

DeepSeek-assisted physics instructional design: An empirical study of high school physics teaching

Dazhen Tong¹, Huihui Wu², A. Y. M. Atiquil Islam^{3,4,*},
Junci Liu², Lei Jiang⁵, and Hongmei Ren^{2,†}

¹*School of Teacher Education, Anhui Normal University, Wuhu, Anhui 241002, China*

²*School of Physics and Electronic Information, Anhui Normal University, Wuhu, Anhui 241002, China*

³*Joint Education Institute of Zhejiang Normal University and University of Kansas,
Zhejiang Normal University, Jinhua, China*

⁴*China-Africa Regional Center for Digital Education Cooperation, Zhejiang Normal University, Jinhua, China*

⁵*Tianjin Academy of Educational Science, Tianjin 300191, China*



(Received 10 December 2025; accepted 27 May 2026; published 30 June 2026)

Generative artificial intelligence (GenAI) has shown growing potential in supporting teacher education, yet little is known about how preservice teachers collaborate with AI during instructional design (ID). This study examined the effects of DeepSeek, a Chinese large language model, on preservice physics teachers' ID performance and cognitive processes. A quasiexperimental design was conducted with 220 participants completing a 45-min ID on Newton's third law, either with or without DeepSeek assistance. Data from IDs, interaction logs, and interviews were analyzed through statistical tests, epistemic network analysis (ENA), and thematic coding. Results indicated that the DeepSeek-assisted group outperformed the control group in analytical and structural aspects of ID, such as teaching content analysis and learner analysis. ENA revealed that high-performing participants exhibited proactive "problem representation–AI interaction–monitoring" patterns, while low-performing participants relied on reactive prompting. Interviews further confirmed that AI was helpful in contextual introductions and experiment design but also revealed issues such as outdated curriculum references and overly absolute expressions that lacked sensitivity to teaching contexts. The findings suggest that GenAI can serve as a cognitive scaffold when used strategically and highlight the need to integrate AI literacy and metacognitive training into teacher education.

DOI: [10.1103/bxql-gzfk](https://doi.org/10.1103/bxql-gzfk)

I. INTRODUCTION

With the rapid development of generative artificial intelligence (GenAI), its potential in education has been increasingly explored. From generating learning resources to conducting learning analytics, and from automated grading to personalized feedback [1–4], GenAI has shown substantial promise in supporting teaching and learning practices. However, existing studies have primarily focused on student-AI interaction, examining how learners use AI to enhance performance in tasks such as writing, drawing, and problem solving [5–7]. In contrast, how teachers collaborate with AI—especially within instructional design (ID), a core component of teachers' professional competence—has received far less empirical attention.

ID involves not only setting learning objectives, organizing instructional activities, and choosing assessment methods but also ensuring the scientific rigor and practical feasibility of classroom teaching [8]. For preservice physics teachers, developing ID competence is a key aspect of professional growth [9]. Yet, due to limited experience, they often face challenges in defining course objectives, organizing learning activities, and addressing students' preconceptions and misconceptions [10–12]. In this context, exploring how AI can help preservice teachers improve their ID quality has become a relevant issue.

DeepSeek, a generative large language model developed in China, provides new opportunities for supporting ID in Chinese educational settings. Beyond generating general lesson frameworks, it can offer insights into experiment design, contextual introductions, and learner analysis. However, empirical evidence remains limited regarding DeepSeek's actual role in ID, the interaction patterns that occur during its use, and preservice teachers' perceptions of its usefulness and limitations. This research gap restricts our understanding of teacher-AI collaboration and hinders the effective application of AI in teacher education. Based

*Contact author: prof.atiquil@gmail.com

†Contact author: hmren@ahnu.edu.cn

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

on this, the present study explores the following three research questions:

- RQ1: Are there significant differences in ID quality between preservice teachers with and without DeepSeek support?
- RQ2: What interaction patterns emerge during DeepSeek-assisted ID?
- RQ3: How do preservice teachers perceive the usefulness and challenges of using DeepSeek in ID?

By integrating comparative experiments, interaction log analysis, and interview data, this study addresses these questions from the perspective of teacher education. It aims to provide empirical evidence and theoretical insights for understanding and promoting teacher-AI collaboration in the era of GenAI.

