

Can ChatGPT support prospective teachers in physics task development?

Stefan Küchemann^{1,*}, Steffen Steinert¹, Natalia Revenga¹, Matthias Schweinberger¹,
Yavuz Dinc¹, Karina E. Avila², and Jochen Kuhn¹

¹*Chair of Physics Education, Faculty of Physics, Ludwig-Maximilians-Universität München
(LMU Munich), Geschwister-Scholl-Platz 1, 80539 Munich, Germany*

²*Department of Mathematics, RPTU Kaiserslautern-Landau,
Paul-Ehrlich-Str. 14, 67663 Kaiserslautern, Germany*



(Received 24 April 2023; accepted 2 August 2023; published 11 September 2023)

The recent advancement of large language models presents numerous opportunities for teaching and learning. Despite widespread public debate regarding the use of large language models, empirical research on their opportunities and risks in education remains limited. In this work, we demonstrate the qualities and shortcomings of using ChatGPT 3.5 for physics task development by prospective teachers. In a randomized controlled trial, 26 prospective physics teacher students were divided into two groups: the first group used ChatGPT 3.5 to develop text-based physics tasks for four different concepts in the field of kinematics for 10th-grade high school students, while the second group used a classical textbook to create tasks for the same concepts and target group. The results indicate no difference in task correctness, but students using the textbook achieved a higher clarity and more frequently embedded their questions in a meaningful context. Both groups adapted the level of task difficulty easily to the target group but struggled strongly with sufficient task specificity, i.e., relevant information to solve the tasks was missing. Students using ChatGPT for problem posing rated high system usability but experienced difficulties with output quality. These results provide insights into the opportunities and pitfalls of using large language models in education.

DOI: [10.1103/PhysRevPhysEducRes.19.020128](https://doi.org/10.1103/PhysRevPhysEducRes.19.020128)

I. INTRODUCTION

The ability to develop tasks is an essential skill for teachers in schools, particularly for creating formative and summative assessments. High-quality tasks enable teachers to diagnose students' conceptual understanding and difficulties, monitor progress, and evaluate the efficacy of pedagogical methods. Utilizing well-designed assessments, teachers can implement timely interventions and support student development, directly impacting student performance. By creating and implementing diverse tasks that accommodate varying learning preferences and prior knowledge levels, teachers can foster inclusive classrooms and account for student diversity, directly impacting educational quality. Consequently, the American Federation of Teachers, the National Council on Measurement in Education, and the National Education Association established standards for teacher competence in student assessment in 1990 [1], and the competence was recognized as a

relevant component of pedagogical content knowledge by Gess-Newsome *et al.* [2] and of classroom assessment by Airasian [3]. In spite of its acknowledged relevance, research indicates that teachers in the United States feel inadequately prepared to develop tasks assessing students' performance [4,5] despite receiving undergraduate assessment training. This feeling of inadequate preparation aligns with the sentiment of limited assessment literacy noted by Popham [6]. According to Popham, limited assessment literacy includes knowledge of (i) "educational measurement's imprecision," (ii) "ways to judge the suitability of an educational test," and (iii) "proper uses of certain kinds of tests." It is therefore essential to identify methods for training prospective and in-service teachers in task development, particularly with digital tools, since, according to Eyal, they are a central part of 21st-century pedagogical approaches and offer additional opportunities for educational assessment [7].

In addition to numerous other applications, recent advancements in large language models, such as ChatGPT, present opportunities for the automated generation of assessments utilizing well-crafted prompts [8]. This raises questions regarding the quality of tasks developed by prospective teachers using ChatGPT compared to traditional methods, such as textbook assistance. By enabling teachers to specify prompts and task creation criteria, ChatGPT may

*s.kuechemann@lmu.de

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

facilitate the streamlining and enhancement of the task development process.

This study compares the quality of tasks developed by prospective physics teachers utilizing ChatGPT (intervention group) to those created using a textbook (control group). The research focuses on three primary questions:

RQ1: What is the quality and type of physics tasks developed by prospective teachers using ChatGPT in comparison to a textbook?

RQ2: What improvements do prospective physics teachers make to ChatGPT-generated and textbook tasks?

RQ3: How do prospective physics teachers rate ChatGPT's usability, perceived usefulness, and output when creating physics tasks?

In the following, we underpin our research questions by summarizing the current state of research on assessments in high school physics, teachers' ability to create assignments, and the rapidly evolving field of large language models in education. Our analysis will focus on the parameters outlined in Sec. II A to determine if using ChatGPT leads to improved quality of physics assessment tasks created by prospective teachers.

II. THEORETICAL BACKGROUND

A. Tasks in high-school-level physics

At its core, physics education involves the ability to effectively approach and solve a wide array of tasks. As defined by Fischer and Kauertz [9], these tasks are oral or written instructions to perform a specific action, serving to organize student learning, monitor progress, and measure performance. Task resolution is essential for comprehending physics, despite its perception as an onerous prerequisite for written performance measurement [9–12]. Students must engage with a multitude of tasks to successfully understand physics, requiring the differentiation between learning and performance tasks [9], as well as an understanding of when errors are permissible [13,14]. Tasks may be presented orally or in written form, with test tasks predominantly written. Text-based assignments necessitate reading proficiency, the application of physical concepts and mathematical procedures, and critical reflection [15].

The process of developing an effective task or adapting it to the varying conditions of a classroom is a creative and iterative endeavor. It involves a cycle of trial and error, followed by revisions to the task. [9]. Examination tasks should assess individual competencies; multiple-choice tasks are appropriate and are prone to be one-dimensional. Such one-dimensional tasks are characteristic and a quality feature of tests, as it can be statistically evaluated to provide a clear overview of student competencies [9]. However, texts utilized in testing must be unambiguous to prevent misinterpretation. Failure in such tasks is typically attributed to text misinterpretation or insufficient knowledge or foundational understanding [9]. Traditional tasks often

neglect student learning processes, impeding the development of resilient concepts [16]. These tasks primarily assess reading or mathematical competencies rather than physical conceptual thinking.

Furthermore, competency-based tasks typically feature a strong contextual orientation and are often preceded by text or material containing both relevant and irrelevant information to provide contextual framing. Visual aids such as pictures, graphics, or newspaper clippings may also be used as part of this setting. According to Kuhn and Müller [17], the integration of real-life contexts into task design can significantly improve student engagement, understanding, and knowledge transfer in science education.

The most challenging tasks are open-ended experimental complex problem-solving tasks [18], which allow for multiple reasonable solutions and may not have a single unambiguous answer. In the case of open-ended tasks, the underlying question must typically be clarified before the task can be solved.

Although tasks are a crucial component of physics education and are routinely used, the selection and creation of appropriate tasks should not be underestimated. Tasks must effectively capture intended learning goals, engage a typically diverse learning group, and accommodate individual understanding of physical concepts [9].

