Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning

Tong Wan[*] and Zhongzhou Chen[†]

Department of Physics, University of Central Florida, Orlando, Florida 32816, USA

(Received 10 November 2023; accepted 23 May 2024; published 13 June 2024)

Instructor's feedback plays a critical role in students' development of conceptual understanding and reasoning skills. However, grading student written responses and providing personalized feedback can take a substantial amount of time, especially in large enrollment courses. In this study, we explore using GPT-3.5 to write feedback on students' written responses to conceptual questions with prompt engineering and few-shot learning techniques. In stage I, we used a small portion ($n = 20$) of the student responses on one conceptual question to iteratively train GPT to generate feedback. Four of the responses paired with human-written feedback were included in the prompt as examples for GPT. We tasked GPT to generate feedback for another 16 responses and refined the prompt through several iterations. In stage II, we gave four student researchers (one graduate and three undergraduate researchers) the 16 responses as well as two versions of feedback, one written by the authors and the other by GPT. Students were asked to rate the correctness and usefulness of each feedback and to indicate which one was generated by GPT. The results showed that students tended to rate the feedback by human and GPT equally on correctness, but they all rated the feedback by GPT as more useful. Additionally, the success rates of identifying GPT's feedback were low, ranging from 0.1 to 0.6. In stage III, we tasked GPT to generate feedback for the rest of the students' responses ($n = 65$). The feedback messages were rated by four instructors based on the extent of modification needed if they were to give the feedback to students. All four instructors rated approximately 70% (ranging from 68% to 78%) of the feedback statements needing only minor or no modification. This study demonstrated the feasibility of using generative artificial intelligence (AI) as an assistant to generate feedback for student written responses with only a relatively small number of examples in the prompt. An AI assistant can be one of the solutions to substantially reduce time spent on grading student written responses.

DOI: 10.1103/PhysRevPhysEducRes.20.010152

I. INTRODUCTION

Developing conceptual understanding is one of the key learning goals in many introductory physics courses. There has been extensive research investigating students' conceptual understanding and reasoning in introductory physics courses [1]. This body of research is often aligned with constructivist learning theory, which assumes that prior knowledge sets the foundation for new knowledge. Thus, effective teaching often requires eliciting and building on students' existing ideas [2].

Conceptual questions on homework assignments in free-response format provide students with great opportunities to practice articulating reasoning and justifying conclusions. In addition, the free-response format allows instructors to gain deep insights into student reasoning so that instructors can provide useful feedback and refine instruction when needed.

Instructor feedback plays a critical role in student learning. According to Ericsson's framework of "deliberate practice" [3], frequent feedback from an expert in addition to repeated and targeted practice is essential in acquiring expert performance. Moreover, empirical studies have confirmed that frequent feedback leads to substantial learning gains [4,5].

Although free-response questions provide opportunities for deliberate practice, grading, and providing personalized feedback to students can be extremely time-consuming, especially in large enrollment courses. Institutions may not have sufficient resources to hire enough teaching assistants to grade homework assignments in large enrollment courses. Even when there is a sufficient number of graders, the quality of grading and feedback can be inconsistent among individual graders without proper training.

[*]Corresponding author: tong.wan@ucf.edu
[†]Corresponding author: zhongzhou.chen@ucf.edu

In response to these challenges and constraints, instructors may choose to give students credit based on completion rather than on correctness of response. However, without personalized feedback, students may not realize their mistakes even with the instructor's solutions provided [6], let alone how to improve.

One possible way to overcome the challenge of limited resources and providing feedback on free-response questions is to offload some of the grading tasks to generative artificial intelligence (GenAI), in particular large language models (LLMs). There has been a rapid recent increase in innovations using GenAI technology in education [7]. In particular, LLMs have been applied to a wide range of areas, such as personalized learning, intelligent tutoring, content creation, and essay grading (for a recent comprehensive review, see Ref. [8]).

Earlier efforts using LLMs to grade or provide feedback to short answers or essays often utilize smaller pretrained language models such as Bidirectional Encoder Representations from Transformers (BERT) [9–12]. However, those earlier models require a significant amount of fine-tuning for satisfactory performance, a process that demands a significant amount of technological expertise, computational resources, and large amounts of existing labeled text data (i.e., text data that are already classified). Those requirements prevented the technology from being widely adopted in teaching. With the rapid development in capability and enhanced accessibility of LLMs, recent studies have focused on using more powerful LLMs, such as Generative Pretrained Transformer 3 (GPT-3). Models such as GPT are capable of achieving high performance with only simple prompt engineering (i.e., the process of developing and refining the prompt for LLMs to get the desired output [13]).

We are only aware of one study that uses GPT to provide feedback on physics tasks. Steinert *et al.* [14] examined ChatGPT's capacity to provide formative feedback to students on a task that asks students to explain the underlying physics principle of an experiment. The study demonstrated that ChatGPT can be prompted to provide feedback that focuses on different theoretical aspects of learning, such as cognition, metacognition, and motivation. However, this study included only a single example of student response, and it did not evaluate the quality of feedback.

In this paper, we report our initial attempt at exploring the feasibility of using GPT-3.5 as a grading assistant for providing feedback on students' written responses to one physics conceptual question. By combining our existing knowledge of students' preconceptions with a simple prompt-engineering technique, we tested if it would be possible to turn GPT-3.5 into an adequate grading assistant with only a few examples of graded student responses. This approach could potentially save a significant amount of instructor grading effort and provide useful feedback to students.

II. PRIOR RESEARCH ON USING LLMS IN EDUCATION

LLMs such as GPT are being created to process and generate natural language. They are complex neural networks with billions of parameters that are pretrained using a large corpus of text. A human user interacts with an LLM by inputting a piece of text, which is frequently referred to as a "prompt" in AI literature. The LLM generates text output by predicting the most likely words to follow the prompt.

Multiple prior studies have tested LLM's capacity to answer disciplinary questions. Kortemeyer [15], for example, tasked ChatGPT to complete all the course assignments like a human student would in an introductory physics course. With all the assignments graded and counted toward a final course grade, Kortemeyer found that ChatGPT could barely pass the course. Dahlkemper *et al.* [16] evaluated students' perceptions of linguistic quality and scientific accuracy of ChatGPT responses to physics comprehension questions. Students' ratings of linguistic quality were essentially the same between ChatGPT and human-written solutions, but the ratings of scientific accuracy were much higher for human-written solutions than for ChatGPT responses. Moreover, the perceived difference in scientific accuracy decreased as the difficulty level of the questions increased.

