Introductory physics students' valued features in AI feedback generated from self-crafted and engineered prompts

Amogh Sirnoorkar and N. Sanjay Rebello

Department of Physics and Astronomy and Department of Curriculum and Instruction, Purdue University, West Lafayette, Indiana 47907, USA

(Received 11 September 2025; accepted 29 April 2026; published 16 June 2026)

Generative artificial intelligence (AI) has shown promising potential in education, particularly in providing real-time, customized, student-facing feedback in large-enrollment courses. A crucial skill that mediates effective use of AI is the systematic structuring of natural language instructions to AI models, commonly referred to as prompt engineering. This study has two objectives: (i) to investigate the sophistication of student-generated prompts when seeking feedback from AI and (ii) to examine trends associated with students' preferences and perceptions of features in AI feedback generated from self-crafted prompts vs prompts incorporating prompt-engineering techniques and principles of effective feedback. Results indicate that student-generated prompts typically reflect only a subset of foundational prompt-engineering techniques. We also identified 12 distinct features that students attribute the usefulness of AI-generated feedback, spanning four broader themes: (i) evaluation, (ii) content, (iii) presentation, and (iv) depth. Finally, students overwhelmingly prefer feedback generated from structured prompts, particularly those combining prompt-engineering techniques with principles of effective feedback. Implications of these results—such as integrating the principles of effective feedback in design and delivery of feedback through AI systems, and incorporating prompt engineering in introductory physics courses—are discussed.

DOI: 10.1103/v3pj-8491

I. INTRODUCTION

Providing student-facing feedback is a critical component of pedagogical practice. Feedback is any information provided by an agent about one's performance [1]. The sources of feedback can be either external or internal. In the context of student learning, peers or teachers can act as external sources, whereas self-monitored observations can act as internal sources of feedback. This information can take various forms such as corrective, directive, alternative, or affective [2]. Effective feedback plays a crucial role in improving students' learning along with enhancing their skills and academic performance [3]. It is also instrumental for motivating students to actively engage with content knowledge, with qualitative feedback being more effective than quantitative feedback [4]. Researchers thus have called for students' active involvement in the feedback process by facilitating them in interpreting, internalizing, and acting on the provided information [5].

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

The process of human- or technology-based feedback generation broadly involves two steps: (i) rubric-based grading of student responses, followed by (ii) communication of results through performance reports [6,7]. In large-enrollment courses, this process can be challenging due to constraints associated with human expertise, material costs, timely delivery of information, and incorporation of effective design features. To mitigate these constraints, researchers have explored ways to reliably automate the design and delivery of static feedback [8,9]. However, provision of customized and personalized feedback by building on student responses remains a challenge.

Recent advancements in generative artificial intelligence (AI) have promised huge potential to facilitate customized feedback. This potential stems from AI's ability to process multimodal data and generate real-time, coherent, and personalized responses based on user inputs (or "prompts") [10–13]. In physics education, considerable attention has been devoted to grading student responses (the first step of the feedback generation process described earlier) using AI models. Studies in this domain have leveraged AI to grade students' handwritten solutions [14], including strategies on answering conceptual questions [15–17] and numerical problems [18]. This strand of research broadly notes that the evolving AI models are becoming increasingly effective

at grading physics solutions. As for the generation of performance reports (the second step) using AI, studies have primarily focused on highlighting ideal answers [19], correctness of student responses [20,21], alignment of responses with grading rubric [22], strengths, weaknesses, and guidance to improve responses [23], and content-based feedback within word limits [24]. In addition, these studies also note the sensitivity of the AI models to user instructions in generating the desired results.

We contribute to these ongoing efforts by exploring students' prompting approaches to AI models when seeking feedback about their solutions to a physics problem. We also unpack the broad trends in students' preferences and valued qualitative features in AI feedback generated from self-crafted prompts and provided prompts combining prompt-engineering techniques and principles of effective feedback. In doing so, we answer the following research questions:

RQ1: To what extent do students' self-crafted prompts evidence the features of prompt-engineering techniques?

RQ2: What are students' preferences in AI feedback generated from (a) self-crafted prompts, (b) prompts incorporating prompt-engineering techniques, and (c) prompts combining prompt-engineering techniques with principles of effective feedback?

This manuscript is structured as follows: in the next section, we review and identify gaps in the literature on AI-generated feedback in education. In Sec. III, we discuss the background on prompt engineering and principles of effective feedback. In Sec. IV, we discuss the study's data, context, and analysis. In the following section, we present the results before discussing them in Sec. VI. We then conclude with study's implications, limitations, and directions for future research in Sec. VII.

II. LLM-MEDIATED FEEDBACK IN EDUCATION

In this section, we review the literature on AI-mediated feedback in education research, beginning with language education, followed by science, technology, engineering, and mathematics (STEM), and finally, physics education. We then discuss the literature specifically associated with use of prompt engineering and principles of effective feedback in design and delivery of feedback through AI models. In the end, we conclude by discussing the gaps in this literature and how the current study addresses these gaps.

A. Language education

A major section of work on AI-mediated feedback in education literature revolves around the context of learning English as a foreign language. Methodologically, studies have employed experimental designs with students in treatment groups receiving AI-mediated feedback, while control groups interacting with feedback from teachers, peers, or experts.

Comparisons between effectiveness of AI- and teacher-generated feedback to facilitate students' language learning have revealed mixed results. In studies favoring AI, large language models have been observed to provide detailed feedback by focusing on content, vocabulary, organization, and linguistic accuracy [25]. As a result, students have demonstrated improvements in their revised texts by spending more time processing the inputs [26]. Additional reported benefits include gains in students' writing skills, engagement, motivation to revise, organizational proficiency, coherence, grammar, vocabulary, perceived competence, and reduced anxiety [27–33].

On the other hand, several studies have highlighted human feedback to be equally effective [34,35] or even better than AI-generated feedback [36–38]. When provided with feedback from both experts and AI, researchers have observed students struggling to identify the source and rating the feedback as less favorable upon learning its source as AI [36]. Furthermore, as compared to AI, human raters have been observed to provide clearer, specific, and corrective inputs through supportive tone [38,39]. In a comparative analysis of narrative writing, Tabari *et al.* [40] observed that feedback from teachers better supported improvements in accuracy and lexical complexity, while feedback from AI models was effective for syntactic complexity. When compared with peer feedback, AI tended to be more descriptive about how an essay was written, whereas peers focused more on identifying errors within the text [41].

Given these varied findings, researchers have advocated for hybrid approaches that blend the strengths of human and AI-generated feedback to enhance students' language learning [42]. Such hybrid models, grounded in active "human-in-the-loop" principles, have been shown to address a broader range of error types by increasing students' uptake of feedback and promote higher-order writing skills such as organization, critical thinking, and sentence variety [43,44]. Additionally, such hybrid models have been noted to enhance preservice teachers' confidence in their ability to teach and assess writing [45].

B. STEM education

Within STEM, research on AI-mediated feedback involves multiple disciplinary contexts, including physics, mathematics, programming, and bioengineering. Studies in these areas have primarily examined the comparative effectiveness of AI- vs human-generated feedback, students' perceptions of AI feedback, and instructional approaches that promote active teacher involvement.

Similar to language education literature, studies comparing human- and AI-generated feedback in STEM contexts too reflect mixed patterns of effectiveness. Ghoochani *et al.* [46] found that nearly half of their participants in an introductory programming and data science course

could not reliably distinguish between AI- and peer-generated feedback on their individual programming projects. Çağlar-Özhan *et al.* [47] and Bauer *et al.* [48] reported no significant differences between AI- and teacher-generated feedback in improving their student performances in project design and statistics, respectively. While Çağlar-Özhan *et al.* [47] found that students rated both feedback types similarly in terms of quality, Bauer *et al.* [48] observed that static teacher feedback was more supportive of students' interest in statistics. Zhang *et al.* [49] observed students in their programming course, highlighting their preference for specificity, clarity, and corrective nature of the feedback from AI models. Furthermore, AI feedback tended to emphasize explanations for why answers were correct or incorrect, with fewer instances offering deeper conceptual insights, learning strategies, or self-regulatory advice [50].

Conversely, studies have also observed AI-mediated feedback to outperform human inputs, particularly in providing direct suggestions and rewritten responses [51]. Despite concerns over consistency, first-year computing students have perceived AI feedback as a catalyst for learning and self-improvement [52]. Furthermore, AI feedback has been shown to better support students in solving problems, enhance learning in programming, improve performance outcomes, and reduce the frequency of students exceeding time limits in completing programming tasks [53–55]. Additionally, Steinbach *et al.* [56] observed that students who engaged with AI feedback containing hallucinations demonstrated greater learning gains compared to those who received feedback without hallucinations.

C. Physics education

In physics education, research on AI-generated feedback revolves around three broad themes: (i) feasibility of generating feedback using open-source AI models, (ii) students' perceptions of AI-generated feedback, and (iii) the impact of AI-generated feedback on student learning. Studies addressing the first theme, i.e., feasibility, highlight several affordances and constraints. For instance, El-Adawy *et al.* [19] compared human- and AI-generated feedback to students' responses to introductory mechanics problems. The authors observed that while human feedback often focused on highlighting key errors and effectively managing "edge cases," feedback from GPT-3.5 provided elaborate, personalized, and detailed descriptions using examples. Similarly, Chen and Wan [22] reported instructors rating feedback from GPT-4o highly when responding to students' reasoning explanations to physics problems. Meyer *et al.* [20] further demonstrated that GPT-4o achieved an accuracy of 85% in characterizing high school students' responses to physics problems, with strong interrater agreement on feedback quality.

