frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 16 October 2024
DOI 10.3389/feduc.2024.1427502

OPEN ACCESS

EDITED BY
Carina Soledad González González,
University of La Laguna, Spain

REVIEWED BY
Lyubka Aleksieva,
Sofia University, Bulgaria
Paula Miranda,
Instituto Politecnico de Setubal (IPS), Portugal

*CORRESPONDENCE
Roy Meissner
✉ roy.meissner@uni-leipzig.de
Alexander Pögelt
✉ alexander.poegelt@htwk-leipzig.de

†These authors have contributed equally to this work and share first authorship

RECEIVED 03 May 2024
ACCEPTED 18 September 2024
PUBLISHED 16 October 2024

CITATION
Meissner R, Pögelt A, Ihsberner K, Grüttmüller M, Tornack S, Thor A, Pengel N, Wollersheim H-W and Hardt W (2024) LLM-generated competence-based e-assessment items for higher education mathematics: methodology and evaluation. *Front. Educ.* 9:1427502. doi: 10.3389/feduc.2024.1427502

LLM-generated competence-based e-assessment items for higher education mathematics: methodology and evaluation

Roy Meissner[1*†], Alexander Pögelt[2*†], Katja Ihsberner[2], Martin Grüttmüller[2], Silvana Tornack[3], Andreas Thor[3], Norbert Pengel[1], Heinz-Werner Wollersheim[1] and Wolfram Hardt[4]

[1]Institute of Educational Sciences, Leipzig University, Leipzig, Germany, [2]Faculty of Computer Science and Media, Leipzig University of Applied Sciences (HTWK), Leipzig, Germany, [3]Faculty for Digital Transformation, Leipzig University of Applied Sciences (HTWK), Leipzig, Germany, [4]Department of Computer Engineering, Chemnitz University of Technology, Chemnitz, Germany

In this article, we explore the transformative impact of advanced, parameter-rich Large Language Models (LLMs) on the production of instructional materials in higher education, with a focus on the automated generation of both formative and summative assessments for learners in the field of mathematics. We introduce a novel LLM-driven process and application, called ItemForge, tailored specifically for the automatic generation of e-assessment items in mathematics. The approach is thoroughly aligned with the levels and hierarchy of cognitive learning objectives as developed by Anderson and Krathwohl, and takes specific mathematical concepts from the considered courses into consideration. The quality of the generated free-text items, along with their corresponding answers (sample solutions), as well as their appropriateness to the designated cognitive level and subject matter, were evaluated in a small-scale study. In this study, three mathematical experts reviewed a total of 240 generated items, providing a comprehensive analysis of their effectiveness and relevance. Our findings demonstrate that the tool is proficient in producing high-quality items that align with the chosen concepts and targeted cognitive levels, indicating its potential suitability for educational purposes. However, it was observed that the provided answers (sample solutions) occasionally exhibited inaccuracies or were not entirely complete, signalling a necessity for additional refinement of the tool's processes.

KEYWORDS

large language models, e-assessment, mathematical item generation, higher education mathematics, generative pretrained transformer, artificial intelligence, educational technologies, competence-orientation

1 Introduction

Recent advancements in Artifical Intelligence (AI) and Natural Language Processing have substantially impacted education and personalized learning (Kasneci et al., 2023), particularly through the use of LLMs to design and generate educational content. LLMs offer the potential to scalable generate tailored, high-quality, and challenging materials that align with modern pedagogical approaches, like constructive alignment (Biggs, 1996), responding to

diverse learning needs (Das et al., 2023; Laverghetta Jr and Licato, 2023; Kumar et al., 2023). In response to contemporary challenges such as a scarcity of specialized personnel, limited time, and general resource constraints, LLMs may support educators and respective institutions in developing versatile and adaptive content, like assessment materials, with the simultaneous potential to elevate their quality and enhance the educational experience (Kasneci et al., 2023). By providing tailored, pertinent, and challenging content, it is possible to foster the development of critical, scientific and entrepreneurial skills, essential for sustainable lifelong learning in a society. However, these technical advancements also present a dual-edged scenario, offering both opportunities and challenges regarding the content's quality and pertinence due to missing LLMs domain skills and their tendency to hallucinate[1] (Kasneci et al., 2023; Zhai and Nehm, 2023).

The present article examines the appraisal of LLM generated competency-oriented mathematical assessment items[2] for formative and summative purposes. The overarching purpose of this work is to estimate the limitation to which the technology is suitable for the precise creation of high-quality assessment items that correspond to the Intended Learning Outcomes (ILOs) of respective courses, and to which degree educators might be supportable or relievable in the process. Benefits are expected to be time savings and the precise generation of high-quality items that correspond to the ILOs of respective courses. The temporally equalised and individual generation of items (of sufficient quality) may also benefit learners, aligning with personalized learning situations, individual needs or based on desired outcomes, utilizable through adaptive learning systems (Sok and Heng, 2024; Zhai and Nehm, 2023).

From a procedural perspective, the generation of mathematical items through LLMs was predicated on a high-quality corpus of mathematical literature, encompassing higher-education textbooks, research articles, and academic lecture notes of a bachelor grade's university course. This corpus was systematically encoded into a vector representation and stored within a vector-database, used to direct the LLM prompting. Two LLMs were facilitated: GPT-3.5 for extracting relevant data pertaining to distinct mathematical concepts, which were thereafter transformed into comprehensible items by GPT-4. Thus, we built upon a Retrieval Augmented Generation approach, drawing on Multiple Agent and Chain of Thought prompting strategies, and implemented ItemForge, that utilizes the LLM capabilities in a directed and reliable way. Through this approach, item creation skills can be examined more precisely, which significantly increases result validity compared to frequently observed scenarios with ChatGPT (Zhai, 2023; Sok and Heng, 2024; Lee, 2024).

The evaluation of the generated items was conducted by experienced higher-education mathematicians. They audited 240 generated items against multiple criteria: alignment to the mathematical concept (or topic), correctness and completeness of item tasks, and validity and completeness of corresponding sample solutions. Items were also rated with levels from Anderson & Krathwohl's taxonomy of learning objectives (Anderson and Krathwohl, 2001), allowing to report on the LLMs capabilities to generate items targeting specific cognitive processes and knowledge levels. These criteria enable to delineate the strengths and weaknesses of the generated items and thus approach, thereby yielding constructive insights into LLM item generation skills, usable for progressive refinement of AI-driven educational tools.

The findings from this evaluation enhance the comprehension of harnessing the potential of LLMs within educational contexts for the generation of high-quality and targeted learning resources. Additionally, they open the opportunity for founded dialogues regarding the prospective incorporation of AI within educational institutions and the imperative need for balance between automated processes and human expertise, as well as shared effort in the development of educational content. Such insights are pivotal for elaborating the function of AI within educational context, utilizing its potential, reducing the resource load on educators and enhancing the contents quality.

This article begins with a characterization of our procedural approach for item generation and a description of the developed tool—ItemForge—in Section 2, before characterizing the executed study in Section 3. The outcomes thereof are discussed as of Section 3.2, including an expert reviews, before concluding on the subject with Section 5. Additionally, an overview of related work, covering recent advancements in the field, and a discussion on study limitations is given in Section 4.

