Generative AI for feedback and collaborative knowledge construction in preservice physics teacher education

Yuchen Jiang,[1] Xiu-Mei Feng,[1,*] Yuting Liu,[1] Yiting Wang,[1] Li Xie,[1,†] and Lei Bao[2,‡]

[1]*College of Physical Science and Technology, Central China Normal University, Wuhan, Hubei 430079, China*

[2]*School of Biological Science and Medical Engineering, Research Center for Learning Science, Key Laboratory of Child Development and Learning Science (MOE), Southeast University, Nanjing 211189, China*

(Received 4 November 2025; accepted 26 January 2026; published 6 February 2026)

Generative artificial intelligence (GenAI) offers strong potential to transform teaching practices; however, empirical research on its classroom effectiveness remains limited. This study explores how GenAI can be integrated into preservice physics teacher training by replacing two traditional instructor roles in an instructional design course: providing evaluative feedback and facilitating collaboration. A GenAI-based assistant was developed to assess instructional designs according to predefined criteria and to provide personalized feedback. A total of 110 preservice physics teachers received feedback from one of the three sources: a teaching expert, a teaching assistant, or GenAI to revise their instructional designs. They subsequently collaborated with GenAI for further refinement. Results indicate that GenAI evaluations closely aligned with those of human experts and teaching assistants. In blinded evaluations, preservice physics teachers rated GenAI feedback as comparable to human feedback. Furthermore, collaboration with GenAI resulted in significant improvements in the overall quality of instructional design. This progress was particularly most evident in the structured, content-expandable components. While these findings demonstrate GenAI's feasibility from a functional perspective, challenges related to preservice teachers' perceptions also emerged. Although design quality improved through collaboration with GenAI, independent design skills, measured by a final paper-and-pencil test, showed no similar improvement. Preservice physics teachers acknowledged the need to filter and revise GenAI-generated content, but often directly adopted its suggestions due to the convenience of GenAI. Notably, feedback was rated lower once participants realized it was AI-generated, even if the actual source was not, revealing trust-related concerns. This study moves beyond technical evaluations of GenAI, offering insights into the pedagogical challenges and limitations that arise when GenAI assumes the roles that are traditionally performed by human educators.

DOI: 10.1103/hm13-jv98

I. INTRODUCTION

The emergence of generative artificial intelligence (GenAI) tools is reshaping the landscape of teaching and learning in science education [1,2]. GenAI can be defined as a technology that leverages deep learning models to generate or create new content [3]. The public availability of GenAI tools such as ChatGPT, a conversational AI system powered by a large language model, enables users to obtain relevant and meaningful content in response to natural-language prompts. This form of human-machine interaction has increased the accessibility of GenAI tools and broadened their reach and user base [4]. With its intuitive interface and adaptive content generation capabilities [5], GenAI has the potential to redefine the roles and responsibilities of physics teachers [6]. It can support instruction by generating adaptive teaching strategies [4,7], producing customized learning materials, and assisting in the development of subject-specific tasks, such as physics problems [8]. In addition, GenAI can grade unstructured responses and provide real-time feedback, thereby reducing teachers' workloads and improving instructional efficiency [9–11].

While teachers generally recognize the potential of GenAI in education, many of them report limited use due to inadequate preparation [12,13]. Preservice teacher education is therefore a key context for introducing GenAI

*Contact author: xiumeifeng@mail.ccnu.edu.cn
†Contact author: shirlyxieli@ccnu.edu.cn
‡Contact author: lei.bao@outlook.com

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.

in ways that are directly connected to teachers' instructional work and the professional development of future physics educators [13–16]. Preservice teachers often express strong intentions to integrate digital technologies into teaching, yet effective use typically benefits from structured support and authentic, low-stakes opportunities to try them [17]. In this respect, embedding GenAI in authentic instructional work within teacher education is especially valuable [7,18]. Such practice-based experiences can enable preservice teachers to collaborate with GenAI and to reflect critically on how AI-generated suggestions shape their instructional decisions.

However, existing research has primarily concentrated on the technical performance of GenAI tools and their perceived effects on educational outcomes, often overlooking the human element, including how teachers adapt to, collaborate with, or are replaced by these technologies, and how such changes influence their professional agency in classrooms [19]. Studies that do address teachers' perspectives tend to focus on general attitudes toward GenAI adoption [20] without examining the specific challenges and opportunities that arise when GenAI takes on instructional roles. Moreover, much of the literature adopts a one-size-fits-all view of technology integration, disregarding the different levels and forms of GenAI use across educational contexts and the implications these differences have for teacher roles [21]. In some subject domains such as science education, research has concentrated on GenAI's technical capabilities rather than the pedagogical consequences of delegating core teaching tasks to GenAI [21,22]. These gaps leave the field with insufficient understanding of how to support physics teachers, and how to alter our teacher education programs to better prepare future physics teachers for success in the GenAI-rich educational environments.

Instructional design is a central component of preservice physics teacher education, linking pedagogical knowledge with classroom practice [8]. However, preservice physics teachers often face challenges in this area due to their limited pedagogical knowledge and experience [23,24]. In higher education, the lack of expert guidance and resources often prevents preservice teachers from receiving timely, personalized feedback and the support they need. These characteristics make instructional design courses a particularly relevant context for investigating how GenAI can support preservice physics teachers by offering continuous feedback and promoting collaborative development of their instructional designs [15].

To address these gaps, the present study examines the practical implications of using GenAI to assume instructional roles that are traditionally carried out by instructors in higher education physics teacher preparation. Rather than treating GenAI merely as a supplementary tool, we investigate its potential to perform core instructional functions in the instructional design class, such as providing evaluative feedback and engaging in collaborative design, that are central to teacher-student interactions. We aim to capture preservice physics teachers' perceptions of this substitution, exploring both their acceptance of and resistance to GenAI in roles that directly shape learning processes. This approach allows us to move beyond technical evaluations of GenAI, focusing instead on the pedagogical, relational, and identity-related dimensions that emerge when GenAI mediates tasks traditionally reserved for human educators.

II. LITERATURE REVIEW

A. Instructional design and challenges for preservice physics teachers

Teachers need to engage in thorough instructional design before the course begins to ensure effective teaching. Instructional design serves as both a conceptual framework and a mental model that guides teachers in systematically approaching teaching tasks [25,26]. It translates principles of learning and instruction into concrete plans and activities, with learning theories providing the foundation for this translation [27]. Accordingly, instructional design frameworks reflect different views of learning and how instruction should be organized. For example, the Kemp model is often described as objective oriented [28], the Smith and Ragan model is linked to cognitive perspectives that foreground learner analysis [29], and the Dick and Carey model is typically viewed from a systems perspective [30].

Although instructional design models differ in their theoretical foundations and specific structures, they converge on three fundamental questions: what the goals of teaching and learning are, how instruction should be organized to facilitate learning, and how the effectiveness of teaching and the outcomes of learning should be evaluated [27,28]. To address these questions, a typical instructional design incorporates seven essential elements: (a) setting clear and measurable learning objectives; (b) providing a thorough analysis of the teaching content, structure, and logical flow; (c) evaluating learner characteristics such as prior knowledge and potential barriers; (d) identifying important or difficult content that may hinder learning; (e) selecting appropriate instructional methods responsive to student needs; (f) designing instructional activities and materials; and (g) implementing systematic evaluation on instructional effectiveness and learning achievement. Together, these elements form a coherent framework that connects principles of learning and instruction with practical decisions in classroom teaching and student learning.

However, preservice physics teachers with limited experience in pedagogy and classroom management often require additional support during teaching and greater clarity about the steps of instructional design [31]. For example, when analyzing learners' knowledge structures

and cognition ability, they typically lack knowledge of student misconceptions and how students conceptualize ideas. This gap makes it difficult for them to design teaching materials or apply appropriate teaching strategies within their instructional designs [24].

Given these challenges, systematic training in instructional design and expert guidance are essential. Instructional design improvement is often revised iteratively through the complementary and ongoing readjustment of reflection and feedback [23]. Feedback is essential to learning, and the most effective feedback is immediate rather than delayed [32]. With immediate, evaluation-based feedback, experienced experts can promptly pinpoint physics-teaching problems in preservice teachers' instructional designs and offer actionable guidance. Preservice physics teachers then use this feedback to revise their instructional designs. Such training can refine their ability to design effective, content-specific lessons.

However, due to the class size and limited time, experts cannot provide timely feedback and sustained support to every preservice teacher in teacher education. Innovative technologies and approaches could offer a solution to this issue.