Research within the second theme, i.e., student perceptions of the AI-generated feedback in physics contexts, broadly highlights positive outcomes. Mills *et al.* [23] observed that despite concerns of occasional inaccuracies, undergraduate students perceived AI-generated feedback on their physics lab reports as actionable, clearly written, and useful. Similarly, Allen *et al.* [24] observed that introductory physics students regarded AI-generated feedback on their arguments as useful, with these perceptions remaining consistent regardless of solution correctness or task difficulty. Wan and Chen [21] further reported that students rated AI-generated feedback as comparable to human-generated feedback in terms of correctness, while rating AI-generated feedback highly on usefulness.

Finally, researchers have also explored the varying effectiveness of AI-mediated feedback on students' learning (the third theme) based on usage patterns. Dai *et al.* [57] highlight that the effectiveness of AI feedback depends on students' agency: mandatory use improved performance among low-achieving students but hindered medium- and high-achieving students' performance and self-regulation, whereas optional use benefited high-achieving students but reduced self-regulation in low-achieving students.

D. Use of prompt engineering and principles of effective feedback

The design of feedback through AI models is often guided through carefully crafted instructions or "prompts" which dictate the desired content, structure, and tone of the feedback statements. Thus, studies discussed in the previous subsections have explored various prompting techniques in crafting instructions to AI models to yield desired feedback. These include role prompting (instructing AI models to assume roles such as course instructor) [22,29,43,58,59], few-shot prompting (providing exemplar feedback) [55,58], zero-shot (no examples) [40], retrieval augmentation (accompanying prompt with information source) [57], and chain-of-thought (asking models to reason "step by step") [20,46,57]. Furthermore, to better facilitate student engagement with the feedback, researchers have also focused on crafting prompts with emphasis on formatting feedback reports [20,38,59–61] and specifying word limit [20,46,61,62].

In addition, a subset of studies have also leveraged principles of effective feedback from education research. These principles include focusing on strengths of student responses, clarifying errors or areas for improvement, and providing effective recommendations [38,61,63]. Other features such as including the use of supportive tone [38] and professional language [59] have also been adopted to better facilitate uptake of the feedback. Details pertaining to prompt engineering and principles of effective effective feedback are discussed in Sec. III A.

E. Gaps in the literature and the current study

As discussed in the previous subsections, literature on AI-mediated feedback in education research has primarily

focused on designing feedback through researcher-crafted (or expert) prompts and exploring its efficacy on student learning. Research on investigating student approaches toward crafting prompts, particularly in the context of seeking feedback from AI platforms on their physics solutions, remains relatively underexplored. Unlike language learning, providing feedback in physics reasoning contexts often involves analyzing the coherence between multiple external representations such as equations, graphs, and words in student solutions and comparing it with desired results. This feature contributes to added complexity in how students seek and interpret feedback from AI models in physics contexts. Along the same lines, research on student perspectives and preferences on AI-mediated feedback generated from prompts varying in features based on prompt-engineering techniques and principles of effective feedback—particularly in physics education—remains scarce.

In the current study, we address these gaps by exploring students' prompting approaches while seeking feedback about their arguments to a physics problem. We also explore students' preferences and their valued features in feedback generated through their self-crafted prompts, prompts entailing the prompt-engineering techniques, and those combining the prompting techniques and principles of effective feedback.

III. BACKGROUND

A. Prompt engineering

Prompt engineering is an empirical and emerging domain that deals with the "systematic design and optimization of input prompts to guide the responses of large language models, ensuring high levels of accuracy, relevance, coherence, and usability in the generated output" [64]. With the growing capability and accessibility of AI models, prompt engineering has been increasingly recognized as an essential competency for the future workforce in STEM contexts.

Broadly, prompt-engineering techniques can be categorized as foundational and advanced. Foundational techniques emphasize clarity, structure, and organization of prompts to optimize outputs. Advanced techniques, on the other hand, involve more sophisticated and dynamic interactions with the responses (such as directing the model "to think step by step") to yield precise, accurate, and high-quality outputs. In this study, we primarily focus on and describe the foundational techniques and recommend Ref. [64] for further details. Below, we briefly outline the foundational techniques adopted in this study.

1. *Providing instructions.* This involves designing directives that guide the AI model toward the intended and relevant output. For example, consider the objective of understanding the significance of an object's resistance (R) with respect to temperature (T). A prompt such as "relation between R and T" may yield broad and diverse results with no clear contextual information in the statement. Additionally, since R and T could correspond to notations for numerous physical quantities (e.g., molar gas constant or time period), the prompt may invoke diverse responses.
2. *Being clear and precise.* Crafting clear and precise instructions helps the model generate accurate and relevant output. In contrast, general instructions can often lead to overly broad or ambiguous responses. For example, if the objective is to gain a basic understanding of electrical resistance, the prompt "*Explain the meaning of R in science*" may yield broad or contextually irrelevant results.
3. *Role-based prompting.* This feature assigns a role to the AI model to yield task-specific output. Within this prompting technique, two subtypes can be distinguished: static role prompting and dynamic role-play prompting. Static role prompting involves assigning a fixed role to the model to enhance contextual accuracy and task-specific performance. ExpertPrompting is one such approach, where the AI is explicitly defined as a domain expert. For example, in the context of grading numerical physics problems, static role prompting (aligned with ExpertPrompting) might take the form: "You are an expert physics instructor tasked with grading student responses …"

   Dynamic role-play prompting, in contrast, varies the assigned role to the AI model based on user interaction. Following up with the grading example above, role-play prompting could extend the static role by supplementing the instruction with: "You are an expert physics instructor tasked with grading student responses …" Now given the response, as a mathematics instructor, expand on the mathematical principle and convey the geometric implications of the results." Such dynamic variation within a single interaction is likely to enhance the output's accuracy.
4. *Use of delimiters.* When the task is complex and the prompt has multiple components, delimiters such as triple quotes or custom symbols (*) are used to separate the components. Clear demarcation between the prompt components helps AI generate more coherent and structured output.
5. *Trying several times.* Also referred as "resampling," this technique involves running the model multiple times and selecting the best candidate output. It is particularly beneficial given the nondeterministic nature of AI models in generating outputs.
6. *Zero-shot, one-shot, or few-shot prompting.* This feature corresponds to providing none (zero-shot), single (one-shot), or multiple examples (few-shot prompting) to draw from and generate responses.

In zero-shot prompting, the model relies on its large training data for generating responses.

7. *Variations of model settings.* Varying the parameter settings of the AI model, such as "temperature," can also influence the outputs. The temperature parameter regulates the degree of randomness of the generated output. A lower temperature parameter can lead to higher deterministic outputs and vice versa. However, for the recent AI models, this hyperparameter tuning cannot be performed at the user level.

In the current study, we adopt a subset of these prompting techniques for two purposes: (i) to craft prompt templates that students use to seek feedback about their arguments to a physics problem and (ii) to analyze students' self-crafted prompts while seeking feedback. Among the seven, we adopt five techniques by omitting the fifth and the seventh from the list. Furthermore, we adopt zero-shot prompting (sixth in the list) by not embedding any examples in our prompt templates. Table I, particularly the prompts corresponding to feedback B and C reflects the five techniques operationalized in our study's context. As a side note, the presence of all the above techniques does not necessarily ensure the sophistication of prompts. Indeed, it is possible that a sophisticated prompt may lack a few techniques described above.

B. Principles of effective feedback

Feedback is any information provided by an agent (people or material artifacts) focused on aspects of one's performance [1]. As an integral component of both student learning and teaching practices, feedback serves to bridge the gap between current performance and desired goals. The effectiveness, however, depends on the design, content, and structure of the feedback. Effective feedback entails the following three components [1,2,5]:

1. Desired performance (*"Where is the learner going?"*): This feature involves explicitly and clearly specifying the goals associated with a given task or performance. Clear communication of goals helps students set appropriate self-directed objectives, thereby providing a strong impetus for learning. This aspect of feedback is often referred to as "Feed Up." For example, in the context of physics problem solving, this might correspond to clarifying the expected product (e.g., the correct answer) or the process (e.g., use of relevant mathematical formalisms).
2. Current performance (*"How is the learner going?"*): Clarifying the current state of performance, often in relation to the desired performance. Also referred to as "Feed Back," it includes highlighting both strengths as well as gaps in student understanding through probing questions. Such a "praise and question" approach can support students' self-efficacy, enhance their motivation to engage with feedback, and enable them to develop as self-regulated learners.
3. Concrete approaches to fill the gaps (*"Where to next?"*): Prescriptions about addressing gaps between current and desired performances. It may include cognitive elements such as task descriptions, specific areas for improvement, actionable suggestions, and, where possible, plans for implementing changes. This component is often referred to as "Feed Forward."

In addition to these components, the timing of feedback plays a crucial role in its effectiveness. Immediate task-related feedback can be beneficial for simpler tasks, whereas delayed, process-oriented feedback may be more effective for complex or challenging tasks.

In the present study, one of the prompt templates provided to students for eliciting feedback from AI ("Feedback C" in Table I) incorporated the above three components. Furthermore, the AI-generated real-time feedback addressed the timing factor.

TABLE I. The three types of feedback (first column) and corresponding prompt statements (second column) embedding the foundational prompt- engineering techniques and principles of effective feedback discussed in Sec. III.

Feedback	Prompt
Feedback A	[Student-generated prompt].
Feedback B	"""You are an expert physicist and your objective is to give feedback on my answer which is presented as an argument with a claim, evidence and reasoning about a physics problem.""" """The problem involves identifying the physical significance of a graph's slope with the gravitational potential energy of a satellite (U) plotted on y-axis and the satellite's distance from the planet's center (r) on x-axis.""" """The following is my answer (argument): [INSERT YOUR ARGUMENT HERE].""" """Provide relevant and useful feedback for my argument."""
Feedback C	[Prompt for feedback B (except for the last sentence)] + """Provide feedback by explicitly highlighting the correct answer, my provided answer (including strengths and limitations), and gaps (if any) between them within 200 words. Also suggest three potential ways through a bulleted list which can help me improve my performance on similar questions in future."""