2 Framework for automated generation of mathematics assessment items

The manual creation of high-quality mathematical problems and sample solutions is a laborious task that demands expertise from professionals in mathematical education. Generative AI tools, such as ChatGPT, have impressively shown potential in aiding the creation of mathematical problems, yet frequently yield poor or inappropriate results. Many studies demonstrated the potential to fine-tune LLMs, either through adjustments to learned parameters or via strategic prompting and focusing on contextual information, with impressive improvements in addressing the targeted issue. Thus, LLMs might be capable of generating high-quality mathematical items and the following section describes our prompting strategy and approach to contextual information, developed with GPT-4.

2.1 LLM-based mathematics item generator—ItemForge

A Python-based web application, named ItemForge, was developed to create competence-oriented mathematics items through an iterative process, utilizing mathematical concepts,

Abbreviations: AI, Artifical Intelligence; FAISS, Facebook AI Similarity Search; ILO, Intended Learning Outcome; LLM, Large Language Model.

1 The LLMs' ability to present fact-like statements containing false or misleading information.

2 A specific task, question, or activity designed not only to evaluate but also to provide formative feedback on an individual's knowledge, skills, or abilities in a particular subject or area of study.

retrieved concept summaries, retrieved ideal student knowledge, and an instructional description of the cognitive process- and knowledge dimensions (Anderson and Krathwohl, 2001)[3]. This tool serves as a resource for instructors and aids in generating learning materials and assessment items. ItemForge is built with the Streamlit[4] Python framework for its visual interface, the LangChain[5] framework for programmatic utilization of LLMs, and the Facebook AI Similarity Search (FAISS)[6] library.

To create items using ItemForge, users need to choose from a list of provided mathematical concepts (or topics), which were extracted from a bachelor grade's mathematical university course, called *Mathematics for Computer Scientists I—Fundamentals, Linear Algebra, Analysis, and Differential Calculus* (translated from German) the proof of concept was built for. Based on the chosen concept, a matching knowledge text is generated through information retrieval and summarization (see section 2.2 for an explanation on the available hyperparameters), based on extensive higher-education textbooks for the respective course. The retrieved knowledge text is used as a basis for concept-related item generation by the employed LLM and should be checked by the user to not contain any mistakes and to be relevant (see left half of Figure 1). After a successful review of the generated knowledge text, users can proceed to generate concept-related items either for a selected taxonomy level (a specific process- and knowledge level from these dimensions), or for all 24 taxonomy levels. Depending on the chosen and reviewed options (concept, knowledge text, single or all taxonomy levels), a LLM prompt is dynamically constructed and inherited by a LangChain chain, which is lastly issued against the chosen LLM to generate a set of one to 24 items, displayed to the user on completion.

In the instructional prompt, already containing the mentioned options from above, there is a requirement to formulate a mathematical item corresponding to a specific concept and taxonomy level, accompanied by guidelines and rules for the educational design of these items. These encompass a contextual background specifying the target audience and purpose of the items, along with a concise summary of the anticipated learner knowledge extracted from the course's lecture notes. Furthermore, the instructional prompt entails a mathematically framed and instructional description of taxonomical dimensions, as outlined in Section 2.3, which all were aligned with educationalist experts.

A Langchaing JSON OutputParser is utilized to convert the response results of the Chain into a specific Python Class format for automated response processing. In addition to translation and matching logic, the OutputParser includes a response format specification in the prompt, requiring the LLM to adhere to the specified JSON format when generating the requested item[7]. The resulting item(s) are presented to the user with a sample solution and can be saved as a JSON file for external processing. In addition to this main LLM prompt to generate items, users got the possibility to use additional prepared Chains and prompts to automate error checks and corrections for items or adjust the perceived difficulty level of the generated item(s). The full workflow is depicted with Figure 1.

Thus, the applied LLM prompting includes generated knowledge (Liu et al., 2022) to execute a Retrieval Augmented Generation (RAG) (Lewis et al., 2020) chain, resulting in a fine-tuned *Zero Shot Chain-of-Thought prompting* approach (Wei et al., 2022a; Kojima et al., 2022), which incorporates a prompt template and injection of specially prepared knowledge (Martino et al., 2023), augmented by a simplified Multiple-Agent approach (Du et al., 2024). The highest-level prompt template is depicted within **Listing 1**.

Listing 1 Prompt for generating competency-oriented and concept-related mathematics items. Variables are represented with {...}.

```
Create a math exercise on the mathematical
concept {concept} representing the process
dimension {pd} and the knowledge dimension
{kd}.

### Instructions ###

1. The items should contain mathematical
annotations and formulas in latex format!
2. The items should be formulated in German!
3. The items should contain random numbers
and not just enumerated numbers (1,2,3,4,5,
...)!
4. The items should not provide any hints,
help, examples or explanations!
5. Think step by step about how to generate
this item correctly

### Context ###

The generated items are intended for first-
semester Bachelor students in computer
science, who are to be individually supported
and challenged so that they satisfy the
demands of the mathematics course.
The students are presented with the generated
items by an automatic recommendation system
and can respond to them in text form.
The items are selected personalised for the
students according to their level of
competence (represented by a knowledge
dimension and process dimension) and
currently relevant concepts.
Therefore, it is extremely important that
the items only fit the specified mathematical
concept, the process dimension and the
knowledge dimension!

{lecture_and_student_information}
```

3 Containing six process- and four knowledge levels.

4 Streamlit Framework Homepage: https://streamlit.io/.

5 LangChain Framework Homepage: https://www.langchain.com/.

6 Facebook AI Similarity Search Documentation: https://ai.meta.com/tools/faiss/.

7 Which is a OpenAI supported functionality, also supported by other LLM providers.

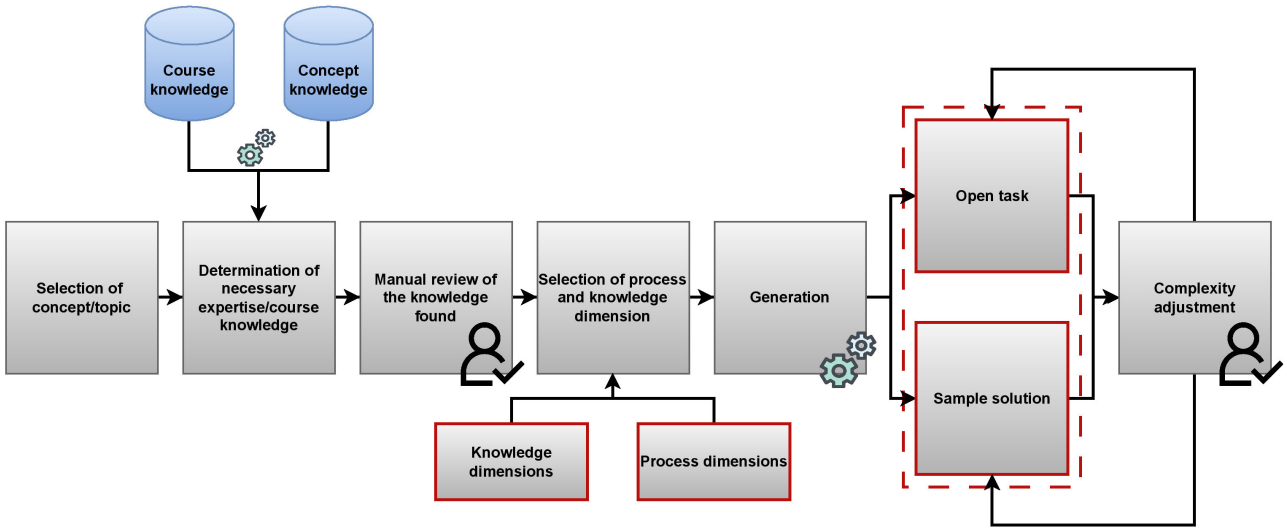

FIGURE 1
Item generation workflow, implemented within ItemForge.

```
### Knowledge ###

Use the following knowledge (concept
information) to create the item:
    {concept_information}

Items on the process dimension {pd} follow
these rules:
    {pd_information}

Items on the knowledge dimension {kd} follow
these rules:
    {kd_information}

{outputparser_format_instructions}
```

2.2 Retrieval and integration of mathematical concept information

While developing Itemforge (see Section 2) to generate mathematical items, it became evident that the focused LLMs GPT-3.5 and GPT-4 already possess broad knowledge of general mathematical concepts. Nevertheless, they frequently exhibit inaccuracies, especially when addressing more complex topics in higher education. The utilization of these models by learners poses a risk, as they may uncritically accept the models' seemingly convincing responses as accurate. These inaccuracies also hinder the generation of items, which are expected to be accurate and correct, particularly in the context of an automatic recommendation system, we are aiming for.