B. GenAI for feedback and collaboration in instructional design

Current research has established the technical feasibility of applying GenAI to automated assessment and timely feedback generation. In grading tasks, GenAI has demonstrated its ability to effectively assess open-ended responses [33]. By using methods like Chain-of-Thought Prompting and Few-shot learning, GenAI can deliver consistent and reliable assessments that closely resemble human grading [9,10]. Furthermore, through accessible prompt-engineering techniques, GenAI can provide personalized and timely feedback articulated in coherent, humanlike language, even when trained on relatively few examples [11,34].

Nevertheless, practical research on the use of GenAI to generate evaluative feedback for instructional design in teacher education remains limited. Moreover, although GenAI assessments demonstrate accuracy in certain domains, preservice physics teachers may still prefer feedback from human experts and exhibit limited trust in AI-generated evaluations [35]. Such perceptions can diminish GenAI's effectiveness in supporting the development of advanced design skills, as they may constrain preservice teachers' willingness to use and act upon its feedback [36]. Further research, from the perspective of preservice teachers, is therefore needed to evaluate both the quality of GenAI feedback and its impact on instructional design within teacher education.

GenAI also demonstrates potential for supporting collaboration in instructional design. Its accessible interfaces allow novice teachers to refine their designs with greater creativity while receiving personalized guidance [14,37,38]. When provided with explicit requirements, it can generate tailored and well-structured instructional designs that improve student performance [39,40]. Despite these affordances, several challenges constrain its educational value. The effectiveness of collaboration depends heavily on how preservice teachers engage with GenAI [41].

Another concern is that GenAI outputs are prone to hallucinations and may therefore lack scientific accuracy, which is particularly concerning in physics education [24]. Training a model from scratch for a specific educational context is prohibitively complex and expensive [42]. Accordingly, recent studies have increasingly leveraged custom chatbots to mitigate hallucinations and adapt them to achieve the desired performance in a given application. For instance, Genie-on-Demand coupled GPT-3.5-turbo with a custom knowledge base to support student learning and teacher guidance [43]. Beyond bespoke development, several platforms now allow teachers and students to configure custom chatbots with relatively low effort. On OpenAI, these custom chatbots are designated as GPTs, and users can specify desired behaviors through a configuration prompt and supply subject- or situation-specific resources by uploading relevant files [44]. Recent work has configured such custom GPTs to support learning about proportional relationships in mathematical and physical contexts [44] and has developed dialogic-learning agents such as the Dialogic Science Teacher to support science learning [45]. However, custom chatbots specifically designed to support physics instructional design remain relatively underexplored.

These challenges also underscore the need for critical thinking about GenAI use [46,47], yet preservice physics teachers often lack sufficient subject expertise to identify such flaws. As a result, they remain vulnerable to uncritical acceptance of inaccurate content or even unintentional plagiarism [24]. This problem is compounded by prevailing evaluation practices, which typically assess skills through the quality of final products created with GenAI [24,48]. Such product-based assessments obscure whether preservice teachers engaged critically with GenAI outputs during collaboration and fail to capture the extent of durable skill development.

These limitations point to the need for assessment approaches that capture actual skill development. They also underscore the importance of examining preservice teachers' perspectives to understand both their acceptance and resistance to GenAI's collaborative roles in instructional design.

C. Research questions

Building on these considerations, this study examines the role of GenAI in enhancing preservice physics teachers' instructional design skills, with particular attention to feedback quality, skill development, and user perceptions. Rather than simply comparing human and AI contributions,

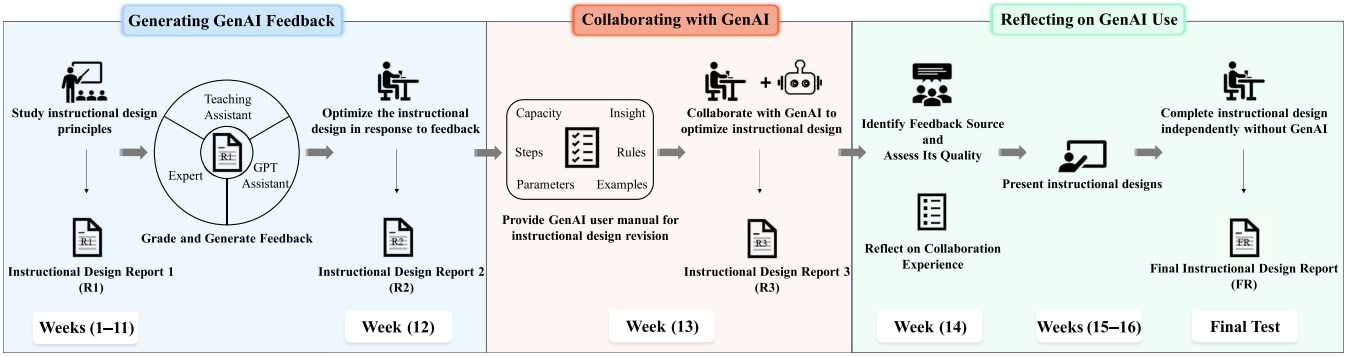


FIG. 1. Course workflow and experimental procedure.

the focus is on understanding whether GenAI can function as a credible and effective partner in instructional design, and how trust and reliance influence its educational value. Accordingly, this study addresses the following research questions:

RQ1: To what extent can GenAI replicate expert judgment when providing evaluative feedback to preservice physics teachers? This question is further explored through the following subsidiary questions:

RQ1a: How consistent are GenAI-generated scores compared to human evaluations?

RQ1b: Can preservice physics teachers distinguish GenAI-generated feedback from human feedback? How do they perceive and evaluate the feedback content?

RQ2: How do preservice physics teachers' instructional design skills change during collaboration with GenAI?

RQ3: How do preservice physics teachers perceive and evaluate the use of collaboration with GenAI in instructional design?

III. METHOD

A. Experimental procedure

The research project integrated GenAI into the secondary physics pedagogy course for preservice physics teachers. The course was primarily lecture-based, with the final two sessions devoted to student presentations of their instructional designs. It systematically addressed core topics including instructional design, secondary physics curriculum standards, student psychology, teaching methods, instructional procedures, assessment practices, and five representative examples of secondary physics teaching. The course aimed to equip students with a solid theoretical foundation in secondary physics teaching, develop their capacity to design and present instructional plans, and foster fundamental skills in teaching evaluation.

To address the research questions, GenAI was utilized in two key functions: providing evaluative feedback and facilitating collaboration. Upon completion of the experimental stages, preservice teachers were invited to provide reflections on their experiences with GenAI. Figure 1 illustrates the experimental procedures and the key steps of course implementation.

Initially, preservice physics teachers attended lectures on secondary physics teaching theories and instructional design. As the course progressed, they by themselves gradually shaped their initial instructional designs based on the course content, culminating in the submission of the first instructional design report (R1). The reports were randomly divided into three groups, and each group was graded by the GPT assistant, an expert professor, or a teaching assistant (TA). The corresponding grader then provided feedback based on predefined evaluation criteria. To prevent any bias in evaluating feedback quality, preservice teachers were not informed that the feedback came from three different types of evaluators. They were only given the feedback content and instructed to revise their designs accordingly, leading to the submission of the second instructional design report (R2).

After submitting R2, preservice teachers were provided with a GenAI user manual during revision and instructed to collaborate with the AI chatbot to refine their designs. Following the manual, they used the system's chat-based interface to describe their instructional design needs and iteratively request targeted revision suggestions across multiple turns. The GenAI system used in this phase was Doubao, a conversational AI platform developed by ByteDance based on the Yunque Transformer model, which is widely used in China [49]. The manual provided basic operational instructions for the platform and introduced the CRISPE framework [41] to scaffold preservice teachers in formulating effective prompts. Manual details are provided in Supplemental Material [50]. This phase supported the refinement of instructional designs and enabled participants to evaluate the strengths and limitations of collaborating with the chatbot through direct experience.

After collaborating with GenAI, preservice teachers refined their designs, resulting in the third instructional design report (R3). To ensure grading consistency, all versions of the instructional design report were evaluated by the same TA using identical criteria.

Upon submitting R3, preservice teachers were required to complete two additional tasks. They were asked to identify the source of the feedback they had previously received (whether human- or AI-generated) [51] and to evaluate its quality. Feedback quality was rated on a 1–5 Likert scale across four dimensions: Accuracy of Facts, Relevance, Language Quality, and Constructiveness. These dimensions were informed by universal intellectual standards [52] and previous research on the quality of feedback [11,51]. See Supplemental Material for the questionnaire [50].

Then participants submitted a reflective report documenting their experience collaborating with GenAI to optimize their instructional design. The researchers posed the following five questions, encouraging participants to add any additional thoughts:

1. Apart from using AI for this instructional design optimization, have you used AI in previous reports?
2. Do you think the AI helped improve the quality of your instructional design this time?
3. What are the advantages of using AI for instructional design?
4. What challenges did you encounter when interacting with the AI? How did you resolve them?
5. If physics teachers were to use AI for instructional design in the future, what role do you think AI would play?