II. LITERATURE REVIEW

A. The role and challenges of ID in physics education

ID is commonly regarded as a systematic process for facilitating learners' acquisition of knowledge and skills. More specifically, Dick defined ID as "the process of systematically applying instructional theory and empirical findings to the planning of instruction" [13], and it has also been described as a process carried out "in a consistent and reliable fashion" [14]. As an independent discipline, ID not only involves the formulation of instructional objectives, the organization of learning activities, and the design of assessment methods but also emphasizes how to enhance teaching quality and efficiency through holistic optimization. Shulman's framework of pedagogical content knowledge reveals the core of ID: Teachers must develop the ability to transform and reorganize knowledge on the basis of both content knowledge and pedagogical knowledge in order to create instructional plans that align with students' cognitive characteristics [15].

In the context of Chinese education, this requirement has evolved in close connection with national curriculum reforms. Over the past two decades, China has gradually shifted from emphasizing "basic knowledge and skills" to adopting "three-dimensional objectives," and more recently, to establishing a curriculum system guided by "core competencies" [16,17]. The concept of core competencies highlights students' key abilities and essential qualities for lifelong development. In the field of physics education, these competencies are manifested in four dimensions: physical concepts, scientific thinking, scientific inquiry, and scientific attitude and responsibility. These dimensions are not only built upon previous instructional goals but also echo the internationally emphasized ideas of disciplinary literacy and 21st-century skills, thereby posing higher demands on teachers' ID abilities.

Nevertheless, tensions remain between curriculum-level expectations and classroom practices. Although core competency-oriented IDs have become increasingly common,

teachers often struggle to effectively integrate dimensions such as physical concepts and scientific attitudes into classroom instruction [18]. This challenge is closely tied to the inherent characteristics of physics. Topics such as mechanics, energy, and electromagnetism are highly abstract, and students frequently harbor deep-seated misconceptions and alternative conceptions [11,12]. To foster accurate scientific understanding and cultivate positive scientific attitudes, teachers must be able to identify and address these cognitive barriers through their ID—an ability that requires deep content knowledge and an understanding of students' learning processes [9]. For preservice teachers, these challenges are particularly pronounced. When setting instructional goals, organizing learning activities, and designing assessments, they often exhibit fragmented or one-sided approaches that fail to balance scientific rigor, student accessibility, and practical feasibility [10]. ID is inherently a dynamic and iterative process, requiring continuous reflection and revision in practice. However, for preservice teachers who lack both experience and effective feedback mechanisms, such reflective cycles are often difficult to achieve [19].

Therefore, a critical issue arises: How can preservice teachers be provided with timely support and external resources to help them achieve a balance between scientific rigor and practical operability in their ID?

B. Applications of GenAI in physics education

The rapid development of GenAI has opened up new possibilities for addressing the challenges described above. Unlike traditional instructional tools that rely on fixed resources and single functions, GenAI features high flexibility and generative capacity, enabling it to respond dynamically to diverse learning and teaching needs. Powered by deep learning and natural language interaction, GenAI can generate explanations of complex physics concepts, design varied exercises and contextualized cases [20], and use visualization techniques to enhance student understanding [1].

Moreover, GenAI can assist teachers in refining learning objectives, analyzing learner characteristics, and optimizing instructional procedures [3,4]. Even preservice teachers without programming skills can use GenAI to rapidly create well-structured IDs or develop interactive simulation experiments [2]. In assessment and testing, GenAI can generate item banks and scoring rubrics [3,4] and, when combined with optical character recognition tools, can automatically grade handwritten responses—significantly reducing teachers' workload [21–23].

In the realm of personalized learning, GenAI-based assistants or platforms can deliver immediate feedback and differentiated support to students while allowing teachers to monitor and adjust content accuracy to align with curriculum standards [24]. At the same time, AI technologies have demonstrated value beyond instruction,

assisting in both administrative tasks [3,4,25] and academic research, such as large-scale qualitative data analysis [26]. These applications not only save teachers' time but also broaden the methodological toolkit available for educational research.

Overall, GenAI is transforming physics education, shifting it from content delivery to intelligent support and from static design to dynamic generation. It not only expands students' ways of learning but also provides teachers with unprecedented instructional resources and tools. For preservice teachers in particular, GenAI can serve as a powerful external support mechanism, helping them overcome the limitations of limited experience and improve the completeness and creativity of their ID.

