Exploring Large Language Models as Formative Feedback Tools in Physics

Shams El-Adawy, Aidan MacDonagh, and Mohamed Abdelhafez
Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139

Significance of formative assessments and feedback is well-established in physics education, yet implementation in large enrollment physics courses poses substantial challenges such as scalability of timely personalized feedback. As part of efforts to productively incorporate large language models (LLMs) into physics education, we use a mixed method approach to compare human and artificial intelligence (AI) feedback to students on conceptual synthesis questions in an introductory mechanics course. We present our preliminary analysis showcasing the promising results and current limitations of tailored numerical and written AI feedback. We found that with physics instructors' guidance, AI provides relevant and timely written feedback to students. Nevertheless, AI struggles with edge cases and with specificity to students' answers, both of which are better handled by humans. Future work will investigate improving feedback quality by using rubrics to prompt AI, with the goal to enhance its potential utility to the physics education community.

2024 PERC Proceedings edited by Ryan, Pawl, and Zwolak; Peer-reviewed, doi.org/10.1119/perc.2024.pr.El-Adawy
Published by the American Association of Physics Teachers under a Creative Commons Attribution 4.0 license.
Further distribution must maintain the cover page and attribution to the article's authors.

I. INTRODUCTION

To facilitate students' learning of physics concepts and refinement of their problem solving skills, it is essential to provide them with timely and specific feedback at different stages of the learning process. Extensive research in physics education has focused on improving our understanding of students' problem-solving abilities and enhancing support for the development of these skills [1–4]. The resources framework in PER has shown that there are multiple and complicated ways in which conceptual, epistemological, and affective dimensions can interact as students learn to problem solve and make meaning of physics concepts [5]. Physics instructors play a crucial role in supporting these multifaceted ways of meaning-making in physics through interactions and feedback they provide to students. That feedback motivates learning and supports knowledge and problem-solving skills development [6]. While instructors often provide feedback to individual students in face-to-face interactions, the opportunity to provide timely and specific feedback becomes challenging in large enrollment physics courses. The development and effectiveness of digital feedback tools that provide this type of feedback remain under-researched.

The most prevalent digital tools that support immediate feedback often function as answer checkers rather than support deep qualitative conceptual reasoning. Over the years, much software has been developed within physics to serve as online tutorial, homework, and assessment tools, including *Mastering Physics* [7] and *WebAssign* [8]. Although these existing physics-specific digital tools are growing in sophistication, they still largely serve as mathematical solution verification or provide step-by-step solutions to specific physics problems. The principal limitation of these tools is the extent to which they can be programmed to individual instructors' prompts and provide qualitative conceptual reasoning feedback to students. The latter requires a deep understanding of content, context, intent, and nuance, which are features that are difficult to fully automate in existing physics-specific digital tools. Nevertheless, with the proliferation of artificial intelligence (AI) tools speficifically Large Language Models (LLMs) such as OpenAI's GPT series, there is an opportunity to leverage these models to support students learning by providing timely personalized feedback to students.

Efforts have been made to characterize the capabilities of LLMs, assessing their potential and pitfalls [9–15]. Factors such as AI's ability to solve physics problems and facilitate understanding are being explored. For example, physicists found that ChatGPT could narrowly pass a standard lecture-based introductory physics course, but with its rapid development, it has the potential to become a subject matter expert [16]. Researchers also explored AI's ability to grade physics problems and found that formative feedback is possible [17], but it is recommended to have humans assist with final evaluations as scientific accuracy is not yet guaranteed by AI [12]. Although these studies constitute seminal work in AI's integration within physics education, much remains unexplored around how AI can be leveraged to transform interactive and adaptive learning in physics.

In parallel, assessment literature often discusses feedback through formative and/or summative assessments. We define formative feedback, the focus of this paper, to be constructive feedback provided after assignments. Formative feedback can be characterized into two types: verification and elaboration [6, 18]. Verification refers to feedback that indicates whether an answer is correct. Elaboration feedback, in contrast, is more than merely an answer checker: it can address the response, discuss particular errors, provide supplementary worked examples, and/or give guidance [6]. There are many key features that characterize best practices in both types of feedback, including timing, specificity, and length of feedback [6]. In both cases, purposefully designed feedback significantly benefits student learning [19]. Despite feedback being at the core of learning, scaling individualized timely formative feedback is logistically and intellectually challenging for physics instructors.