Then a two-week instructional design presentation and in-time comments from the course teacher was conducted, followed by a final closed-book examination. This exam included an instructional design task (FR), with task format and grading criteria consistent with those used in the course training. The resulting score was recorded as a measure of each participant's final instructional design skills. Upon completion of the experimental procedures, preservice physics teachers were informed the source of the first feedback. To maintain fairness for those who preferred human over AI judgment, course TA and expert offered consultations and human feedback on instructional designs.

B. Participants

A total of 121 students enrolled in the "Secondary Physics Pedagogy" course participated in the study. After excluding incomplete submissions, the final dataset comprised 110 students (56 male and 54 female), all of whom were third-year undergraduates aged 19–20 years. Participation was voluntary, and informed consent was obtained from all students following a detailed briefing on the data collection procedures.

C. Development of the criteria for instructional design skills assessment

Building on the commonly identified elements of instructional design, the present study developed scoring criteria adapted to the context of preservice physics teacher education. The criteria consist of seven components: learning objectives, teaching content analysis, learner analysis, important or difficult content, instructional methods, instructional procedures, and board writing design. Evaluation was integrated into instructional procedures, as assessment is inseparable from teaching content and organization. In addition, board writing design was added to emphasize the clarity and sequencing of instructional presentation.

The total score is 100, with each component of the criteria assigned different weights based on its importance in instructional design. Instructional procedures carry the greatest weight, as they capture how all elements of instructional design are enacted in practice. Scores from the seven components are used to evaluate preservice teachers' skills in these key aspects of instructional design. The specific components and their corresponding evaluation descriptions are outlined in Table I, which details the

TABLE I. Description of expected tasks and standards for each component of preservice physics teachers' instructional design.

Components	Description	Score weight (%)
Learning objectives	Describe clear, measurable learning goals using appropriate action verbs, covering the four core competency dimensions.	10
Teaching content analysis	Provide a thorough and accurate analysis of the textbook's content, structure, and logical flow.	10
Learner analysis	Offer a comprehensive analysis of students' cognitive foundations and potential learning barriers, including prior knowledge.	10
Important or difficult content	Identify important or difficult content based on the teaching material.	10
Instructional methods	Identify appropriate instructional methods that emphasize important content and address challenges in the learning process.	5
Instructional procedures	Provide a structured account of the teaching process, including the sequencing of content, teacher and student activities, and evaluation.	50
Board writing design	Organize and present board writing clearly, emphasizing important content and their relationships effectively.	5

tasks and performance expectations that preservice teachers are expected to meet for each component.

Based on these descriptions, the components are categorized into different levels to assess preservice physics teachers' skill levels in these key aspects of instructional design. This classification ensures alignment with the requirements of the physics pedagogy course, curriculum standards, and instructional design theories. The instructional design criteria was developed and refined through expert review. Specific criteria and the development process are provided in Supplemental Material [50].

The goal of this process was not to define a fixed standard, but to create a clear, structured, and reliable evaluation framework aligning with teachers' rough criteria for evaluating instructional design based on their experience, ensuring the feasibility of subsequent research on the consistency between human and GenAI scoring.

D. Designing prompts for generating scores and feedback content

OpenAI's platform offers users the ability to create customized versions based on different models, providing flexibility and emphasizing personalization. This functionality, known as "explore GPT," allows continuous optimization based on specific needs [44,45]. We selected the ChatGPT-4 model, the newest model available at the time of study planning, to build a personalized GPT for scoring and feedback generation, which we refer to as the instructional design GPT assistant (GPT assistant).

The GPT assistant is tasked with two main objectives: scoring preservice teachers' instructional designs based on established grading criteria and providing feedback with improvement suggestions based on the scoring results. To reduce hallucinations and improve scoring and feedback accuracy, we provided detailed explanations for specific terms and concepts in the grading criteria, derived from the physics curriculum standards issued by China's Ministry of Education. These explanations, titled "Explanation of GenAI Criteria," were included beneath the grading annotations for each section and were fully integrated into the GPT assistant's knowledge base.

Additionally, we developed detailed guidelines for the feedback format. The feedback begins by informing preservice teachers of their score, which is divided into four levels: excellent (85–100), good (70–84), satisfactory (60–69), and unsatisfactory (below 60). The feedback then identifies areas for improvement in each of the seven sections and provides targeted suggestions for enhancement. Finally, feedback was capped at 350 words to standardize length across responses and keep it concise and actionable [53].

To enhance scoring consistency, each instructional design was evaluated component by component, with criteria applied independently [33]. We adopted a prompt engineering strategy [9], instructing the GPT assistant to evaluate the instructional designs sequentially according to the seven components. This ensured that the GPT assistant completed the evaluation of one section before moving on to the next, strictly adhering to the grading criteria. Missing sections were assigned a score of zero. Grading consistency was assessed through stratified random sampling of 10 instructional designs (approximately 10%), revealing a statistically significant agreement between the GPT assistant and the TA (ICC $= 0.807$, $F = 5.169$, $p = 0.011$).

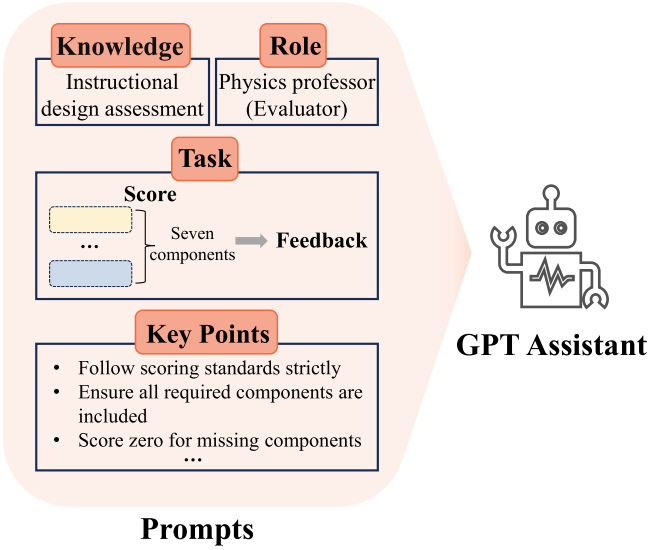


FIG. 2. Key components of the instructional design GPT assistant prompts.

The prompts were used to develop the GPT assistant, with the main sections outlined in Fig. 2. The instructional design assessment criteria were integrated into the system's knowledge base. Additionally, the prompts clearly defined the GPT assistant's role, including scoring instructional designs and generating feedback. Detailed prompts are provided in Supplemental Material [50].

E. Data collection and analysis

This study utilized diverse data sources and applied both quantitative and qualitative analysis to comprehensively assess the research questions.

To evaluate the effectiveness of GenAI in replacing traditional grading, 110 instructional designs from R1 were randomly assigned to three groups: 36 designs were assessed by the GPT assistant, 37 by a physics education expert, and 37 by a TA. Interrater reliability among the three evaluators was assessed using the average measures intraclass correlation coefficient (ICC) with a two-way random effects model, measuring consistency across various components. Interpretation followed the criteria established by Koo and Li [54]: <0.5 (poor), 0.5–0.75 (moderate), 0.75–0.9 (good), >0.9 (excellent), with 95% confidence intervals.

Next, the quality of feedback generated by the GenAI was assessed. To determine whether preservice teachers could distinguish between GenAI-generated and human feedback, they completed a questionnaire where they were asked to identify the source of the feedback and justify their judgment. To explore whether feedback quality assessments were influenced by the source, participants rated feedback across four dimensions: Accuracy of Facts, Relevance, Language Quality, and Constructiveness, using a 1–5 scale (5 being the highest score) [11,52]. Internal consistency of the feedback quality measure was acceptable, with Cronbach's alpha equal to 0.712. Deleting any single item did not increase Cronbach's alpha above 0.712. The four ratings were analyzed as facets of perceived feedback quality.

As the questionnaire data did not meet the assumption of normality, the Mann-Whitney U test was used to compare feedback quality ratings between those who identified the feedback as AI- or human-generated. The Kruskal-Wallis H test was applied to compare feedback quality across the actual sources: GenAI, expert, and TA. If significant differences were found, *post hoc* Dunn's procedure with Bonferroni correction was conducted for pairwise comparisons.

To examine whether GenAI improves instructional design skills, scores from R1, R2, R3, and FR were analyzed. Since these scores were not normally distributed, the Friedman test was used instead of repeated measures ANOVA to assess changes across the four time points. Kendall's W was reported as the effect size (small = 0.10, medium = 0.30, large = 0.50) [55]. If significant differences were found, *post hoc* analyses were conducted using the Wilcoxon signed-rank test, with significance values adjusted for multiple comparisons using the Bonferroni correction.