C. Preservice teachers and GenAI in human-AI collaboration

As AI continues to evolve from large-scale data processing toward complex decision making and creative collaboration, human-AI interaction (HAI) has emerged as a central focus in contemporary research. HAI emphasizes a user-centered approach and investigates how humans and AI engage in dynamic relationships of cooperation, conflict, and coexistence [27]. Within the educational domain, student-AI interaction (SAI) has become a major subfield of HAI, with a growing body of studies demonstrating its profound impact on both learning processes and outcomes.

In terms of learning performance, AI involvement has been shown to enhance students' ability to solve scientific problems [7]. Regarding interaction patterns, students with higher levels of AI literacy often exhibit more strategic and collaborative engagement with AI tools, leading to higher-quality outcomes [6,28]. In academic writing, GenAI provides support for idea generation and drafting, enabling students to iteratively refine their work through continuous interaction with AI [29,30]. These findings suggest that AI has deeply permeated and transformed multiple aspects of the learning process. However, some scholars have also noted potential risks: When faced with AI's hallucinations, biases, and opacity, students may develop overreliance and experience diminished critical thinking skills [31]. Overall, despite these concerns, there is a broad consensus that AI exerts a profound influence on learning.

In contrast to the rapid development of SAI research, studies examining how teachers collaborate with AI remain relatively limited. A small number of empirical studies have begun to explore how preservice teachers use AI to support ID. For example, Lee and Zhai investigated 29 preservice science teachers' use of ChatGPT in curriculum design. Their findings revealed that preservice teachers were able to integrate ChatGPT across multiple disciplines and employ diverse instructional methods [32]. Nevertheless, most of the existing research remains focused on technical integration, with insufficient attention paid to the interactional processes and the quality of ID outcomes.

Preservice physics teachers occupy a dual identity as both learners and future educators. As learners, they rely on AI to deepen their understanding of physics concepts and reflect on their own learning processes; as future teachers, they must integrate disciplinary knowledge and pedagogical strategies within their ID. While prior research has primarily centered on SAI, little is known about how this particular group utilizes GenAI in their dual roles.

Addressing this research gap, the present study systematically examines the roles and limitations of GenAI in supporting preservice physics teachers' ID, thereby uncovering the unique value GenAI may contribute to teacher education from a human-AI collaboration perspective.

III. METHODOLOGY

A. Participants and context

The participants were 220 second-year preservice physics teachers (135 males and 85 females, aged 20–22) from a comprehensive normal university in eastern China. They were enrolled in a compulsory course, *Teaching Design and Implementation in Secondary Physics*, which aims to cultivate basic competence in ID in alignment with national curriculum standards. By the middle of the semester, the students had been introduced to fundamental concepts of ID, including the formulation of learning objectives, the identification of instructional challenges, and the structuring of teaching procedures. However, their experience with GenAI tools such as DeepSeek was limited, and a significant proportion had never used such tools before.

As part of the course requirements, each student was assigned an individual formative task: to design a complete 45-min high school physics lesson on Newton's third law. This ID activity was carried out during class time in a structured and supervised environment.

B. Data collection

This study was approved by the university's institutional review board and conducted in accordance with ethical guidelines. All participants were informed of the purpose of the study and provided consent.

A quasiexperimental design was employed to examine the impact of DeepSeek support on the ID quality of preservice physics teachers. The participants were two intact classes of second-year preservice physics teachers ($N = 220$), assigned to the experimental group ($N = 110$) and the control group ($N = 110$). Although the two cohorts shared the same grade level, major, and course background, to further control potential differences, participants' grades in *Physics Teaching Theory* from the previous semester were collected as pretest scores and treated as covariates in subsequent analyses.