To ensure the generation of reliably correct items, particularly for advanced concepts in higher education, a process was devised, drawing inspiration from the *Generated Knowledge Prompting* (Liu et al., 2022) and *Knowledge Injection*(Martino et al., 2023) methodologies. The devised process entails the creation of a knowledge text, which serves as an explanation of the targeted mathematical concept. The knowledge text is produced through a retrieval-based *question-answering chain* facilitated by LangChain, leveraging information from two mathematics textbooks[8] and the course's lecture script. This mechanism enables concept queries to be addressed by a LLM utilizing a customized retriever, to build upon secured knowledge rather than encoded knowledge of the LLM. For ItemForge, the retriever comprises a FAISS vector database, housing the vectorized versions of the mathematical textbooks and lecture script, initialized during application launch via a locally executed text embedding model with multi-stage contrastive learning (Li et al., 2023)[9]. The selection of the aforementioned embedding model was based on its superior performance in the respective HuggingFace benchmark[10] at the time of application development (July 2023). The described knowledge sources are conventionally extracted, separated, and processed on a page-by-page basis.

As part of the specific query formulation (see **Listing 2**), prompt engineering techniques were utilized to optimize responses and direct the LLM in its summarization to rely solely on the found sources rather than pre-trained knowledge. This involved employing *Chain-of-Thoughts* techniques (Wei et al., 2022b), encouraging the model to answer the question step by step and to think through the provided steps incrementally. Additionally, the query was structured in a *Zero-Shot* (Wei et al., 2022a) fashion to

8 ISBNs: 978-3-662-63313-7 and 978-3-662-64389-1.

9 For embedding purposes, the HuggingFace Model *gte-large* was used - https://huggingface.co/thenlper/gte-large.

10 Massive Text Embedding Benchmark Leaderboard: https://huggingface.co/spaces/mteb/leaderboard.

avoid biases on expected style and contained information of the generated knowledge text.

Listing 2 Prompt for the inference of concept-specific knowledge

```
Explain the mathematical concept "{selected_
concept}" in a comprehensive section.
Only give an example, if one is presented in
the provided document and refer to the
documents only.
Let's think step by step about how to solve
this task. Here are the provided documents
that might help you. Say ''I don't know''
if you are unable to answer based on the
documents provided.
Answer solely based on the documents provided.
Describe the question in detail before
attempting to answer.
Answer in German.
```

Upon execution of the concept explanation chain, the FAISS database is queried for text-pages similar to the selected mathematical concept. The user can adjust the quantity of pages to be retrieved and the search algorithm utilized to fine-tune the knowledge retrieval process, based on prior knowledge or observed results. From the two selectable search algorithms, the "Maximal-Marginal-Relevance" algorithm aims to minimize redundant information while retaining relevance to previously assessed pages, resulting in a higher probability of diverse content retrieved from the vector database. The "Similarity-Score-Threshold" algorithm identifies pages most similar to the specified concept, based on a defined threshold, leading to the selection of pages with similar content.

The identified pages are subsequently utilized by the LLM to generate a textual representation of a selected mathematical concept. Depending on a user-chosen method, the identified pages are handed to the LLM in different fashions and either explained as of a single prompt (Stuff-Chain), or as multiple prompts, transferring responses of former prompts as input data to the next prompt (Map-Reduce-, Map-Rerank-, Refine-Chain), provided by the LangChain framework[11]. We found that the type of chain and the quantity of retrieved documents impact the quality of knowledge texts generated, showing no consistent regression pattern.

2.3 Item alignment with competencies and intended learning outcomes

To ensure the generation of high-quality items, it is essential to convey to the LLM not only the explanation of mathematical concepts and content, but also the principle of constructing competence-oriented items in the context of Constructive Alignment (Biggs, 1996). By precisely defining course-specific ILOs and their corresponding taxonomic and cognitive requirement levels, the development and validation of teaching, learning, and assessment design is facilitated. In this regard, the taxonomy of the cognitive domain by Anderson and Krathwohl (2001) was selected, due to its wide applicability across diverse domains, robust theoretical foundation, and endorsement by the German Rectors' Conference (Gröblinghoff, 2015). This taxonomy provides a structured framework that can seamlessly be integrated within the context of Constructive Alignment to precisely describe ILOs, which in turn, enables the systematic development of course content derived from ILOs and design of appropriate, as well as ILO and course-content matching assessment items.

One of the taxonomy's features involves the precise allocation of each item to a particular ILO, facilitating a precise assessment of the level of requirements in terms of cognitive processes and knowledge needed to address the item. This allocation ensures that the item's requirement level offers direct insights into the individual achievement of the corresponding ILO. Requirement levels are two-dimensional for Anderson and Krathwohl (2001), consisting of a process dimension and knowledge dimension. The process dimension delineates six specific cognitive processes required for learners to address an item or problem, while the knowledge dimension, consisting of four knowledge types, delineates the different types of needed knowledge respectively. Thus, 24 different requirement levels can be utilized to describe assessment items.

Listing 3 Adapted description of the factual knowledge type, providing guidance to the LLM in item construction [translated from German, building upon (Anderson and Krathwohl, 2001)]

```
Through items of this dimension level,
students are expected to demonstrate
their understanding of the basics
required to engage with a specialised
discipline (e.g., Mathematics) or solve
specific problems.
You should choose one of the
following categories:

1. Knowledge of disciplinary terminology

    Examples: Discipline-specific
    vocabulary (definitions, terms),
    mathematical symbols (logical
    connectors, summation and
    integral symbols,
    mathematical constants
    such as e and pi.

2. Knowledge of components and specific
details.

    Examples: Important or standard examples
    (e.g., the sequence 1/n, geometric series
    with limit behaviour, properties of sine,
    cosine, exponential function, standard
    normal distribution, determinant).
```

To instruct the LLM effectively through a taxonomical description of ILOs, the necessity for tailoring the taxonomy levels to the domain-specific processes and knowledge of mathematics

11 See https://js.langchain.com/docs/modules/chains/ for an overview of available chains.

became evident. This involved formulating specific cognitive processes and knowledge types as item construction guidelines to ensure the generation of items aligned with a particular taxonomy level, like visible for the factual knowledge type with **Listing 3**. The transformation process involved a consulted mathematics expert, who transposed the abstract description for the process and knowledge dimensions, along with examples outlined by the German Rectors' Conference (Gröblinghoff, 2015), to the field of mathematics. Subsequently, an expert in educational sciences transferred the transposed descriptions into construction instructions for LLMs, while validating their integrity and accuracy in comparison to the original description by Anderson and Krathwohl.

