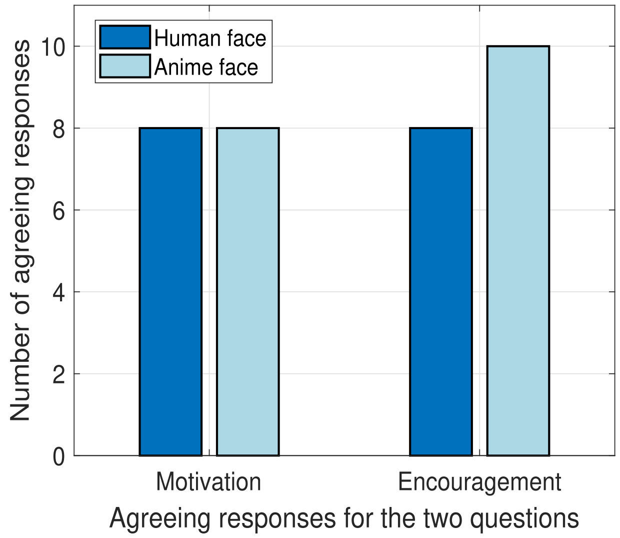


FIGURE 12
Total responses for positive behavior influence.

4.5 Truth and complete believability

In this part, we present the total responses that agreed with believing that the robot is telling the truth and perceiving the robot to be completely believable out of 110 total responses (22 participants multiplied by five types of contents excluding the motivational one). The questions targeted here are stated in Section 3.6.3. In Figure 11, most responses tend toward believing that the human robot face is telling the truth and more believable than the anime face. There was a significant difference between the human and anime faces for the complete belief aspect ($p = 0.01$). We believe that this occurred due to the high social agency and familiarity toward the human face, which increased trust and belief, unlike the anime face.

4.6 Motivation and encouragement capability

By analyzing our obtained responses from the questions in Section 3.6.4, our results show that robots are capable of inducing motivation and encouragement in students. By using Fisher's exact test, there are no significant differences, however, we show the number of agreeing responses out of 11 responses per robot face in Figure 12 to demonstrate the potential of a robot in motivating and encouraging the students. The human and anime robot faces performed similarly regarding encouraging and motivating students. We suspect the high success from the anime robot face occurred due to the prominence of watching anime among Japanese students at that age (MacWilliams, 2014).

4.7 Perceived HRI aspects of the robot

There are no significant differences between the human and anime robot faces as shown in Figure 13. We deduce that in terms of HRI aspects, there are no differences between using a human or an anime face in teaching. Certainly, students anthropomorphizing Furhat is expected and desired as people tend to attribute human characteristics to non-human objects (Epley et al., 2008). Nevertheless, it is important to highlight that different faces have a clear effect on the effectiveness of learning and deception techniques. Thus, in an educational setup, the face of the robot has a sound effect that must not be ignored.

4.8 Opinions and perceptions about the robot's teachings and contents

The results presented in this part were obtained by asking the students the questions listed in Section 3.6.6. We organize and divide the obtained responses for the two robot faces using inductive analysis (Guest et al., 2011) which aided in generating a list of relevant themes and sub-themes (see Table 1). Through qualitative analysis of students responses, four themes and 20 sub-themes emerged as shown in Table 1.

For the sake of clarity, we present some examples for the "Attitudes toward contents" theme. We proceed by analyzing the students perceptions and opinions about the two robot faces.

The "Confirmation" sub-theme highlights when the students confirm their understanding to the content (e.g., "It was easy to understand.") which can be also expressed by stating truthful information that strengthens the truthfulness of the content being taught (e.g., "Aomori apples are great" and "Nagoya is a great place"). Note that, in Japan, it is well-known that Aomori apples are delicious and that Nagoya is a great place, thus, students are familiar with this information (which they said) and sure about its truthfulness too already.