B. Criteria for the rating of tasks

In computer science, AI models have previously been used to automatically generate tasks in education [19,20]. To achieve this, AI models must be trained on datasets containing tasks and their corresponding labels, which are characteristics typically assigned by human raters. In this manuscript, our rating of physics tasks partially builds on certain previous tasks labels. For instance, Rodriguez-Torrealba *et al.* used a transformer model to generate multiple-choice tasks and rated them using human raters [21]. The raters were asked to evaluate “how well formed do you consider these questions and answers to be” and specifically consider aspects such as spelling, syntax, clarity, and meaning. While spelling and syntax were related to common errors made by large language models at that time, human-generated tasks may also lack clarity. As such, we included clarity and nonmisleadingness in our rating categories. We also added the aspect of meaning, which is similar to the concept-relevance aspect included in Chen *et al.* [19]'s question rating. Here, we included this aspect in the category relevance to map the target concept, as the participants were specifically asked to design tasks for specific concepts. Additionally, the authors included the context-complete category, indicating that sufficient context is provided to solve the tasks. In our rating, this aspect was divided into the aspect of specificity, in line with Bhat *et al.* [20], and context, indicating whether context is provided or not. Moreover, Raina and Gales [22] designed a framework for automated tasks, suggesting four quality

aspects of multiple-choice tasks: grammatical fluidity, answerability, diversity, and complexity. While some of these aspects are already included in the aspects mentioned above, we included the aspect of complexity which measures the amount of reasoning required to solve the question. In our work, the participants were asked to develop tasks for 10th-grade high school students; thus, we translated this aspect into an adequate difficulty for the target group. Eventually, we considered the categories of correctness and overall quality in addition to the previously mentioned categories because they are relevant aspects that were not included in previous considerations.

C. Teachers' ability to create tasks

Extensive research has been conducted on the pedagogical and domain-specific skills required by teachers to effectively facilitate student learning, resulting in the proposal of several models describing fundamental teacher competencies [23,24]. Recent studies in mathematics and science indicate that teachers' pedagogical content knowledge (PCK) significantly impacts student achievement [25]. Shulman emphasizes the importance of PCK, which enables teachers to make the subject matter accessible to students by combining subject knowledge and teaching skills [24,26,27]. Subject knowledge presupposes the possession of problem-solving skills among educators. However, to effectively employ these skills for didactic purposes through selection, adaptation, or development, educators must possess a thorough understanding of task didactic analysis [28].

Educators are expected to possess the ability to develop and reformulate tasks in order to facilitate meaningful learning environments for students [29]. This activity, referred to as problem posing by Silver [30], encompasses both the generation of novel problems and the reformulation of existing tasks. Problem posing is crucial for both students and educators, as it fosters creative thinking among students [31] and provides educators with insight into students' thought processes and conceptual understanding [32]. For instance, educators have successfully identified mathematical misconceptions among students through the use of problem-posing tests and student self-posed problems [33].

The ability of educators to apply and teach problem posing significantly impacts students' conceptual understanding and problem-solving skills. As such, it is crucial for educators to possess the skills and knowledge necessary to design and reformulate problems in order to facilitate similar learning activities for their students [29]. Lowrie (2002) highlights the close correlation between problem-posing and problem-solving processes [34]. As a result, problem posing constitutes a critical component of educators' responsibilities, as they are tasked not only with presenting problems to students but also with guiding them

toward becoming proficient problem posers in their own right [35].

Despite its importance, problem posing presents several challenges. For example, Crespo and Sinclair (2008) found that prospective teachers experienced difficulties engaging with problem posing due to their lack of familiarity with it [36]. Additionally, the quality of posed problems is not always high, as demonstrated by Cai and Hwang (2002) [37] and Silver and Cai (1996) [38]. These findings underscore the need for further research into how educators learn to incorporate problem posing into their instruction and how they can improve the quality of their own task sets in order to effectively integrate them into their teaching [39].

D. Large language models in education

The ChatGPT system [40] used in this work is based on a Large Language Model (LLM). LLMs are neural networks for natural language processing (NLP) that are trained on extensive text datasets and capable of generating humanlike text for a variety of language-related tasks [8]. Models such as GPT [41], BERT [42], and RoBERTa [43] have revolutionized the field of NLP and expanded the possibilities for research and applications. In education, these models can be integrated with chatbots to create adaptive and personalized learning experiences for students while supporting educators in their roles [44,45]. LLMs have been employed to assist students in numerous ways, including content generation [8,22], improving question-asking skills [46], generating code explanations [47], automating assessments [48], and providing feedback in language learning [49]. Chatbots can serve as conversational partners, supporting students experiencing foreign language anxiety [50] or low communication readiness [51]. The integration of LLMs and chatbots can facilitate more engaging learning experiences and aid students in expressing curiosity and comprehending complex concepts [46]. Research indicates that educators hold positive attitudes toward AI in education [52], with factors such as perceived usefulness, ease of use, and trust in AI-based tools influencing their acceptance [53].

Despite the potential of AI and chatbots in education, several challenges and open questions remain to be addressed. The responsible integration of AI into education will require the collaborative efforts of diverse communities, including educators, researchers, and policymakers [54]. Further research is required to investigate the effectiveness of LLMs and chatbots in various educational contexts, evaluate their impact on learning outcomes, and examine potential ethical concerns and biases [8].

The version of ChatGPT evaluated in this study, released on January 30, demonstrated some limitations. Based on GPT-3.5, this version was unable to process images. Prior to the study, according to its own information as well as the information on the website of OpenAI [55], it exhibited several limitations including a lack of common-sense knowledge, the potential for biased output, limited

conversational context, and difficulty with abstract reasoning and creativity (see Fig. 5 in the Appendix and [56]).

On February 27, upon further inquiry regarding the limitations of the ChatGPT version based on GPT-3.5, additional limitations were reported. These included time-limited knowledge, the potential for ambiguous or unclear responses, the generation of plausible but incorrect or misleading answers, sensitivity to input phrasing, verbosity, the potential for inappropriate content, a lack of common sense, an inability to ask clarifying questions, and limited consideration of longer conversational contexts (see Fig. 6 in the Appendix).

After the completion of the study, a version of ChatGPT based on GPT-4 became available for testing [57,58], which provided information about the differences between its limitations and those of the GPT-3.5-based version. The limitations of ChatGPT versions based on GPT-4 and GPT-3.5 are largely similar, as both models share common issues in their design and training. However, GPT-4 can potentially provide more current information than GPT-3.5, deliver higher quality responses, handle ambiguous questions more effectively, be better at maintaining context, and have a potentially lower frequency of inappropriate content generated by the model (see Fig. 7 in the Appendix).

It has been announced that GPT-4 will be able to process images [57], opening up new possibilities for educational applications.

LLMs and chatbots have the potential to revolutionize education by providing adaptive and personalized learning experiences. These technologies can assist learners in acquiring knowledge and support teachers in their roles, thereby enhancing the engagement and effectiveness of education. However, it is crucial to further investigate their capabilities and address the challenges associated with their integration to ensure their responsible and ethical use in educational settings.

In this work, we investigated the potential of using a large language model to support teachers during task development, as we intended to identify ways to support teachers who may struggle with this process, particularly those with low PCK. Large language models seemed to be a promising technology for providing such support due to their extensive range of capabilities. Furthermore, ChatGPT and other language have rapidly become integrated into daily life and professional practice. As a result, it is likely that teachers will utilize ChatGPT for task development even without any specialized training. Therefore, it is important to study how prospective physics teachers interpret the output generated by ChatGPT, their success in developing tasks with it, and the challenges associated with using ChatGPT for this purpose.

III. METHODS

A. Participants

The aim of this study is to examine the potential of chatbots in assisting physics educators in the development

of appropriate assessment tasks for high school students. Given that the challenges associated with task development are often related to experience, the focus of this investigation is on undergraduate prospective physics teacher students and graduate prospective physics teachers who do not regularly teach physics.

A total of 26 prospective physics teachers (13 female and 13 male) from LMU Munich, with a median age of 23 years, participated in this study. Of these participants, 80% had one year or less of teaching experience and 95% of this subgroup (76% of the total sample) reported having little to no experience in creating assessment tasks. Only 15% of participants reported having moderate experience in this area, while a mere 7% claimed to have extensive experience, with teaching experience ranging from 3.5 to 6 years. Participation was voluntary and uncompensated. Prior to the commencement of the study, the local ethics board reviewed and approved the performance of the study.