Finally, we conducted a qualitative content analysis of the reflective reports. To assess the effectiveness of GenAI-assisted instructional design, preservice teachers submitted written reflections on their experiences collaborating with GenAI to improve their instructional designs. After completing the course, 6 of the 110 submissions were blank and 5 addressed unrelated topics. The final dataset therefore comprised reflective reports from 99 preservice physics teachers.

Following initial coding protocols, two independent coders engaged in discrepancy resolution dialogues and executed a secondary coding iteration to establish analytical consistency. Intercoder reliability assessment yielded a Cohen's kappa coefficient of 0.82, indicating substantial interrater agreement. The final coding themes are presented in Table VIII. This qualitative analytical framework complements quantitative findings by delineating the specific contextual manifestations encountered by educators during GenAI integration practices.

All statistical analyses were performed using SPSS version 27 and Nvivo 14, with the detailed analysis presented in the Results section.

IV. RESULTS

A. The effectiveness of GenAI in replacing expert and TA grading and feedback

1. RQ1a: Human-AI grading consistency

To assess the interrater reliability among the three evaluators, both the TA and the GPT assistant independently reevaluated the 37 instructional designs assigned to the expert group using the instructional design grading criteria.

The consistency between the expert and TA ratings was high across all components (ICC > 0.7, $p < 0.001$), indicating substantial agreement. This suggests that both the expert and the TA shared a similar understanding of the grading criteria.

The total instructional design scores showed high consistency across the GPT assistant, the expert, and the TA. Specifically, as shown in Table II, the ICC value for the total instructional design score between the GPT assistant and the expert was 0.764 ($F = 4.234$, $p = 0.000$), while the ICC value between the GPT assistant and the TA was

TABLE II. Consistency of ratings between the GPT assistant, expert, and the TA.

	Intraclass correlation coefficient (ICC)		95% confidence interval		F test with true value 0			
					AI vs TA		AI vs expert	
Component	AI vs TA	AI vs expert	AI vs TA	AI vs expert	Value	p	Value	p
Total instructional design score	0.800	0.764	0.611–0.897	0.541–0.878	4.996	0.000	4.234	0.000
Learning objectives	0.388	0.346	−0.188–0.685	−0.269–0.663	1.635	0.073	1.530	0.103
Teaching content analysis	0.801	0.805	0.614–0.898	0.622–0.900	5.031	0.000	5.133	0.000
Learner analysis	0.499	0.414	0.027–0.742	−0.138–0.698	1.996	0.021	1.707	0.057
Important or difficult content	0.318	0.375	−0.325–0.649	−0.214–0.678	1.465	0.128	1.599	0.082
Instructional methods	0.981	0.950	0.963–0.990	0.902–0.974	52.860	0.000	19.896	0.000
Instructional procedures	0.785	0.728	0.582–0.889	0.472–0.860	4.646	0.000	3.676	0.000
Board writing design	0.838	0.852	0.686–0.917	0.714–0.924	6.184	0.000	6.779	0.000

TABLE III. Preservice teachers' perceived source of feedback vs actual source.

	GPT assistant	Expert	TA
AI	23 (64%)	14 (38%)	22 (59%)
Human	13 (36%)	23 (62%)	15 (41%)
Total	36	37	37

0.800 ($F = 4.996$, $p = 0.000$), both indicating significant consistency.

However, lower consistency (ICC < 0.5) was observed for learning objectives, learner analysis, and important or difficult content. These discrepancies can be attributed to the GPT assistant's tendency to extrapolate implicit meanings from the content, which leads to greater variability in components that require precise and concise evaluation. Additionally, since instructional design reflects an individual's cognitive process rather than simple content matching, the GPT assistant may misinterpret a student's reasoning or introduce unsubstantiated information, leading to assessment bias [9].

Overall, the total instructional design score consistency between the instructional design GPT assistant and human raters is relatively high, suggesting its potential to provide a consistent standard and match the understanding of human raters, making it a valuable tool for assisting with instructional design evaluations.

2. RQ1b: The evaluation of feedback content by preservice physics teachers

After completing the instructional design courses, 110 preservice physics teachers completed a questionnaire to report their predictions on whether the feedback on their reports were provided by AI or human raters, which are shown in Table III. The teachers were also asked to explain their reasoning.

To test whether preservice physics teachers could distinguish the feedback source, the actual sources were collapsed into two categories: GPT assistant feedback coded as AI-generated, and expert and TA feedback coded as human-generated. The chi-square analysis showed that perceived source judgments did not differ significantly across these two actual-source categories ($p = 0.133$). This result provides no evidence that preservice physics teachers could reliably distinguish AI-generated from human-generated feedback.

To examine whether perceived source varied across grader types, a chi-square test of independence was conducted between perceived source and actual grader type. The association did not reach conventional significance ($p = 0.057$). Descriptively, GPT assistant and TA feedback was more often labeled as AI-generated, whereas expert feedback was more often labeled as human-generated.

A qualitative analysis was conducted to explore the reasons behind the misclassification of feedback sources, based on the participants' justifications. Table IV presents the typical reasons cited by preservice teachers for

TABLE IV. Reasons for preservice teachers' judgments.

Judgments	Actual results	Typical reasoning for judgment
AI	TA	• The scores and grades are given first, followed by point-based feedback. The structure is too clear, resembling machine interaction. • The formatting and language are very rigid. Moreover, some evaluations seemed inaccurate, and I even doubted whether they were truly evaluated by experienced teachers when I first received them.
	Expert	• Considering the facts, providing such comprehensive feedback for a large number of assignments is a big challenge for teachers. • The overall description is quite abstract, and it is unclear where the specific issues in the instructional design lie, with no clear direction for revisions.
	GPT assistant	• Each section starts with praise, followed by pointing out weaknesses. If it were a teacher, they might be more critical. • The feedback language is formal and official, with a somewhat procedural tone.
Human	GPT assistant	• Photos of the board design were also provided as feedback, which AI should not be able to give. • For each section of my instructional design, detailed revision suggestions were given, and I found them very reasonable, not something AI could write.
	Expert	• The feedback was brief and to the point. • The feedback suggested adding innovative teaching components, something that AI generally struggles to address in terms of innovation. • The feedback was accurate and professional, and I believe AI is not capable of such judgment at this level.
	TA	• The feedback was concise and targeted, with no traces of AI language, and very reasonable. • The feedback was written in sections and was quite detailed, and I think AI isn't that smart.

TABLE V. Differences in feedback quality by preservice teachers' classification. Note: Values are reported as mean (SD) for readability. Group differences were tested using the Mann-Whitney U test. $*p < 0.05$, significant; $**p < 0.01$, highly significant.

Perceived source	Count	Accuracy of facts	Relevance	Language quality	Constructiveness
AI-generated	59	3.75 (0.76)	3.76 (0.80)	3.80 (0.71)	3.78 (0.77)
Human-generated	51	4.14 (0.63)	4.04 (0.80)	4.10 (0.78)	4.06 (0.76)
Z statistic		2.661	1.942	2.070	2.007
p value		0.008**	0.052	0.038*	0.045*

classifying feedback as either AI-generated or human-generated. The primary reason for misidentifying TA feedback as AI-generated was its structured language style and relatively lower quality. Preservice teachers who made this misclassification perceived the feedback as overly organized and rigid, lacking the depth and professionalism expected from human evaluators.

On the other hand, expert-generated feedback was sometimes mistakenly identified as AI-generated, but for different reasons. Preservice teachers who made this error assumed that experts would not have the capacity to review a large volume of instructional designs within a short period, leading them to believe that the feedback was produced by an AI system with greater efficiency. Additionally, they found expert feedback to be concise and abstract, often lacking specific, actionable revision suggestions, which reinforced their belief that it was AI-generated.

Conversely, some feedback generated by the GPT assistant was misclassified as human-generated. This misperception primarily arose from misunderstandings regarding the current capabilities and limitations of GenAI. For instance, some preservice teachers believed that AI could neither accurately recognize elements like blackboard writing in instructional designs nor provide detailed instructional guidance.

To assess the quality of feedback, participants' responses to the feedback quality questionnaire were categorized based on whether they classified the feedback as AI-generated or human-generated. The results are shown in Table V. Significant differences were identified in the scores for Accuracy of Facts ($z = 2.661$, $p = 0.008$), Language Quality ($z = 2.070$, $p = 0.038$), and Constructiveness ($z = 2.007$, $p = 0.045$). Although the difference in Relevance did not reach conventional significance ($z = 1.942$, $p = 0.052$), mean ratings were consistently higher for feedback perceived as human-generated across all four dimensions.

This pattern is further supported by the visualized score distributions across the four dimensions. As shown in Fig. 3, the mean rank values for feedback perceived as human-generated are consistently higher, and the number of participants assigning a full score of 5 (represented by the blue bars) is noticeably greater compared to AI-generated feedback (red bars).