3 Evaluation of item generation quality

In this study, we investigate into the performance of a specifically prompted LLM in generating mathematical e-assessment items (refer to Section 2.1). The focus lies on the capabilities to generate appropriate, correct and complete items and corresponding sample solutions according to subject-specific concepts, cognitive processes, and knowledge required by students in accordance with constructive alignment (Biggs, 1996; Biggs and Tang, 2007). Through systematic variation of parameters, our goal is to gain insights into the strengths and limitations of the LLM's mathematical item generation capabilities. Generated items were evaluated by three domain experts independently, who's results are used as a qualitative benchmark.

We initially outline our methodology in Section 3.1, followed by an exposition of the study's execution and the analytical method employed. The findings are presented in Section 3.2, while the study's constraints are discussed in Section 4.2.

3.1 Methodology and evaluation criteria

The study aims to evaluate e-assessment items, including sample solutions, generated by LLMs in terms of appropriateness, correctness and completeness, as well as alignment with the specified generation parameters—subject-specific concept, cognitive process, and knowledge to be utilized (refer to Section 2.1).

As the used taxonomy - the taxonomy of the cognitive domain by Anderson and Krathwohl (2001)—consists of 24 levels (four knowledge levels by six process levels), 24 items are generated per selected concept. To reliably measure for level accuracy and recall, 10 concepts from a higher education mathematics course[12] were selected, yielding 240 items to be evaluated (refer to Section 2.1). The concept selection process consisted of examining the generated knowledge texts (refer to Section 2.1) for surface-level errors and excluding concepts lacking proper source information. This precaution aimed to prevent the measurement of fabricated or hallucinated information (Zhang et al., 2024; Martino et al., 2023) that could be erroneous or inaccurate, thus producing low-quality items. ILOs were not incorporated when generating the 240 items, as they would have introduced a bias in generating items for specific taxonomical levels and circumscribed the variety of possible results for a selected concept. In order to provide more general results on item generation, inclusion of ideal invidual knowledge, which is an optional feature for item generation (see Section 5.1), was not selected for creating the 240 items.

Three mathematics experts evaluated all 240 generated items to avoid measuring for the opinion of a single rater. These domain experts were guided by a manual (see Section 5.1) to appropriately categorize and rate the generated items regarding the generation-parameters and qualitative measures. The manual outlined the study's objectives, used scales, interpretation guidelines, the taxonomy used, and included selected domain-specific example items and their ratings, which were not part of the questionnaire.

The manual, as well as the bespoke questionnaire, comprising 240 items with seven questions per item, were reviewed by a mathematics expert, two computer-science experts and two educational sciences experts to ensure its quality. Questions were tailored to focus on specific item aspects, as existing and reviewed questionnaires were deemed too detailed for efficiently assessing 240 items based on high-level aspects within a practical timeframe. Consequently, we chose create a bespoke questionnaire and to have it evaluated by domain experts.

For the cognitive process and knowledge dimensions of an item, a selection component was provided to select the most appropriate level of the dimension. The remaining qualitative measures—appropriateness to the mathematical concept, completeness and correctness of the task, and completeness and correctness of the sample solution—were captured by a five-point Likert scale (Likert, 1932), allowing the raters to indicate tendencies, albeit potentially leading to central ratings. An additional comment section allows expressing thoughts per item individually.

3.1.1 Sample and large language models

The questionnaire was taken by three mathematical higher education experts independently. The three experts consisted of the teaching professor and his two associates, which are particularly well-suited due to their mathematical education (diploma, Ph.D., or professorship in mathematics) and their teaching experience through teaching several mathematics courses for several years. These experts underwent advanced training in competence assessment, possess expertise in creating suitable mathematical problems, and demonstrate a profound understanding of the study's concepts.

From an item generation perspective, both GPT-3.5 and GPT-4 by OpenAI were incorporated when generating the 240 items to be evaluated. Specifically, GPT-3.5-Turbo (0613) was used to produce the required knowledge text for the selected concept, used for all 24 related items, due to its ability to produce sufficient knowledge texts. GPT-4 (0613, non-turbo) was used to generate all 240 items, guided by the various parameters and the developed template-prompt (refer to Section 2), as it demonstrated superior item quality with fewer inaccuracies compared to GPT-3.5. Both LLMs offer the functionality to adhere to a specified response

12 Called *Mathematics for Computer Scientists I - Fundamentals, Linear Algebra, Analysis, and Differential Calculus* (translated from German).

format (now alled JSON mode[13]) and were selected based on their extensive evaluation and adoption within academia (Haverkamp, 2023; Chang et al., 2024; Sok and Heng, 2024), along with their provision as an Application Programming Interface to address the absence of suitable LLM infrastructure.

3.1.2 Execution

The questionnaire was prepared as Google Spreadsheet documents, allowing to rate items in rater chosen batches. Each rater received an individual and unique document to maintain their independence, along with a detailed manual (refer to Section 3.1). Prior to rating, all raters were instructed to read the manual thoroughly and to address any uncertainties by asking questions for clarification, where questions and answers were shared among all raters.

After training completion, the raters were allotted a period of 21 days to evaluate all 240 items. Two raters successfully completed the evaluation within this time frame. However, one rater was unable to assess all aspects of the 240 items within the given timeframe, leading to a reduction in the questionnaire scope to focus solely on the cognitive process and knowledge dimensions. Consequently, there were three raters for the evaluation of all 240 items in relation to cognitive processes and knowledge, and two raters for the qualitative aspects of the items.

3.1.3 Analytical method

The questionnaire results were examined in terms of two main dimensions: (1) alignment of process and knowledge dimensions, and (2) qualitative evaluation of the generated items.

ItemForge, and in general, an item generator, can be treated as a prediction engine in the context of predicting items. A common analytical method for predicted values involves calculating a confusion matrix and a fitting F-Measure to assess precision and recall of the prediction engine (1) (Kelleher et al., 2020). Due to the non-numerical and discrete nature of the two dimensions under consideration, an averaging approach for differing values from multiple raters is not feasible. When values from multiple raters vary, a discrete range is established based on the specified levels by each rater and the prediction engine is referred to as appropriate if its parameters are inside this range and deviating, if outside the range.

The five qualitative aspects (2), namely concept appropriateness (ca), task completeness (tcm), task correctness (tcr), solution completeness (scm), and solution correctness (scr), are evaluated through statistical analysis. These aspects are represented using box plots, as these provide a visual summary of the distribution of data for each qualitative aspect while displaying key statistical measures such as medians, standard deviations, quartiles, and outliers (DuToit et al., 2012). By providing five box plots side by side, it is possible to compare the distributions of the different aspects simultaneously, aiding to understand how these vary and whether any consistent patterns or differences arise.

13 LLM JSON response mode documentation by OpenAI: https://platform.openai.com/docs/guides/text-generation/json-mode.

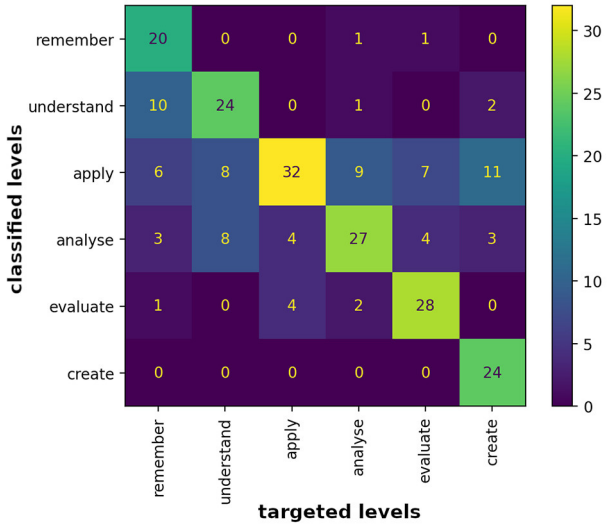

FIGURE 2
Manual item-process classification in contrast to targeted levels of the cognitive process dimension as a confusion matrix.