Thus, leveraging technological advancements and addressing current limitations in providing timely individualized feedback to students are the motivation behind this research project. This methodological paper focuses on evaluating AI capabilities in providing formative physics feedback to student responses to conceptual synthesis physics questions. In particular, we aim to describe our methodology and preliminary analysis to address the following research question: **How does artificial intelligence perform compared to humans in providing scalable formative feedback on students' responses to synthesizing physics questions?**

Note on terminology: Typically, "Large Language Models (LLMs)" refer to specific types of artificial intelligence models designed primarily for understanding and generating human-like text based on machine learning. ChatGPT is an example of an LLM. "Artificial Intelligence (AI)" refers to a broader concept that encompasses not only large language models but also systems capable of performing tasks that would normally require human intelligence, such as making decisions and recognizing speech. For simplicity in this paper, we will use the terms LLMs, ChatGPT, and AI interchangeably.

II. CONTEXT

In this study, we focus on our introductory calculus-based classical mechanics course, which took place in Fall 2023. This large enrollment course, which had 700 students, is a general requirement course often taken during students' first year at our institution. The course follows a blended learning model where in-person instruction is combined with a delivery of online video instruction [20]. Students watch online videos and complete formative assignments pre-class, work on activities during in-person instruction and complete formative weekly assignments post-class. Students also take three summative exams during the semester. In preparation

for these exams, we provide students with practice problems and questions to help synthesize ideas and problem-solving strategies. Nevertheless, we find it challenging to provide individualized timely constructive feedback to the whole class during these review activities. Therefore, as part of this study, we crafted questions to help students connect ideas of the course and receive formative feedback from ChatGPT 3.5. Questions were embedded within our review sequence in our online learning management system (more technical details in section III B). Our goal was threefold:

(a) Encourage students to qualitatively draw connections among physics concepts, and provide timely feedback supporting their exam preparation;
(b) Evaluate AI's effectiveness in delivering feedback to students' physics responses;
(c) Enhance students' critical thinking skills by having them compare and contrast their answers with those provided by AI.

In this paper, we focus on the preliminary analysis of goal (b), where we analyze differences and similarities between AI and humans in providing feedback to students responses.

III. METHODOLOGY

A. AI Activity General Description

We developed two synthesizing questions to provide to students pre-exams where they input answers and get instant feedback from AI. Both questions were optional and included alongside practice problems. Although we assessed responses for this study, none of the scores were used for credit towards their performance in the course. The two synthesizing questions were the following:

> **Question 1**: There are two main topics we discussed in the course so far: 1- Kinematics 2- Newton's laws (especially $\vec{F} = m\vec{a}$). Briefly describe any connections and differences you see between the two topics. Provide an example of a situation where we need to use both ideas in the same problem.
> **Question 2**: Give concrete examples of problems that should be solved using energy-work vs. momentum-impulse. Building on that, explain general themes that energy and momentum problems have, separately.

Question 1 was given to students pre-exam 1, and Question 2 was given pre-exam 2. When engaging with these questions, students had a text box to enter their answer, then a text box with AI's feedback would emerge. Subsequently students had the opportunity to rate how useful the feedback was for them. The following disclaimer was also included at the top of the text boxes: *Please note the bot may sometimes generate partially inaccurate responses so that's why we advise you to be critical and provide feedback to us*.

B. AI Activity Technical Description

We used ChatGPT-3.5 [21], which we integrated into our learning management system through an Application Programming Interface (API). For each question, we (instructors) provided a prompt to ChatGPT to give feedback to students. Instructors' prompt was fixed and programmed into our learning management system. Students only saw the question and responded to it once, with no knowledge of the prompt. Each student's answer was appended to the instructor prompt, forming a complete query that was sent to ChatGPT for processing. The instructor's prompt guided ChatGPT on key aspects such as the focus of the feedback, evaluation criteria, and how to offer constructive feedback to students. Following guidelines about prompt engineering [22], here is how we decomposed the prompt into smaller parts for ChatGPT to complete and assemble:

> **Part 1:** This step asked ChatGPT to play the role of a physics instructor: *You are a physics instructor in an introductory calculus-based college-level class receiving students' responses to the following question:*
> **Part 2:** The same question that students were asked was included at this step.
> **Part 3:** This step guided ChatGPT's feedback to students. The instructor incorporated details about what constitutes a good response, including one or two detailed examples of what an ideal response would look like and specific points students must mention. Based on the instructor's own experience, guidelines on how ChatGPT should use these examples to identify and fill gaps in the students' responses were given. Additionally, the instructor explained to ChatGPT what structure and tone to give its feedback, and the instructor provided instructions on how the scoring scale 0-5 should be used in relation to the students' answers.