B. Design and materials

In this study, participants were tasked with developing four conceptual tasks to assess common concepts in Newtonian physics for 10th-grade German high school students (aged 15–16 years). The concepts included the relationship between velocity and acceleration and the first, second, and third laws of motion. These concepts were selected due to their fundamental importance in the high school physics curriculum. Half of the participants used ChatGPT 3.5 [40] (based on the January 30 version of ChatGPT 3.5) as a support to create physics tasks (intervention group), while the other half had access to a digital standard high school physics textbook as a support to create the physics tasks (control group) [59]. Physics textbooks for high schools often contain tasks that teachers may use in class. However, to prepare for an exam, teachers often need to modify textbook tasks to test students' problem-solving skills and conceptual understanding without relying on memorization of the results. To account for this practice and to allow for a fair comparison between the two groups, participants in the control group were asked to make significant changes to the textbook tasks (to change more than just the given values). Moreover, the participants in the intervention group did not have access to the physics textbook and the control group did not have access to ChatGPT.

Prior to task creation, intervention group students were asked to enter three given prompts into ChatGPT (1. "Create a question about the book Goethe's Faust," 2. "Create a question on Hermann Hesse's book *Demian* about the role of Demian for 11th grade students in a high school," and 3. "Create a multiple-choice question on the importance of Quidditch for Harry Potter in the book *Harry Potter and the Sorcerer's Stone*.")) to familiarize themselves with ChatGPT's sensitivity and flexibility in responding. After completing these prompts, students were given the

opportunity to experiment with their own inputs before proceeding to the conceptual task creation phase.

In each group, participants were instructed to provide both their own created or adapted questions and the original questions from either the textbook or ChatGPT that inspired their task development or adaptation, if applicable. This serves two purposes: first, to ensure that participants in the control group do not simply copy questions directly from the textbook without making modifications. Such a result would not be valid for our study, as our aim is to evaluate participants' ability to create original assessment tasks. Second, we aim to observe to what extent participants modify the tasks suggested by ChatGPT. This enables us to evaluate both ChatGPT's ability to generate high-quality assessment tasks, and the total workload required for the educator when using the tool. The chat history has been saved to verify students' inputs. Moreover, the students worked individually on creating the tasks, i.e., each task was designed by one student. This means that each student evaluated and modified either the tasks in the textbook or the output of ChatGPT. For example, in the textbook group, each student first selected a task in the textbook and then modified the task or created a new task without using the textbook for each of the four topics. In comparison, in the ChatGPT group, the students first designed a prompt to make ChatGPT create a task for a certain concept, and then they either improved the prompt and made ChatGPT create another task or they made ChatGPT modify the task by entering comments about the tasks. If they were satisfied with the tasks created by ChatGPT or preferred to make changes to the tasks by themselves, they continued to do so and then worked on the next task.

As an educator's ability to create suitable assessments is influenced by their own conceptual understanding of physics, participants' prior knowledge was evaluated using the half-length Force Concept Inventory (FCI) version 2 [60]. The test was selected because it covers the aforementioned concepts for which students developed the tasks. A good understanding of the subject matter might be crucial to compensate for the potential limitations of ChatGPT in generating consistent tasks (see the Appendix). It should be noted that standard textbooks typically do not suffer from this limitation as they undergo regular quality checks and revisions. However, even in this case, it is also essential for educators to possess a strong grasp of the assessed concepts and task quality when it comes to modifying textbook tasks for final exams.

Finally, given that large language models are a relatively new tool and may be associated with user-related challenges, a usability survey was conducted with intervention group participants. This survey included ten questions from the System Usability Scale (SUS) [61] and six questions regarding perceived usefulness and output quality from the Technology Acceptance Model 2 (TAM2) [62]. Additionally, an extra question was included to assess

TABLE I. Interrater reliability of the rating of developed tasks.

Category	Cohen's Kappa
Specificity	0.43
Clarity	0.52
Correctness	0.56
Not misleading	0.35
Adequate difficulty	0.76
Context	0.66
Average	0.55

whether unexpected slowdowns in ChatGPT's average response time affected the user experience (see Supplemental Material [63]). Each question was answered on a five-point Likert scale: strongly disagree, do not agree, neutral, agree, strongly agree. At the end of the study, participants were asked to provide their demographic data.

C. Qualitative data analysis of developed tasks

In total, participants developed $N = 103$ assessment tasks ($N = 51$ in the intervention group and $N = 52$ in the control group). Based on previous literature, eight categories were identified for evaluating the developed assessment tasks: specificity (scale: 0 = does not apply; 1 = applies), clarity (scale 0/1), correctness (scale 0/1), not misleading (scale 0/1), adequate difficulty (scale 0/1), context (scale 0/1), relevance to map the target concept (five-point Likert scale: absolutely irrelevant, rather irrelevant, about the same number of irrelevant as relevant aspects, rather relevant, very relevant), and overall quality (five-point Likert scale: very low, rather low, medium, rather high, very high; see Supplemental Material [63] for an explanation of these categories).

Two independent raters with over six years of experience in teaching prospective physics teachers, conducting physics education research, and developing conceptual questions for high school and university-level physics rated each task according to these eight categories. The interrater reliability Cohen's κ between the two raters was determined for the first six categories (see Table I), while the other two categories were averaged (see Sec. IV).

According to Landis and Koch, these values of Cohen's κ can be interpreted as indicating fair agreement ($\kappa = 0.35$) in the lowest case of whether a task's phrasing is misleading, up to substantial agreement ($\kappa = 0.76$) in the category of whether a task has adequate difficulty for a 10th-grade high school physics class [64]. On average, we find a moderate agreement between the two raters ($\kappa = 0.55$). Disagreements in ratings were resolved through discussion.

IV. RESULTS

A. Quality of generated tasks

After resolving disagreements, we compared the differences between the two groups across the eight

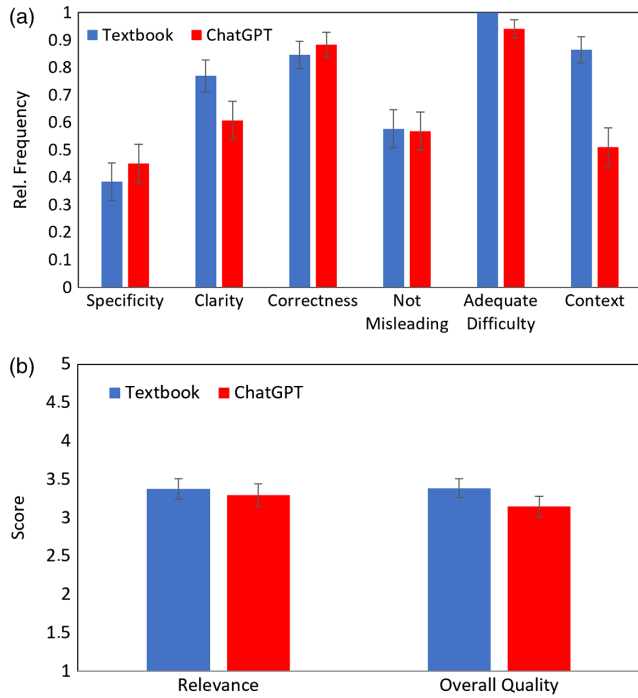


FIG. 1. Differences between the control group (textbook) and intervention group (ChatGPT) in (a) six categories rated on a scale from 0 (does not apply) to 1 (applies), and (b) categories rated on a Likert scale from 1 (very low) to 5 (very high).

categories (see Fig. 1). A *t* test revealed no significant difference in task specificity, correctness, whether phrasing was misleading, or whether tasks had adequate difficulty (each $p > 0.05$). It is interesting to note that both groups

achieved (nearly) perfect scores in task adequacy for the target group and very high levels of task correctness. In contrast, both groups achieved only low values in task specificity of around 0.4, indicating that students in both groups had difficulty providing sufficient information for tasks to be solvable. Apart from that, we found significant differences in task clarity ($p = 0.04$) with a small effect size of Cohen’s $d = 0.38$ and task context ($p = 6 \times 10^{-5}$) with a large effect size of $d = 1.04$ between the two groups. In both cases, students who worked with the textbook achieved higher scores.