Other studies have examined the opportunities and limitations of LLMs to generate disciplinary questions. Küchemann *et al.* [17] compared the quality of physics questions developed by preservice teachers who used ChatGPT and those who used textbooks. They found that the correctness and specificity (i.e., all relevant information to answer a question is present) of the questions were about the same between the two groups, but the clarity was much better in the questions developed by students who used a textbook. Additionally, pre-service teachers who used ChatGPT reported that ChatGPT is easy to use, but they felt neutral about the usefulness and quality of ChatGPT's output.

A. Prompt engineering

The quality of the output from an LLM may vary significantly depending on the quality of the prompt. Prompt engineering is the process of developing and refining a prompt in order to get the desired output [13]. For example, Polverini and Gregoricic have shown that the performance of GPT on conceptual physics tasks can be significantly improved by using prompt engineering techniques [18].

As detailed in Polverini and Gregoricic, designing an effective prompt requires some understanding of how LLMs work, including their strengths and weaknesses. Since GPT is sensitive and responsive to context, the performance can be improved by providing the relevant

context for response in the prompt. Providing context often includes specifying the domain (e.g., specific topics or concepts) and specifying how to act (e.g., act like a physics teacher or an undergraduate student). However, Polverini and Gregoricic also cautioned that LLMs' context sensitivity and responsiveness can also be seen as a weakness. Shi *et al.* demonstrated that an LLM can be distracted by excess details resulting in decreased performance [19].

In addition, it has been widely documented that LLMs can sometimes generate outputs that are factually incorrect or contextually implausible (which is termed "hallucination" in AI research) [20–22]. Hallucination in LLMs can be effectively reduced by either employing better prompting engineering techniques or through domain-specific fine-tuning of the model (see below for definition). Yet because the output of LLMs is probabilistic, hallucinations of LLMs cannot be completely prevented, which makes it critical to establish human-in-the-loop procedures to minimize its potential negative impact [8].

Moreover, research in other disciplines also indicates that prompt engineering is an essential skill for learners. Woo *et al.* [23] explored English as a foreign language (EFL) students' prompt engineering pathways to writing using ChatGPT. The results suggest that prompt engineering is an important emergent skill for EFL students to improve their writing. Heston and Khun discussed the necessities challenges and concerns of using prompt engineering techniques in medical education [24].

B. Few-shot learning or prompting

Few-shot learning [25] is a framework in machine learning. It allows the pretrained model to learn the underlying pattern from a few examples so that the model can generalize to new scenarios. Few-shot learning is often achieved by including a few examples in the prompt to demonstrate how LLMs should respond to a similar task. A notable recent example of using few-shot prompting in the science education context is by Zong and Krishnamachari [26], who used few-shot prompting to train GPT to solve math problems. They found that GPT was able to solve the problems with high accuracy, and the accuracy improved with an increased number of examples.

C. Chain-of-thought prompting

When the task requires multiple intermediate steps, chain-of-thought prompting [27] can be used to improve the accuracy of the LLMs' response. A chain-of-thought prompt instructs an LLM to first generate a chain of intermediate reasoning steps before it generates the final answer. Chain-of-thought prompting can be combined with few-shot prompting to further enhance the LLM's performance. When combined, the examples provided in the prompt should include a chain of necessary intermediate reasoning steps, such as intermediate steps toward solving a math word problem.

D. Fine-tuning

Fine-tuning is the process in which the parameters of an existing pretrained LLM model are updated based on a large corpus of domain-specific data, for the purpose of better performance on tasks within that specific domain [28]. For example, Latif and Zhai [29] found that a fine-tuned GPT is effective at automatically scoring student written responses to science questions. Zong and Krishnamachari [26] found that on math problem generation tasks, a fine-tuned GPT model outperformed the native GPT model plus prompt engineering methods. However, fine-tuning can be expensive as it requires large amounts of domain-specific, labeled data.

III. RESEARCH QUESTIONS

In this study, we use few-shot prompting to task GPT to generate feedback on students' written responses to one physics conceptual question. We investigate both student researchers' and physics instructors' perceptions of the quality of GPT-generated feedback. For students, we are interested in how they compare AI-generated feedback to human-written feedback, in terms of both perceived correctness and perceived usefulness. We are also interested in whether students can notice if a feedback statement is generated by AI. For physics instructors, we are investigating their perceptions of how much modification to the feedback, if any, is needed before they deem the feedback ready for students. Answering those questions could provide insights into the extent to which an AI-based assistant could potentially save instructors' time and effort on grading and writing feedback. Specifically, we address the following three research questions (RQs):

RQ1: How do student researchers rate the correctness and usefulness of GPT-generated feedback messages on student written responses to a physics conceptual question?

RQ2: Can student researchers distinguish between AI-generated and human-written feedback?

RQ3: How do instructors rate the level of editing needed to make GPT-generated feedback messages satisfactory?

IV. METHODS

A. GPT-3.5 turbo complete mode

We used GPT-3.5 turbo in "complete" mode through Azure OpenAI Studio. The complete mode is different from the more popular "chat" mode used in applications such as ChatGPT. In complete mode, GPT functions by treating the input prompt as an unfinished piece of text or a document and generates output by predicting the most probable words and sentences that would follow the prompt text. The complete mode is well suited for tasks that require a single, well-structured response following instructions or prior examples. In contrast, the chat mode is optimized for multiple rounds of conversation with a human, in which

GPT treats the prompt as a transcript of a conversation with a human and tries to predict the response to the human. In chat mode, GPT is much more likely to generate text such as "Sure! Here is what you requested." Since our purpose is to task GPT to only generate feedback on individual student written responses rather than having a conversation with the grader, the complete mode was chosen for our study.

B. Question selection and feedback design

The student responses to the one conceptual question were collected from an introductory physics course taught by the first author with 99 students. The course was taught in studio mode, in which lecture, recitation or tutorial, and lab were integrated. During one class meeting, students were tasked to complete a tutorial that was adapted from the University of Maryland Open Source Tutorials [30]. The tutorial targeted Newton's second and third laws for a multiobject system. Students received credit for completion, but their answers were not graded for correctness. Students submitted their work (either a scan of their handwritten work or work done in an electronic file) through an online portal. A total of 85 student responses were collected from the class. The first author (also the instructor of record) typed out students' written responses before they were provided to GPT.