C. Data collection and analysis

We implemented a mixed-method approach [23], where we concurrently collected quantitative and qualitative data to be able to optimize analysis of generative AI capabilities in feedback generation. Data collected concurrently from our online learning management system included student responses to the questions, feedback from ChatGPT, criteria provided to ChatGPT (described in section III B), and students' feedback about AI's feedback. Subsequently, we distributed to physics instructors and graduate teaching assistants students' responses and the criteria given to ChatGPT to provide feedback. All feedback included a score ranging from 0 to 5, and written feedback.

For analysis, we began by examining both types of data separately. For the numerical scores, using descriptive statistics, we summarized and interpreted the basic patterns. For the written feedback from ChatGPT and humans, we used thematic analysis [24]. Given the vast amount of data, we

Fall 2023	Number of student responses	Pearson correlation between AI and human scores	Mean difference between scores
Question 1	103	0.43	0.2/5
Question 2	76	0.37	0.63/5

TABLE I. Descriptive Statistics of AI and Human Grading Scores Across Review Questions

focused on four components, commonly used in elaborate feedback [6, 18], which are: content detail, specificity, relevance and consistency. We began by examining the emergent themes related to these components across the AI feedback, looking within each question and across all student responses. Afterwards, we examined the same features across human feedback, again within each question and across all student responses. Subsequently, we compared differences and similarities across AI and human feedback. Lastly, we went back and forth between data types to answer our research question. This analysis, conducted by the authors of this paper, is still ongoing.

IV. FINDINGS

Table I provides a breakdown of the descriptive statistics between AI and human scores to students' responses. We observe that the agreement between AI and human graders is low to moderate, with a positive correlation of 0.43 for Question 1 and 0.37 for Question 2. Moreover, the average difference in scores is a mere 0.2 out of 5.0 for Question 1, demonstrating a high degree of consistency between AI and human grading for this question. For Question 2, the average difference in scores is higher at 0.63. This disparity in average score between the two questions may be due to how these two questions differ in their complexity. Kinematics and dynamics are well documented online, so we anticipate that ChatGPT will have relevant information about the topic, even with minimal guidance from the instructor creating the prompt. However, Question 2 addresses more complex topics within our course context, such as the rocket equation and continuous mass transfer. In addition, Question 2 involves deeper problem-solving elements, making it inherently more sophisticated than the first. Thus, despite prompt engineering [22], ChatGPT may find it more challenging to understand and evaluate students' responses in the second question.

A deeper dive into the numerical scores shows that, in Question 1, AI never assigns a score of 0/5 and gives a score of 1/5 only twice out of 103 responses. In contrast, human graders assign a 0/5 nine times and 1/5 twenty-two times for the same set of responses. These edge cases are the main source of discrepancy between the Pearson correlation of 0.43, which is moderate and the small mean difference of 0.2/5. This result is also explained by the standard deviation of AI scores being about half that of human scores. As shown in histograms in Figures 1 and 2, the standard deviation is 0.79 versus 1.59 for Question 1 and 0.84 versus 1.41 for Question 2. While AI generally positively correlates with human grading, AI results in less variation among students' responses by opting for scores around the average.