Table II shows some examples of tasks developed by students in the ChatGPT and in the textbook group. In the first example, the tasks from both groups reached an overall high quality in the ratings in the eight categories, e.g., they are correct, clear, concise, embedded in a meaningful context, and map the target concept. In the second example, the task in the ChatGPT group misses the time required for a car to accelerate from a standstill to a speed of 100 km/h, thus it is not sufficiently specified as the acceleration of the car cannot be determined. Similarly, in the second task in the textbook group, the participant mentioned that there is no friction occurring but failed to specify if another force is acting on the body. In the third example, both groups’ tasks lack context.

Furthermore, we investigated whether the scores achieved by students in each group were related to their prior knowledge. The intervention group participants achieved an average FCI score of 0.62 ± 0.05 , while the control group participants achieved an average FCI score of 0.69 ± 0.05 (overall average: 0.65 ± 0.05), but the difference between

TABLE II. Specific examples of tasks generated using ChatGPT 3.5 or a textbook that demonstrate high quality, lack of specificity, or missing context.

Aspect	ChatGPT 3.5	Textbook
Overall high quality	A car accelerates from an initial speed of 20 m/s to a final speed of 60 m/s in a period of 10 s. Calculate the acceleration of the car within this time period assuming that the acceleration is constant.	Justify why space debris in Earth orbit can be very dangerous for a space station like the ISS. Argue with Newton’s first law.
Lack of specificity	A car accelerates from a standstill to a speed of 100 km/h. 1. Explain the difference between the physical quantities acceleration and velocity and explain the units. 2. Calculate the acceleration of the car in m/s^2 .	A body moves frictionless with the velocity v_0 . What is the velocity of the body after 10 s? Justify your decision using Newton’s laws.
Missing context	How does Newton’s 1st law describe the behavior of a body in terms of its motion or rest?	How does the average velocity of a constant positive acceleration process from a standstill relate to the terminal velocity? What tools can you use to do this?

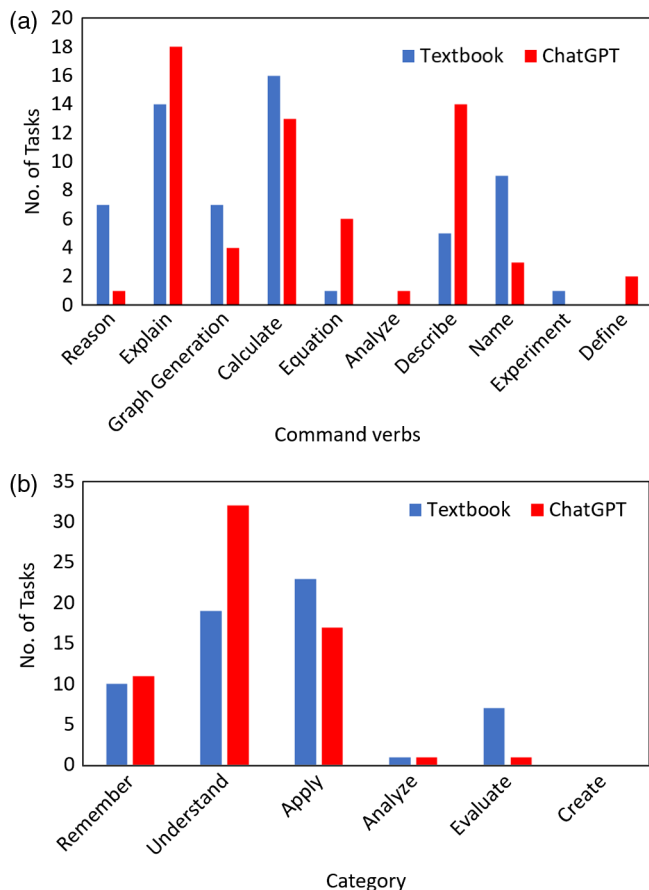


FIG. 2. Type of tasks created either with the help of ChatGPT or a textbook in terms of (a) used command verbs in the tasks and (b) the categories in the revised Bloom's taxonomy [65].

the two groups was not significant ($p > 0.05$). To this end, we performed a logistic regression between students' FCI scores and the ratings of the tasks they created in the first six categories, which were rated on a scale from 0 to 1. We found no significant relationship between FCI scores and ratings in the first six categories. Additionally, we performed a linear regression between students' FCI scores and the ratings of the last two categories, which were rated on a scale from 1 to 5. We also found no linear relationship between these two quantities. Consequently, in this study, prior knowledge had no significant influence on the quality of the tasks created.

Additionally, we analyzed the type of tasks that were created by both groups based on the command verbs that were used in the tasks. We added a table in the Appendix that shows how certain command verbs or interrogative pronouns were summarized into one category. We found that most of the tasks in both groups asked students to calculate or explain a phenomenon. In the ChatGPT group, there was also a significant number of tasks that required students to describe. The students in the textbook group created notably more reasoning tasks and list tasks (with the command verb "name"), whereas the ChatGPT group created more items with the task to provide an equation or a

description (see Fig. 2). Both groups also created a similar amount of graph generation tasks. Overall, there were no ranking tasks in both groups [66]. Translated to the revised version of Bloom's taxonomy [65], it is noticeable that the ChatGPT group creates more tasks in the "Understand" category, whereas the students who worked with a textbook created more tasks in the categories "Apply" and "Evaluate."

B. Adaptations to textbook and ChatGPT-generated tasks

In addition to evaluating the quality of tasks in the eight categories, we also assessed the changes students made to textbook tasks compared to ChatGPT-generated tasks. We evaluated whether these changes led to an improvement (+1, positive modification) or a decrease (−, negative modification) in task quality. Overall, one participant in the ChatGPT group did not use ChatGPT as a support tool for one task (2% of all tasks in this group) and instead created the task without any support. In contrast, 42 tasks (81% of all tasks in this group) were developed by participants in the textbook group without using the textbook as a resource. Consequently, participants in the textbook group used the textbook for ten tasks (19% of all tasks in this group) and adapted the given questions. In comparison, participants in the ChatGPT group used ChatGPT for 50 tasks (98% of all tasks in this group) and adapted ChatGPT tasks in 12 cases (24% of all tasks in this group). This means that students in the ChatGPT group used 38 tasks (75% of all tasks in this group) as provided by ChatGPT.