Figure 1 shows three questions A, B, and C from the Maryland tutorial, in which question C was chosen for the current study. The scenario involves a student pushing two adjacent boxes with a force of 200 N. The first two questions asked students to consider whether the accelerations of the boxes are the same and then calculate the acceleration of the boxes. Question C asked whether the contents of box B were in danger of breaking given that they would break if the box experienced a force greater than 200 N. The question explicitly asked students not to do any calculations but to answer intuitively and explain their reasoning.

In the tutorial, question C was designed to elicit students' preconceptions about forces, while the remainder of the tutorial guided students to answer question C using Newton's second and third laws without doing calculations. Specifically, it prompted students to draw a free-body diagram for each box and then asked students to tentatively assume that the force exerted by box A on box B equals 200 N and see where the assumption leads. By working through the tutorial, students were expected to recognize that this assumption would lead to the incorrect and inconsistent conclusion of the net force on box A (as well as its acceleration) being zero, and thus they should reject this assumption. At the end of the tutorial, students were asked to refine their intuition about question C.

We chose question C because it elicited students' ideas as to whether a force can be "transmitted." We expect students' ideas to be rich in variety and thus the question is well suited as a test case for a potential GenAI-based grading assistant that can provide feedback.

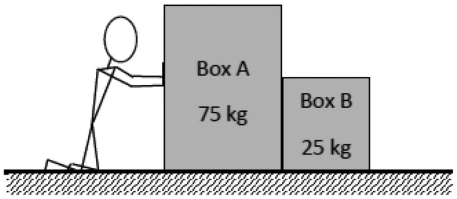
The feedback that we intended for GPT to generate includes a judgment statement on whether the students' conclusion and reasoning are correct. If either the conclusion or reasoning was incorrect or missing, then the feedback will directly point out what was incorrect and give a hint toward an alternative direction of thought for the students to consider, which is the acceleration of the boxes (see the Appendix).

It is worth pointing out that this type of feedback is not designed to be given to students who are in the middle of completing the tutorial. Rather, a more likely use case would be when a question similar to question C is given again on following exams or homework, which will serve as an assessment of whether students have developed intuitions that are aligned with Newtonian physics after having completed the tutorial. Therefore, we designed the feedback to directly hint at Newtonian physics-aligned reasoning, rather than trying to attend to and build on students' existing ideas in their response. Another reason for this choice is that usually a multiround dialogue

A student pushes two boxes, one in front of the other, as shown in the diagram. Box A has mass 75 kg, while box B has mass 25 kg. Fortunately for the student, the boxes are mounted on tiny rollers and slide with negligible friction. The student exerts a 200 N horizontal force on box A.

A. Without doing any calculations, state whether the acceleration of block A is greater than, less than, or equal to that of block B. How do you know?

B. Using any method you want, find the acceleration of the blocks. (Hint: It's possible to do this quickly and the blocks move together.)



C. Box B contains kitchen stuff, including some poorly packed glassware that might break if the force pushing on the side of the box approaches 200 newtons. Recall that the student pushes on box A with a force of 200 newtons. Is that force "transmitted" to box B? In other words, is the glassware in the box in danger of breaking? Don't do any calculations; answer intuitively, and explain your thinking.

FIG. 1. The first three questions in the tutorial. We used students' responses to question C (in bold) in our study.

between the instructor and the student is required to build on students' existing ideas (for example, see studies based on the resource framework [31]). This type of multiround conversation is significantly more complicated to develop using GenAI and out of the scope of the current exploratory study.

C. Study design

The study consisted of three stages. In stage I, four student responses and human-written feedback pairs were used as examples to develop a prompt for the GPT model using few-shot prompting. The prompt was then refined to generate satisfactory feedback for another 16 student responses. In stage II, four student raters were asked to rate both human-written and GPT-generated feedback for the 16 responses. In stage III, the same prompt was used to generate feedback for the rest of the 65 responses, and four instructors were asked to rate the 65 feedback messages based on perceived modifications needed.

1. Stage I: GPT feedback generation with few-shot prompting

To prepare examples to include in few-shot learning, we first manually categorized all student responses based on whether the conclusion (as to whether or not the glassware in the box is in danger of breaking) is correct, and whether the explanation is correct, incomplete, incorrect, or not present. We ended up with four categories: *correct conclusion with correct explanation*, *correct conclusion and incomplete/incorrect explanation, incorrect conclusion, and no explanation.* We did not see any students who gave correct explanation but arrived at an incorrect conclusion. Also, for those who did not give an explanation, we did not divide them based on whether their conclusion was correct or not.

In each of the four categories, we selected five responses that we considered quite dissimilar to one another. The authors of the paper wrote feedback to those 20 responses as if we were writing feedback to students in our own classes. One student response plus feedback from each category was included in the prompt given to GPT. We then tasked GPT to generate feedback to the remaining 16 responses.

The prompt we developed includes the following elements: the context (i.e., an instructor is giving feedback on student written responses), the physics problem, an expert response, physics concepts and principles involved, feedback instructions (i.e., what the feedback should look like), and four examples of student response with human-written feedback.

To use GPT to generate a new feedback message, a new student response is appended to the end of the prompt, following the same structure as the four previous human-written response-feedback examples, one for each category. GPT then attempts to complete the text by predicting the most likely text that appears next, which is the feedback. Unlike ChatGPT which is trained to respond in a chat format, GPT in complete mode generates the feedback only (with the setting of a proper "stop sequence," which are specific character(s), such as a "new line" character, that signals GPT to stop generating following text).

The feedback generated by GPT in the first round had some obvious mistakes that resemble some of the common preconceptions identified in the PER literature. In an earlier study, it was also documented that ChatGPT's outputs reflect student preconceptions [32]. Therefore, as contextual information, we included in the prompt those common preconceptions that are well documented in PER literature, such as "force can be transmitted through objects" and "force can be divided between objects" [33]. In addition to addressing preconceptions, we also added numbers to the physics concepts and principles in this section and specified what "not to do" in the feedback instructions. The prompt engineering process went through several iterations to optimize performance on the remaining 16 selected responses. The final version of the prompt is shown in the Appendix.

All 16 GPT-generated feedback messages were deemed to be correct by both authors [34]. By correct, we mean there were no apparent or indisputable mistakes, such as a misjudgment of the correctness of the student's response or a statement that resembles a student's preconception. This is different from judging the feedback as potentially helpful to the students.