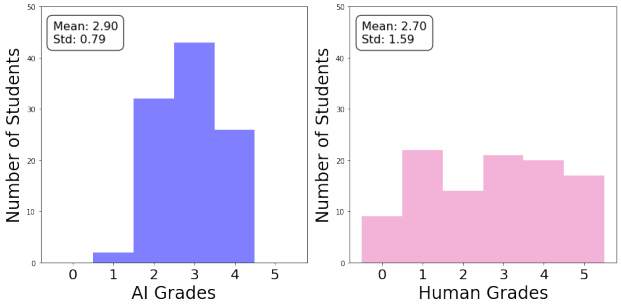

FIG. 1. Distribution of AI and human scores for Question 1

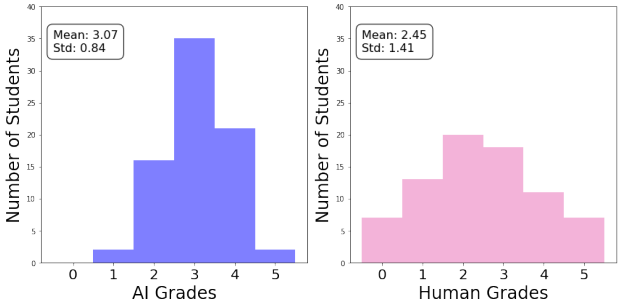

FIG. 2. Distribution of AI and human scores for Question 2

Our qualitative analysis is still in progress, but patterns are emerging in terms of content detail, specificity to students' response, relevance, and consistency of feedback. For content detail, AI, when programmed accordingly, can provide elaborate feedback that includes worked examples and guidance to students on how to attain stronger conceptual reasoning. In contrast, human feedback tends to be of the verification type and mainly highlights key errors. In terms of specificity, human feedback clearly and concisely identifies what went wrong in each student response, whereas AI is not as specific to students' answers. Regarding relevance, AI consistently provides feedback that is relevant to the topic, regardless of the depth or scientific accuracy of students' responses, adhering to the criteria it was programmed to follow. Human feedback also maintains relevance but shows greater adaptability in handling edge cases, such as when students write the question as their answer. In such cases, humans tend to recognize and note that no genuine attempt was made, while AI still provides topic-specific feedback. As for consistency, unlike AI, human instructors' feedback tends to become less thorough as the number of responses being assessed increases. Overall, we have yet to observe differences between Questions 1

and 2 in terms of the above patterns.

Our emergent qualitative themes highlight that human feedback still exhibits greater specificity and adaptive relevance to student responses compared to AI-generated feedback. However, AI consistently provides relevant feedback in accordance with the programmed instructions. Physics concepts are correctly discussed across questions, without any type of hallucination. This is promising, as it contrasts with the scientific accuracy issues that emerged in previous research that used ChatGPT without guided prompting from instructors [12, 16]. Additionally, as expected, AI feedback is more extensive in length and is generated immediately, whereas human feedback is brief and required a few weeks to complete. Though timing is just one of many aspects of useful feedback, we believe this rapid feedback is particularly beneficial for students' learning.

In summary, three key takeaways emerge from our preliminary analysis of human versus AI feedback:

Theme 1: Despite detailed guidance from instructors, the reliability of AI scores decreases as question complexity increases. Unlike human graders, the sophistication of context, intent and nuance is more challenging to prompt engineer in AI.

Theme 2: AI generally assigns scores around the average and avoids assigning perfect or zero scores, even when directed to do so for responses that are either ideal or irrelevant. This may be due the probabilistic mechanisms intrinsic to LLMs [25] or due to the fact that the LLM was fine-tuned to not be extreme in its responses [26]. In contrast, human instructors are much more adept at recognizing when a student response merits a perfect or null score.

Theme 3: When prompted with specific text-based criteria, AI can be reliable and consistent in providing scalable written feedback to students. Yet, the specificity to students' responses remains much more evident in human feedback.

In tandem, while not the primary focus of this paper, students' perspectives are important to consider when evaluating the value and effectiveness of using AI for feedback. During our data collection, students provided written feedback about AI's feedback and an illustrative quote of their perception is:

> *The AI was very accurate in evaluating my response. I understood in what ways my response contained inaccuracies and important missed information.I knew my answer was right-ish but lacked detail. The AI chatbot helped me fill in the blanks and its response appears to be correct. This is so cool and super helpful, I genuinely love this feature.*

The written feedback is accompanied by students' rating of the AI feedback, which is at an average of 4.03 stars out of 5. Interpreting the meaning of this rating requires further exploration, as we lack a comparable metric for rating feedback from human instructors. However, the rating does indicate that students were actively engaging with the AI feedback, as the opportunity to rate only appeared after the AI feedback was generated. While our analysis has centered on comparing the quality of feedback from AI versus humans, students' positive perceptions warrant further investigation into how AI feedback can support student learning. Particularly, since research has shown that students' assessments of the scientific accuracy of ChatGPT responses varies depending on their knowledge levels [12].

V. DISCUSSION

LLMs could transform how students engage with and understand physics concepts and develop problem solving skills. Despite the rapid advancement of AI, there remain numerous open questions and unexplored opportunities in their implementation in physics education. Our study underlined some preliminary nuances and patterns in using AI as a tool to support physics instructors in providing formative feedback to students. Most notably, AI's ability to provide reliably relevant and consistent written feedback at scale appears promising, particularly when the feedback is very specifically prompted by physics instructors and tailored to particular questions. However, unlike humans, AI tends to gravitate towards assigning scores around the average. This tendency makes it challenging to constructively use AI's numerical assessments, which is concurrent with recent literature [17].

Our analysis presents several limitations. First, our data collection and analysis is specific to two synthesis physics questions and prompting of ChatGPT-3.5 in Fall 2023. As others discussed [16], this limitation is inherent to any study that uses generative AI, as new versions are projected to perform better than previous ones. Second, our data is limited to a subset of our students who engaged in these optional review question activities. There may be unforeseen effects that only emerge as we scale these AI activities further.