Figure 3 shows the number of improvements students made in seven categories. A positive value indicates that

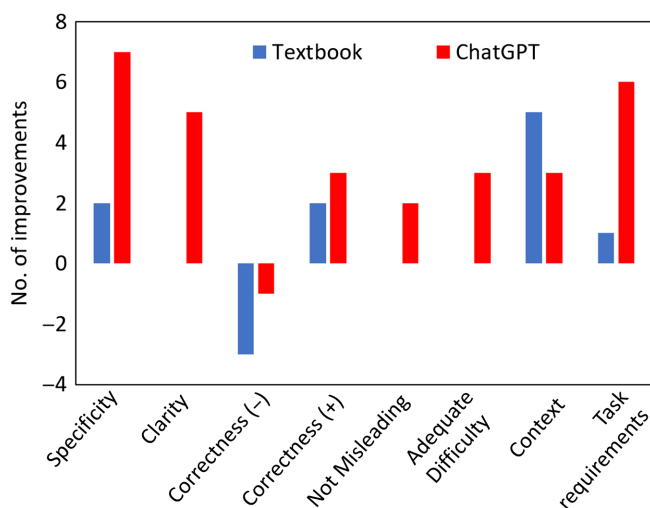


FIG. 3. Number of improvements made by students to textbook or ChatGPT-generated tasks within the seven categories. The category "correctness" was split up into negative [correctness (−)] and positive [correctness (+)] modifications. Negative modifications indicate that a participant's adaptation decreased the task correctness.

TABLE III. Examples of participants' modifications of the tasks that were either developed by ChatGPT or extracted from the textbook.

Example no.	Group	ChatGPT-generated or textbook task	Adapted task by the participant
1	ChatGPT	A spaceship has a mass of 2000 kg and is in orbit around the planet X. The gravitational force of planet X on the spacecraft is 20 000 N. <ol style="list-style-type: none"> 1. Calculate the acceleration of the spacecraft with respect to planet X. 2. What is the velocity of the spacecraft in orbit? 3. What is the size of the trajectory of the spacecraft when the speed of the spacecraft is changed? 	A spaceship has a mass of 2000 kg and is in orbit around the planet X. The gravitational force of planet X on the spacecraft is 20 000 N. <ol style="list-style-type: none"> 1. Calculate the acceleration of the spacecraft with respect to planet X. 2. What is the velocity of the spacecraft in orbit? 3. Describe the trajectory of the spacecraft when the speed of the spacecraft is changed.
2	ChatGPT	A car accelerates from an initial speed of 20 m/s to a final speed of 60 m/s in 10 s. Calculate the acceleration of the car.	A car accelerates from an initial speed of 20 m/s to a final speed of 60 m/s in 10 s. Calculate the acceleration of the car within this time assuming that the acceleration is constant.
3	Textbook	The reason for the propulsion of a rocket, as for an inflated balloon that you release, is the interaction principle. Explain the statement.	In order for a Boeing 747 to get enough thrust to fly, Newton's law of interaction works much like letting go of an inflated ball of air. Explain this principle in the context of the Boeing 747 flying (or taking off).
4	Textbook	Selma ($m_1 = 60$ kg) takes a running jump onto a stationary sled ($m_2 = 12$ kg). Both then continue to travel at $u = 2.5$ m/s. <ol style="list-style-type: none"> 1. Calculate the velocity of Selma, with which she landed on the sled. 2. Prove that the kinetic energy is not conserved during the process. 	Selma ($m_1 = 60$ kg) takes a running jump onto a moving sled ($m_2 = 12$ kg, $v_2 = 2$ m/s). Both continue to travel together. <ol style="list-style-type: none"> 1. Calculate the joint final velocity, assuming no friction and no gradient. 2. Calculate the joint momentum at impact.

more changes led to task improvement, while a negative value indicates that more changes led to a decrease in the task quality. The first six categories were identical to those used in the quality rating, with the exception of the category regarding relevance to map the concept, which was omitted. The overall quality was also omitted and a new category reflecting changes in task requirements was added. In this new category, we evaluated whether prospective physics teachers adapted to the type of the task or what the task the students required to do. Table III presents four instances of such task modifications made by participants using either ChatGPT-generated tasks or tasks extracted from the textbook.

In Table III, example 1 illustrates an adaptation of the task type by a participant in the ChatGPT group. In this instance, the prospective physics teacher failed to recognize that the

velocity of the spacecraft in part 2 of the task was indeterminate given the available information. Nevertheless, the participant shifted their focus to modifying the task requirements in part 3 from a quantitative to a descriptive task. Figure 3 shows that participants working with ChatGPT made a greater number of improvements compared to those in the textbook group. This is expected as participants in the textbook group primarily developed tasks independently of the textbook, resulting in no recorded changes to textbook tasks. In contrast, students in the ChatGPT group primarily altered task specificity and requirements. For instance, in Table III, example 2 demonstrates a modification of task specificity by a participant in the ChatGPT group after ChatGPT generated the task. In this instance, it is impossible to determine the acceleration of the car due to the ambiguity surrounding whether its acceleration is constant. The

participant recognized this limitation and subsequently modified the task accordingly, resulting in a well-defined and specific final task.

The textbook group primarily focused on modifying the context of the tasks. In Table III, example 3 illustrates that both tasks are correct with only a minor alteration in context. However, there were instances where participants made incorrect changes to textbook tasks. For instance, in Table III, example 4 demonstrates a failure to address the fact that the task does not focus on the relationship between velocity and acceleration, but rather on the conservation of momentum during an inelastic collision. Additionally, the participant neglected to include Selma's initial velocity. As a result, the final velocity in part 1 and the joint momentum in part 2 of the task can no longer be determined. Moreover, as mentioned above, the students in the textbook group only used the textbook in ten tasks, and in three cases (30%), these changes led to incorrect tasks. In comparison, in the ChatGPT group, the modifications only led once to an incorrect task (2%).

C. Usability of ChatGPT 3.5 to create physics tasks

To evaluate how participants perceived the output they received from ChatGPT and the ease or difficulty of receiving a response from the system of sufficient quality, we asked participants to rate the usability, perceived usefulness, and output (Fig. 4). On average, participants rated system usability at 4.1, indicating agreement that the system is easy to use, functions are well integrated, and they were confident in using it. Regarding perceived usefulness in developing physics tasks, participants were neutral (average value 3.24). This implies that they did not feel more productive or effective but were also not slowed down in their efforts to develop physics tasks. This scale also included a question

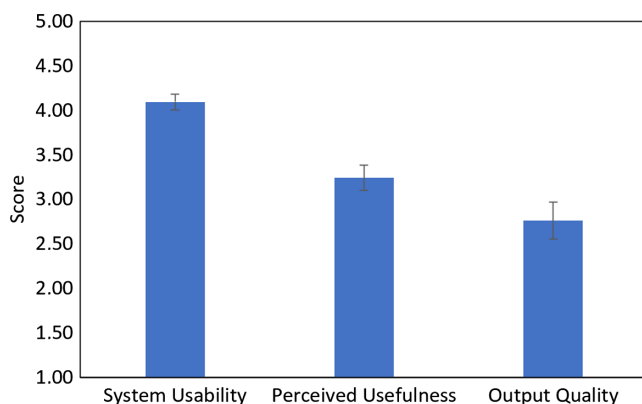


FIG. 4. Rating of participants regarding the ChatGPT's usability, its perceived usefulness, and its output quality on a five-point Likert scale: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree. For the analysis, the items with negative phrasings have been switched so that high values reflect a positive statement of the students.

about whether the time until receiving a response from ChatGPT complicated system use. On average, students did not agree (average value 2.09) that the response time complicated system use. Additionally, participants rated ChatGPT's output quality as neutral but with the lowest average value of 2.76, indicating that participants tended to be critical of the system's output.