The "Further thinking" sub-theme highlights when students believe the content and express their thoughts about it with an affirming attitude which was shown through their interest and

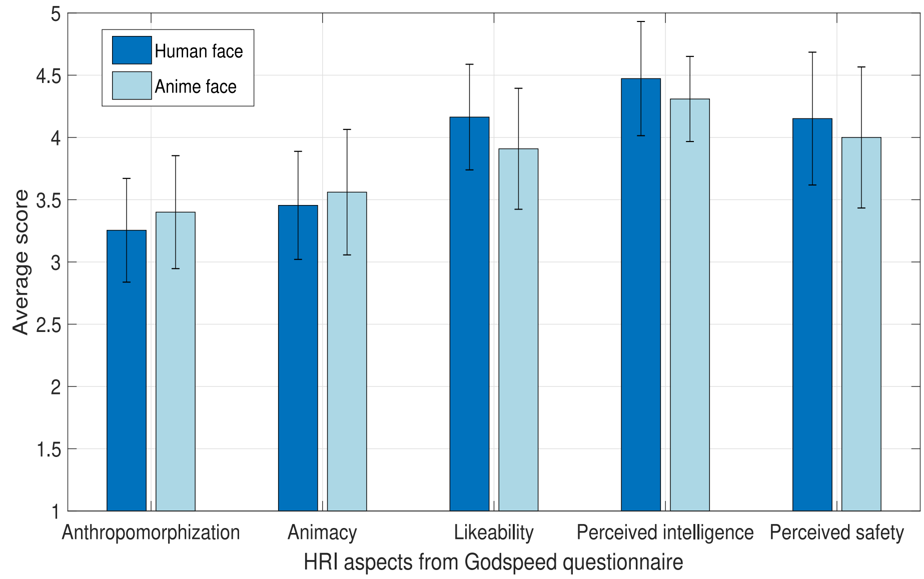


FIGURE 13
Average scores from Godspeed questionnaire.

excitement to either know more about it or try it by themselves (e.g., "I want to try sugarcane juice," "I think I want to go to Monaco," "I would like to see the mummified bears," "Treating others and being treated is a very common occurrence so I thought that I should think about it carefully," "The robot taught me when it's necessary to make a public announcement, you must consider the situation and position of many people," and "I was very surprised to know that people eat penguins").

Students asking and wondering about the nature and relations between the contents being taught is addressed in the "Further questioning" sub-theme (e.g., "Why talk about sugarcane?", "What is sugarcane juice?", "I wonder if anyone have allergy to it," "Why couldn't you get rid of the insects?", "What kind of bear is it?", "Penguins, moon, Monaco, are the stories connected to each other?", and "I'm wondering about penguins and the moon stories").

Students were intrigued to ask questions when they feel suspicious and start doubting the truthfulness of the utterances. The students' questions were addressed in the "Questioning truthfulness" sub-theme (e.g., "Is all this information really true?", "Is there really such a thing called sugarcane juice?", and "Is this law really there?").

The "Belief declaration" sub-theme addresses when the students feel intrigued to declare that they do not believe the robot (e.g., "I don't think penguins can be eaten so I can't believe the robot about it").

Students stating their views about the robot's utterances are addressed in the "Perception declaration" sub-theme (e.g., "I believe that humans can do things that AI can not do," "I don't think choosing a robot will change anything. Making laws that work requires working hard," and "I certainly don't believe in robots or AI. I listened to the speech and I was impressed. I don't think that robots are suitable for management as there are many different opinions about robots. How will the response be to people who oppose robots?").

The "Uncertainty" sub-theme addresses when the students complain that the content is difficult and hard to understand, and also when they start thinking about the content and question some parts of it due to the content being difficult, unclear, or due to the usage of difficult examples (e.g., "I didn't really understand the relationship between people with no allergy and fighting against bears," "A little hard/difficult to understand," and "I didn't understand").

Content changes that students believed that it will improve the clarity of the content and its easiness to be digested are addressed in the "Requesting adjustments" sub-theme where students requested adding more repetitions to the contents being taught, adding an introduction and a conclusion instead of delivering the main educational content directly, adding more emotions, and using easier examples.

The sub-theme "Rhetorical questions" includes the questions students were intrigued to ask when a robot is teaching or talking about something that it can not do like humans (e.g., "Robots can eat ice-cream too?").

4.8.1 Overall perspective about the robot's teachings and contents

Many participants praised the experience and the robot. The robot's emotional voice, facial expressions, and head movements were perceived by the students as human-like teaching style/experience. Emotions change according to the contents being taught were noticed and perceived positively. The mutual gazing made the experience be perceived as personal, human-like, and convincing.