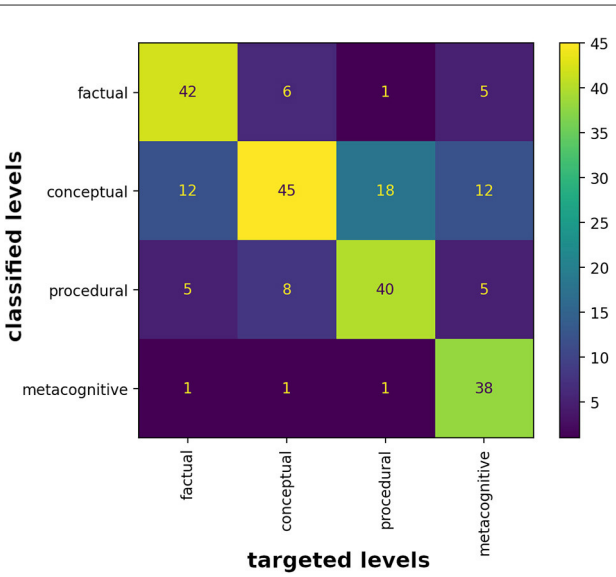

FIGURE 3
Manual item-knowledge classification in contrast to targeted levels of the knowledge dimension as a confusion matrix.

3.2 Findings and analysis

The alignment of encountered cognitive processes and knowledge in contrast to the targeted ones are analysed using confusion matrices for each dimension separately. Human ratings serve as the reference point for evaluating the fit of item generation parameters, used by ItemForge. The confusion matrices are illustrated in Figures 2, 3.

Figure 2 illustrates the outcomes related to the cognitive process dimension, which is one human selectable generation parameter in our approach. Noticeable is a strong alignment with the matrix's

major diagonal, indicating a close match between the generated items and the selected levels by the raters. Each column represents 40 items generated for the respective process level. The levels *apply* (32 items, 0.8 Prec., 0.44 Rec.) and *evaluate* (28 items, 0.7 Prec., 0.8 Rec.) demonstrate the highest conformity (Prec.), with only a small deviation of 8 and 12 items, respectively, from the expected values. A possible rationale, in the light of the remaining results, might be, that mathematical items are often used to carry out techniques or to evaluate a technique or result. The performance of the LLM in these levels may be attributed to the existence and usage of such items in the LLM's training phase.

The level *analyse* (27 items, 0.68 Prec., 0.55 Rec.) shows a similar pattern, with 13 items deviating, mainly in two nearby levels. Notably is a significant amount of 9 items in the nearby level *apply*, as well as a general tendency to categorize items toward or within the two middle levels (*apply* and *analyse*, refer to the general highlighting of these rows in Figure 2). The origin of this phenomenon remains ambiguous, as it is uncertain whether it is due to raters' inclinations or biases toward these levels, or whether there is tendency of the LLM to generate items within the levels *apply* and *analyse*. According to our appraisal, it is probable that the raters got a tendency toward these levels, as raters found interpretive overlaps in the description of the taxonomy levels, raising difficulties in clearly classifying items.

The levels *understand* (24 items, 0.6 Prec., 0.65 Rec.) and *create* (24 items, 0.6 Prec., 1.0 Rec.) are mid-performing, where *understand* spreads up to three nearby levels and *create* to three more distant levels. With a recall value of 1.0 (see Table 1), the generator seems to create items matching the level *create* more reliable than for other levels, except for *remember*, gaining the second best recall value of 0.91. *Remember* (20 items, 0.5 Prec., 0.91 Rec.) is the overall worst performing level, spreading noticeably over four levels and having the lowest precision of all levels.

Judging from those numbers, the LLM seems to reliably being able to generate items meeting the lower boundaries of the levels *understand* and *apply*, as well as the upper boundaries of *analyse* and *evaluate*. Even though meeting certain lower and upper boundaries of specific levels, there remain inconsistencies toward the middle levels of the dimension. On a macro-level, the average precision is rated with 0.65 and the average recall as 0.72, resulting in an overall F_1-score of 0.66. These findings demonstrate a tendency toward correct item generation for the requested process level, but also a tendency to occasionally generate items in mostly two nearby levels. In contrast, the low F_1-scores per level indicate that either the LLM is currently not well-equipped for the given task, needs to be prompted differently or further tuned by the available hyperparameters of our current approach.

Figure 3 illustrates the outcomes related to the knowledge dimension, which is the other human selectable generation parameter in our approach. Similar to the process dimension (see Figure 2), there is a noticeable strong alignment toward the major diagonal of the matrix, indicating a close fit between the generated items and the selected levels by the raters. As 240 items now account for four levels, each column sums up to 60 items. Thus, *conceptual knowledge* (45 items, 0.75 Prec., 0.52 Rec.) is best met, with the highest precision value. The levels *factual knowledge* (42 items, 0.7 Prec., 0.78 Rec.) and *procedural knowledge* (40 items, 0.67 Prec., 0.69 Rec.) show a similar distribution, with 18 to 20 items missing the correct category by mainly one level. *metacognitive knowledge* (38 items, 0.63 Prec., 0.93 Rec.) was missed by 22 items, spread over all three remaining levels with a peak of two levels away (in *conceptual knowledge*). Striking is a rater tendency toward *conceptual knowledge* (see respective row in Figure 3) and two overlapping 4x4 squares (factual to conceptual and conceptual to procedural). The two squares indicate a certain fuzziness in the considered levels, either on the side of the item generator or raters. Similar to the row highlighting for the process dimension (see Figure 2), there exists a visible highlighting of the middle two rows for the knowledge dimension. According to our appraisal, it probably got a similar rationale, which is the raters tendency toward these levels, as raters found interpretive overlaps in the description of the taxonomy levels factual and conceptual, raising difficulties in clearly classifying items. Noticeable is a upper boundy for the levels factual to procedural, which rarely leaks by one item only into the cognitive level.

In terms of the knowledge dimension, the LLM seems to perform better than for the process dimension, as higher precision (overall Prec.: 0.69) and recall (overall Rec.: 0.73) values were achieved (see Table 2). *conceptual* performs worst from a F_1-score perspective, due to low recall values. *conceptual* also performs best from a precision perspective, indicating its possible use in automatisms, followed by *factual*. *metacognitive* shows the highest recall value of 0.93, indicating that if an item was generated as metacognitive, it is probably also rates as such.

In terms of qualitative aspects (see Figure 4), most items apply to the chosen concept quite well, with a median value of 4.5 (out of 5), no lower quantile and a upper quantile of 0.5, indicating a tendency toward a perfect match. There are a few outliers ranging down to 2.5, still indicating a tendency to matching concepts. Task completeness and task correctness are rated equally (median 4.5, tendency to complete tasks, outliers down to 1.0). On close inspection, a few tasks seem to show problems in terms of completeness, as most needed values are given to solve the task, but there seem to exist better suited values according to the rater opinions. Just a few tasks lack the necessary information or are incorrect in itself, yielding a strong tendency toward correct and complete task generation.