These results suggest that feedback classified as human-generated was evaluated more favorably overall, with statistically significant differences observed between the two groups.

To further examine whether quality judgments were influenced by assumptions about the feedback source, we categorized the feedback into three groups based on its actual source: GPT assistant-generated feedback, expert-generated feedback, and TA-generated feedback. We then analyzed whether significant differences existed among these groups across the four dimensions of feedback quality.

The results are presented in Table VI. As shown, no significant differences were found in Accuracy of Facts ($H = 1.762$, $p = 0.414$), Relevance ($H = 4.037$, $p = 0.133$), and Language Quality ($H = 2.199$, $p = 0.333$). However, significant differences were observed in the Constructiveness dimension ($H = 14.429$, $p = 0.001$), prompting pairwise comparisons. The description for Constructiveness is: *The feedback provides specific, actionable suggestions for improvement that are helpful for refining the instructional design*. No significant differences were found between the expert feedback group and the TA feedback group ($z = -0.272$, $p = 1.000$). However, the GPT assistant group received significantly higher scores than both the expert feedback group ($z = 3.424$, $p = 0.002$) and the TA feedback group ($z = 3.153$, $p = 0.005$). This suggests that GPT assistant-generated feedback may help address some limitations of human feedback, such as the abstractness of expert feedback and the lack of professionalism in TA feedback.

In summary, preservice teachers' evaluations of feedback quality are strongly influenced by whether they perceive the feedback as GenAI-generated or human-generated. However, in practice, there are generally no significant differences in their assessments based on the actual feedback source. Preservice teachers seem to face trust issues with GenAI-generated feedback.

B. Instructional design skills improve with the assistance of generative artificial intelligence

To examine the impact of GenAI on preservice teachers' instructional design skills, this study compares scores across seven components in three versions of instructional design: R1, R2, and R3. R1 scores reflect initial abilities,

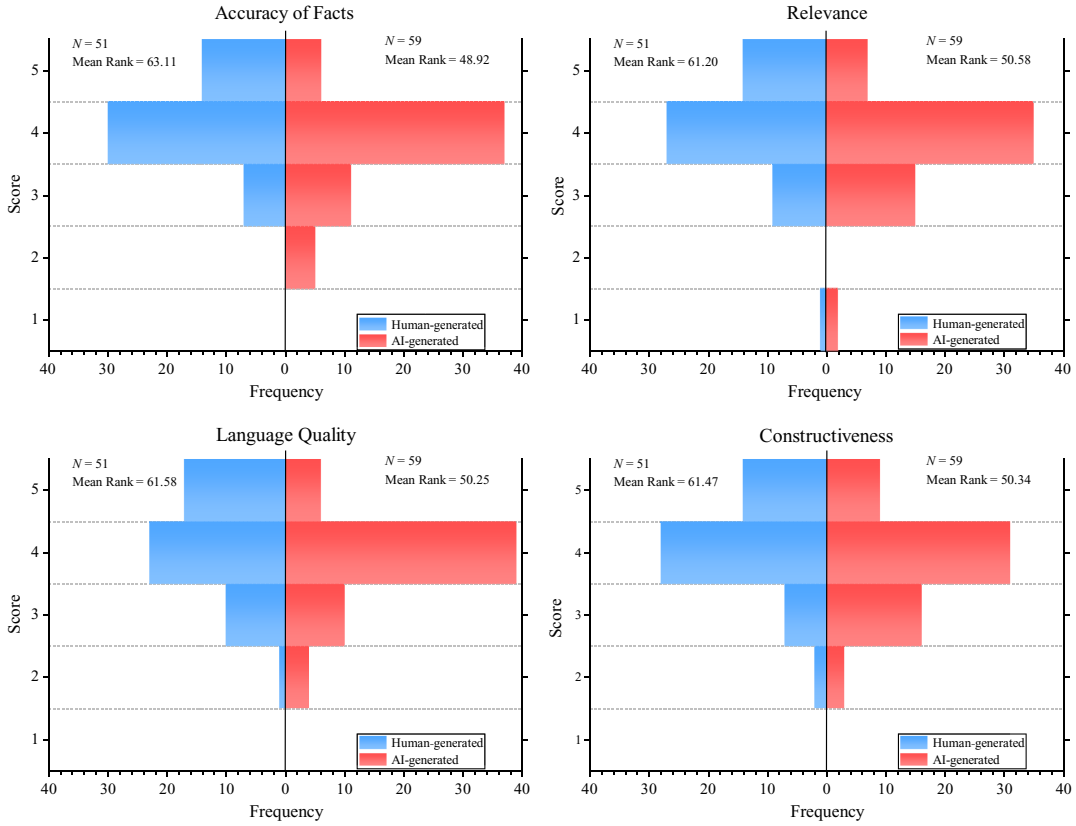


FIG. 3. Differences in feedback quality across dimensions by preservice teachers' classification.

R2 scores capture the effects of feedback, and R3 scores demonstrate the added value of GenAI support. At the course's conclusion, a closed-book exam assessed students' ability to apply knowledge in a practical instructional design task, accounting for 20% of the total test score. The scoring criteria for this task were consistent with those used throughout the course, ensuring alignment between the training process and the final assessment. The raw score was converted into a percentage (score/20 × 100) and recorded as the final report score (FR). Since only the total score was reported, individual component scores were not included.

To assess preservice teachers' actual instructional design skills, Friedman test and pairwise comparisons were conducted on the total scores at each time point. As shown in Fig. 4, significant improvements were observed in total instructional design scores from R1 to R2, reflecting the impact of feedback, and from R2 to R3, reflecting the impact of GenAI support. Final report (FR) scores were significantly lower than R3 but higher than R1 and were comparable to R2 scores. These results suggest that preservice teachers' skills might be closer to the R2 level, with the improvement seen in R3 likely inflated due to reliance on GenAI-generated content.

The Friedman test revealed significant differences across the seven components for R1, R2, and R3, with effect sizes ranging from moderate to large (see Table VII). Pairwise *post hoc* comparisons revealed that, compared to the feedback intervention from R1 to R2, the transition from R2 to R3 showed stronger significance in the components of learning objectives, teaching content analysis, and instructional procedures. This finding aligns with

TABLE VI. Differences in feedback quality based on actual feedback grouping. Note: Values are reported as mean (SD) for readability. Group differences were tested using the Kruskal-Wallis H test. $^{*}p < 0.05$, significant; $^{**}p < 0.01$, highly significant.

Actual source	Count	Accuracy of facts	Relevance	Language quality	Constructiveness
GPT assistant	36	4.03 (0.61)	4.11 (0.52)	4.08 (0.65)	4.31 (0.62)
Expert	37	3.95 (0.82)	3.81 (1.02)	3.92 (0.76)	3.70 (0.74)
TA	37	3.81 (0.74)	3.76 (0.76)	3.81 (0.85)	3.73 (0.80)
H statistic		1.762	4.037	2.199	14.429
p value		0.414	0.133	0.333	0.001**

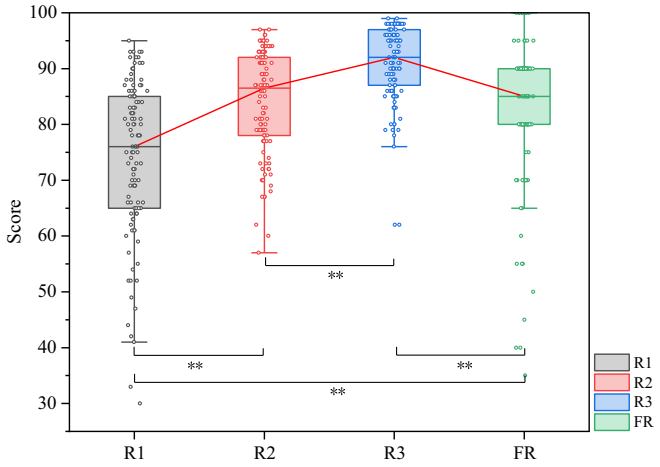


FIG. 4. Comparison of final report scores (FR) across different reports of instructional design (R1, R2, R3), $*p < 0.05$, significant; $**p < 0.01$, highly significant.

challenges noted in the literature, where preservice teachers often face difficulties in these areas of instructional design [24,56]. Even when feedback highlights issues in these components, preservice teachers frequently struggle to make improvements on their own. At this stage, GenAI support provides opportunities for multiple interactions and personalized feedback to help them improve these components. In contrast, feedback was more effective in areas such as important or difficult content, instructional methods, and board writing design. Specifically, during the GenAI-collabrated phase (R2 to R3), there were no statistically significant gains in instructional methods or board writing design.This may be due to feedback being better aligned with specific challenges in content and method selection.