2. Stage II: Student researcher evaluations

The 16 student responses and the corresponding human-written and GPT-generated feedback messages were given to four student raters who were involved in physics education research (PER). Three of the students were undergraduates and one was a first-year graduate student. All three undergraduate students had already completed the calculus-based physics I course and received high grades in the course. They were majoring in aerospace engineering, computer science, and medical laboratory sciences, respectively. One of them was a learning assistant in physics I during this study. All students were asked to evaluate the correctness and usefulness of each feedback message. They were also asked to indicate which one of the two feedback messages for each response they think was generated by AI. The survey questions are shown in Fig. 2.

We note that the order of the responses was randomized rather than organized based on the categories we developed in stage I. The order of the feedback messages was also randomized such that GPT-generated feedback messages were not always listed first or second.

3. Stage III: Instructor evaluations

In stage III, we tasked GPT to generate feedback to the 65 responses that were not used in stage I, based on the

1. Which of the feedback statements do you think is scientifically correct?
   a. A is correct.
   b. B is correct.
   c. Both are correct.
2. Which of the feedback statements do you think is more useful?
   a. A is more useful.
   b. B is more useful.
   c. Both are equally useful.
3. Which feedback do you think is generated by Generative AI (GPT)?
   a. A
   b. B
   c. Not sure

FIG. 2. Survey questions given to student researchers to rate the 16-student responses with human-written and GPT-generated feedback.

same prompt as we developed in stage I with no modification. The 65 feedback outputs were rated by four instructors, two of whom were the authors of this paper, and the other two were non-PER faculty members who had ample experience teaching this course.

The goal of instructor ratings is to examine the potential of GPT to reduce the instructors' grading effort. Therefore, the scoring criteria are based on the instructor's perceptions of how much modification is needed. All four instructors rated the feedback on a scale of 0–3. A score of 3 means that the instructor would give the feedback message to a student without any modifications. This requires that the feedback is not only correct but also addresses the student's specific response. However, the feedback does not necessarily need to address all the incorrect ideas in the response. A score of 2 means the feedback needs some quick modifications, such as some minor wording changes or partial deletion. A score of 1 means the feedback needs major revisions that often require deliberation and would take a longer time to write than a feedback message of score 2. Finally, a score of 0 means the feedback needs to be rewritten completely. Those scoring conditions were fully communicated with the instructor raters.

V. RESULTS

A. Student researcher ratings

We report student researchers' ratings on perceived scientific correctness and usefulness for both GPT-generated and human-written feedback on the 16 responses. We also show results for success rates for identifying the GPT-generated feedback.

1. Perceived correctness

Figure 3 shows the distribution of correctness ratings of GPT-generated and human-written feedback by all four student researchers. Recall that all the feedback messages were considered correct (i.e., no apparent and indisputable mistakes) by the authors. Student A seemed to slightly favor feedback generated by GPT. This student rated both feedback messages as being correct in approximately half of the cases; for the other half, the student rated more GPT-generated feedback messages as being correct. In contrast, student B appeared to favor the feedback written by a human instructor. There was only one instance in which student B rated both feedback messages as being correct; the human-written feedback messages were rated by student B as correct more frequently. Both students C and D showed no clear preference. They rated both human-written and GPT-generated feedback messages as being correct for three-quarters (or above) of the responses. Overall, we did not find a clear trend for favoring either human-written or GPT-generated feedback regarding correctness. It seemed that both human-written and GPT-generated feedback messages were equally correct as

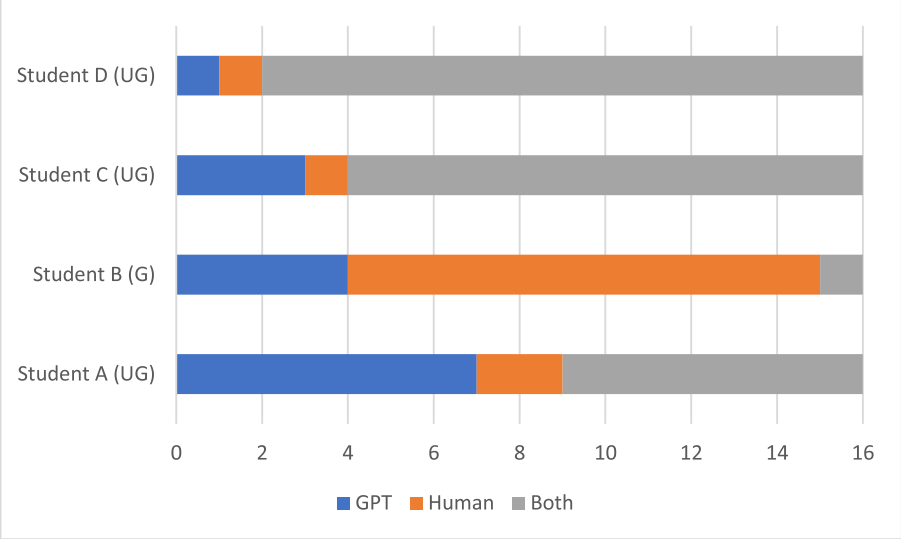


FIG. 3. Distribution of correctness for GPT-generated and human-written feedback for all four student researchers.

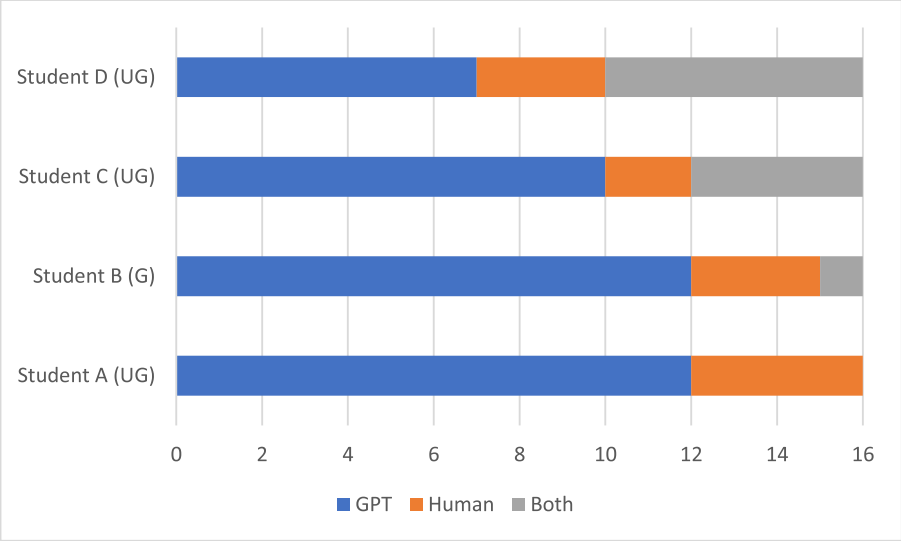


FIG. 4. Distribution of usefulness for GPT-generated and human-written feedback for all four student researchers.

perceived by the student researchers. The graduate student (student B) seemed to favor human-written feedback.