V. DISCUSSION

The aim of the manuscript was the study of three research questions. The first research question targeted the quality of tasks generated by participants using ChatGPT compared to tasks created by students using a textbook. Both groups achieved very high task correctness, exceeding the average score in the concept test. Assuming that the difficulty of generating a correct assessment task is comparable to the average difficulty in the half-length force concept inventory, this finding suggests that ChatGPT is able to compensate for conceptual difficulties that prospective physics teachers may have, reducing their likelihood of translating to assessment tasks.

Apart from that, it was noticeable that both groups achieved low task specificity, indicating that information was missing for the pupil to be able to complete the task. This difficulty among prospective physics teachers could not be compensated by either ChatGPT 3.5 or the textbook. It would be interesting to see whether future versions of large language models can overcome this shortcoming. In any case, it must be given greater consideration in teacher training. Additionally, tasks created by participants using ChatGPT were significantly less frequently embedded in an appropriate context. However, some tasks generated by ChatGPT had context if participants explicitly prompted it. Therefore, the low score in the ChatGPT group was not caused by the system but rather by students' insensitivity to the value of the context in a physics assessment task for 10th-grade high school students. Students in the textbook group achieved a high score in this category regardless of whether they modified a given textbook task or developed a physics task without inspiration from a textbook task.

The second research question focused on the number and type of changes students made to ChatGPT-generated and textbook tasks. It is relevant to note that 81% of tasks in the textbook group were not developed based on a given textbook task but were instead created by students without using this resource. This observation suggests that it was less effort for students in this group to include tasks from memory or their own efforts rather than going through the tasks in the textbook. Such behavior likely depends on students' familiarity with the textbook and how many assessment tasks in the same context they had already created in which they implemented their ideas. The results show that some participants were aware of low specificity in tasks created by ChatGPT and improved it, but most participants did not have this awareness.

Furthermore, the results demonstrate that some students modified preexisting textbook tasks which resulted in incorrect tasks. Although this observation occurred only in three tasks (6% of all tasks in this group), and the correctness of all tasks in this group was comparable to the ChatGPT group, such a shortcoming may potentially be compensated by the additional use of large language models to solve a given task before giving it to students.

The third research question addressed the usability, perceived usefulness, and quality of ChatGPT’s output. We found that participants rated the usability of ChatGPT 3.5 high but judged the quality of physics tasks generated by ChatGPT slightly below a neutral level. It would be interesting to see if the judgment of output quality improves with future versions of ChatGPT and whether this judgment would change if students received training to create prompts for ChatGPT to generate physics tasks.

In future research, it would be interesting to study how in-service teachers with different experience levels use and rate the usability as well as the output quality of ChatGPT for the development of assessment tasks.

VI. CONCLUSION

In this study, we evaluated the quality of tasks developed by prospective physics teachers using ChatGPT 3.5 compared to using a textbook. The correctness, the frequency of tasks with an adequate difficulty, and the overall quality of the tasks created by ChatGPT were comparable to those created by participants who could use a textbook. This is a remarkable achievement for a large language model, as developing physics assessment tasks using textbooks is the current state-of-the-art and, in general, developing physics tasks requires teachers to have assessment knowledge, conceptual knowledge, and problem-solving skills [67]. At the same time, the field of large language models is currently quickly evolving and they are likely to become more proficient in such discipline-specific exercises as task development soon.

We found that some shortcomings in tasks, such as specificity, could not be compensated by either the textbook or ChatGPT 3.5, and other aspects needed to be explicitly prompted by the user of ChatGPT, such as the context of the task.

In summary, we demonstrated that advanced large language models such as ChatGPT 3.5 can effectively support physics teachers in their common practice of developing assessment tasks. It would be beneficial to include specific training for physics teachers in education programs to design prompts for the effective use of large language models.

ACKNOWLEDGMENTS

This research was supported by LMUexcellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder.

APPENDIX A: TYPES OF TASKS

To analyze the categories the tasks were framed in and the requirements of the tasks for the students, we identified the command verbs in the tasks or, in case no command verb was used, we assigned the phrasings in the tasks to a certain command verb (Table IV) in line with previous assignments of command verbs to categories in Bloom’s taxonomy [65].

APPENDIX B: LIMITATIONS OF ChatGPT

The dialogue in Fig. 5 contains statements of ChatGPT 3.5 regarding its limitations on February 9. In comparison, ChatGPT 4.0 lists the limitations of ChatGPT 3.5 in Fig. 6 on February 27 and mentions the differences in limitations between ChatGPT 4.0 and 3.5 in Fig. 7.

TABLE IV. Assignments of task phrasing to command verbs and to categories in the revised version of Bloom’s taxonomy.

Category in Bloom’s taxonomy	Command verbs	Question phrasing
Remember	Define	“How do you define”
	Name	“Name,” “Which”
	Equation	“Provide an equation,” “How can ... be expressed in equation”
Understand	Describe	“Describe,” “How do you describe”
	Explain	“Explain,” “How do you explain,” “How does ... behave”, “How is the relation”, “How does ... change”, “How does ... depend on,” “What is”
Apply	Calculate	“Calculate,” “How do you calculate,” “What is”
Analyze	Graph generation	“Draw ... in a graph,” “Visualize,” “Make a diagram”
	Analyze	“Analyze”
Evaluate	Experiment	Experiment description with an assignment what to do
	Reason	“Reason,” “Justify,” “Why”

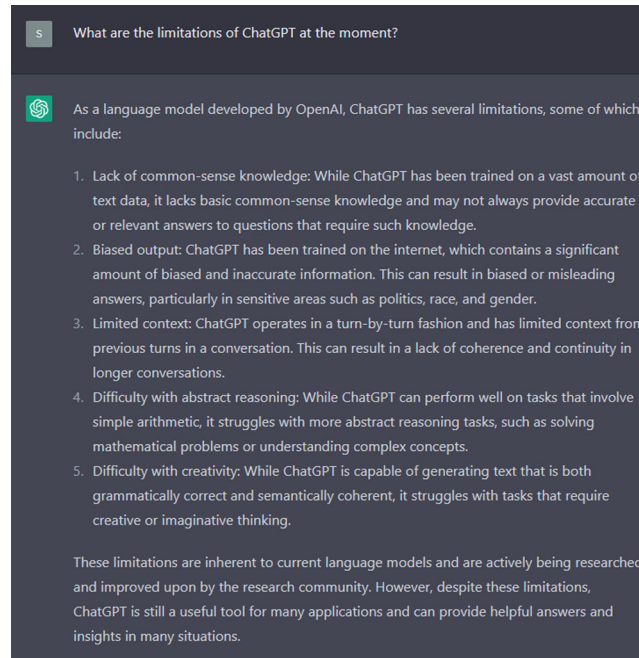


FIG. 5. Online dialogue with ChatGPT regarding its limitations on February 9, 2023.