C. Preservice physics teachers' perceptions of instructional design through human-AI collaboration

1. Advantages

Table VIII presents preservice physics teachers' perceptions of collaboration. By the end of the course, nearly all participants (96%) reported improvements in the quality of their instructional design work and expressed a willingness to continue using GenAI in future educational settings. The majority (61%) indicated that GenAI helped them overcome cognitive constraints and spark diverse ideas for instructional design. In addition, 31% noted that it made their instructional design language more professional and concise, and 51% reported that it enabled rapid production of needed materials, significantly enhancing the efficiency of their instructional design.

> PST16: The efficiency is striking. A revision that used to take an entire afternoon now takes about thirty minutes with AI. It reinforces my belief that the value of science and technology is to ease human effort, lighten study and work, and spare us a great deal of low-value busywork.

2. Challenges

Preservice physics teachers also reflected on the challenges they faced when using GenAI. A majority (82%) of

TABLE VII. Differences of instructional design scores by components for each report. Note: (a) The Friedman test (d.o.f. = 2) was conducted as a non-parametric test to examine overall differences across time points. (b) Pairwise comparisons were performed using the Wilcoxon signed-rank test, with significance values adjusted for multiple comparisons using the Bonferroni correction. $*p < 0.05$, significant; $**p < 0.01$, highly significant.

Component	Score weight (%)	Mean (SD) R1	Mean (SD) R2	Mean (SD) R3	χ^2	p value	Effect size (Kendall's W)	Pairwise comparison
Learning objectives	10	7.6 (1.90)	8.4 (1.46)	9.4 (1.00)	124.87	<0.001	0.57	R1 < R2* R2 < R3**
Teaching content analysis	10	7.5 (2.47)	8.3 (1.48)	9.6 (0.88)	118.99	<0.001	0.54	R1 < R2 R2 < R3**
Learner analysis	10	7.5 (2.48)	8.7 (1.49)	9.6 (1.19)	119.32	<0.001	0.54	R1 < R2** R2 < R3**
Important or difficult content	10	8.3 (1.61)	9.5 (0.63)	9.9 (0.41)	122.18	<0.001	0.56	R1 < R2** R2 < R3*
Instructional methods	5	3.1 (2.00)	4.5 (1.25)	4.8 (0.83)	114.17	<0.001	0.52	R1 < R2** R2 < R3
Instructional procedures	50	37.7 (7.23)	40.7 (5.11)	43.5 (5.15)	74.32	<0.001	0.34	R1 < R2 R2 < R3**
Board writing design	5	2.5 (2.11)	4.2 (1.38)	4.4 (1.23)	89.65	<0.001	0.41	R1 < R2** R2 < R3

TABLE VIII. Interview coding themes.

Category	Description	Percentage (%)
1. Prior experience with GenAI		
No prior use	Had not used GenAI in earlier reports.	57
Prior use	Had used GenAI in earlier reports.	43
2. Perceived effectiveness		
Effective	Reported that collaborating with GenAI improved the quality of their instructional designs.	96
Not effective	Reported no improvement from collaborating with GenAI.	4
3. Advantages		
Professional language	Helped render design language more professional and concise.	31
Idea generation and variation	Generated diverse ideas and design alternatives that inspired instructional design.	61
Resource synthesis	Synthesized information and integrated relevant resources, examples, and methods.	27
Efficiency and convenience	Produced usable content quickly, increasing preparation efficiency and reducing workload.	51
4. Challenges		
Reliability of content	Questioned the accuracy and classroom feasibility of GenAI outputs, noting that claims require further verification.	57
Overreliance risk	Worried that dependence on GenAI limits the development of pedagogical skills and independent problem solving.	8
Prompting skill	Found it difficult to ask precise questions.	82
5. Future roles		
Irreplaceable teacher role	Stated that GenAI should not replace teachers because current outputs often lack the human-centered responsiveness needed to engage and adapt to students.	61
Procedural tasks	Preferred GenAI to handle mechanical and procedural tasks.	54

preservice physics teachers emphasized that the clarity and specificity of their prompts significantly affect the quality of GenAI-generated responses. They struggled with providing precise instructions, which led to misinterpretations and less relevant results. Through iterative refinement of their questioning strategies, some of them were able to generate more accurate and practical feedback. As one preservice teacher noted:

> PST72: At the beginning of my interaction with AI, my input was not specific enough, leading to broad and less targeted responses that did not fully meet my expectations. Later, following the template provided by my instructor and refining my requirements based on my own needs, I was able to make my prompts more precise. As a result, the AI-generated contents became more accurate and practical.

Concerns about the reliability of GenAI-generated content were raised by 57% of preservice teachers, particularly in the context of physics.

> PST5: Sometimes it gives answers that aren't grounded in evidence. They can sound convincing at first, but a closer look shows they are not scientific. As physics teachers, we work in a discipline built on experimental facts, so using such flawed material in class is a serious mistake.

While the information is factually correct, it may not always align with current educational practices due to regional or institutional differences. As one preservice teachers remarked:

> PST29: AI's data updates may lag behind actual developments in education, resulting in recommendations and instructional designs that lack innovation or fail to meet contemporary teaching requirements.

Additionally, while GenAI-generated instructional designs may seem theoretically sound, their feasibility in actual classrooms requires further validation. These concerns highlight the importance of critically assessing and adapting GenAI-generated instructional designs before implementing them in practice. One preservice teacher noted:

> PST49: Some suggestions seem reasonable in theory but may be impractical due to constraints such as time, equipment availability, and student engagement. For instance, complex experimental designs may require extended class time, which is often unfeasible.

While the powerful functions and efficiency of GenAI are impressive, 8% of preservice teachers expressed concerns about the overreliance on GenAI, noting that it could undermine independent thinking. Directly incorporating GenAI-generated content into instructional designs may limit professional growth and affect teaching quality. As another preservice teacher shared:

> PST75: I have significant concerns about the increasing dependence on it. If AI becomes overly relied upon in future teaching, it may hinder progress and growth. This has led me to reflect on how to balance the convenience AI offers while maintaining the independence and creativity essential to teaching.

3. Future roles

Preservice teachers emphasized that GenAI cannot replace the role of physics teachers (61%). They felt that GenAI-generated instructional designs often lack human-centered responsiveness. Unlike human teachers, who can adapt in real time to students' reactions and provide encouragement through language, facial expressions, and gestures, GenAI struggles with emotional understanding and empathy. Consequently, it is less effective in fostering students' curiosity or intrinsic motivation to learn physics. Instead, preservice teachers expressed a preference for GenAI to assist with the more mechanical and procedural aspects of teaching, viewing it primarily as a tool to enhance teaching effectiveness (54%).

> PST86: I want AI to handle the heavy but unavoidable parts of instructional design, like searching for materials and polishing text. These are low-complexity tasks that demand little mental engagement, and letting AI do them saves a lot of time and effort.

V. DISCUSSION

This study investigates the impact of GenAI replacing physics teacher educators in providing evaluative feedback and collaborating on instructional design. Over one semester, 110 preservice physics teachers participated, and findings from empirical data and interviews reveal both its pedagogical value and the perceptions surrounding this role substitution.

A. The effectiveness of GenAI in providing feedback and collaborating

Although GenAI has been widely applied in education, it is not inherently designed for teacher instruction. Empirical research is therefore necessary to investigate its practical application as an educational tool and assess its performance within pedagogically grounded contexts [39]. This study examines whether GenAI can generate evaluation scores and feedback comparable to those of human experts when applying the same instructional design assessment criteria. A GenAI user manual for instructional design revision was then provided to all preservice teachers, and subsequent changes in design quality were measured. The overarching aim is to determine whether GenAI can effectively support preservice teachers in addressing challenges encountered during instructional design.

Through prompt engineering and sectional evaluation, the GPT-4 model-based assistant produces total instructional design scores consistent with those given by human experts and TAs. It effectively evaluates preservice teachers' instructional designs based on established criteria, demonstrating the potential of GenAI as an effective tool for assessing students' written texts. While differences with human evaluations persist in three component level, Jauhiainen suggests that these can be mitigated by refining the prompts or eliminating unnecessary evaluation criteria that contribute to grading discrepancies [9]. GenAI-generated feedback was sufficiently natural and humanlike that preservice teachers often could not distinguish it from expert feedback, with many perceiving it as more detailed and actionable. This may be because expert feedback tends to be more abstract and concise due to workload-induced fatigue, whereas TA feedback is more rigid and lacks depth [57].