Solution completeness performs slightly worse, having also a median value of 4.5, but gaining a lower quantile of 0.5 and an additional lower whisker up to 2.5. According to the rater opinions, solutions contain all to most relevant information, with a tendency to hinting about the correct solution instead of fully providing it. Solution correctness performs worst of all qualitative aspects, with a median of 4.0, upper and lower quartiles of size 1 and a lower whisker ranging to as low as 1. Thus, solutions are still probably correct, but might come with some inaccuracies or drawbacks. Because of the lower whisker, solutions must be checked properly for their correctness or hinted as beeing possibly incorrect, to not lead learners to false information. As the system described in Section 2 was not tuned for correct and complete solutions, the quality of the provided solutions is already impressive. Possibly incorrect solutions might even trigger students to think about their answer twice, producing a positive side-aspect if hinted appropriately.

Table 1 Precision, recall, F_1-score and support for the cognitive process dimension.

	Remember	Understand	Apply	Analyse	Evaluate	Create
Precision	0.5	0.6	0.8	0.68	0.7	0.6
Recall	0.91	0.65	0.44	0.55	0.8	1.0
F1-score	0.65	0.62	0.57	0.61	0.75	0.75
Support	22	37	73	49	35	24

Table 2 Precision, recall, F_1-score and support for the knowledge dimension.

	Factual	Conceptual	Procedural	Metacognitive
Precision	0.70	0.75	0.67	0.63
Recall	0.78	0.52	0.69	0.93
F1-score	0.74	0.61	0.68	0.75
Support	54	87	58	41

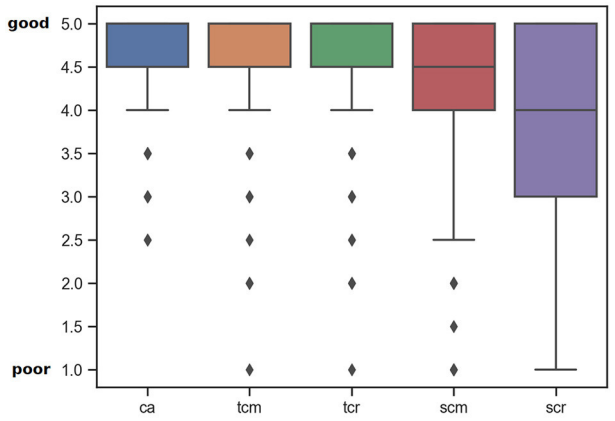

FIGURE 4
Qualitative aspects of the generated items, visualizing concept appropriateness (ca), task completeness (tcm), task correctness (tcr), solution completeness (scm), and solution correctness (scr) using box plots on a 5 point Likart-Scale for which 5.0 corresponds to good, 1.0 to poor.

3.3 Mathematical expert feedback

According to the feedback from the three mathematics experts (see Section 3.1.1), the linguistic quality of both the generated tasks and the solutions is consistently very high. The targeted concepts are always addressed by the generated items, often done in connection with related concepts, for which the focus sometimes changes to the related one instead of the actually targeted one. It may thus be helpful to specify (sub-)concepts more precisely in future iterations. The task formulations occasionally suggest the existence of certain mathematical prerequisites which, however, are not fulfilled, and learners may not always be aware that when incorrect assumptions are made, the actual task becomes irrelevant, as they might expect the task to be properly designed.

When creating tasks of for higher levels of the process dimension, numerous subtasks are often presented, rendering these tasks unnecessarily complex. As an expert, one might prefer formulating shorter, more targeted tasks instead. In contrast, the highest process level - create - can often be met with simple answers, and it would occasionally be sensible to impose additional requirements to truly achieve the desired high process level.

The generated solutions, on the other hand, exhibit a noticeably lower mathematical quality compared to the generated tasks themselves. Frequently, while the actual mathematical reasoning was sound, there were significant computational errors that detracted from the overall impression. Based on the experience of the mathematics experts, these errors resemble those commonly made by beginners with insufficient mathematical background knowledge or unfocused students.

4 Discussion

In addition to our own research and findings, relevant literature is examined and deliberated in Section 4.1, and the methodological and interpretive constraints of our study are addressed in Section 4.2.

4.1 Related work

Previous research about automated generation of mathematical assessments, problems, or tasks and related sample solutions denotes a common methodological foundation through the use of structured approaches to systematically generate these. Ahmed et al. (2013) presented a method involving forward solution search for solution generation and backward generation of exercise problems in natural deduction, considering all potential inference rule applications over small propositions. Through their approach, they were able to find new problems to given solutions, usable in exercises. A similar pattern was used by Singh et al. in algebraic problem generation, where proof problems are identified. By syntactically generalizing proof problems into abstract queries and automatically exploring the problem space, problems-solution combinations became generatable. This approach enables

educators to design exam-oriented exercise tasks and enables learners to create personalized exercise problems derived from learning-materials (Singh et al., 2012). Not related to mathematics, but to structured item generation, Faizan and Lohmann (2018) presented an approach to derive multiple-choice questions from slide-extracted keywords and semantic querying of a pre-designed knowledge-base to create varied questions and appropriate answer options. Xu et al. introduced an approach in elementary mathematics education for problem generation, comprised of two phases: initial formulation of abstract mathematical problems via a template-driven approach, followed by a rule-based generation of exercises, distractors, and visual aids using a multi-language adaptive pipeline. This method ensures the creation of relevant and solvable problems customized for primary school learners and resulted in significant time savings for educators (Xu et al., 2021) in creating items. All these approaches share varying sets of restrictions, based on: topic restrictions to comply with methodological requirements, necessitating predefined formal abstractions up to whole knowledge-bases, and adherence to pattern- and rule-based frameworks, becoming increasingly complex for advanced mathematical concepts.

Even though the presented approaches are limited as outlined above, their application among learner-adaptive assessment and exercise systems, particularly within structured domains such as mathematics, became popular. Such systems allow tailoring items to a learner's skill level and to generate item instances on-demand and as needed (Tvarožek et al., 2008). A recent study revealed that learner-adaptive systems, utilizing AI and providing adaptive exercises and assessments, haven been found to significantly improve on the learning-performance (Das et al., 2023).

Recent advancements in AI resulted in LLMs, which rapidly became widespread and often applied among educators, but also learners. Application among learners was partly acknowledged as a serious threat to academic integrity, leading to research on AI-generated content detection (Orenstrakh et al., 2023). On the other hand, LLMs may be utilized to create educational content and personalize learning resources, possibly enhancing student engagement and interaction (Kasneci et al., 2023). In the field of mathematics, LLMs are predominantly applied for problem-solving activities, as seen with several recent investigations (He-Yueya et al., 2023; Imani et al., 2023), rather than for the creation of (personalized) educational materials, including assessments and exercises. The study at hand specifically targets the latter aspect and seeks to address found limitations from the structured methodologies mentioned earlier.

4.2 Limitations

Out of the 240 items assessed across seven key aspects, only two human experts evaluated all aspects for each item, while three raters assessed the process and knowledge dimensions. Despite their expertise, discrepancies in opinion on the generated items were evident among the raters. The assessment by two, up to three raters does not definitively determine the quality of item generation, but rather indicates trends and offers some level of reliability based on their expertise. However, the task of rating 240 items requires a significant time commitment of 20 to 50 hours per rater, posing a challenge in recruiting an adequate number of raters.