IV. METHODS

A. Data and context

Our data comes from introductory students' responses to an extra-credit assignment at a Midwestern land-grant R1 university in the United States. The assignment was part of a large-enrollment, calculus-based physics course for future engineers and was administered through Qualtrics [65] during Spring 2025. The course follows the "*Matter and Interactions*" textbook by Chabay and Sherwood [66]. It also includes assignments in which students constructed arguments using the "claim, evidence, and reasoning (CER)" structure [67,68].

The assignment focused on the variation of a satellite's potential energy with respect to its distance from a planet's center. Hypothetical data corresponding to the satellite's altitude from the planet's surface, its distance from the planet's center, and associated potential energy were provided in a table (Fig. 1). The task consisted of three parts. The first part focused on determining the planet's radius and magnitude of the acceleration due to gravity on its surface. Students were asked to submit their written solutions in PDF or image file formats. The task's second part focused on interpreting the physical significance of the slope of a graph with gravitational potential energy plotted against the satellite's distance from its center. Students were asked to communicate this significance through arguments structured as CER.

The final part required students to seek feedback from AI platforms of their choice on their arguments about the graph's slope (second task). The task was intentionally designed to encourage students to include relevant details in their prompts rather than copying and pasting the problem statement. This requirement to craft detailed prompts allowed us to gather data and explore how students constructed prompts through the lens of prompt-engineering techniques. Students interacted with AI through three prompt types: (i) self-crafted prompts, (ii) a provided prompt incorporating foundational prompting techniques (Sec. III A), and (iii) a provided prompt combining foundational prompting techniques with principles of effective feedback (Sec. III B). The corresponding feedback from AI to these three prompts are henceforth referred to as "Feedback A," "Feedback B," and "Feedback C." Thus, all students received the three feedback types in the same order and they could only read subsequent prompts after completing the preceding activity. For the provided prompts, students were required to insert their arguments in highlighted areas. No specific AI platform was recommended, allowing students to turn to the platform of their choice and enabling more authentic observations of their interactions. The survey did not allow students to go back and modify their prompts or their responses.

Information regarding both their self-crafted and provided prompts along with corresponding responses from AI (feedback statements) were collected. In the final stage of the task, participants were asked to rank the three feedback types by perceived usefulness and provide detailed justifications. They were also asked to list up to three things (if any) they had learned from the activity and to reflect on its usefulness for their learning. Our analysis focuses on this third part i.e., students' seeking of feedback, their rankings, and justifications.

You are an intern in an astrophysics laboratory, monitoring the motion of a satellite of mass $500kg$ around a planet (with mass $6.42 \times 10^{23}\ kg$) in a nearby galaxy. You receive data the following from the observatory highlighting the satellite's gravitational potential energy with respect to its distance from the planet's center.

Altitude 'h' (km)	Distance 'r' (m) $\times 10^6$	Potential energy 'U' (J)
100	3.49	−6.11
200	3.59	−5.94
300	3.69	−5.78
400	3.79	−5.62
500	3.89	−5.48

(1). Based on the provided data, produce a detailed written solution (including the concepts, principles, assumptions and approximations, and your strategy) to determine the values of the planets: (a) Radius (b) Magnitude of acceleration due to gravity on its surface.

Submit your detailed written solution in a pdf or image file format.

(2). Consider a plot of U (y-axis) vs r (x-axis). What would be the physical significance of the graph's slope? Detail your argument by explicitly highlighting your Claim, Evidence, and Reasoning using relevant equations, diagrams, and physics principles.

FIG. 1. Statement of the entire task. Problem 2 was the specific task for which students sought feedback from AI.

B. Analysis

Of the 2044 students enrolled in the course during the Spring 2025 semester, 1661 responded to the extra-credit activity (81% response rate). The data were downloaded into a spreadsheet, and incomplete responses were removed, resulting in 1235 qualified responses. To address the first research question, students' self-crafted prompts were analyzed through the lens of foundational prompting techniques, with a particular focus on the first four (*Providing instructions*, *Being clear and precise*, *Role-based prompting*, and *Use of delimiters*). The remaining three techniques: "trying several times," "one-shot/few-shot

prompting," and "variation of model settings" were neither included in the provided prompt design nor in the analysis of student-generated prompts. We characterize sophistication of student-crafted prompts by the extent to which the prompts entail the four techniques. As acknowledged in Sec. III A, a good prompt can be devoid of a subset of these techniques, which forms a crucial limitation of our results. However, given the objective and scope of this study (students seeking feedback about their physics arguments), the four techniques provide a reasonable benchmark to capture students' prompting approaches.

The first four techniques were coded dichotomously ("y" or "n"), with y indicating the presence of evidence and n indicating its absence. Prompts that included image or PDF attachments (17.6%) were reviewed and coded separately. Evidence of the first technique, *providing instructions* was coded as present if students' prompts included expectations from the AI model, such as seeking feedback or evaluating their arguments. Any instruction or request in students' prompts about their arguments was considered evidence for this technique. The second technique, *being clear and precise* was coded into two subcategories: (i) explicit description of the task's objective as identifying the slope of a graph of gravitational potential energy vs distance from the planet's center, and (ii) explicit specification of the student's argument or their final claim. Both subcategories were coded separately on a binary scale (y or n) and only if students evidenced both components, the prompt was coded as evidencing the second technique. Evidence for the third technique, *role-based prompting,* was coded as present if the prompt assigned a role to the AI model or else was marked as absent. In a similar fashion, the evidence for the final technique, *use of delimiters,* was also coded. A total of 84 student prompts (around 7%) were excluded since they were either unanswered or sought feedback for earlier parts of the task.

For the second research question, we unpacked trends in students' ranking of the three feedback types through the use of basic spreadsheet functions such as "COUNTIF." Thereafter, the first author employed emergent coding approach [69] to analyze students' justifications for their preferred feedback types. Justifications that were left unanswered or were difficult to interpret were removed, leaving 1160 valid responses from the initial 1235. These responses were analyzed in two stages. In the first stage, emergent codes were identified by noting the features that students referred to while justifying their choice. A "code sheet" was then created documenting each feature along with its corresponding meaning (Table III). In the second stage, each student's justification was coded in binary ("yes" or "no") based on the features identified in the first stage to capture their prevalence across the dataset. Two independent researchers reviewed the codebook and coded 25 random responses each. Their coding was compared and discussed until full agreement was reached. No major additions or deletions were made to the original codes identified in the first stage after the discussion.

V. RESULTS

In the rest of this section, we detail the results of the two research questions outlined in Sec. I. The first question examines the extent of sophistication evidenced in student-generated prompts through the lens of foundational prompt-engineering techniques (Sec. III A). The second question focuses on students' ranking of the feedback types and the emergent themes of the feedback features that students attributed to the perceived usefulness of the feedback. Below, we discuss each research question.

A. Prompt sophistication

RQ1: To what extent do students' self-crafted prompts to AI evidence the features of foundational prompt-engineering techniques?

As discussed in Sec. IV A, we evaluated students' self-crafted prompts in terms of prompt-engineering techniques, focusing on *providing clear instructions*, *being clear and precise*, *role-based prompting*, and *using delimiters*. The distribution of these features is presented in Figs. 2 and 3, while Table II provides exemplar prompts evidencing each technique.

1. Providing clear instructions

Overall, 90.2% responses evidenced the technique of providing clear instructions in their prompts by explicitly seeking feedback for their arguments. These instructions were framed as requests ("Please review and provide feedback for the argument"), commands ("Give me feedback on this argument"), or questions ("Can you let me know if it makes sense and if I explained the idea clearly?"). The remaining 9.8% prompts provided only arguments (with no explicit instruction to provide feedback), and yet

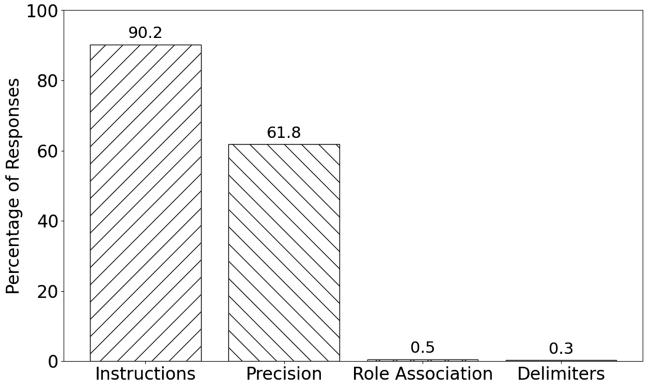


FIG. 2. Graph highlighting the percentage of student-crafted prompts reflecting the foundational prompt-engineering techniques. Each column correspondingly represents the following techniques: "Providing instructions," "Being clear and precise," "Role-based prompting," and "Use of delimiters."