Instructional designs optimized with GenAI showed significant improvement in five of the seven evaluated components, including the overall score of the final designs. Guided by the GenAI-assisted optimization manual, preservice teachers reported that they could quickly generate diverse teaching ideas and methods, overcoming limitations in their thinking [8,16]. Furthermore, GenAI offered systematic support in structuring knowledge, providing comprehensive frameworks for aligning textbook content with learner needs. This effectively addressed the issues of preservice teachers' insufficient understanding of the depth of teaching materials and students' needs [24]. These findings indicate that a GenAI-integrated evaluation platform could improve teaching efficiency, deliver timely feedback, and collaboratively support preservice physics teachers in refining their instructional designs.

B. Challenges and limitations of GenAI in teacher role substitution

Despite the potential, further use of GenAI in role substitution faces challenges, with one key issue being a lack of trust. Preservice teachers often assume that human-generated feedback is inherently superior to GenAI-generated content, even though no significant difference in quality exists between the two. This lack of trust could present a substantial barrier to the widespread adoption of GenAI-based grading and feedback assistants. Trust plays a crucial role in the adoption of unfamiliar systems [36].

Low trust often leads to skepticism about new systems, making it less likely to be used [58]. In future implementations where GenAI directly provides feedback and evaluations, preservice teachers may be reluctant to revise their work based on AI-generated suggestions, questioning the fairness and accuracy of GenAI assessments [35]. Such hesitation could limit the willingness to fully integrate GenAI into educational practices.

In their experiences, some preservice teachers voiced concerns about becoming overly reliant on technology. While they acknowledged the need to filter and revise the generated content, the tool's convenience often led them to directly adopt the instructional designs it produced. This reliance was reflected in the final exam, where preservice teachers scored significantly lower on independently completed instructional design tasks than on those developed with GenAI. Their final performance more closely resembled the postfeedback level, suggesting that the apparent gains from collaboration were inflated by reliance on GenAI-generated content. These findings align with previous research, where some participants regarded GenAI as a helpful assistant, while others perceived it as a form of cheating [59,60]. Therefore, they need to critically assess the GenAI-generated content, applying evaluative skills and critical thinking [24,47] to judge the scientific accuracy, appropriateness, and practical feasibility of the results [6,61]. In this process, preservice teachers play a dual role as both users of technology and decision makers, ensuring that GenAI applications align with pedagogical objectives and ethical standards [62].

Additionally, preservice teachers continue to face challenges in integrating this technology into collaborative practices. A key difficulty lies in their limited ability to formulate effective questions when interacting with GenAI, which is essential for obtaining relevant and high-quality results that address their instructional needs [41]. Moreover, building on the differential gains observed across the seven dimensions from R1 to R3, the findings indicate that the collaboration with GenAI is most helpful for structured, content-expandable components, including learning objectives, teaching content analysis, and instructional procedures. However professional judgment by teachers remains essential for context-sensitive decisions, such as identifying important or difficult content, selecting instructional methods, and board writing design. These latter feedbacks require a holistic grasp of physics subject matter, knowledge of common misconceptions and learning progressions, and sensitivity to classroom contingencies, which enables more professional and situated recommendations. Accordingly, the question of when to involve GenAI, in what form of support, and for which content targets remains central.

To address these limitations, the introduction of elective courses focused on a comprehensive understanding of GenAI's strengths and weaknesses could encourage preservice teachers to engage more critically and objectively with AI-generated content. Such courses should aim to foster deeper awareness of GenAI's functions, its developers, its beneficiaries, and its broader societal implications [7,16]. Moreover, teacher educators in universities must continue to explore more effective collaborative models that help preservice teachers overcome dependency, fostering genuine enhancement of their professional skills.

VI. CONCLUSION

The findings indicate that generative artificial intelligence holds significant potential in enhancing preservice physics teachers' instructional design skills through both evaluative feedback and collaborative assistance. However, the study reveals that preservice physics teachers' assessments of GenAI-generated content are closely linked to their understanding in intelligent technology. This group exhibits a contradictory tendency, while they recognize the need for independent thinking and skepticism toward GenAI-generated content, aspiring to be decision makers in instructional design, they also rely excessively on GenAI due to its convenience. This paradox may stem from their limited experience in education, which makes their well-intentioned efforts less effective. These findings underscore the importance of focusing not only on the feasibility of GenAI technology but also on the user experience of preservice physics teachers. Such factors directly influence the effectiveness and real-world applicability of GenAI tools in educational contexts.

VII. LIMITATIONS AND FUTURE WORK

This study has several limitations. First, it was conducted exclusively with preservice teachers in physics education and focused solely on instructional design development, meaning the results may not be generalizable to other groups of preservice teachers. In addition, the instructional design components and weighting scheme were developed for this specific secondary physics context in China. Both were established through expert consensus and should be validated in future studies across broader samples and instructional contexts.

Second, the study could not track the specific actions and behaviors of the participants as they used GenAI to modify their instructional designs, offering only an assistant manual for guidance. And when examining changes in instructional design skills, we conducted an initial investigation by comparing preservice physics teachers' instructional design reports before and after GenAI collaboration phase revision within the same cohort. Future studies should incorporate a control group to enable more rigorous experimental comparisons.

Third, feedback quality was assessed using a concise, theory-informed rating scheme that is suitable for capturing preservice teachers' perceived differences between AI- and human-generated feedback in this instructional design

context. However, this measure should be interpreted as an exploratory operationalization and may not capture the full breadth of feedback quality perceptions. Future work should refine and expand the item set and validate the resulting instrument across tasks and contexts to enable more systematic comparisons and stronger inferences about the mechanisms underlying teachers' feedback perceptions.

Finally, the sequence of feedback and GenAI collaboration was not considered. Future research should isolate the effects of phase order through controlled comparisons that vary the sequence of feedback and collaboration. Because all participants engaged in the collaboration phase, future studies could also examine whether perceptions of collaboration differ by feedback group and whether collaboration experiences shift teachers' interpretations and evaluations of AI-generated feedback. Such evidence would advance understanding of how preservice physics teachers make sense of technology-mediated instructional materials and how these judgments evolve through use.

ACKNOWLEDGMENTS

This work was supported by the Cross Research Project of Fundamental Research Funds for Central Universities of Central China Normal University in 2025, "Advanced Detection and Artificial Intelligence at the Frontiers of Physics-AI and Physics Education" (Grant No. CCNU25JCPT014).

Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

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] G. Cooper, Examining science education in ChatGPT: An exploratory study of generative artificial intelligence, J. Sci. Educ. Technol. **32**, 444 (2023).

[2] Y. Feldman-Maggor, R. Blonder, and G. Alexandron, Perspectives of generative AI in chemistry education within the TPACK framework, J. Sci. Educ. Technol. **34**, 1 (2025).

[3] J. A. Bowen and C. E. Watson, *Teaching with AI: A Practical Guide to a New Era of Human Learning* (Johns Hopkins University Press, Baltimore, Maryland, 2024).

[4] M. Bower, J. Torrington, J. W. M. Lai, P. Petocz, and M. Alfano, How should we change teaching and assessment in response to increasingly powerful generative artificial intelligence? Outcomes of the ChatGPT teacher survey, Educ. Inf. Technol. 0 (2024).

[5] S. A. D. Popenici and S. Kerr, Exploring the impact of artificial intelligence on teaching and learning in higher education, Res. Pract. Technol. Enhanced Learn. **12**, 22 (2017).

[6] H. Jang and H. Choi, A double-edged sword: Physics educators' perspectives on utilizing ChatGPT and its future in classrooms, J. Sci. Educ. Technol. **34**, 267 (2025).

[7] I. Celik, M. Dindar, H. Muukkonen, and S. Järvelä, The promises and challenges of artificial intelligence for teachers: A systematic review of research, TechTrends **66**, 616 (2022).

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

[9] J. S. Jauhiainen and A. Garagorry Guerra, Generative AI in education: ChatGPT-4 in evaluating students' written responses, Innovation Educ. Teach. Int. **62**, 1377 (2025).

[10] A. V. Y. Lee, Supporting students' generation of feedback in large-scale online course with artificial intelligence-enabled evaluation, Stud. Educ. Eval. **77**, 101250 (2023).

[11] 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).

[12] I.-A. Chounta, E. Bardone, A. Raudsep, and M. Pedaste, Exploring teachers' perceptions of artificial intelligence as a tool to support their practice in Estonian K-12 education, Int. J. Artif. Intell. Educ. **32**, 725 (2022).

[13] G. Cooper, K.-S. Tang, and A. Fitzgerald, Intersections of mind and machine: Navigating the nexus of artificial intelligence, science education, and the preparation of preservice teachers, J. Sci. Educ. Technol. **34**, 1255 (2025).

[14] R. Hall, Generative AI and re-weaving a pedagogical horizon of social possibility, Int. J. Educ. Technol. High. Educ. **21**, 12 (2024).