2. Perceived usefulness

Figure 4 shows the distribution of usefulness ratings for GPT-generated and human-written feedback for all four student researchers. It appears that all the student researchers rated the GPT-generated feedback as more useful. Student A showed the strongest preference toward GPT-generated feedback: three-quarters of the GPT-generated feedback messages were rated more useful than human-written feedback messages. Even student D, who had the highest frequency of rating both human-written and GPT-generated feedback messages as being equally useful, tended to favor GPT-generated over human-written feedback. Interestingly, student B, who rated more human-written feedback statements as being correct in the previous task, still rated significantly more GPT-generated feedback statements as more useful.

To gain insights into characteristics of the GPT-generated feedback that were perceived as more useful, we examined the feedback messages that were rated as more useful by all four student researchers. As shown in Table I, there were five instances in which the GPT-generated feedback messages were rated as more useful by all four student researchers, while only one instance in which the human-written feedback was unanimously rated as more useful.

At a quick glance, all five GPT-generated feedback messages that were perceived as more useful are much longer than the corresponding human-written feedback messages. Taking a closer look, GPT-generated feedback addresses students' responses to a greater extent. The five student responses happened to come from either the category "*correct conclusion with correct explanation*" or the category "*no explanation*." The feedback statements generated by GPT always prompted students to take a step further in their reasoning when the explanation was correct; it gave a hint even when students did not provide an explanation.

The GPT-generated feedback message, which was perceived as less useful, appeared to address the student's response to the same extent as the human-written feedback. Two of the student researchers rated that only the human-written feedback message was correct, and the other two rated both messages as correct. We infer that the lower rating in the usefulness of the GPT-generated feedback might have been due to the lower perceived correctness.

3. Which feedback messages were generated by AI?

Interestingly, although all student researchers tended to rate the GPT-generated feedback messages as more useful, the success rates of correctly identifying AI-generated feedback were much lower than chance, as shown in Table II. The graduate student researcher had a much higher success rate (63%) than the three undergraduate students (ranging from 13% to 31%), but it is unclear to us why this was the case.

It is worthwhile to point out that the feedback messages in this study are rather short and follow a rather rigorous pattern. Therefore, it can be challenging for most people to distinguish between human-written and AI-generated feedback. The purpose of this question is to test if the feedback created by AI seems natural and humanlike to students at first glance. If students could easily tell AI-generated feedback from human-written feedback, it might cause them to perceive the AI-generated feedback as untrustworthy.

TABLE I. List of GPT-generated feedback messages perceived as more useful or less useful by all four student researchers. The corresponding student responses and human-written feedback messages are also provided.

	Student response	Human-written feedback	GPT-generated feedback
GPT-generated feedback rated unanimously as more useful	No, because the force of box A on box b would not be 200 because box b has a smaller mass and would not take 200 N to move.	Both your conclusion and explanation are correct.	Your conclusion is correct and your explanation is on the right track. However, you may want to explain why smaller mass means smaller force. Think about an equation that relates force, mass and acceleration.
	Yes, I think the glassware is in danger of breaking.	Your conclusion is incorrect. Please give an explanation so that we can provide some feedback on your reasoning.	Your conclusion is incorrect. You should think about why the force on box B may not be equal to the applied force of 200 N by the student on box A. Hint: Think about the mass and acceleration of the boxes.
	Box B will be safe because it has a lower mass and requires less force to accelerate at the same rate as box A.	Both your conclusion and explanation are correct.	Your conclusion and explanation are both correct. To further strengthen your answer, think about the relationship between the force, mass, and acceleration in this scenario.
	Yes	Your conclusion is incorrect. Please give an explanation so that we can provide some feedback on your reasoning.	Your conclusion is incorrect. Even though the student pushes on box A with 200 N, it does not mean that the box B receives a 200 N force. Think about the mass and acceleration of the two boxes and explain why the force on box B must be smaller than the 200 N applied force.
	Yes, the glassware could become in danger of breaking.	Your conclusion is incorrect. Please give an explanation so that we can provide some feedback on your reasoning.	Your conclusion is incorrect. The force on box B is not necessarily equal to the force that the student applies to box A. Think about the mass and acceleration of these two boxes and consider what the force on box B must be to maintain the same acceleration.
Human-written feedback rated unanimously as more useful	Yes, the force would be transmitted because they are moving together at 200 N.	Your conclusion is incorrect. The boxes moving together only means they have the same acceleration, not necessarily the same force. The 200 N force is only applied to box A, not box B. Will the force on box B be greater than or smaller than 200 N given its mass and acceleration?	Your conclusion is incorrect. It is true that the two boxes are moving together with the same acceleration, but the force on box B does not necessarily need to be 200 N. Consider how the masses of the two boxes impact the force acting on them.

B. Instructor ratings

The distribution of instructor ratings is shown in Fig. 5. Since the scoring was based on instructors' perception of how much modification the feedback needed, we did not expect the instructors to reach an agreement on the individual feedback messages. Overall, the distribution across all four instructors appeared similar. Although the instructors were not quite aligned with the number of feedback messages that could be readily used with no modifications (score 3), they seemed to be consistent on the number of feedback messages that required either no modification (score 3) or only a minor modification (score 2). The percentages of feedback messages that required minor or no modification ranged from 68% to 78%. In addition, the average rating across all four instructors was 2.06 (on a scale of 0–3), which suggests that on average, the feedback messages only required minor modifications.

There was one instance in which all four instructors agreed to use the feedback without any modification. The student stated, "Yes, because both boxes will experience the same amount of force." GPT's feedback stated,

TABLE II. Distribution of GPT-generated feedback statements being indicated correctly and incorrectly.

	Correct	Incorrect	%Correct
Student A (UG)	5	11	31
Student B (G)	10	6	63
Student C (UG)	2	14	13
Student D (UG)	3	13	19

"Your conclusion is incorrect. Both boxes may experience the same acceleration, but they need to have different forces depending on the mass. Think about this in terms of net force and the mass of box B relative to box A." GPT correctly pointed out that the acceleration was the same, but the net forces are different because the masses of the boxes are different.