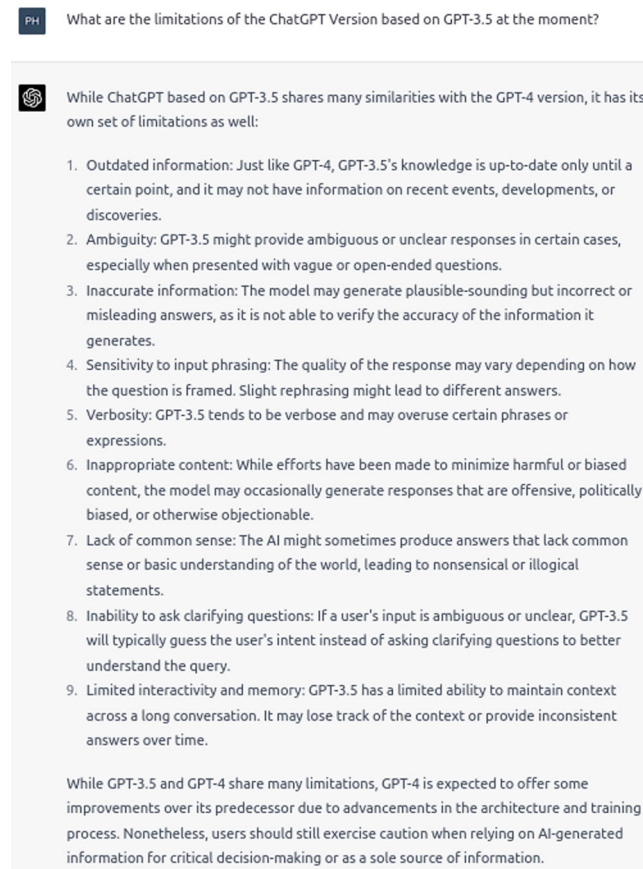


FIG. 6. Online chat with ChatGPT based on GPT4 about the limitations of its version based on GPT3.5 on February 27, 2023.

PH What are the differences between the limitations of the ChatGPT Version based on GPT-4 and the Version of the ChatGPT Version based on GPT-3.5 at the moment?

The limitations of ChatGPT versions based on GPT-4 and GPT-3.5 are largely similar, as both models share common issues inherent to their design and training. However, some differences may arise due to improvements in the GPT-4 architecture and training process. These differences could lead to variations in the extent or frequency of certain limitations:

1. Knowledge freshness: Although both versions have outdated information, GPT-4 has a more recent knowledge cutoff, potentially providing more up-to-date information compared to GPT-3.5.
2. Response quality: GPT-4 may provide better-quality responses or handle ambiguous questions more effectively due to improvements in architecture, training data, or model size. This could result in a reduction of some limitations, such as generating misleading or inaccurate information, though not necessarily eliminating them entirely.
3. Context understanding: GPT-4 might be better at maintaining context or providing more consistent answers over time compared to GPT-3.5, thanks to advancements in the architecture or training techniques.
4. Content safety: Efforts to minimize harmful or biased content could have been improved in GPT-4, leading to a potential decrease in the frequency of inappropriate content generated by the model.

It's important to note that while GPT-4 may offer some improvements over GPT-3.5, the limitations still exist to varying degrees. Users should continue to exercise caution when using either version of ChatGPT for critical decision-making or as a sole source of information.

FIG. 7. Online chat with ChatGPT based on GPT4 about the difference in limitations of its version based on GPT3.5 and GPT4 on February 27.

-
- [1] American Federation of Teachers, National Education Association. American Federation of Teachers, National Council on Measurement in Education & National Education Association (1990), Standards for teacher competence in educational assessment of students, *Educ. Meas.* **9**, 30 (1990).
- [2] J. Gess-Newsome, J. A. Taylor, J. Carlson, A. L. Gardner, C. D. Wilson, and M. A. Stuhlsatz, Teacher pedagogical content knowledge, practice, and student achievement, *Int. J. Sci. Educ.* **41**, 944 (2019).
- [3] P. W. Airasian, *Classroom Assessment: Concepts and Applications* (ERIC, Blacklick, OH, 2001).
- [4] C. A. Mertler, *Classroom Assessment Practices of Ohio Teachers* (ERIC, Bowling Green, OH, 1998).
- [5] R. J. Stiggins, Are you assessment literate?, *High Sch. Mag.* **6**, 20 (1999).
- [6] W. J. Popham, Seeking redemption for our psychometric sins, *Educ. Meas.* **22**, 45 (2003).
- [7] L. Eyal, Digital assessment literacy—The core role of the teacher in a digital environment, *J. Educ. Technol. Soc.* **15**, 37 (2012).
- [8] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günemann, E. Hüllermeier *et al.*, ChatGPT for good? On opportunities and challenges of large language models for education, *Learn. Individ. Diff.* **103**, 102274 (2023).
- [9] H. E. Fischer and A. Kauertz, Physics tasks, in *Physics Education* (Springer, New York, 2022), pp. 231–267.
- [10] J. Leisen, Qualitätsteigerung des Physikunterrichts durch Weiterentwicklung der Aufgabenkultur, *Math. Naturwiss. Unterr.* **54**, 401 (2001).
- [11] M. Pozas, P. Löffler, W. Schnotz, and A. Kauertz, The effects of context-based problem-solving tasks on students' interest and metacognitive experiences, *Open Educ. Stud.* **2**, 112 (2020).
- [12] P. Löffler, M. Pozas, and A. Kauertz, How do students coordinate context-based information and elements of their own knowledge? An analysis of students' context-based problem-solving in thermodynamics, *Int. J. Sci. Educ.* **40**, 1935 (2018).
- [13] K. Neumann, T. Viering, W. J. Boone, and H. E. Fischer, Towards a learning progression of energy, *J. Res. Sci. Teach.* **50**, 162 (2013).
- [14] J. F. Osborne, J. B. Henderson, A. MacPherson, E. Szu, A. Wild, and S.-Y. Yao, The development and validation of a learning progression for argumentation in science, *J. Res. Sci. Teach.* **53**, 821 (2016).
- [15] V. Villarroel, D. Boud, S. Bloxham, D. Bruna, and C. Bruna, Using principles of authentic assessment to redesign written examinations and tests, *Innovations Educ. Teach. Int.* **57**, 38 (2020).
- [16] P. Sinaga and S. Feranie, Enhancing critical thinking skills and writing skills through the variation in non-traditional writing task, *Int. J. Instr.* **10**, 69 (2017).
- [17] J. Kuhn and A. Müller, Context-based science education by newspaper story problems: A study on motivation and learning effects, *Perspect. Sci.* **2**, 5 (2014).