The methodology outlined in Section 2.1 enables the configuration of multiple hyperparameters to refine the information retrieval and generation procedures. These parameters were employed with either their default settings or settings found to work, but not with thoroughly evaluated setting combinations, which is why adjusting these settings could potentially enhance, but also lower the outcomes under investigation. Due to the absence of a baseline for comparison (which was initially created with this study), a systematic and scientifically rigorous tuning process was not feasible.

In terms of evaluation techniques, various techniques were explored, resulting in slightly varied confusion matrices. A mid-strict approach was selected, by focusing on whether the generated item matches any of the rater judged levels. An alternative and more relaxed technique involves interpolating taxonomy levels among raters, leading to slightly improved performance metrics (Precision, Recall, F_1-score), yet may pose challenges when raters largely differ in their opinions. Conversely, a stricter strategy entails focusing solely on the highest level rating, resulting in notably worse outcomes by disregarding the possibility that raters could be wrong.

Reasoning of the confusion matrices and of the small number of three raters, we can only highlight our appraisal on whether a finding is related to insufficient skills of the LLM, rater tendencies, or methodological shortcomings. The evaluations in Section 3.2 primarily address rater tendencies over LLM skills, as raters encountered challenges in categorizing items due to interpretative complexities of the taxonomy and their historical emphasis on *apply, analyse, and evaluate* tasks. Methodological shortcomings also emerge as a plausible explanation, with educators facing challenges in implementing Anderson and Krathwohl's taxonomy and, according to our knowledge, often resorting to simpler alternatives, such as the CELG taxonomy recommended by the Free University of Berlin[14].

The LLMs GPT-3.5-turbo (0613) and GPT-4 (0613) were employed (see Section 3.1.1), which are proprietary models available for a currently undefined, but limited period. Consequently, the study's exact reproducibility is constrained within a narrow timeframe. Given the primary objective of this study to validate the LLMs' ability for competency-based e-assessment item generation from a qualitative standpoint, the utilization of proprietary models is not deemed a drawback. For broader and reproducible findings, the study must be replicated using diverse open models to ensure future result reproducibility. To facilitate this process, we have made our rater statements, manuals, item sets, and outcomes accessible online in a reusable format (refer to Section 5.1).

14 The Free University of Berlin recommends the CELG taxonomy, providing four process and three knowledge levels: https://wikis.fu-berlin.de/display/eexamathome/Lernzieltaxonomien.

5 Conclusion and future directions

In conclusion, this study has shown the viability of LLMs as building blocks in constructing mathematical assessment items, that closely align with domain concepts and largely adhere to predefined taxonomy levels based on Anderson and Krathwohl (2001). The assessment items were evaluated by experts in higher mathematical education, encompassing 240 unique items spanning the entire taxonomy and selected domain concepts. The results yielded encouraging findings, highlighting the linguistic precision, overall correctness and completeness of the items tasks, and alignment with the intended taxonomy levels. However, the degree of alignment presents opportunities for progressive refinement, e.g. by enhancing the preciseness of taxonomy level alignment through algorithm and prompt fine-tuning. Sample solutions, generated alongside the items, warrant an additional layer of expert or crowd verification to ensure their completeness and validity.

An interesting insight from this study is the identification of the *Apply* and *Conceptual Knowledge* taxonomy levels, having the highest generation precision of all levels (0.8 and 0.75). Automated item generation within those levels seems viable and holds promise regarding direct and possibly ad-hoc application. For levels beyond these, the study points for an integrated approach where the developed tool and thus utilized LLMs serve as an assistive technology rather than a standalone solution. Even though the precision for the remaining taxonomy levels is lower, falsely generated items distribute closely to their targeted levels and mostly cluster around the center of the corresponding dimension. Consequently, the generated items typically establish a strong foundation, serving as either realized concepts or useful starting points that streamline the laborious process of item development.

In summary, the presented approach opens a route to significantly lessen the burden of creating assessment items, enabling educators to reallocate their time to more personalized and impactful teaching and mentoring activities. Still, continuous educator involvement with LLM-generated content remains crucial to preserve the nuances of higher-level cognitive tasks and to ensure the created items' suitability and validity. Our research offers a basis for an informed debate on AI's role in education and deepens understanding of leveraging LLMs for creation of high-quality, tailored educational materials. We demonstrated that LLMs hold promise for augmenting the educational experience but underscore the need for a balance between automation and the irreplaceable depth of human expertise. Thus, our findings can serve as a point of reference in enabling a founded discourse on collaborative educational resource development, guiding institutions in integrating LLMs into their pedagogical processes, while adhering to academic standards and enriching the educational journey for both educators and students.

5.1 Refinement and future work

Building upon our methodology and insights, we can pinpoint four key directions for future research, each with the potential to amplify the utility and precision of LLMs in educational content creation, specifically assessment item creation.

Firstly, an in-depth investigation into optimizing algorithm hyperparameters, LLM prompting, and splitting methods of source material for information retrieval could be a promising next step. By adjusting these parameters and prompts, it may be feasible to improve the accuracy and relevance of both the knowledge text and item generation processes, thus enhancing the quality of the assessment items generated.

Secondly, the exploration of advanced techniques for improving the generation of sample solutions, such as the integration of computer algebra systems, shows potential for further investigation. This approach holds promise in automating and enhancing the correctness and completeness of sample solutions for mathematical tasks, ensuring alignment with the corresponding task.

Expanding the tool's scope to include closed-ended item creation would enhance its versatility, offering educators a wider range of assessment options for formative and summative purposes. These might be complemented by researching the adaptability and transferability of the approach and tool to different domains, providing valuable insights into its performance and flexibility. This cross-disciplinary approach would offer insights into the broader applicability of LLMs in education and the potential need for domain-specific adaptations.

Lastly, in the rapidly advancing landscape of LLMs, it is crucial to assess the performance of diverse LLMs, encompassing proprietary and open-source alternatives. A comparative evaluation of various LLMs can elucidate their strengths and limitations, guiding educators and developers toward the most effective models for item generation.

As we continue to advance the role of AI in educational settings, the focal points highlighted for future work illustrate a commitment to enhance the symbiosis between technology and pedagogy. By pursuing these areas of research, we can refine AI-based tools to not only adhere to educational standards, but to also enrich the educational journey for educators and learners alike. Continuous improvement and evaluation of these tools are crucial, emphasizing the quality of educational content and the potential for AI to enhance teaching and learning. The future of AI-supported education relies on the interaction between the innovative capabilities of LLMs and the valuable insights of educators, promoting constructive collaboration and mutual enhancement in this evolving landscape.

Data availability statement

ItemForge is distributed as Open-Source Software, with its source code accessible at https://gitlab.com/Tech4Comp/assessment-llm-studies/. The evaluation processes were conducted using Python scripts to analyse the generated items and rater information, and to create confusion matrices and box plots. These scripts can be accessed from the aforementioned repository as well. The datasets generated and examined in this research can be accessed at https://gitlab.com/Tech4Comp/itemforge-study-documents, including documentation and the instructional manual intended for item raters.

Author contributions

RM: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. AP: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. KI: Writing – review & editing, Resources, Methodology, Investigation, Conceptualization. MG: Writing – review & editing, Supervision, Resources, Methodology, Investigation, Funding acquisition, Conceptualization. ST: Writing – review & editing, Validation, Investigation, Conceptualization. AT: Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition. NP: Validation, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization, Writing – review & editing, Writing – original draft. H-WW: Writing – review & editing, Supervision, Methodology, Funding acquisition. WH: Writing – review & editing, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The research is part of a governmental funded research project tech4compKI, funded by the Federal Ministry of Education and Research in Germany (16DHB2206 and 16DHB2211).