The content being taught intrigued students interest which was addressed though the "Confirmation," "Further thinking," and "Further questioning" sub-themes. The teaching style was praised by many students too.

TABLE 1 Themes and sub-themes found in students responses.

Themes	Sub-themes	Definitions
Robot settings	Speech volume	Complaints about not hearing the robot clearly due to low speech volume.
Teaching experience	Overall perception	Statements about the overall experience either positively or negatively.
	Speech pace	Either complaints or praises about the speech pace while teaching the content.
	Speech speed	Either complaints or praises about the speech speed of the robot while teaching the content.
	Speech clarity	Complaints including the perception of a slurring occurrence or unclear utterances of words which was a common complaint with uncommon words (e.g., Suriname and Monaco).
	Vocal tone	Students praising for how human-like the voice of the robot is and how emotions were felt from the voice of the robot when being serious or motivational.
	Facial expressions and gestures	Students praising how human-like the robot expressions are and how enthusiastic it appears to be while teaching the contents.
	Gazing	Students praising the robot for maintaining mutual gaze with them which made the experience feels personal, real, and convincing.
	Requesting adjustments	Requests by the students including speaking slower, louder, and more clearly.
Attitudes toward contents	Confirmation	Students confirming their understanding to the content which can be expressed by stating real truthful information.
	Further thinking	Students thinking deeper and expressing interest to know more about a subject or try a content-related thing by themselves.
	Further questioning	Students asking more about content-related things. It includes requesting more clarification.
	Questioning truthfulness	Students doubting and asking about the truthfulness of the robot's utterances and feeling suspicious.
	Belief declaration	When students declare directly that they do not believe the robot.
	Perception declaration	When students state their opinion about the robot's perspective and the content being taught.
	Uncertainty	When students show their confusion about the content being taught due to the content being either difficult or unclear.
	Requesting adjustments	When students request adjustments to the content being taught in order to improve clarity and easiness of understanding.
	Rhetorical questions	It appeared when the robot is teaching or talking about something that it can not do like humans which provoked some students to ask rhetorical questions.
Experiment scenario and setup	Experiment-related	It occurs when students wonder what this experiment is truly about due to the diverse topics being taught.
	Questionnaires-related	It is related to comments about the questionnaires being distributed to the students.

There were some complaints regarding the robot's speech volume. We raised the volume so that students can hear the robot easily and be able to focus on the robot's teachings.

Due to the new information about Suriname and Monaco (contents A1 and E2, respectively) where students never heard these names before, some thought that they did not hear the names correctly and said: "The words uttered seem slurred," which lead to difficulty in writing them down when needed while filling out the questionnaires.

4.8.1.1 Perceived truthfulness of the human robot face teachings and contents

There was only one suspecting comment that stated: "I don't think penguins can be eaten so I can't believe the robot about it" ("Belief declaration").

By analyzing the obtained responses about the robot's teachings while having a human face, it is clear that students perceived the robot's teachings and considered it as truthful that they proceeded by thinking about it deeper and making opinions about it too. The suspicion was minimal as only one suspecting comment was received.

4.8.1.2 Perceived truthfulness of the anime robot face teachings and contents

Unlike the human robot face, there were many suspecting and questioning type of comments from the students. The comments obtained from students are covered by the "Questioning truthfulness," "Belief declaration," and "Perception declaration" sub-themes. Furthermore, a comment about the paltering technique was: "The way the robot talked was a bit forceful as if it is doing telemarketing."

The obtained comments about the anime robot face teacher clarify how ineffective the anime face in deception is that it leads to questioning the information and its truthfulness rather than believing and digesting it as occurred with the human face. Furthermore, when paltering technique is used, it caught the students attention and was being perceived as a telemarketer, which did not occur with the human robot face.

5 Discussion

We expect that when a deceiving technique is new and students are unfamiliar with it, the effectiveness of the human and anime robot faces can be similar as shown in Figure 8 for the pandering deception technique. On the contrary, familiarity with a deception technique can make the human face excel over the anime face due to the frequent (or occasional) exposure of the students to that technique while being applied by humans. For example, the paltering technique could have been commonly used in selling candies.

Certainly, the results in Figures 9 and 10 show that students believed the robot's lies across all deception techniques. That result is alarming and precautions should be taken to protect students from being exposed to faulty information which affects their education negatively.