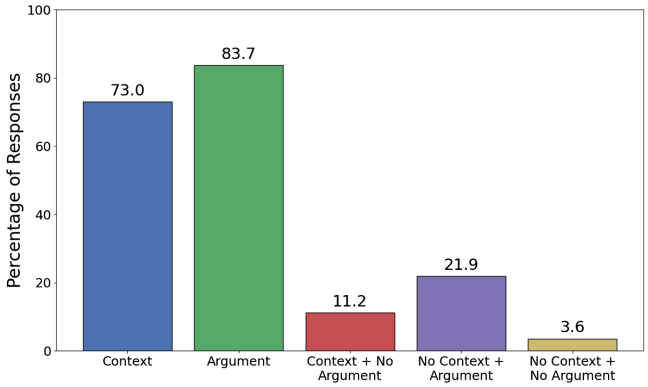


FIG. 3. Graph highlighting the details of student-crafted prompts reflecting the foundational technique of "being clear and precise." This technique was coded when students described the problem's context and also attached their arguments.

the model consistently responded with feedback. For instance, for one of the student's prompt with no explicit instruction for feedback:

> The slope of the gravitational potential energy U vs r graph is equal to the negative of the gravitational force acting on the satellite. From the equation $U = -GMm/r$ the derivative is $dU/dr = GMm/r^2$. The slope of the U vs r graph tells us how fast potential energy changes with distance. The force is the negative rate of change of potential energy. The slope of this graph represents gravitational force.

The AI responded with feedback:

> Your argument demonstrates a strong conceptual understanding of the relationship between gravitational potential energy and gravitational force. You correctly identified that the slope of the U vs r graph represents the derivative and you clearly linked it to the gravitational force.

2. Being clear and precise

To capture the second technique, *being clear and precise*, we analyzed students' prompts across two components: (i) explicit description of the task's context by specifying the slope of gravitational potential energy vs distance graph and (ii) explicit specification of their arguments. If students evidenced both components, prompt was coded as evidencing this technique. While 73% of students clearly described the graph's context and 83.7% included their arguments, only 61.8% demonstrated both together. In contrast, 11.2% provided context without arguments, 21.9% included arguments with unclear context, and 3.3% provided neither. These trends are represented in Fig. 3.

TABLE II. Exemplar student-crafted prompts reflecting the foundational prompt-engineering techniques. Section of the prompt highlighting the technique has been underlined.

Prompting technique	Example prompt
Providing instructions	*The slope of a graph (U vs r, gravitational potential energy vs distance) is gravitational force because Newton's universal gravitation law tells us that $F = GMm/r^2$ and the derivative U with respect to r gives us the same equation. Can you give me feedback on this claim?*
Being clear and precise	*Hey! I'm working on a physics assignment and had to explain what the slope of a gravitational potential energy vs distance (U vs r) graph means. I wrote my answer using a Claim, Evidence, and Reasoning format. Can you let me know if it makes sense and if I explained the idea clearly?* *Claim: [I think the slope of the graph shows….]* *Evidence: [From the data, I saw that…]* *Reasoning: [If the slope is steep, it means….]*
Role-based prompting	*You are a specialized physicist on the gravitational energy field. I had a task about finding the physical significance of the slope in a plot that compared the gravitational potential energy (U) in the y-axis, with the radius (r) in the x-axis. I need you to give me brief feedback (keep it to a maximum of 200–300 words) of my response to this question. My response has been attached.*
Use of delimiters	*Hey ChatGPT! Based on the following definition of claim, evidence, and reasoning, please give feedback about the argument written in claim evidence and reasoning format!* *The definitions of claim, evidence, and reasoning will be pasted between asterisks (*) and the argument presented by a student will be pasted between quotations (").* *Please give feedback based both on the validity of the argument and the adherence to the formatting of claim, evidence, and reasoning.* *Thanks!* **A Claim is every decision […]** *"The slope of GPE vs r represents […]"*

Given these variations, we further examined how AI responses varied when prompts lacked clear or complete context descriptions. Of the 293 such cases, AI correctly interpreted the context and generated feedback in 285 instances. In only eight cases, AI sought additional clarifications or recommend significant rewording of the prompt. For example, for one student's prompt with incomplete context regarding the type of energy (potential vs gravitational potential):

> Provide me with feedback about my claim, evidence, and reasoning regarding the potential energy and distance relationship. Claim: As the distance increases, the potential energy increases. Evidence: There is a positive slope between the distance and potential energy. Reasoning: Therefore there is a positive and linear relationship between potential energy and distance. However, the magnitude of the potential energy decreases as distance increases.

AI responded by suggesting clarity in the argument:

> Your claim, evidence, and reasoning contain important observations, but there are a couple of points that could be clarified for better consistency and accuracy. Claim: The statement "As the distance increases, the potential energy increases" is straightforward, but it assumes a specific scenario. Clarify whether this refers to gravitational potential energy, elastic potential energy, or another type of potential energy, as the relationship can vary depending on the context.

3. Role-based prompting and use of delimiters

The third and fourth techniques (*role-based prompting* and the *use of delimiters*) were the least evidenced. Of the 1151 responses, only six demonstrated role-based prompting, and only three included explicit use of delimiters such as quotes and asterisks.

B. Preferences and perceptions of AI-feedback

RQ2: What are students' preferences in AI feedback generated from (a) self-crafted prompts, (b) prompts incorporating prompt-engineering techniques, and (c) prompts combining prompt-engineering techniques with principles of effective feedback?

We address this question in three segments by exploring (i) trends in students' rankings of the feedback types, (ii) themes characterizing the features students value in AI-generated feedback, and (iii) the prevalence of these themes across the different feedback types. Each of these segments are discussed below.

1. Trends in students' preferences

As noted in Sec. IV, students used AI models of their choice to generate three distinct feedback statements through three different prompts: (i) self-crafted prompts (resultant feedback referred to as "feedback A"), (ii) provided prompts that entailed features of foundational techniques of prompt engineering (feedback B), and (iii) provided prompts that entailed the combination of foundational techniques and the principles of effective feedback (feedback C). Students then ranked the three feedback statements generated by AI based on their perceived usefulness. The associated ranking trends are highlighted in Fig. 4.

Across 1235 responses, around 61.4% of students ranked feedback C as the most useful, followed by B (29.7%) and A (8.9%). For their least preferred feedback, a major section of students ranked feedback A (61.2%), then C (30.2%), and finally B (8.6%). These results highlight three overarching trends about student preferences. First, a substantial section (91.1%) of students preferred feedback generated using structured prompts (combined first choices across B and C). Within this preference, a majority of students preferred the feedback generated using the prompt with combined features of prompt engineering and effective feedback (C). Second, this top choice was either liked or disliked by students. This is evident from 61.7% indicating C as their top choice and 30.1% preferring it as their least one. Only 8.2% however highlighted C as their second choice. Finally, students' least preferred feedback corresponded to the one generated using their self-crafted prompts.

2. Themes associated with AI-generated feedback

Based on the emergent coding of students' justifications of their preferred feedback types, we identify the following four themes associated with students' valued features in AI-generated feedback: (i) evaluation, (ii) content, (iii) presentation, and (iv) depth. Table III presents the description of

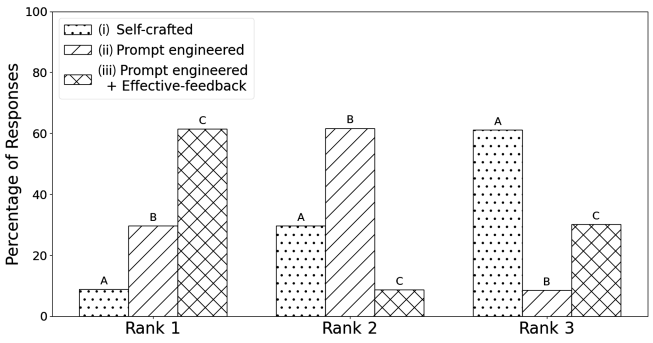


FIG. 4. Plot highlighting students' ranking of the three types of feedback generated from prompts that (i) were self-crafted, (ii) entailed features of prompt engineering, and (iii) entailed features of prompt engineering along with principles of effective feedback.

TABLE III. Themes and corresponding features associated with the features of feedback evidenced in students' descriptions of their preferred feedback type. Themes are highlighted in the bold text. Numbers in the parenthesis represent the percentage of student responses which evidenced at least one of the associated features.

Theme/code	Description
Evaluation (69.4%)	**Strengths and shortcomings in arguments or suggestions for improvement**
E1: Affirmation	Appreciation or acknowledgment of the correct sections of the argument
E2: Critique	Pointing out errors or miscalculations in the argument
E3: Guidance	Suggestions for improvement or pointing to specific areas of deliberate practice
Content (55.2%)	**Exemplar correct answers or physics/mathematical principles or formulas or generic argument or argument in the form of claim, evidence, reasoning**
C1: Correctness	Exemplar "correct" solutions OR the "right approach"
C2: Conceptual	Physical principles or mathematical relations
C3: Schema	Focus on argument or on claim, evidence, and reasoning structure
Presentation (47.8%)	**Feedback's clarity or structured formatting or brevity or ease of understanding**
P1: Clarity	Clear communication of feedback information or description of the task
P2: Structure	Well-formatted structure of feedback or feedback with distinct sections
P3: Conciseness	Less on "wordiness" but more on useful information
P4: Intelligibility	Easy-to-understand or simple phrasing
Depth (44.2%)	**Detailed description or detailed solution or task-related specificity of feedback**
D1: Detail	Detailed account of student solution or detailed feedback
D2: Specificity	Task-specific feedback or feedback with strong focus on student solution

each theme along with the associated features. Table IV provides exemplar student quotes that reflect each feature. Below, we describe each theme and associated features.

1. Evaluation (E)

The first major theme corresponds to features of feedback that critically evaluate the strengths and shortcomings of student solutions along with a clear description of ways for improvement to better facilitate future performance on similar tasks. Three features are associated with this theme: (E1) affirmation (acknowledgment of the solution's correctness or strengths), (E2) critique (highlighting errors or shortcomings of the solutions), and (E3) guidance (actionable suggestions for improvements). At least one of these features was evidenced in the justifications of 65.2% of student descriptions. The following is one of the quotes that evidences all three features:

> I found Feedback B the most useful because it gave a clear evaluation of my argument and highlighted areas of improvement without being vague. It confirmed what I did well (identifying the slope as the negative of the force) and then pointed out small gaps, like not fully explaining the curvature of the graph and not connecting the shape to the inverse-square nature of gravity. This helped me reflect more deeply on physics, not just the formula. I also appreciated the suggestion to use real-world examples, which would make my explanation stronger and more relatable.