- [18] N. Akben, Effects of the problem-posing approach on students' problem solving skills and metacognitive awareness in science education, *Res. Sci. Educ.* **50**, 1143 (2020).
- [19] G. Chen, J. Yang, C. Hauff, and G.-J. Houben, LearningQ: A large-scale dataset for educational question generation, in *Proceedings of the International AAAI Conference on Web and Social Media* (2018), Vol. 12, <https://doi.org/10.1609/icwsm.v12i1.14987>.
- [20] S. Bhat, H. A. Nguyen, S. Moore, J. Stamper, M. Sakr, and E. Nyberg, Towards automated generation and evaluation of questions in educational domains, in *Proceedings of the 15th International Conference on Educational Data Mining* (2022), Vol. 701, <https://dev.stamper.org/publications/2022EDM-posters85.pdf>.
- [21] R. Rodriguez-Torrealba, E. Garcia-Lopez, and A. Garcia-Cabot, End-to-end generation of multiple-choice questions using text-to-text transfer transformer models, *Expert Syst. Appl.* **208**, 118258 (2022).
- [22] V. Raina and M. Gales, Multiple-choice question generation: Towards an automated assessment framework, [arXiv:2209.11830](https://arxiv.org/abs/2209.11830).
- [23] T. Rowland, P. Huckstep, and A. Thwaites, Elementary teachers' mathematics subject knowledge: The knowledge quartet and the case of Naomi, *J. Math. Teach. Educ.* **8**, 255 (2005).
- [24] L. S. Shulman, Paradigms and research programs in the study of teaching: A contemporary perspective, in *Handbook of Research on Teaching*, edited by M. C. Wittrock (Macmillan, London, 1986), pp. 3–36.
- [25] J. Gess-Newsome, J. Taylor, J. Carlson, A. Gardner, C. Wilson, and M. Stuhlsatz, Teacher pedagogical content knowledge, practice, and student achievement, *Int. J. Sci. Educ.* **41**, 944 (2017).
- [26] T. Kleickmann, D. Richter, M. Kunter, J. Elsner, M. Besser, S. Krauss, and J. Baumert, Teachers' content knowledge and pedagogical content knowledge, *J. Teach. Educ.* **64**, 90 (2012).
- [27] L. S. Shulman, Knowledge and teaching. Foundations of the new reform, *Harv. Educ. Rev.* **57**, 1 (1987).
- [28] R. Edwards, R. Hyde, M. C. O'Connor, and J. Oldham, The importance of subject knowledge for mathematics teaching: An analysis of feedback from subject knowledge enhancement courses, in *British Society for Research into Learning Mathematics Day Conference* (2015), <https://eprints.soton.ac.uk/387221/>.
- [29] Y. Lee, R. Capraro, and M. Capraro, Mathematics teachers' subject matter knowledge and pedagogical content knowledge in problem posing, *Int. Electron. J. Math. Educ.* **13**, 75 (2018).
- [30] E. Silver, On mathematical problem posing, *Lern. Math.* **14**, 19 (1994).
- [31] C. Bonotto, Artifacts as sources for problem-posing activities, *Educ. Stud. Math.* **83**, 37 (2013).
- [32] J. Cai and R. Leikin, Affect in mathematical problem posing: Conceptualization, advances, and future directions for research, *Educ. Stud. Math.* **105**, 287 (2020).
- [33] B. Koichu, G. Harel, and A. Manaster, Ways of thinking associated with mathematics teachers' problem posing in the context of division of fractions, *Instr. Sci.* **41**, 681 (2013).
- [34] T. Lowrie, Designing a framework for problem posing: Young children generating open-ended tasks, *Contemp. Issues Early Child.* **3**, 354 (2002).
- [35] S. Crespo, Learning to pose mathematical problems: Exploring changes in preservice teachers' practices, *Educ. Stud. Math.* **52**, 243 (2003).
- [36] S. Crespo and N. Sinclair, What makes a problem mathematically interesting? Inviting prospective teachers to pose better problems, *J. Math. Teach. Educ.* **11**, 395 (2008).
- [37] J. Cai and S. Hwang, Generalized and generative thinking in US and Chinese students' mathematical problem solving and problem posing, *J. Math. Behav.* **21**, 401 (2002).
- [38] E. A. Silver and J. Cai, An analysis of arithmetic problem posing by middle school students, *J. Res. Math. Educ.* **27**, 521 (1996).
- [39] J. Cai and S. Hwang, Learning to teach through mathematical problem posing: Theoretical considerations, methodology, and directions for future research, *Int. J. Educ. Res.* **102**, 101391 (2020).
- [40] OpenAI, Openai chat (2023).
- [41] L. Floridi and M. Chiriatti, GPT-3: Its nature, scope, limits, and consequences, *Minds Mach.* **30**, 681 (2020).
- [42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, [arXiv:1810.04805](https://arxiv.org/abs/1810.04805).
- [43] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, [arXiv:1907.11692](https://arxiv.org/abs/1907.11692).
- [44] M. Zhu, O. L. Liu, and H.-S. Lee, The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing, *Comput. Educ.* **143**, 103668 (2020).
- [45] J. P. Bernius, S. Krusche, and B. Bruegge, Machine learning based feedback on textual student answers in large courses, *Comput. Educ.* **3**, 100081 (2022).
- [46] R. Abdelghani, Y.-H. Wang, X. Yuan, T. Wang, H. Sauz on, and P.-Y. Oudeyer, GPT-3-driven pedagogical agents for training children's curious question-asking skills, [arXiv:2211.14228](https://arxiv.org/abs/2211.14228).
- [47] S. Sarsa, P. Denny, A. Hellas, and J. Leinonen, Automatic generation of programming exercises and code explanations using large language models, in *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1* (2022), pp. 27–43, <https://doi.org/10.1145/3501385.3543957>.
- [48] Q. Jia, J. Cui, Y. Xiao, C. Liu, P. Rashid, and E. F. Gehring, All-in-one: Multi-task learning Bert models for evaluating peer assessments, [arXiv:2110.03895](https://arxiv.org/abs/2110.03895).
- [49] J. Jeon, Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis, *Comput. Assist. Lang. Learn.* **0**, 1 (2021).
- [50] M. Bao, Can home use of speech-enabled artificial intelligence mitigate foreign language anxiety—investigation of a concept, *Arab World Eng. J. (AWEJ) Special Issue on CALL* (2019), [10.2139/ssrn.3431734](https://doi.org/10.2139/ssrn.3431734).
- [51] T.-Y. Tai and H. H.-J. Chen, The impact of Google assistant on adolescent EFL learners' willingness to communicate, *Interact. Learn. Environ.* **31**, 1485 (2023).

- [52] S. Polak, G. Schiavo, and M. Zancanaro, Teachers' perspective on artificial intelligence education: An initial investigation, in *Proceedings of CHI Conference on Human Factors in Computing Systems Extended Abstracts* (2022), pp. 1–7, <https://doi.org/10.1145/3491101.3519866>.
- [53] R. Chocarro, M. Cortinas, and G. Marcos-Matás, Teachers' attitudes towards chatbots in education: A technology acceptance model approach considering the effect of social language, bot proactiveness, and users' characteristics, *Educ. Stud. Math.* **49**, 295 (2023).
- [54] W. Holmes, M. Bialik, and C. Fadel, *Artificial Intelligence in Education* (Globethics Publications, Geneva, 2020).
- [55] OpenAI, Introducing ChatGPT, "<https://openai.com/blog/chatgpt#fn-1>" (2022).
- [56] L. Gao, J. Schulman, and J. Hilton, Scaling laws for reward model overoptimization, [arXiv:2210.10760](https://arxiv.org/abs/2210.10760).
- [57] OpenAI, GPT-4: Advanced ai language model (2023) [accessed March 28, 2023].
- [58] OpenAI, GPT-4 technical report, [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
- [59] R. Dietrich and F. Finkenberger, *Physik 10, Gymnasium Bayern* (C.C. Buchner, Bamberg, 2022).
- [60] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig, Dividing the force concept inventory into two equivalent half-length tests, *Phys. Rev. ST Phys. Educ. Res.* **11**, 010112 (2015).
- [61] M. Schaefer, B. Xu, H. Flor, and L. G. Cohen, Effects of different viewing perspectives on somatosensory activations during observation of touch, *Hum. Brain Mapp.* **30**, 2722 (2009).
- [62] V. Venkatesh and F. D. Davis, A theoretical extension of the technology acceptance model: Four longitudinal field studies, *Manage. Sci.* **46**, 186 (2000).
- [63] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.19.020128> for user experience Likert scale.
- [64] J.R. Landis and G.G. Koch, The measurement of observer agreement for categorical data, *Biometrics* **33**, 159 (1977).
- [65] L.W. Anderson and D.R. Krathwohl, *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives* (Longman, New York, 2021).
- [66] T.L. O'Kuma, D.P. Maloney, and C.J. Hieggelke, *Ranking Task Exercises in Physics* (Prentice Hall, Upper Saddle River, NJ, 2000).
- [67] S.M. Brookhart, Educational assessment knowledge and skills for teachers, *Educ. Meas.* **30**, 3 (2011).