[15] R. Blonder, Y. Feldman-Maggor, and S. Rap, Are they ready to teach? Generative AI as a means to uncover preservice science teachers' PCK and enhance their preparation program, J. Sci. Educ. Technol. **34**, 1301 (2024).

[16] P. MacDowell, K. Moskalyk, K. Korchinski, and D. Morrison, Preparing educators to teach and create with generative artificial intelligence, Can. J. Learn. Tech. **50**, 1 (2024).

[17] S. Duan, M. Exter, and Q. Li, In their ideal future, are preservice teachers willing to integrate technology in their teaching and why?, TechTrends **68**, 734 (2024).

[18] Q. Tan, Reimagining teacher development in the era of generative AI: A scoping review, Teach. Teach. Educ. **168**, 105236 (2025).

[19] R. F. Kizilcec, E. Huber, E. C. Papanastasiou, A. Cram, C. A. Makridis, A. Smolansky, S. Zeivots, and C. Raduescu, Perceived impact of generative AI on assessments: Comparing educator and student perspectives in Australia, Cyprus, and the United States, Comput. Educ. Artif. Intell. **7**, 100269 (2024).

[20] S. G. Garofalo and S. J. Farenga, Science teacher perceptions of the state of knowledge and education at the advent of generative artificial intelligence popularity, Sci. Educ. **34**, 893 (2025).

[21] X. Zhai, Transforming teachers' roles and agencies in the era of generative AI: Perceptions, acceptance, knowledge, and practices, J. Sci. Educ. Technol. **34**, 1323 (2024).

[22] X. Zhai, M. Nyaaba, and W. Ma, Can generative AI and ChatGPT outperform humans on cognitive-demanding problem-solving tasks in science?, Sci. Educ. **34**, 649 (2025).

[23] R. Zhang, X. Liu, Y. Yang, J. Tripp, and B. Shao, Preservice science teachers' instructional design competence: Characteristics and correlations, Eurasia J. Math. Sci. Technol. Educ. **14**, 1075 (2018).

[24] T. J. Gurl, M. P. Markinson, and A. F. Artzt, Using ChatGPT as a lesson planning assistant with preservice secondary mathematics teachers, Digital Exp. Math. Educ. **11**, 114 (2025).

[25] S. G. Magliaro and N. Shambaugh, Student models of instructional design, Educ. Technol. Res. Dev. **54**, 83 (2006).

[26] W. B. Rouse and N. M. Morris, On looking into the black box: Prospects and limits in the search for mental models, Psychol. Bull. **100**, 349 (1986).

[27] M. Molenda, C. M. Reigeluth, and L. M. Nelson, Instructional design, in *Encyclopedia of Cognitive Science*, 1st ed., edited by L. Nadel (Wiley, New York, 2006).

[28] J. E. Kemp, *The Instructional Design Process* (HarperCollins Publishers, New York, NY, 1985).

[29] P. L. Smith and T. J. Ragan, *Instructional Design* (John Wiley & Sons, New York, 2004).

[30] W. Dick, L. Carey, and J. O. Carey, *The Systematic Design of Instruction* (Scott Foresman, Glenview, IL, 1985).

[31] C.-C. Liu, H.-J. Wang, D. Wang, Y.-F. Tu, G.-J. Hwang, and Y. Wang, An interactive technological solution to foster preservice teachers' theoretical knowledge and instructional design skills: A chatbot-based 5E learning approach, Interact. Learn. Environ. **32**, 6698 (2024).

[32] M. C. Scheeler, K. McKinnon, and J. Stout, Effects of immediate feedback delivered via webcam and bug-in-ear technology on preservice teacher performance, Teach. Educ. Spec. Educ. **35**, 77 (2012).

[33] 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).

[34] 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).

[35] M. F. Shahzad, S. Xu, and H. Zahid, Exploring the impact of generative AI-based technologies on learning performance through self-efficacy, fairness & ethics, creativity, and trust in higher education, Educ. Inf. Technol. **30**, 3691 (2025).

[36] G. Bilquise, S. Ibrahim, and S. M. Salhieh, Investigating student acceptance of an academic advising chatbot in higher education institutions, Educ. Inf. Technol. **29**, 6357 (2024).

[37] F. Kehoe, Leveraging generative AI tools for enhanced lesson planning in initial teacher education at post primary, Ir. J. Technol. Enhanced Learn. **7**, 172 (2023).

[38] C. K. Lo, What is the impact of ChatGPT on education? A rapid review of the literature, Educ. Sci. **13**, 410 (2023).

[39] B. Hu, L. Zheng, J. Zhu, L. Ding, Y. Wang, and X. Gu, Teaching plan generation and evaluation with GPT-4: Unleashing the potential of LLM in instructional design, IEEE Trans. Learn. Technol. **17**, 1445 (2024).

[40] M. R. Karaman and İ. Göksu, Are lesson plans created by ChatGPT more effective? An experimental study, Int. J. Technol. Des. Educ. **7**, 107 (2024).

[41] M. Wang, M. Wang, X. Xu, L. Yang, D. Cai, and M. Yin, Unleashing ChatGPT's power: A case study on optimizing information retrieval in flipped classrooms via prompt engineering, IEEE Trans. Learn. Technol. **17**, 629 (2024).

[42] G. Kortemeyer, Tailoring chatbots for higher education: Some insights and experiences, arXiv:2409.06717.

[43] V. G. A. Hakim, N. A. Paiman, and M. H. S. Rahman, Genie-on-demand: A custom AI chatbot for enhancing learning performance, self-efficacy, and technology acceptance in occupational health and safety for engineering education, Comput. Appl. Eng. Educ. **32**, e22800 (2024).

[44] J. Lademann, J. Henze, and S. Becker-Genschow, Augmenting learning environments using AI custom chatbots: Effects on learning performance, cognitive load, and affective variables, Phys. Rev. Phys. Educ. Res. **21**, 010147 (2025).

[45] K.-S. Tang and G. B. S. Putra, Generative AI as a dialogic partner: Enhancing multiple perspectives, reasoning, and argumentation in science education with customized chatbots, J. Sci. Educ. Technol. (2025), 10.1007/s10956-025-10240-1.

[46] M. Abbas, F. A. Jam, and T. I. Khan, Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students, Int. J. Educ. Technol. Higher Educ. **21**, 10 (2024).

[47] G. Van Den Berg and E. Du Plessis, ChatGPT and generative AI: Possibilities for its contribution to lesson planning, critical thinking and openness in teacher education, Educ. Sci. **13**, 998 (2023).

[48] G.-G. Lee and X. Zhai, Using ChatGPT for science learning: A study on pre-service teachers' lesson planning, IEEE Trans. Learn. Technol. **17**, 1643 (2024).

[49] P. Jia and C. Stan, Artificial intelligence factory, data risk, and VCs' mediation: The case of ByteDance, an AI-powered startup, J. Risk Finance Manage. **14**, 203 (2021).

[50] See Supplemental Material at http://link.aps.org/supplemental/10.1103/hm13-jv98 for the specific criteria, prompts, manual, questionnaire, and results of subscale analyses.

[51] 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).

[52] R. Paul, The state of critical thinking today, New Dir. Community Coll. **2005**, 27 (2005).

[53] 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. High. Educ. **21**, 23 (2024).

[54] T. K. Koo and M. Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, J. Chiropr. Med. **15**, 155 (2016).

[55] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (L. Erlbaum Associates, Hillsdale, NJ, 1988).

[56] D. Schiering, S. Sorge, M. M. Keller, and K. Neumann, A proficiency model for pre-service physics teachers' pedagogical content knowledge (PCK)—What constitutes high-level PCK?, J. Res. Sci. Teach. **60**, 136 (2023).

[57] W. Dai, J. Lin, H. Jin, T. Li, Y.-S. Tsai, D. Gašević, and G. Chen, Can large language models provide feedback to students? A case study on ChatGPT, in *Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies (ICALT)* (IEEE, New York, 2023), pp. 323–325.

[58] W. Chang and J. Park, A comparative study on the effect of ChatGPT recommendation and AI recommender systems on the formation of a consideration set, J. Retail. Consum. Serv. **78**, 103743 (2024).

[59] G. Kortemeyer and W. Bauer, Cheat sites and artificial intelligence usage in online introductory physics courses: What is the extent and what effect does it have on assessments?, Phys. Rev. Phys. Educ. Res. **20**, 010145 (2024).

[60] N. Şimşek, Integration of ChatGPT in mathematical story-focused 5E lesson planning: Teachers, and pre-service teachers' interactions with ChatGPT, Educ. Inf. Technol. **30**, 11391 (2025).

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

[62] M. Bearman, J. Tai, P. Dawson, D. Boud, and R. Ajjawi, Developing evaluative judgement for a time of generative artificial intelligence, Assess. Eval. Higher Educ. **49**, 893 (2024).