The student's reference to feedback acknowledging what was done well, i.e., identifying the graph's slope as the negative of the gravitational force between the planet and satellite, illustrates the "affirmation" feature. In addition, the feedback's identification of minor gaps in the solution, along with recommendations to include real-world examples in the arguments, reflects the "critique" and "guidance" features, respectively.

2. Content (C)

This theme reflects feedback features that predominantly focus on content knowledge. These include the elucidation of the "correct" approach to solving the physics problem or the specification of the "correct answer." The theme also encompasses the application of physics and mathematics principles, along with relevant equations. In addition, it involves explicit discussion of content knowledge embedded within the argumentation structure, i.e., claim, evidence, and reasoning. We identified three features associated with this theme: (C1) correctness (focus on the correct answer), (C2) conceptual (discussion of physics or mathematics principles), and (C3) schema (integration of content knowledge within the argument structure). Overall, 51.8% of responses evidenced at least one of these features. The following quote illustrates all three features associated with the content theme:

> Feedback C was the most helpful because it provided mathematical explanations with clear equations illustrating any mistakes in my argument. Additionally, it explicitly highlighted the correct answer and allowed me to compare it to my answer, easily allowing me to understand any flaws. Lastly, it provided a list of tips to improve my argument, which was far more useful than the other two arguments, which only provided at most 1–2 areas for improvement.

TABLE IV. Exemplar student responses highlighting each feedback feature. Underlined text in the student responses reflect the corresponding feature.

Code	Example
E1: Affirmation	*I found the third feedback to be the most useful. […] It allowed me to see what I did well and what I should continue to do.*
E2: Critique	*Feedback C, as it actually pointed out that I was wrong, others felt like they were half-hearted explanations and were just glazing my answer but feedback C actually pointed out I was wrong which made me feel more eager to learn my mistake even though I felt that all 3 of the answers I put in, I felt unclear about all but only feedback C made me realise the error and gave me a better answer from which I could learn off.*
E3: Guidance	*Feedback C was the most useful because it gave me ways to improve in the future instead of just correcting what I did wrong like A and B.* […]
C1: Correctness	*A is the most useful because it provides the right answer. The other 2 say that we got a negative slope, which is wrong.*
C2: Conceptual	*I liked feedback B the most and found it most useful because it helped me understand the overall shape and behavior of the gravitational potential energy vs distance graph. By mentioning that the curve is hyperbolic and that the slope decreases due to the inverse square law. It gave me a clear visual and conceptual understanding of how gravity weakens with distance. This made it easier for me to interpret the graph and connect it to real-world gravitational behavior, rather than just thinking in terms of equations. It deepened my conceptual grasp, not just my ability to calculate.*
C3: Schema	*C was the most helpful because it specifically highlighted which parts of my arguments were strong and which weren't.*
P1: Clarity	*I found B the most useful. It provided helpful feedback and had a lot of explanation. It was clear.* […]
P2: Structure	*I liked B the best because to me it seemed to make the most sense it organized its response in a way that made sense to me.* […]
P3: Conciseness	*I found C the most useful because it felt that it was the most concise.* […]
P4: Intelligibility	*I liked A the most as it was much more easier to understand compared to the others.*
D1: Detail	*I found the third and last feedback to be the most effective and useful since it went more in depth about all of the information that I needed.*
D2: Specificity	*I found C the most useful because it did not really run in circles, the rest kind of reiterated the same point over and over again, but c was straight to the point.*

The reference to "explicitly highlighting the correct answer" evidences the correctness feature. In addition, the inclusion of mathematical explanations with clear equations reflects the conceptual feature, while the use of an enumerated list with recommendations to strengthen the argument illustrates the schema feature.

3. Presentation (P)

The third theme encompasses features related to the presentation of feedback that are independent of the task's content. These include the clarity of the provided information, the external structure or formatting (e.g., use of an enumerated list), the feedback's length or brevity, and its overall ease of comprehension. Four features capture this theme: (P1) clarity, (P2) structure, (P3) conciseness, and (P4) intelligibility. Approximately 45% of student justifications referenced at least one of these features. The following quote illustrates all four features associated with this theme:

> I found Feedback C to be the most useful and my favorite among the three because it offered a clear, structured, and detailed evaluation of my argument. What stood out was how it explicitly separated the correct answer, my response, and any gaps between the two. This side-by-side comparison made it easy to understand not only what I did well but also what specific areas could be improved. Additionally, the three bulleted suggestions at the end were especially helpful because they were clear, actionable, and directly applicable to future problems. The word limit also made the feedback feel focused and concise without losing depth. […]

The response's repeated references to clarity of evaluation of the argument (first sentence) and actionable suggestions (fourth sentence) evidence the "clarity" feature. Along the same lines, the response's multiple references to structured evaluation of the argument, i.e., "side-by-side" comparison of what was expected and what was written along with bulleted suggestions reflect the "structure" feature of the feedback. The last sentence on the feedback being short yet specific, evidences the "conciseness" feature. Finally, the feedback's ease of comprehension, and the strengths and weaknesses of the argument, mark the final "Intelligibility" feature.

4. Depth (D)

The last theme corresponds to the extent of detailed discussion of the student's solution within or beyond the

purview of the task. This theme, however, does not correspond to the "quantity" or "wordiness" of the feedback but rather the quality of detailed discussion of the students' solution. Two features are associated with this theme: (D1) detail (detailed discussion of the solution) and (D2) specificity (feedback's primary focus on the solution with reference to the task with no generic discussions). Overall, 44.2% of students' descriptions evidenced at least one of these two features. An example highlighting the four features is provided below.

> I think that [feedback] B gave more direction in the best way to explain the prompt in a detailed manner. It gave specific answers as to what was missing from my response and how to improve it, where as the others were more generalized.

In the above quote, the first sentence referring to the feedback feature to explain the task in a detailed fashion evidences the "Detail" feature. On the other hand, the next sentence referring to feedback's specificity about the shortcomings and ways to improve the solution mark the second feature, i.e., "Specificity."

3. Themes and features across the feedback types

Figure 5 presents the trends among the 12 feedback features across the three feedback types as well as the entire data. Figure 6 presents trends across the four themes. Across 1160 responses, students most frequently valued features associated with the evaluation theme (69.4%), followed by content (55.2%), presentation (47.8%), and depth (44.2%). Within evaluation, critique and guidance were the most valued features. Within content, conceptual and schema were rated most highly. For presentation, conciseness was emphasized most, while for depth, the detail feature was most frequently valued. Notably, no single feature was consistently valued across all three feedback types.

Students who ranked feedback from self-crafted prompts (feedback A) most frequently emphasized conceptual arguments (the conceptual feature) in their justifications. Specifically, 45.4% of responses in this group evidenced the conceptual feature, compared to 33.9% for feedback B and 26.5% for feedback C. This emphasis on conceptual reasoning aligns with the broader pattern observed in Fig. 6, where students who ranked feedback A as most useful also valued the content theme more than any other group (60.8%). A similar, though less pronounced trend was observed for the clarity feature, with relatively low percentages across all groups (16.5% for A, 7.6% for B, and 17.5% for C).

Students who preferred feedback generated from prompts incorporating features of prompt engineering (feedback B) predominantly valued the "content," "depth," and to a slightly lesser extent the "evaluation" themes