Acknowledgments

The authors thank Martin Grüttmüller, Katja Ihsberner, and Benjamin Schmidt for the effort of rating 240 generated assessment items in terms of program parameters and qualitative aspects. GPT-3.5-turbo (v0125), a LLM developed by OpenAI Inc., was utilized to explore alternative phrasings of paragraphs and enhance on the grammatical accuracy of the texts in this article.

Conflict of interest

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

Publisher's note

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

References

Ahmed, U. Z., Gulwani, S., and Karkare, A. (2013). "Automatically generating problems and solutions for natural deduction," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (Beijing: AAAI Press), 1968–1975. Available at: https://dl.acm.org/doi/abs/10.5555/2540128.2540411#core-collateral-info

Anderson, L. W., and Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: a Revision of Bloom's Taxonomy of Educational Objectives: Complete Edition*. North York, ON: Addison Wesley Longman, Inc.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Educ.* 32, 347–364. doi: 10.1007/BF00138871

Biggs, J., and Tang, C. (2007). "Outcomes-based teaching and learning (OBTL)," in *Why is it, How do We Make It Work* (Hobart). Available at: https://talic.hku.hk/wp-content/uploads/2016/08/OBTL_what_why_how1.pdf

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., et al. (2024). A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* 15:3641289. doi: 10.1145/3641289

Das, A., Malaviya, S., and Singh, M. (2023). The impact of AI-driven personalization on learners' performance. *Int. J. Comp. Sci. Eng.* 11, 15–22. doi: 10.26438/ijcse/v11i8.1522

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. *arXiv* [Preprint] arXiv:2305.14325v1. doi: 10.48550/arXiv.2305.14325

DuToit, S. H., Steyn, A. G. W., and Stumpf, R. H. (2012). *Graphical Exploratory Data Analysis*. Cham: Springer Science & Business Media.

Faizan, A., and Lohmann, S. (2018). "Automatic generation of multiple choice questions from slide content using linked data," in *Proceedings of the 8th International Conference on web Intelligence, Mining and Semantics* (New York, NY: Association for Computing Machinery), 1–8. doi: 10.1145/3227609.3227656

Gröblinghoff, F. (2015). "Lernergebnisse praktisch formulieren," in *Nexus impulse für die Praxis* (Hochschulrektoren-Konferen). Available at: https://www.hrk-nexus.de/fileadmin/redaktion/hrk-nexus/07-Downloads/07-02-Publikationen/Lernergebnisse_praktisch_formulieren_01.pdf

Haverkamp, W. (2023). *Uptake and Dissemination of Chatgpt in the Academic World as Reflected in the Web of Science: a Bibliometric Analysis of the First 6 Months After its Release*. doi: 10.13140/RG.2.2.16254.77121

He-Yueya, J., Poesia, G., Wang, R., and Goodman, N. (2023). "Solving math word problems by combining language models with symbolic solvers," in *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23*. Available at: https://openreview.net/forum?id=m7m14acWQi

Imani, S., Du, L., and Shrivastava, H. (2023). "Mathprompter: Mathematical reasoning using large language models," in *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models, Vol. 5*, eds. S. Sitaram, B. Beigman Klebanov, and J. D. Williams (Toronto, ON: Association for Computational Linguistics), 37–42. doi: 10.18653/v1/2023.acl-industry.4

Kasneci, E., Sessler, K., Khemann, 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

Kelleher, J. D., Mac Namee, B., and D'arcy, A. (2020). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. Cambridge, MA: MIT Press.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2022). "Large language models are zero-shot reasoners," in *Proceedings of the 36th International Conference on Neural Information Processing Systems, Vol. 35*, eds. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (New Orleans, LA: Curran Associates, Inc.), 22199–22213.

Kumar, H., Rothschild, D. M., Goldstein, D. G., and Hofman, J. (2023). Math education with large language models: peril or promise? [preprint]. doi: 10.2139/ssrn.4641653

Laverghetta, A., and Licato, J. (2023). "Generating better items for cognitive assessments using large language models," in *Proceedings of the 18th Workshop on*

Innovative Use of NLP for Building Educational Applications (BEA 2023), eds. E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, T. Zesch (Toronto, ON: Association for Computational Linguistics), 414–428. doi: 10.18653/v1/2023.bea-1.34

Lee, H. (2024). The rise of chatgpt: Exploring its potential in medical education. *Anat. Sci. Educ.* 17, 926–931. doi: 10.1002/ase.2270

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems, Vol. 33*, eds. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Vancouver, BC: Curran Associates, Inc.), 9459–9474. Available at: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., and Zhang, M. (2023). Towards general text embeddings with multi-stage contrastive learning. *arXiv* [preprint] arXiv:2308.03281. doi: 10.48550/arXiv.2308.03281

Likert, R. (1932). A technique for the measurement of attitudes. *Arch. Psychol.* 22:55.

Liu, J., Liu, A., Lu, X., Welleck, S., West, P., Le Bras, R., et al. (2022). "Generated knowledge prompting for commonsense reasoning," 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), 3154–3169.

Martino, A., Iannelli, M., and Truong, C. (2023). "Knowledge injection to counter large language model (llm) hallucination," in *The Semantic Web: ESWC 2023 Satellite Events*, eds. C. Pesquita, H. Skaf-Molli, V. Efthymiou, S. Kirrane, A. Ngonga, D. Collarana, et al. (Cham. Springer Nature Switzerland), 182–185.

Orenstrakh, M. S., Karnalim, O., Suarez, C. A., and Liut, M. (2023). "Detecting LLM-generated text in computing education: A comparative study for ChatGPT cases," in *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)* (Osaka: IEEE).

Singh, R., Singh, R., Gulwani, S., Gulwani, S., Rajamani, S. K., and Rajamani, S. K. (2012). "Automatically generating algebra problems," in *AAAI Conference on Artificial Intelligence*.

Sok, S., and Heng, K. (2024). Opportunities, challenges, and strategies for using chatgpt in higher education: a literature review. *J. Digit. Educ. Technol.* 4:14027. doi: 10.30935/jdet/14027

Tvarožek, J., Kravčík, M., and Bieliková, M. (2008). "Towards computerized adaptive assessment based on structured tasks," in *Adaptive Hypermedia and Adaptive Web-Based Systems*, eds. W. Nejdl, J. Kay, P. Pu, and E. Herder (Berlin: Springer Berlin Heidelberg), 224–234.

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., et al. (2022a). "Finetuned language models are zero-shot learners," in *International Conference on Learning Representations*. Available at: https://openreview.net/forum?id=gEZrGCozdqR

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., et al. (2022b). "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, eds. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (New York: Curran Associates, Inc.), 24824–24837. Available at: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

Xu, Y., Xu, Y., Smeets, R., Smeets, R., Bidarra, R., and Bidarra, R. (2021). Procedural generation of problems for elementary math education. *Int. J. Serious Games*. doi: 10.17083/ijsg.v8i2.396

Zhai, X. (2023). Chatgpt for next generation science learning. *XRDS* 29, 42–46. doi: 10.1145/3589649

Zhai, X., and Nehm, R. H. (2023). Ai and formative assessment: the train has left the station. *J. Res. Sci. Teach.* 60, 1390–1398. doi: 10.1002/tea.21885

Zhang, M., Press, O., Merrill, W., Liu, A., and Smith, N. A. (2024). "How language model hallucinations can snowball," in *Forty-First International Conference on Machine Learning*. Available at: https://openreview.net/forum?id=FPlaQyAGHu