The lowest-rated GPT feedback statements received three "0"s and a "2," which occurred in two instances. For example, a student stated, "That force will have been transferred and the glassware will break." GPT's feedback stated, "Your conclusion is incorrect, but it seems like you have a good understanding of the concepts. Think more about why the force can't possibly be transferred to box B." One possible explanation might be that GPT "interpreted" the students' expression of "force will have been transferred" as indicating that the force will actually not be transferred to box B.

VI. DISCUSSION

In this study, we tested the feasibility of using GPT to assist in grading and generating personalized feedback on students' written responses to one conceptual question. Taking advantage of existing rich PER literature on student preconceptions of Newtonian mechanics, we included relevant common preconceptions in the prompt to improve the accuracy of feedback. We also categorized student responses and provided GPT one example response-feedback pair from each category to initiate the few-shot learning process.

The results showed that three out of four student researchers perceived GPT-generated and human-written feedback messages to be equally correct. At the same time, all the student researchers rated GPT-generated feedback messages as more useful. This was probably because GPT-generated feedback messages are generally longer and address students' responses to a greater extent, especially when students gave a correct conclusion with a correct explanation or when they did not provide an explanation. In contrast, the authors gave identical feedback to students whose responses were classified in either of those categories. Our justification for the short feedback is that it saves time and the saved time can be used to write more detailed feedback to students who the instructor judged as needing more instructor input, such as those who gave incorrect conclusions or explanations. However, this practice could potentially leave students who gave the correct answer with the feeling of being neglected, which is where GPT can be extremely helpful as it treats every response with equal patience and effort.

Moreover, the percentages of correctly identified GPT-generated feedback messages were overall low among the student researchers, which suggests that the student researchers often perceive GPT-generated feedback as humanlike and perceived brief human feedback as being created by a machine. A possible explanation might be that students expect the human expert to be more helpful and a machine to be more repetitive. It is worth noting that the purpose here is not to evaluate student researchers' ability to distinguish between AI and humans, as student

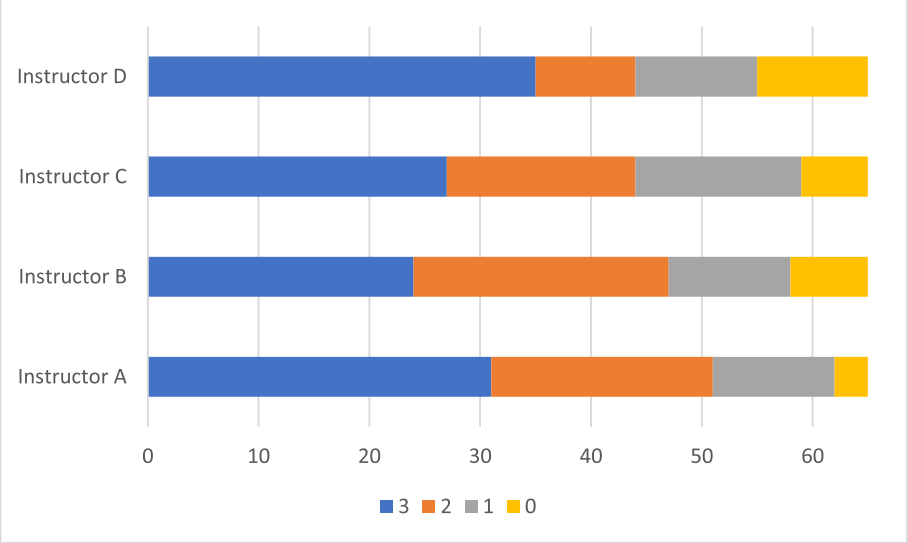


FIG. 5. Distribution of ratings on a scale of 0–3 for all four instructors. The number of GPT-generated feedback messages is 65.

researchers based their judgments on just a few lines of feedback text in each case. The outcomes might be different if the feedback messages were more extensive or a multiturn conversation was involved. For example, Jones and Bergen [35] showed that in a multiturn conversation situation, GPT-3.5 is only misclassified as a human in 14% of the cases. However, the results do suggest that students are unlikely to quickly judge the AI-generated feedback as "artificial" or "machine generated" at the first glance. In addition, our results also speak to one strength of GenAI, which is that it has infinite patience and will always respond to student answers with enough detail. In contrast, human graders' performance can be easily impacted by fatigue and emotion, which can result in short and repetitive feedback that may be perceived as more "machine like."

The four instructors only considered a major modification or rewriting of GPT-generated feedback messages about 30% of the time, indicating a relatively high extent of satisfaction with GPT's performance in the current setting. However, it should be cautioned that the number of student answers is relatively small, so it remains to be seen whether the 70% satisfactory rate can be generalized to more diverse student answers and to situations where more sophisticated feedback is required. There are some studies that show that the performance of LLMs might be artificially boosted by the models relying on certain common keywords that exist in the specific dataset or expected in the output (also known as short-cut learning [36]).

In summary, our results suggest that LLMs such as GPT-3.5 have a promising potential for serving as a grading assistant using only prompt engineering and few-shot learning, and without the need for additional fine-tuning. Even though the feedback is only satisfactory in 70% of the cases, it could still save a significant amount of time and effort required from the instructors. This allows instructors to assign open-response problems to students more often. From students' perspectives, GPT-generated feedback statements are perceived as more useful and sometimes even more "humanlike" than the ones that were actually written by humans. Moreover, GPT-generated feedback has the unique advantage of being highly consistent in the level of detail regardless of the amount of grading workload, which could provide a more equitable learning experience to all students.

VII. LIMITATIONS AND FUTURE WORK

As an explorative initial study, there are many limitations and caveats that are worth discussing. We also propose ideas for future work on how to efficiently use LLMs in generating personalized feedback.

First, one of the authors manually categorized all the student answers and provided GPT with one answer-feedback pair from each category as an example. The manual categorization process can be very time consuming and impractical if LLMs are to be used as an actual grading assistant. Future studies could explore two possible alternate solutions. First, the LLM could randomly select a small number (for example, 10–20) of student answers and ask a human to write feedback on those as examples. Alternatively, one could task the LLM to perform clustering analysis on student responses using text embedding and machine learning (for example, see Ref. [37]). The LLM could then select 1–2 representative student answers from each cluster and require a human to write feedback on those as examples.