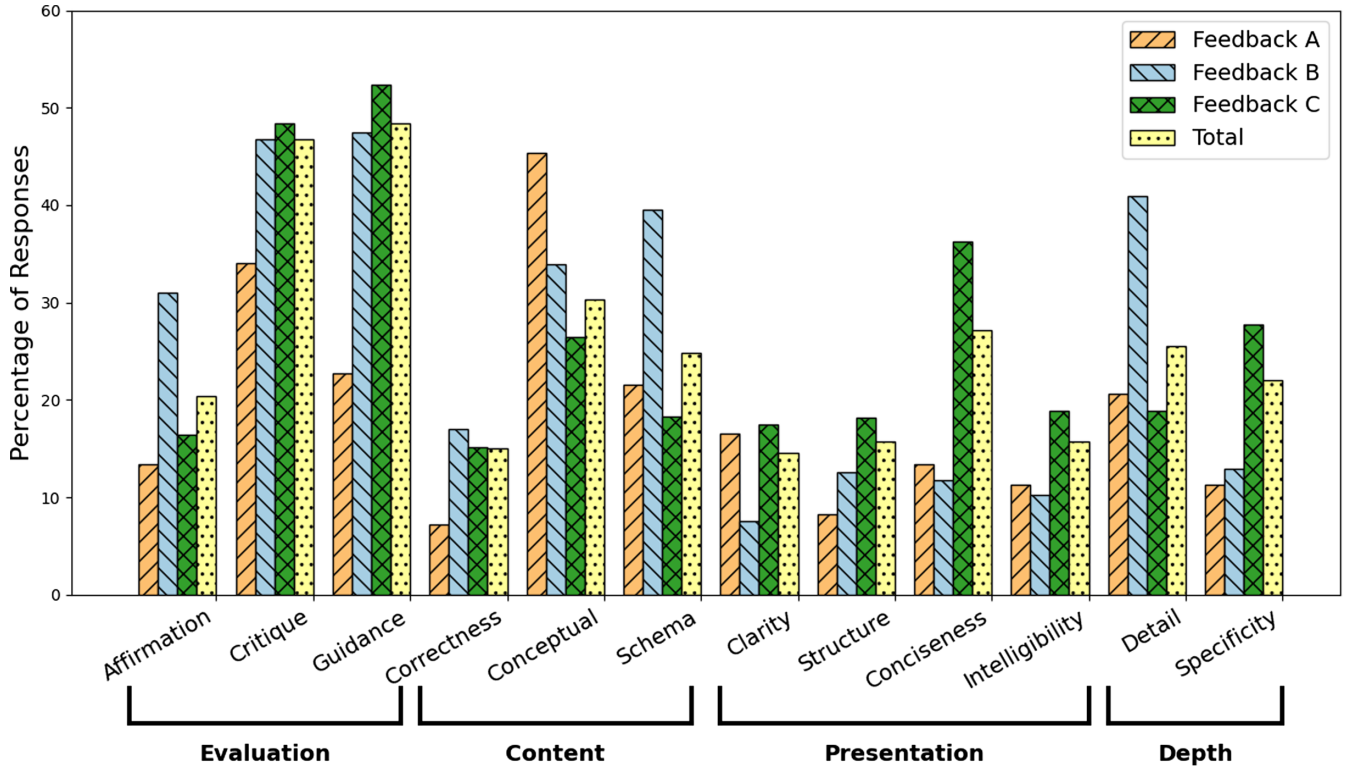


FIG. 5. Graph highlighting the percentage of responses of the total data and that across the three feedback types in terms of the feedback features. Associated themes across all features have been highlighted at the bottom end with bold text.

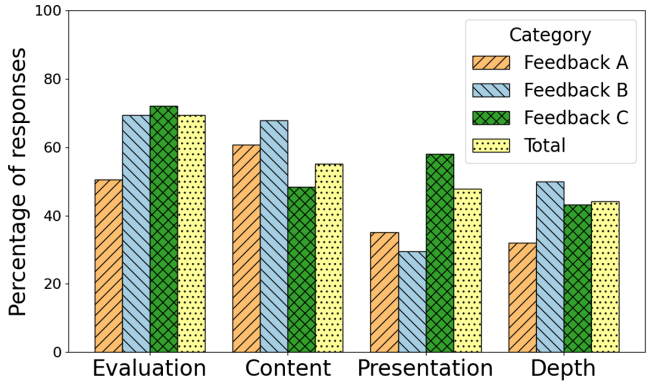


FIG. 6. Graph highlighting the percentage of responses of the total data and that across the three feedback types in terms of the feedback themes.

compared to the other groups. In particular, students who chose feedback B highlighted features associated with affirmation (31%), correctness (17%), schema (39.5%), and detail (40.9%) more than the other groups. They were also a close second in emphasizing critique (46.8%) and guidance (47.4%).

Students who preferred feedback generated from prompts combining foundational prompt-engineering techniques with principles of effective feedback (feedback C) most often emphasized the "evaluation" and "presentation" themes. This is consistent with the design of the associated prompt, which explicitly asked for highlighting of strengths, shortcomings, and addressing gaps (evaluation) as well as specifying a word limit and using a bulleted list (presentation). These students also valued the "depth" theme, though they were a close second to those who ranked feedback B as their top choice. In contrast, they placed the least emphasis on the content theme compared to the other two groups.

Within this group, students most frequently highlighted critique (48.4%), guidance (52.4%), conciseness (36.2%), and specificity (27.7%). Smaller proportions also noted clarity (17.5%), structure (18.2%), and intelligibility (18.9%). Overall, these features were valued more highly by students in this group than by those in the other two groups.

VI. DISCUSSION

In the previous section, we presented results on the sophistication of students' prompts when seeking feedback on their solutions (RQ1). We also discussed students' preferences, feedback features, and associated themes in their preferred feedback choices among three options (RQ2). The three feedback versions were generated using (i) self-crafted prompts (feedback A), (ii) provided prompts incorporating foundational prompt-engineering techniques (feedback B), and (iii) provided prompts combining foundational techniques with principles of effective feedback (feedback C).

The results highlight that student-generated prompts rarely incorporate prompting strategies in combination, as outlined in the prompt-engineering literature. While techniques such as the use of delimiters may be less crucial for the current AI models, our findings underscore the potential of well-structured prompts for maximizing AI's potential in students' learning. Regarding the second research question, students overwhelmingly preferred feedback generated from structured prompts, particularly those embedding principles of effective feedback. Conversely, they showed the least preference for feedback generated from self-crafted prompts. We also identified four themes capturing the features students attribute for the perceived usefulness of AI feedback: evaluation, content, presentation, and depth. The specific features include: (i) affirmation, (ii) critique, (iii) guidance, (iv) correctness, (v) conceptual, (vi) schema, (vii) clarity, (viii) structure, (ix) conciseness, (x) intelligibility, (xi) detail, and (xii) specificity.

Furthermore, those who preferred feedback generated from self-crafted prompts preferred conceptual-related information in feedback the most. Students who preferred feedback generated from prompt entailing prompt-engineering techniques valued detailed discussion of conceptual-related information along with critical evaluation of their solutions. Students who preferred feedback generated from a combination of prompt-engineering techniques and principles of effective feedback valued the features of critical examination of their responses, actionable suggestions along with their communication in a structured way. There was no single feature that was consistently valued by students across the different preferred feedback types.

These findings both complement and extend prior research on students' use of AI in STEM contexts. While earlier studies have examined prompting strategies in domains such as computer science [70] and chemistry problem solving by preservice teachers [71], our study extends this work into the domain of physics argumentation tasks. Sawalah *et al.* [70] found that undergraduates primarily relied on three prompting strategies while seeking solutions in their computer engineering program: directly copying and pasting questions, presenting a partially or completely phrased one-time question, and posing multiple questions. They also observed that 41% of participants used role-based prompting, particularly when posing multiple questions. Somewhat along with same lines, Tassoti [71] reported a predominant reliance of their participants on the "copy-and-paste" approach when interacting with AI. Zhang *et al.* [49] observed programming students preferring feedback derived from prompts that included their code, with feedback features such as specificity, clarity, and corrective nature guiding their preferences.

Observations made in the current study align as well as differ from the above observations in several ways. First, our task's context required students to make arguments

about the physical significance of a graph and then seek feedback from AI. It naturally made the copy-and-paste approach an inefficient way of prompting. As a result, our data did not reflect this practice to the extent observed in the above studies, suggesting a possible relationship between task objectives and prompting approaches. Second, unlike the results from Sawalah *et al.*, only 6 out of 1235 students in our data employed role-based prompting. However, consistent with Zhang *et al.*, our participants preferred feedback from structured prompts that specified their arguments, attributing its usefulness to features such as specificity, clarity, critique, and guidance.

Valeri *et al.* [72] similarly examined upper-secondary students' prompting approaches to ChatGPT. They reported that 61% of participants employed strategies such as specifying output structure (e.g., bulleted lists), requesting detail or concise responses though explicit descriptors (e.g., "brief," "longer," or "medium"), and providing specific contexts to optimize results. Our observations show partial overlap with these findings. While two-thirds of our participants specified context, we did not observe them requesting specific output format of AI responses or requesting detailed/concise responses. Interestingly, these aspects emerged in the students' valued features in AI feedback.

On feedback features, our findings also resonate with contemporary research. Henderson *et al.* [73] qualitatively analyzed university students' AI use for soliciting feedback and identified valued features such as depth (nuanced vs superficial), understandability, and contextualization (sensitivity to assignment requirements). A subset of our features such as specificity, intelligibility, and schema align closely with these observations. Perhaps, the study from Prompiengchai *et al.* [74] stands as the closest counterpart to our study in terms of integrating the principles of effective feedback into AI-generated feedback. The authors discuss the potential of leveraging prompting techniques such as zero-shot and one-shot approaches in integrating the principles of effective feedback for students' formative assessment responses. Our study complements their discussion by not only blending prompting techniques with feedback principles but by also capturing students' perspectives about its efficacy on feedback quality.

VII. CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

There is a growing body of research on leveraging AI platforms to facilitate students' customized learning in physics through personalized feedback. Aligned with these efforts, we unpack the spectrum of features students value in AI-generated feedback and their associated themes. We explore students' preferences across three types of feedback generated from distinct prompt types: self-crafted prompts, provided prompts incorporating foundational prompt-engineering features, and provided prompts combining prompt-engineering principles with those of effective feedback. Finally, we also analyze students' AI interactions by examining their prompts through the lens of foundational prompting techniques.

Each result presented in this study offers a unique contribution to contemporary research on the use of AI systems for physics teaching and learning. As noted in Sec. II E, while numerous studies have examined the efficacy of AI models in solving physics assessments and grading student responses, research on AI-generated feedback remains comparatively sparse. Among the studies that focus on feedback, most employ empirical approaches that provide AI-driven feedback with little to no grounding in the theoretical principles of effective feedback from the education research literature. Noting this trend, there have been calls for blending theory into practice in the context of design and delivery of feedback in education settings. In addition to incorporating the principles, the current study also highlights students' clear preferences for feedback generated from prompts embedding these principles.

While prior research has considered students' perceptions of linguistic quality or scientific accuracy in AI outputs [75], few studies have examined the features students associate with the perceived usefulness of feedback in the context of physics problem solving. We contribute to this literature by identifying 12 distinct features that students value in AI-generated feedback and grouping them into broad themes. Finally, our study also analyzes students' prompts in terms of their sophistication through the lens of foundational prompting techniques.