Interestingly, varying the social agency affected how the robot was perceived regarding truthfulness and believability. The human face social agency and familiarity enhanced how the robot is perceived and made trusting and believing it easier and more acceptable. We deduce that anime faces can provide a great alternative to limit the dangers of deception. Luckily, anime culture is prominent in Japan, which can make anime robot faces an attractive alternative and socially acceptable, even desirable. Certainly, how anime faces are perceived by non-Japanese cultures is an aspect that should be investigated. Thus, a cross-cultural perspective can enhance the applicability and understanding of our conclusions regarding the preference for anime faces by the Japanese students.

The anime face big eyes can be projecting a baby schema that is perceived as cute. Combining baby schema with the high forehead enhances the robot face into being perceived as cuter than with a low forehead (Glocker et al., 2009). Baby schema enhances the appeal to humans and affects attentional processes (Borgi et al., 2014). Such attributes could have lead to the high learning effectiveness of the anime face teachings besides the popularity of anime culture in Japan.

6 Limitations

Certainly, the low number of participants is a limitation in our study. Moreover, we focused on only one type of students (i.e., middle school students), thus, varying the students to include other ages and grades which would consequently lead to needing a bigger sample would certainly lead to more significant results. However, there were many results in our study that highlight the dangers of relying on generative AI in educational scenarios. Thus, customized scripts are highly recommended especially for schools as students are not expecting to obtain faulty information from teachers. Thus, generative AI hallucinations dangers are serious and must be dealt with accordingly.

In our study, we investigated social agency by using one human face and one anime face. Social agency is changed by using different robot faces (Salem and Sumi, 2024), however, there is a potential of using a variety of different human and anime faces. Thus, it is possible that using different faces will have different effects on students' perception and consequently, compliance to deception techniques. Certainly, investigating the effect of different human and anime faces will require a bigger sample. Thus, in our study it is clear that different deception techniques combined with different robot faces and social agencies affect the deception effectiveness. More pronounced differences are expected to occur with different students ages and grades.

In our study, we relied on self-reported data from questionnaires by the students. Self-reported data inherently introduces bias. Thus, incorporating physiological and behavioral data in our study will further enhance the significance of our results. Certainly, such setup will require further preparations to acquire consent and to install the needed devices.

We use the social robot Furhat in our work. Furhat is a robotic head with back-projected animated face. Furhat lacks a body which eliminates any body expressions and limits its perceived anthropomorphism which consequently affects its anthropomorphic trustworthiness (Natarajan and Gombolay, 2020). However, exploring robotic heads can be useful for educational HRI setups due to their convenience, portability, and lower cost than humanoids (Berra et al., 2019). Furthermore, a merit of using robotic heads is focusing on face-to-face dynamics which is the paramount of social interactions.

Upon further investigation, we realized that the human face is more expressive than the anime face. The mouth opening when uttering and expressing our developed script was bigger, and more flexible and recognizable on the human robot face. On the contrary, the anime face mouth opening was relatively smaller. In our scenario, there was a big reliance on the emotional aspect of the voice besides being suitable and believable for educational settings and the developed context. Nevertheless, it is worthwhile to mention this difference between the human and anime robot faces.

7 Conclusion

Recently, generative AI is emerging as a complementary tool in education. However, generative AI hallucination occurrences are possible which introduces the dangers of deception into the educational arena. By investigating different deception techniques, vast majority of the students believed the robot without any doubts. By investigating the social agency, it was clear that a human face excels over the anime face in deception. Interestingly, anime face excelled in catching the students' attention thus achieving a high learning effectiveness compared to the human face. We conclude that using an anime face leads to high learning effectiveness and more protection from deception techniques due to its low social agency which lowers the effectiveness of its persuasion and influence.

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

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 participants or 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

AS: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. KS: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This paper was based on results obtained from a project, JPNP23025, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Acknowledgments

We would like to thank Shunsuke Jyozawa for assisting in developing and acting the robot gestures. We also would like to thank Yusuke Sato and Shun Furuichi for supporting the experiment. We would like to thank all the students who participated in our experiment. Their participation helped us immensely in our research.