Second, the current study only involves one conceptual question. Future studies could evaluate the performance of GPT-based grading assistance on a wider variety of problems. Moreover, a valuable future direction is to investigate whether GPT can generate quality feedback for a class of similar problems involving the same physics principle, given a single, context-independent general prompt. Future studies could also explore the possibility of using LLMs as chatbots and engaging students in multiround conversations that build on students' existing knowledge, based on theoretical frameworks such as the resource model [31].

Third, the feedback was only evaluated by four student researchers. Moreover, all of them were involved in PER projects, and two of them were working on projects that involved LLMs, which could have an influence on their judgments. In future studies, we plan to survey students who are enrolled in introductory physics courses about their perceptions of the AI-generated feedback they receive. We also plan to survey more graduate teaching assistants who do not have prior knowledge of GenAI and evaluate their perceptions of the quality of the AI-generated feedback.

Fourth, we tasked instructors to rate the feedback based on their perceptions of the amount of editing needed before they would deem the feedback ready for students. We did not provide more contextual information regarding the use of the problem and feedback, such as whether the problem was used as part of a homework assignment, an exam, or an in-class activity. Nor did we ask the instructors to rate the quality of the feedback. A content-specific rubric will need to be developed for instructors to rate the quality of the feedback.

Furthermore, the number of student responses in this study was still relatively small, which enabled the study to be conducted using GPT-3.5 turbo (a large 20B parameter model), at a minimum cost to the researcher. For future applications with a much larger student body and many more problems, the usage cost could be significant. Therefore, it is worth investigating whether smaller models pretrained on materials from a specific discipline or a specific course could reach a similar level of performance.

Finally, using GenAI to assist in grading has some potential caveats. For example, AI-generated feedback could be potentially biased [38,39] because LLMs like

GPT are pretrained using a large amount of data from different sources including the Internet and social media. The outputs can be influenced by existing biases contained in the training data. LLMs could also generate an incoherent chain of arguments or even factually incorrect information. Currently, we recommend instructors who intend to use GPT as a grading assistant to check all the feedback messages and make edits if needed before sending the feedback to students. Future research should explore using AI or machine learning methods to automatically suggest the feedback messages that are most likely to be incorrect for instructors to review. Finally, another issue worth considering regarding the use of AI as an assistant in grading is its environmental impact, as training and running AI systems often require substantial computing power and thus significant electricity consumption [40].

VIII. CONCLUSION AND IMPLICATIONS

In conclusion, we believe that GenAI holds a significant potential to serve as a grading assistant for open-response questions. One possible model of using GPT as a grading assistant could be as follows: First, the instructor is asked to write feedback on several student responses chosen by the assistant and given the opportunity to input common student preconceptions from either research literature or experience. Second, the assistant grades and writes feedback on all student responses, using a prompt that incorporates the instructor's example feedback and preconception input. Third, all response-feedback pairs are presented to the instructor, and the instructor edits the feedback messages. Finally, the corrected feedback messages will be presented to students and students will have an opportunity to request the instructor to review the GPT-generated grading outcomes and feedback. Such a process could significantly reduce the grading load for instructors and increase the quality and consistency of the feedback, potentially leading to improved student conceptual understanding.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Xiaofeng Feng and Dr. Richard Jerousek for assisting in rating the feedback. This work is supported in part by the University of Central Florida Digital Curriculum Innovation Initiative and the University of Central Florida College of Sciences Seed Grant Program.

APPENDIX

Final version of GPT prompt

Context

A physics instructor is rating students' answers to the following physics problem:

Physics Problem:

A student pushes two boxes, one in front of the other, as shown in the diagram. Box A has mass 75 kg, while box B has mass 25 kg. Fortunately for the student, the boxes are mounted on tiny rollers and slide with negligible friction. The student exerts a 200 N horizontal force on box A.

Box B contains kitchen stuff, including some poorly packed glassware that might break if the force pushing on the side of the box approaches 200 newtons. Recall that the student pushes on box A with a force of 200 newtons. Is that force "transmitted" to box B? In other words, is the glassware in the box in danger of breaking? Don't do any calculations; answer intuitively, and explain your thinking.

The instructor rates students' answer and gives feedback based on how similar it is to this expert answer:

Expert Response:

No, the 200 N force is not 'transmitted" to box B. For box A, there is a 200 N applied force to the right and a force by box B to the left, thus the net force on box A is less than 200 N. Box A and box B have the same acceleration, but box B has a smaller mass. To maintain the same acceleration, net force on box B needs to be smaller than net force on box A. Therefore, the force on box B must be smaller than 200 N.

Physics principles involved:

1. The net force on an object is equal to the mass times its acceleration.
2. "Force", "energy", and "power" are different concepts.
3. Force is never transmitted from one object to another.
4. Force is never divided among objects.
5. Force is never distributed among objects.
6. Force is never absorbed by an object.
7. Force cannot pass through an object.
8. Two forces acting on different bodies do not "cancel out".
9. Box B has smaller mass, so to maintain the same acceleration, force on box B needs to be smaller, which means that it must be smaller than 200 N.

The feedback should start with whether the answer is correct or incorrect, followed by a justification, and then a follow up question for further thinking. The feedback should not give away the expert answer. Those physics principles should not all be repeated in the feedback to students.

Here are several examples of student answer and instructor feedback:

Answer: No because in order to maintain the same acceleration, the force on box B must be lower than 200 N.

Feedback: Both your conclusion and explanation are correct.

Answer: No, the 200 N push force does not transfer to box B. The push force is only being applied to box A and you cannot assume that the force is transferred. The force acting on box B would be the force of A on B, not from the push force.

Feedback: Your conclusion is correct, but your explanation needs some improvements. You are correct that the force does not transfer to box B, but you will need to explain why force of A on B must be smaller than 200 N. Hint: Think about the mass and acceleration.
Answer: Yes, it will break because the forces are equal and opposite force on each object.
Feedback: Your conclusion is incorrect. Yes, the force by Box B on Box A is equal and opposite to the force by Box A on Box B. However, it does not necessarily mean that those forces are equal to 200 N applied force by the student on Box A. You may want to consider whether the force on box B is greater than or smaller than 200 N based on its mass and acceleration.
Answer: No, it is not in danger there isn't a 200 N force on the box.
Feedback: Your conclusion is correct, but you need to provide an explanation for why the force on box B is not 200 N.
Answer: [new student response goes in here.]
Feedback:

[1] R. Beichner, An introduction to physics education research, in *Getting Started in PER*, edited by C. Henderson and K. Harper (2009), Vol. 2.

[2] L. S. Shulman, Those who understand: Knowledge growth in teaching, Educ. Res. **15**, 4 (1986).