Results reported in this study carry several implications for researchers. First, valued feedback features attributed by students varied depending on their preferred feedback type, suggesting a potential complex structure guiding their choices. These results suggest the need for investigations focused on exploring the influence of factors such as task's context, perceived task's difficulty, familiarity with AI models, and attitudes toward AI use in learning environments on students' choices of valued feedback features. We also observed students' prompting strategies often fall short of expectations from the prompt-engineering literature. Future research could examine how students' diligence in crafting prompts depends on factors such as task type (well structured vs ill structured) or participation format (individual vs collaborative).

Reasoning in physics contexts often involves engaging in complex *processes* such as making sense of multiple, equivalent representations (equations, graphs, or laws) and establishing coherence between them. This, in turn, leads to *products* or solutions. End-of-the-activity evaluations or static feedback (similar to the one adopted in this study) thus may end up focusing more on the *products* as compared to providing assistance to students' reasoning *processes*. A key challenge in providing feedback in physics contexts can thus involve providing real-time

inputs as students engage with the task thereby facilitating valued processes in physics education. Thus, exploring students' valued features in a dynamic feedback focused on students reasoning approaches and comparing with valued features identified in the current study also presents potential avenue for future research.

For instructors, our observation that students preferred feedback generated from structured prompts underscores the importance of blending AI literacy (such as prompt engineering) with content knowledge to better support their physics learning. Our results also highlight the need to explicitly embed principles of effective feedback in design and delivery of AI-mediated feedback.

This study has several limitations. First, by design, students were aware of the prompts used to generate the three types of feedback. This may have introduced a bias in their preferences with notions such as *more sophisticated prompts yield more sophisticated feedback*. Along the same lines, all students received the three types of feedback in the same order. Thus, reading the third feedback upon going through other two versions may have induced bias toward selecting it. Second, as noted in Sec. III A, the prompt-engineering techniques through which we analyzed students' self-crafted prompts may have inherent limitations given the evolution of AI models. We acknowledge that sophisticated prompts can be devoid of few of these techniques, and our results to the first research question accompany this inherent limitation. Third, the study involved students' interactions with AI on a single physics task. Future research using a wider range of tasks, varying in type (multiple-choice, open-ended, conceptual, numerical, etc.) and quantity, could yield deeper insights into students' prompting approaches and feedback preferences. Finally, students were restricted to crafting a single prompt for seeking feedback. In practice, AI interactions often involve iterative exchanges to refine responses, which was not captured in this study.

Future work will explore addressing the above limitations along with unpacking the co-occurrence of feedback features across preferred feedback types. Examining student preferences in relation to the correctness of their solutions also remains an avenue for further explorations. Finally, we aim to investigate students' perceptions of the usefulness of integrating AI-related activities into their assignments for their learning experiences.

ACKNOWLEDGMENTS

ChatGPT-4o was used to correct grammar. This study was supported by Purdue University's Innovation Hub and U.S. National Science Foundation with Grant No. 2111138.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available because they contain sensitive personal information. The data are available from the authors upon reasonable request.

[1] J. Hattie and H. Timperley, The power of feedback, Rev. Educ. Res. **77**, 81 (2007).
[2] D. Burke and J. Pieterick, *Giving Students Effective Written Feedback* (McGraw-Hill Education, London, 2010).
[3] R. Azevedo and R. M. Bernard, A meta-analysis of the effects of feedback in computer-based instruction, J. Educ. Comput. Res. **13**, 111 (1995).
[4] D. Giamos, O. Doucet, and P.-M. Léger, Continuous performance feedback: Investigating the effects of feedback content and feedback sources on performance, motivation to improve performance and task engagement, J. Organ. Behav. Manage. **44**, 194 (2024).
[5] D. J. Nicol and D. Macfarlane-Dick, Formative assessment and self-regulated learning: A model and seven principles of good feedback practice, Stud. Higher Educ. **31**, 199 (2006).
[6] J. M. Reddig, A. Arora, and C. J. MacLellan, Generating in-context, personalized feedback for intelligent tutors with large language models, Int. J. Artif. Intell. Educ. **35**, 1 (2025).
[7] T. Fan, J. Song, and Z. Guan, Integrating diagnostic assessment into curriculum: A theoretical framework and teaching practices, Lang. Test. Asia **11**, 2 (2021).
[8] A. Sirnoorkar, A. P. Jambuge, K. D. Rainey, A. Adamson, B. R. Wilcox, and J. T. Laverty, Theoretical approach for providing feedback for instructors through a standardized assessment for undergraduate physics, in *Proceedings of the 17th International Conference of the Learning Sciences-ICLS 2023* (International Society of the Learning Sciences, Montreal, 2023), pp. 1326–1329.
[9] J. T. Laverty, A. Sirnoorkar, A. P. Jambuge, K. D. Rainey, J. Weaver, A. Adamson, and B. R. Wilcox, A new paradigm for research-based assessment development, *presented at PER Conf. 2022, Grand Rapids, MI*, 10.1119/perc.2022.pr.Laverty.
[10] G. Polverini and B. Gregorcic, How understanding large language models can inform the use of ChatGPT in physics education, Eur. J. Phys. **45**, 025701 (2024).
[11] G. Polverini and B. Gregorcic, Evaluating vision-capable chatbots in interpreting kinematics graphs: A comparative study of free and subscription-based models, Front. Educ. **9**, 1452414 (2024).
[12] A. Sirnoorkar, D. Zollman, J. T. Laverty, A. J. Magana, N. S. Rebello, and L. A. Bryan, Student and AI responses to physics problems examined through the lenses of

sensemaking and mechanistic reasoning, Comput. Educ. **7**, 100318 (2024).
[13] A. Bralin, A. Sirnoorkar, Y. Zhang, and N. S. Rebello, Mapping the literature landscape of artificial intelligence and machine learning in physics education research, *presented at PER Conf. 2024, Boston, MA*, 10.1119/perc.2024.pr.Bralin.
[14] G. Kortemeyer, J. Nöhl, and D. Onishchuk, Grading assistance for a handwritten thermodynamics exam using artificial intelligence: An exploratory study, Phys. Rev. Phys. Educ. Res. **20**, 020144 (2024).
[15] G. Kortemeyer, Performance of the pre-trained large language model GPT-4 on automated short answer grading, Discover Artif. Intell. **4**, 47 (2024).
[16] G. Kortemeyer and J. Nöhl, Assessing confidence in AI-assisted grading of physics exams through psychometrics: An exploratory study, Phys. Rev. Phys. Educ. Res. **21**, 010136 (2025).
[17] G.-G. Lee, E. Latif, X. Wu, N. Liu, and X. Zhai, Applying large language models and chain-of-thought for automatic scoring, Comput. Educ. **6**, 100213 (2024).
[18] Y. Wei, R. Zhang, J. Zhang, D. Qi, and W. Cui, Research on intelligent grading of physics problems based on large language models, Educ. Sci. **15**, 116 (2025).
[19] S. El-Adawy, A. MacDonagh, and M. Abdelhafez, Exploring large language models as formative feedback tools in physics, *presented in PER Conf. 2024, Boston, MA*, 10.1119/perc.2024.pr.El-Adawy.
[20] A. Meyer, T. Bleckmann, and G. Friege, Automatic feedback on physics tasks using open-source generative artificial intelligence, Int. J. Sci. Educ. 1 (2025).
[21] T. Wan and Z. Chen, Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning, Phys. Rev. Phys. Educ. Res. **20**, 010152 (2024).
[22] Z. Chen and T. Wan, Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy, Phys. Rev. Phys. Educ. Res. **21**, 010126 (2025).
[23] E. Mills, A. Mizouri, and A. Peach, Prompting better feedback: A study of custom GPT for formative assessment in undergraduate physics, Educ. Sci. **15**, 1058 (2025).
[24] W. Allen, A. Shanker, and N. S. Rebello, Students' perceptions to a large language model's generated feedback and scores of argumentation essays, arXiv:2508.14759.
[25] A. Almegren, H. S. Mahdi, A. N. Hazaea, J. K. Ali, and R. M. Almegren, Evaluating the quality of AI feedback: A comparative study of AI and human essay grading, Innov. Educ. Teach. Int. **62**, 1858 (2025).
[26] A. Kinder, F. J. Briese, M. Jacobs, N. Dern, N. Glodny, S. Jacobs, and S. Leßmann, Effects of adaptive feedback generated by a large language model: A case study in teacher education, Comput. Educ. **8**, 100349 (2025).
[27] C. Song and Y. Song, Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students, Front. Psychol. **14**, 1260843 (2023).
[28] K. Guo, M. Pan, Y. Li, and C. Lai, Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability, Internet Higher Educ. **63**, 100962 (2024).
[29] E. Fokides and E. Peristeraki, Comparing ChatGPT's correction and feedback comments with that of educators in the context of primary students' short essays written in English and Greek, Educ. Inf. Technol. **30**, 2577 (2025).
[30] C.-W. Kao and B. L. Reynolds, Timed second language writing performance: Effects of perceived teacher vs perceived automated feedback, Hum. Soc. Sci. Commun. **11**, 1 (2024).
[31] T. Jansen, L. Höft, L. Bahr, J. Fleckenstein, J. Möller, O. Köller, and J. Meyer, Empirische Arbeit: Comparing generative AI and expert feedback to students' writing: Insights from student teachers, Psychol. Erzieh. Unterr. **71**, 80 (2024).
[32] J. Zare, F. Ranjbaran Madiseh, and A. Derakhshan, Generative AI and English essay writing: Exploring the role of ChatGPT in enhancing learners' task engagement, Appl. Linguist. amaf045 (2025).
[33] S. T. S. Chan, N. P. K. Lo, and A. M. H. Wong, Enhancing university level English proficiency with generative AI: Empirical insights into automated feedback and learning outcomes, Contemp. Educ. Technol. **16**, ep541 (2024).
[34] A. H. Alsofyani and A. M. Barzanji, The effects of ChatGPT-generated feedback on Saudi EFL learners' writing skills and perception at the tertiary level: A mixed-methods study, J. Educ. Comput. Res. **63**, 431 (2025).
[35] J. Escalante, A. Pack, and A. Barrett, AI-generated feedback on writing: Insights into efficacy and ENL student preference, Int. J. Educ. Technol. Higher Educ. **20**, 57 (2023).
[36] S. Rüdian, J. Podelo, J. Kužílek, and N. Pinkwart, Feedback on feedback: Student's perceptions for feedback from teachers and few-shot LLMs, in *Proceedings of the 15th International Learning Analytics and Knowledge Conference, Dublin, Ireland* (Association for Computing Machinery, New York, NY, 2025), pp. 82–92.
[37] A. Aljasser, Investigating EFL students' perceptions of feedback: A comparative study of instructor and ChatGPT-generated responses in academic writing, Educ. Inf. Technol. **30**, 22967 (2025).
[38] J. Steiss, T. Tate, S. Graham, J. Cruz, M. Hebert, J. Wang, Y. Moon, W. Tseng, M. Warschauer, and C. B. Olson, Comparing the quality of human and ChatGPT feedback of students' writing, Learn. Instr. **91**, 101894 (2024).
[39] K. ElEbyary and R. Shabara, ChatGPT-generated corrective feedback: Does it do what it says on the tin?, Teach. Engl. Technol. **24**, 68 (2024).
[40] M. Abdi Tabari, A. Kushki, and Y. Wang, Comparing the effects of teacher-and AI-mediated corrective feedback on accuracy, complexity, and quality in L2 written narratives, Comput. Assisted Lang. Learn. 1 (2025).
[41] S. K. Banihashem, N. T. Kerman, O. Noroozi, J. Moon, and H. Drachsler, Feedback sources in essay writing: Peer-generated or AI-generated feedback?, Int. J. Educ. Technol. Higher Educ. **21**, 23 (2024).
[42] S. K. Banihashem, O. Noroozi, H. Khosravi, C. D. Schunn, and H. Drachsler, Pedagogical framework for hybrid intelligent feedback, Innov. Educ. Teach. Int. **63**, 1 (2025).