Our preliminary results show promise in terms of scalability of AI as a resource for formative feedback in a physics instructor's toolkit. Future work will include further thematic analysis of the accuracy of feedback in our qualitative data, its interactions with the numerical scores, and the impact of AI feedback on students' learning experiences. Further research will also draw deeper from chain of thought prompting [27] to improve specificity' of our evaluation criteria (Part 3 in III B) provided to AI. This will include examining robustly how to use detailed rubrics to elicit more specific and adaptive feedback to students.

ACKNOWLEDGMENTS

Thank you to the graduate teaching assistants who assisted with the grading of student responses.

[1] J. L. Docktor, J. Dornfeld, E. Frodermann, K. Heller, L. Hsu, K. A. Jackson, A. Mason, Q. X. Ryan, and J. Yang, Assessing student written problem solutions: A problem-solving rubric with application to introductory physics, Physical review physics education research **12**, 010130 (2016).

[2] E. Burkholder, J. Miles, T. Layden, K. Wang, A. Fritz, and C. Wieman, Template for teaching and assessment of problem solving in introductory physics, Physical Review Physics Education Research **16**, 010123 (2020).

[3] A. Maries and C. Singh, Helping students become proficient problem solvers part I: A brief review, Education Sciences **13**, 156 (2023).

[4] E. Kuo, M. M. Hull, A. Gupta, and A. Elby, How students blend conceptual and formal mathematical reasoning in solving physics problems, Science Education **97**, 32 (2013).

[5] E. F. Redish and E. Kuo, Language of physics, language of math: Disciplinary culture and dynamic epistemology, Science & Education **24**, 561 (2015).

[6] V. J. Shute, Focus on formative feedback, ETS Research Report Series **2007** (2007).

[7] C. Leow, K. Lee, and Y. J. Rho, *Mastering Physics A study of two semesters of an introductory, prerequisite physics course at Penn State University*, Tech. Rep. (Pearson, 2017).

[8] J. Brunsmann, A. Homrighausen, H.-W. Six, and J. Voss, Assignments in a virtual university–the webassign-system, in *Proc. 19th World Conference on Open Learning and Distance Education, Vienna/Austria* (Citeseer, 1999).

[9] F. Kieser, P. Wulff, J. Kuhn, and S. Küchemann, Educational data augmentation in physics education research using chatgpt, Physical Review Physics Education Research **19**, 020150 (2023).

[10] 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, Physical Review Physics Education Research **19**, 020123 (2023).

[11] R. Wink and W. M. Bonivento, Artificial intelligence: New challenges and opportunities in physics education, New Challenges and Opportunities in Physics Education , 427 (2023).

[12] 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, Physical Review Physics Education Research **19**, 010142 (2023).

[13] N. L. Rane, S. P. Choudhary, A. Tawde, and J. Rane, ChatGPT is not capable of serving as an author: ethical concerns and challenges of large language models in education, International Research Journal of Modernization in Engineering Technology and Science **5**, 851 (2023).

[14] 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?, Physical Review Physics Education Research **19**, 020128 (2023).

[15] G. Polverini and B. Gregorcic, How understanding large language models can inform the use of ChatGPT in physics education, European Journal of Physics **45**, 025701 (2024).

[16] G. Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course?, Physical Review Physics Education Research **19**, 010132 (2023).

[17] G. Kortemeyer, Toward AI grading of student problem solutions in introductory physics: A feasibility study, Physical Review Physics Education Research **19**, 020163 (2023).

[18] K. Wancham and K. Tangdhanakanond, Effects of feedback types and opportunities to change answers on achievement and ability to solve physics problems, Research in Science Education **52**, 427 (2022).

[19] C. Cobbold and L. Wright, Use of formative feedback to enhance summative performance., Anatolian Journal of Education **6**, 109 (2021).

[20] P. Bazelais and T. Doleck, Investigating the impact of blended learning on academic performance in a first semester college physics course, Journal of Computers in Education **5**, 67 (2018).

[21] Open AI, ChatGPT-3.5, https://chat.openai.com/ (2023).

[22] Open AI, Prompt engineering, https://platform.openai.com/docs/guides/prompt-engineering (2023).

[23] M. H. Towns, Mixed methods designs in chemical education research (ACS Publications, 2008).

[24] V. Braun and V. Clarke, Using thematic analysis in psychology, Qualitative research in psychology **3**, 77 (2006).

[25] Google for Developers, Introduction to Large Language Models Machine Learning, https://developers.google.com/machine-learning/resources/intro-llms (2023).

[26] J. W. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo, S. Panzeri, G. Manzi, *et al.*, Testing theory of mind in large language models and humans, Nature Human Behaviour , 1 (2024).

[27] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems **35**, 24824 (2022).