[3] K. A. Ericsson, R. T. Krampe, and C. Tesch-Romer, The role of deliberate practice in the acquisition of expert performance, Psychol. Rev. **100**, 363 (1993).

[4] Louis Deslaurier, Ellen Schelew, and Carl Wieman, Improved learning in a large-enrollment class, Science **332**, 862 (2011).

[5] P. Black and D. Wiliam, Assessment and classroom learning, Int. J. Phytorem. **5**, 7 (1998).

[6] J. Larreamendy-Joerns, G. Leinhardt, and J. Corredor, Six online statistics courses: Examination and review, Am. Stat. **59**, 240 (2005).

[7] D. Baidoo-Anu and L. Owusu Ansah, Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning, J. AI **7**, 52 (2023).

[8] E. Kasneci *et al.*, ChatGPT for good? On opportunities and challenges of large language models for education, Learn. Individ. Differ. **103**, 102274 (2023).

[9] Z. Li, C. Zhang, Y. Jin, X. Cang, S. Puntambekar, and R. J. Passonneau, Learning when to defer to humans for short answer grading, in *Artificial Intelligence in Education*, edited by N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, and V. Dimitrova, Lecture Notes in Computer Science() Vol. 13916 (Springer, Cham, 2023), 10.1007/978-3-031-36272-9_34.

[10] C. Sung, T. Ma, T. I. Dhamecha, V. Reddy, S. Saha, and R. Arora, Pre-training BERT on domain resources for short answer grading, in *Proceedings of the EMNLP-IJCNLP 2019–2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Hong Kong, China* (Association for Computational Linguistics, Hong Kong, China, 2019), p. 6071.

[11] A. Ahmed, A. Joorabchi, and M. Hayes, On deep learning approaches to automated assessment: Strategies for short answer grading, in *Proceedings of the 14th International Conference on Computer Supported Education* (SCITEPRESS—Science and Technology Publications, 2022), pp. 85–94.

[12] A. Condor, Exploring automatic short answer grading as a tool to assist in human rating, in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2020), Vol. 12164.

[13] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Comput. Surv. **55**, 9 (2023).

[14] S. Steinert, K. E. Avila, S. Ruzika, J. Kuhn, and S. Küchemann, Harnessing large language models to enhance self-regulated learning via formative feedback, arXiv:2311.13984v2.

[15] G. Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course, Phys. Rev. Phys. Educ. Res. **19**, 010132 (2023).

[16] M. N. Dahlkemper, S. Z. Lahme, and P. Klein, How do physics students evaluate artificial intelligence responses on comprehension questions: A study on the perceived scientific accuracy and linguistic quality of ChatGPT, Phys. Rev. Phys. Educ. Res. **19**, 010142 (2023).

[17] S. Küchemann, S. Steinert, N. Revenga, M. Schweinberger, Y. Dinc, K. E. Avila, and J. Kuhn, Can ChatGPT support prospective teachers in physics task development?, Phys. Rev. Phys. Educ. Res. **19**, 020128 (2023).

[18] G. Polverini and B. Gregorcic, How understanding large language models can inform the use of ChatGPT in physics education, Eur. J. Phys. **45**, 025701 (2023).

[19] F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. Chi, N. Schärli, and D. Zhou, Large language models can be easily distracted by irrelevant context, arXiv:2302.00093v3.

[20] M. Lee, A mathematical investigation of hallucination and creativity in GPT models, Mathematics **11**, 2320 (2023).

[21] M. Zhang, O. Press, W. Merrill, A. Liu, N. A. Smith, and P. G. Allen, How language model hallucinations can snowball, arXiv:2305.13534.

[22] Y. Bang *et al.*, A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity, arXiv:2302.04023v4.

[23] D. J. Woo, K. Guo, and H. Susanto, Cases of EFL secondary students' prompt engineering pathways to complete a writing task with ChatGPT, arXiv:2307.05493.

[24] T. F. Heston and C. Khun, Prompt engineering in medical education, Int. Med. Educ. **2**, 198 (2023).

[25] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, Generalizing from a few examples: A survey on few-shot learning, ACM Comput. Surv. **53**, 1 (2020).

[26] M. Zong and B. Krishnamachari, Solving math word problems concerning systems of equations with GPT-3, in *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI-23)* (2023).

[27] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, arXiv:2201.11903v6.

[28] E. Radiya-Dixit and X. Wang, How fine can fine-tuning be? Learning efficient language models, in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, edited by S. Chiappa and R. Calandra (PMLR, 2020), pp. 2435–2443.

[29] E. Latif and X. Zhai, Fine-tuning ChatGPT for automatic scoring, Comput. Educ. **6**, 100210 (2024).

[30] A. Elby, R. E. Scherr, T. McCaskey, R. Hodges, E. F. Redish, D. Hammer, and T. Bing, Open Source Tutorials in Physics Sensemaking: Suite I (2007), https://www.physport.org/curricula/MD_OST/.

[31] D. Hammer, Student resources for learning introductory physics, Am. J. Phys. **68**, S52 (2000).

[32] S. Wheeler and R. E. Scherr, ChatGPT reflects student misconceptions in physics, presented at PER Conf. 2023, 10.1119/perc.2023.pr.Wheeler.

[33] R. E. Scherr and E. F. Redish, Newton's zeroth law: Learning from listening to our students, Phys. Teach. **43**, 41 (2005).

[34] In one of the student responses with the correct conclusion, there was a minor mistake in the explanation, which we did not realize when we wrote the feedback. Both GPT-generated and human-written feedback stated that the response was correct on both the conclusion and explanation.

[35] C. Jones and B. Bergen, Does GPT-4 pass the turing test?, arXiv:2310.20216.

[36] M. Mitchell and D. C. Krakauer, The debate over understanding in AI's large language models, Proc. Natl. Acad. Sci. U.S.A. **120**, e2215907120 (2023).

[37] P. Tschisgale, P. Wulff, and M. Kubsch, Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory, Phys. Rev. Phys. Educ. Res. **19**, 020123 (2023).

[38] D. Rozado, The political biases of ChatGPT, Soc. Sci. **12**, 148 (2023).

[39] L. Lucy and D. Bamman, Gender and representation bias in GPT-3 generated stories, in *Proceedings of the Third Workshop on Narrative Understanding, Virtual* (Association for Computational Linguistics, 2021), p. 48.

[40] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2021, Virtual Event Canada* (Association for Computing Machinery, New York, NY, 2021), p. 610.