[43] Z. Zhang, S. Aubrey, X. Huang, and T. K. Chiu, The role of generative AI and hybrid feedback in improving L2 writing skills: A comparative study, Innov. Lang. Learn. Teach. 1 (2025).

[44] J. Han and M. Li, Exploring ChatGPT-supported teacher feedback in the EFL context, System **126**, 103502 (2024).

[45] M. Navío-Inglés, J. Guzmán Mora, P. O'Connor-Jiménez, and A. García González, What's next for feedback in writing instruction? Pre-service teachers' perceptions of assessment practices and the role of generative AI, Educ. Sci. **15**, 1534 (2025).

[46] F. Ghoochani, J. Scharfenberger, B. Funk, R. Doublan, M. Jakharabhai Odedra, and B. Etsiwah, From feedback to formative guidance: Leveraging LLMs for personalized support in programming projects, in *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, New York City* (Association for Computing Machinery, New York, NY, 2025), pp. 398–403.

[47] Ş. Çağlar-Özhan, P. Tekeli, and S. Arkün-Kocadere, Comparison of AI-generated and instructor feedback: No significant difference in perceived feedback quality and neither on performance, J. Comput. Assist. Learn. **41**, e70134 (2025).

[48] E. Bauer, C. Richters, A. J. Pickal, M. Klippert, M. Sailer, and M. Stadler, Effects of AI-generated adaptive feedback on statistical skills and interest in statistics: A field experiment in higher education, Br. J. Educ. Technol. **56**, 1735 (2025).

[49] Z. Zhang, Z. Dong, Y. Shi, T. Price, N. Matsuda, and D. Xu, Students' perceptions and preferences of generative artificial intelligence feedback for programming, in *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence* (AAAI Press, Vancouver, 2024), pp. 23250–23258.

[50] N. Ippisch, M. Herklotz, A.-C. Haensch, and C. Schwemmer, Beyond correctness: Evaluating and improving LLM feedback in statistical education, arXiv:2511.07628.

[51] S. Yeung, University students' engagement with generative AI-supported automated writing evaluation (AWE) feedback, J. Second Lang. Writing **68**, 101203 (2025).

[52] R. AlGhamdi, Exploring the impact of ChatGPT-generated feedback on technical writing skills of computing students: A blinded study, Educ. Inf. Technol. **29**, 18901 (2024).

[53] Y. Zhou, M. Pankiewicz, L. Paquette, and R. S. Baker, Impact of LLM feedback on learner persistence in programming, in *Proceedings of the 33rd International Conference on Computers in Education, ICCE 2025, Chennai, India* (Asia-Pacific Society for Computers in Education, Chennai, 2025).

[54] L. V. Navarro, M. M. G. Pérez, P. Barajas-Á lvarez, E. Ortiz-Zavala, E. R. Villegas, and J. M. Sánchez-Lizárraga, Qualitative feedback comparison between professors and AI in STEM education, in *Proceedings of the 2025 IEEE Global Engineering Education Conference (EDUCON), London, United Kingdom* (IEEE, New York, 2025), pp. 1–6.

[55] T. Maillo, C. A. Alexandru, and A. Constantin, Enhancing marked with peer feedback functionality with LLM support, in *Proceedings of the 2025 Conference on UK and Ireland Computing Education Research* (2025), pp. 1–7.

[56] M. Steinbach, S. Bhandari, J. Meyer, and Z. A. Pardos, When LLMs hallucinate: Examining the effects of erroneous feedback in math tutoring systems, in *Proceedings of the Twelfth ACM Conference on Learning Scale, Palermo Italy* (Association for Computing Machinery, New York, NY, 2025), pp. 139–150.

[57] X. Dai, Z. Wen, J. Jiang, H. Liu, and Y. Zhang, How students use AI feedback matters: Experimental evidence on physics achievement and autonomy, arXiv:2505.08672.

[58] Q. Jia, J. Cui, R. Xi, C. Liu, P. Rashid, R. Li, and E. Gehringer, On assessing the faithfulness of LLM-generated feedback on student assignments, in *Proceedings of the 17th International Conference on Educational Data Mining* (International Educational Data Mining Society, Atlanta, 2024), pp. 491–499.

[59] A. Becerra and R. Cobos, Enhancing the professional development of engineering students through an AI-based collaborative feedback system, in *Proceedings of the 2025 IEEE Global Engineering Education Conference (EDUCON), Global Engineering Education Conference (EDUCON)* (IEEE, New York, 2025), pp. 1–9.

[60] L. W. Liebenow, F. T. Schmidt, J. Meyer, and J. Fleckenstein, Self-assessment accuracy in the age of artificial intelligence: Differential effects of LLM-generated feedback, Comput. Educ. **237**, 105385 (2025).

[61] J. Weidlich, F. Gotsch, K. Schudel, C. Marusic-Würscher, J. Mazzarella, H. Bolten, D. Bütler, S. Luger, B. Wohlfehnder, and K. M. Merki, Teacher, peer, or AI? Comparing effects of feedback sources in higher education, Comput. Educ. Open **9**, 100300 (2025).

[62] K. Seßler, A. Bewersdorff, C. Nerdel, and E. Kasneci, Towards adaptive feedback with AI: Comparing the feedback quality of LLMs and teachers on experimentation protocols, arXiv:2502.12842.

[63] L. J. Jacobsen and K. E. Weber, The promises and pitfalls of LLMs as feedback providers: A study of prompt engineering and the quality of AI-driven feedback, Cent. Open Sci. **10**, 433 (2023).

[64] B. Chen, Z. Zhang, N. Langrené, and S. Zhu, Unleashing the potential of prompt engineering for large language models, Patterns **6**, 101260 (2025).

[65] https://www.qualtrics.com/.

[66] R. W. Chabay and B. A. Sherwood, *Matter and Interactions* (John Wiley & Sons, New York, 2015).

[67] S. E. Toulmin, *The Uses of Argument* (Cambridge University Press, Cambridge, England, 2003).

[68] K. L. McNeill and J. Krajcik, Inquiry and scientific explanations: Helping students use evidence and reasoning, *Science as Inquiry in the Secondary Setting*, edited by J. Luft, R. L. Bell, and J. Gess-Newsome (2008), https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=McNeill+and+Krajick%2C+Inquiry+and+scientific+explanations%3A+Helping+students+use+evidence&btnG=.

[69] J. W. Creswell, W. E. Hanson, V. L. Clark Plano, and A. Morales, Qualitative research designs: Selection and implementation, Counselling psychology quarterly **35**, 236 (2007).

[70] G. Sawalha, I. Taj, and A. Shoufan, Analyzing student prompts and their effect on ChatGPT's performance, Cogent Educ. **11**, 2397200 (2024).

[71] S. Tassoti, Assessment of students use of generative artificial intelligence: Prompting strategies and prompt engineering in chemistry education, J. Chem. Educ. **101**, 2475 (2024).

[72] F. Valeri, P. Nilsson, and A.-M. Cederqvist, Exploring students' experience of ChatGPT in STEM education, Comput. Educ. **8**, 100360 (2025).

[73] M. Henderson, M. Bearman, J. Chung, T. Fawns, S. Buckingham Shum, K. E. Matthews, and J. de Mello Heredia, Comparing generative AI and teacher feedback: Student perceptions of usefulness and trustworthiness, Assess. Eval. Higher Educ. 1 (2025).

[74] S. Prompiengchai, C. Narreddy, and S. Joordens, A practical guide for supporting formative assessment and feedback using generative AI, arXiv:2505.23405.

[75] 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).