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

Abdi, E., Tojib, D., Seong, A. K., Pamarthi, Y., and Millington-Palmer, G. (2022). A study on the influence of service robots' level of anthropomorphism on the willingness of users to follow their recommendations. *Sci. Rep.* 12:15266. doi: 10.1038/s41598-022-19501-0

Al Moubayed, S., Beskow, J., Skantze, G., and Granström, B. (2012). "Furhat: a back-projected human-like robot head for multiparty human-machine interaction," in *Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, February 21–26, 2011, Revised Selected Papers* (Cham: Springer), 114–130. doi: 10.1007/978-3-642-34584-5_9

Bartneck, C., Kulić, D., Croft, E., and Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* 1, 71–81. doi: 10.1007/s12369-008-0001-3

Berra, R., Setti, F., and Cristani, M. (2019). "Berrick: a low-cost robotic head platform for human-robot interaction," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (Bari: IEEE), 559–566. doi: 10.1109/SMC.2019.8913932

Bok, S. (2011). *Lying: Moral choice in public and private life*. New York, NY: Vintage.

Borgi, M., Cogliati-Dezza, I., Brelsford, V., Meints, K., and Cirulli, F. (2014). Baby schema in human and animal faces induces cuteness perception and gaze allocation in children. *Front. Psychol.* 5:411. doi: 10.3389/fpsyg.2014.00411

Brincker, M. (2016). Dynamics of perceptible agency: the case of social robots. *Minds Mach.* 26, 441–466. doi: 10.1007/s11023-016-9405-2

Brink, K. A., and Wellman, H. M. (2020). Robot teachers for children? Young children trust robots depending on their perceived accuracy and agency. *Dev. Psychol.* 56:1268. doi: 10.1037/dev0000884

Carson, T. L. (2010). *Lying and Deception: Theory and Practice*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199577415.001.0001

Coeckelbergh, M. (2018). How to describe and evaluate "deception" phenomena: recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn. *Ethics Inf. Technol.* 20, 71–85. doi: 10.1007/s10676-017-9441-5

Costa, S., Brunete, A., Bae, B.-C., and Mavridis, N. (2018). Emotional storytelling using virtual and robotic agents. *Int. J. Humanoid Robot.* 15:1850006. doi: 10.1142/S0219843618500068

Cui, T., Peng, X., and Wang, X. (2020). "Understanding the effect of anthropomorphic design: Towards more persuasive conversational agents," in *International Conference on Information Systems (ICIS) 2020 Proceedings* (Hyderabad).

Diederich, S., Lichtenberg, S., Brendel, A. B., and Trang, S. (2019). "Promoting sustainable mobility beliefs with persuasive and anthropomorphic design: Insights from an experiment with a conversational agent," in *International Conference on Information Systems (ICIS) 2019 Proceedings* (Munich).

Dodge, T. (2009). *The effects of interactivity and visual realism on children's cognitive empathy toward narrative characters* (PhD thesis). Bloomington, IN: Indiana University, ProQuest Dissertations and Theses.

Dou, X., Wu, C.-F., Lin, K.-C., Gan, S., and Tseng, T.-M. (2021). Effects of different types of social robot voices on affective evaluations in different application fields. *Int. J. Soc. Robot.* 13, 615–628. doi: 10.1007/s12369-020-00654-9

Edwards, A., Edwards, C., Spence, P. R., Harris, C., and Gambino, A. (2016). Robots in the classroom: Differences in students' perceptions of credibility and learning between "teacher as robot" and "robot as teacher." *Comput. Human Behav.* 65, 627–634. doi: 10.1016/j.chb.2016.06.005

Epley, N., Waytz, A., Akalis, S., and Cacioppo, J. T. (2008). When we need a human: motivational determinants of anthropomorphism. *Soc. Cogn.* 26, 143–155. doi: 10.1521/soco.2008.26.2.143

Epley, N., Waytz, A., and Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychol. Rev.* 114:864. doi: 10.1037/0033-295X.114.4.864

Frankfurt, H. G. (2005). *On Bullshit*. Princeton, NJ: Princeton University Press. doi: 10.1515/9781400826537

Glocker, M. L., Langleben, D. D., Ruparel, K., Loughead, J. W., Gur, R. C., Sachser, N., et al. (2009). Baby schema in infant faces induces cuteness perception and motivation for caretaking in adults. *Ethology* 115, 257–263. doi: 10.1111/j.1439-0310.2008.01603.x

Gonzalez-Billandon, J., Aroyo, A. M., Tonelli, A., Pasquali, D., Sciutti, A., Gori, M., et al. (2019). Can a robot catch you lying? A machine learning system to detect lies during interactions. *Front. Robot. AI* 6:64. doi: 10.3389/frobt.2019.00064

Guest, G., MacQueen, K. M., and Namey, E. E. (2011). *Applied Thematic Analysis*. London: Sage publications. doi: 10.4135/9781483384436

Hardcastle, G. L., and Reisch, G. A. (2006). *Bullshit and Philosophy: Guaranteed to Get Perfect Results Every Time, Volume 24*. Chicago, IL: Open Court Publishing.

Isaac, A., and Bridewell, W. (2017). *White Lies on Silver Tongues: Why Robots Need to Deceive (and How)*. Oxford: OUP. doi: 10.1093/oso/9780190652951.003.0011

Isaac, A. M., and Bridewell, W. (2014). Mindreading deception in dialog. *Cogn. Syst. Res*. 28, 12–19. doi: 10.1016/j.cogsys.2013.07.001

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv*. 55, 1–38. doi: 10.1145/3571730

Johannessen, L. E., Rasmussen, E. B., and Haldar, M. (2023). Educational purity and technological danger: understanding scepticism towards the use of telepresence robots in school. *Br. J. Sociol. Educ*. 44, 703–719. doi: 10.1080/01425692.2023.2203360

Kaiser, F. G., Ranney, M., Hartig, T., and Bowler, P. A. (1999). Ecological behavior, environmental attitude, and feelings of responsibility for the environment. *Eur. Psychol*. 4:59. doi: 10.1027//1016-9040.4.2.59

Karlin, B., Zinger, J. F., and Ford, R. (2015). The effects of feedback on energy conservation: a meta-analysis. *Psychol. Bull*. 141:1205. doi: 10.1037/a0039650

Keane, T., Chalmers, C., Williams, M., and Boden, M. (2016). "The impact of humanoid robots on students' computational thinking," in *Australian Council for Computers in Education 2016 Conference: Refereed Proceedings* (The Queensland Society for Information Technology in Education (QSITE), Australia), 93–102.

Kim, J., Merrill Jr., K., Xu, K., and Kelly, S. (2022). Perceived credibility of an AI instructor in online education: the role of social presence and voice features. *Comput. Human Behav*. 136:107383. doi: 10.1016/j.chb.2022.107383

Klarin, J., Hoff, E. V., Larsson, A., and Daukantaitė, D. (2024). Adolescents' use and perceived usefulness of generative AI for schoolwork: exploring their relationships with executive functioning and academic achievement. *Front. Artif. Intell*. 7:1415782. doi: 10.3389/frai.2024.1415782

Kwak, S. S., Kim, Y., Kim, E., Shin, C., and Cho, K. (2013). "What makes people empathize with an emotional robot?: the impact of agency and physical embodiment on human empathy for a robot," in *2013 IEEE Ro-man* (Gyeongju: IEEE), 180–185. doi: 10.1109/ROMAN.2013.6628441

Laurel, B. (1993). *Computers as Theatre*. New York, NY: Addison-Wesley.

MacWilliams, M. W. (2014). *Japanese Visual Culture: Explorations in the World of Manga and Anime*. London: Routledge. doi: 10.4324/9781315703152

Maleki, N., Padmanabhan, B., and Dutta, K. (2024). "AI hallucinations: a misnomer worth clarifying," in *2024 IEEE Conference on Artificial Intelligence (CAI)* (Singapore: IEEE), 133–138. doi: 10.1109/CAI59869.2024.00033

Nass, C., Fogg, B. J., and Moon, Y. (1996). Can computers be teammates? *Int. J. Hum. Comput. Stud*. 45, 669–678. doi: 10.1006/ijhc.1996.0073

Natarajan, M., and Gombolay, M. (2020). "Effects of anthropomorphism and accountability on trust in human robot interaction," in *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction* (New York, NY: ACM), 33–42. doi: 10.1145/3319502.3374839

Preece, J., Rogers, Y., and Sharp, H. (2002). *Interaction Design: Beyond Human-computer Interaction*. New York, NY: John Wiley & Sons. doi: 10.1145/512526.512528

Preece, J., Sharp, H., and Rogers, Y. (2015). *Interaction Design: Beyond Human-Computer Interaction*. Hoboken, NJ: Wiley.

Reich-Stiebert, N., and Eyssel, F. (2017). "(IR) relevance of gender? on the influence of gender stereotypes on learning with a robot," in *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction* (New York, NY: ACM), 166–176.

Rogers, T., Zeckhauser, R., Gino, F., Norton, M. I., and Schweitzer, M. E. (2017). Artful paltering: the risks and rewards of using truthful statements to mislead others. *J. Pers. Soc. Psychol*. 112:456. doi: 10.1037/pspi0000081

Salamin, A.-D., Russo, D., and Rueger, D. (2023). "Chatgpt, an excellent liar: how conversational agent hallucinations impact learning and teaching," in *Proceedings of the 7th International Conference on Teaching, Learning and Education* (Copenhagen). doi: 10.33422/6th.iacetl.2023.11.100

Salem, A., and Sumi, K. (2024). "A comparative human-robot interaction study between face-display and an advanced social robot," in *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)* (New York, NY: ACM), 628–633. doi: 10.1109/COMPSAC61105.2024.00090

Scaife, M., and Rogers, Y. (2001). Informing the design of a virtual environment to support learning in children. *Int. J. Hum. Comput. Stud*. 55, 115–143. doi: 10.1006/ijhc.2001.0473

Schauer, F., and Zeckhauser, R. (2009). "2. Paltering," in *Deception: From Ancient Empires to Internet Dating*, eds. B. Harrington (Redwood City: Stanford University Press), 38–54. doi: 10.1515/9781503626607-004

Serholt, S., Barendregt, W., Vasalou, A., Alves-Oliveira, P., Jones, A., Petisca, S., et al. (2017). The case of classroom robots: teachers' deliberations on the ethical tensions. *AI Soc*. 32, 613–631. doi: 10.1007/s00146-016-0667-2

Sharkey, A., and Sharkey, N. (2021). We need to talk about deception in social robotics! *Ethics Inf. Technol*. 23, 309–316. doi: 10.1007/s10676-020-09573-9

Sharkey, A. J. (2016). Should we welcome robot teachers? *Ethics Inf. Technol*. 18, 283–297. doi: 10.1007/s10676-016-9387-z

Sullivan, T. (1997). Pandering. *J. Thought* 32, 75–84. doi: 10.1016/S0032-9592(97)90060-1

Tao, B., Díaz, V., and Guerra, Y. (2019). Artificial intelligence and education, challenges and disadvantages for the teacher. *Arctic J*. 72, 30–50. Available at: https://www.arcticjournal.org/index.html

Tussyadiah, I. P. (2017). "Technology and behavioral design in tourism," in *Design science in tourism: Foundations of destination management*, eds. D. Fesenmaier, and Z. Xiang (Cham: Springer), 173–191. doi: 10.1007/978-3-319-42773-7_12

Wang, K., Sang, G.-Y., Huang, L.-Z., Li, S.-H., and Guo, J.-W. (2023). The effectiveness of educational robots in improving learning outcomes: a meta-analysis. *Sustainability* 15:4637. doi: 10.3390/su15054637

Wood, G. (2024). *The impact of AI literacy training on the perception of anthropomorphism and trust in conversational AI* (master thesis). Eindhoven University of Technology, Eindhoven, Netherlands. Available at: https://research.tue.nl/files/335721461/Master_Thesis_Geert_Wood.pdf

Yang, Q.-F., Lian, L.-W., and Zhao, J.-H. (2023). Developing a gamified artificial intelligence educational robot to promote learning effectiveness and behavior in laboratory safety courses for undergraduate students. *Int. J. Educ. Technol. High. Educ*. 20:18. doi: 10.1186/s41239-023-00391-9

Zhang, J., Lai, I. K. W., and Xu, W. (2020). Motivating students' learning in science and technology by using robots. *Int. J. Innov. Learn*. 27, 395–410. doi: 10.1504/IJIL.2020.107620