Before the task, the experimental group received a 10-min introduction to the DeepSeek platform, accessed via its web-based interface. Participants interacted with the

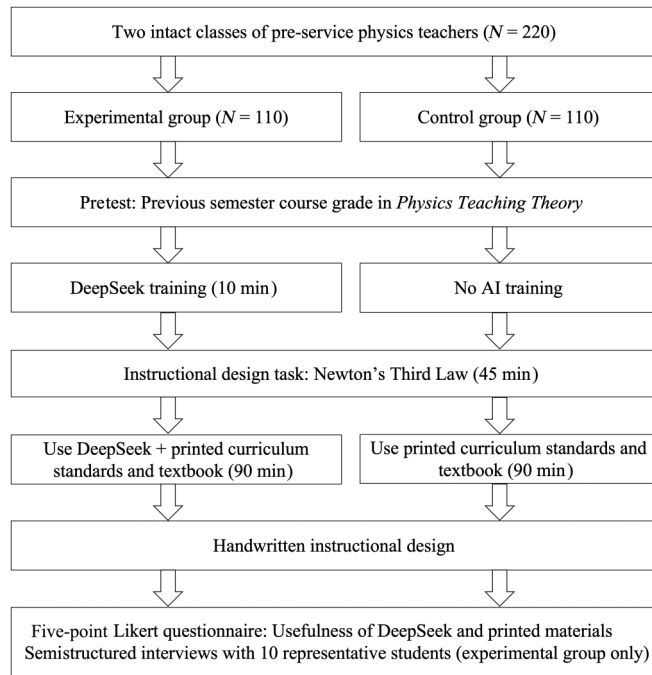


FIG. 1. Research design and data collection flow.

system using a standard browser environment. No technical issues, such as system lag, interruptions, or access failures, were reported during the task. The introduction covered basic operation guidance and general principles for interacting with the system (e.g., clarifying task constraints, refining prompts iteratively, and critically evaluating AI-generated responses), rather than giving instructions to request specific ID components. The experimental group was allowed to use DeepSeek alongside the printed curriculum standards and textbook. The control group completed the same task using only the printed curriculum standards and textbook. Both groups independently produced a handwritten 45-min ID on Newton's third law within 90 min. All submissions were evaluated using a rubric developed with reference to systems approach to ID developed by Dick *et al.* [33] and adapted to the requirements of high school physics ID. In the present study, the evaluation rubric included the following dimensions:

teaching content analysis, learner analysis, instructional objectives, key and difficult points, teaching strategies and methods, instructional materials, teaching procedure, instructional process detailing, and blackboard design.

Three types of data were collected: (i) ID submissions, (ii) interaction logs from the experimental group, and (iii) semistructured interview transcripts from a subset of experimental group participants. In addition, the experimental group completed a five-point Likert questionnaire evaluating the usefulness of DeepSeek and printed materials. The overall procedure is illustrated in Fig. 1.

C. Data measurements

1. Data analysis

The data analysis was divided into three main components: evaluating the quality of lesson designs produced by preservice physics teachers with the support of DeepSeek, analyzing the interaction patterns between preservice teachers and DeepSeek, and investigating preservice teachers' perceptions of DeepSeek-assisted lesson design. The specific data analysis methods are summarized in Table I.

2. ID evaluation

The ID rubric used in this study comprised nine dimensions, covering key aspects of ID. Each dimension was scored holistically; that is, raters assigned a single overall score for each dimension based on qualitative performance descriptors and predefined evaluative criteria. Except for the instructional process detailing dimension, which was assigned a maximum of 20 points, all other dimensions (teaching content analysis, learner analysis, instructional objectives, key and difficult points, teaching strategies and methods, instructional materials, teaching procedure, and blackboard design) were allocated 10 points each, resulting in a total maximum score of 100. The rubric was developed based on the systems approach to ID developed by Dick *et al.* to ensure alignment with the systematic requirements of ID. The instructional process detailing dimension was given a higher weight because it plays a central role in overall ID, reflecting the completeness, coherence, and operability of classroom procedures

TABLE I. Data analysis method.

Data type	Data	Analytical methods	Analytical purpose
Performance data	Final ID scores (9 dimensions + total score)	Descriptive statistics; independent samples <i>t</i> test; ANCOVA	To compare ID quality between experimental and control groups (RQ1)
Log data	Interaction records with DeepSeek	Content coding; ENA	To analyze patterns and types of AI-assisted ID interactions (RQ2)
Interviews and questionnaires	Semistructured interview transcripts; evaluation questionnaires	Thematic analysis; frequency analysis	To explore preservice teachers' perceptions, experiences, and cognitive to AI support (RQ3)

and thus contributes more substantially to the evaluation of instructional quality.

Before the formal scoring process, a calibration session was conducted to ensure scoring consistency. Two raters independently evaluated a set of five sample IDs, selected to represent a range of quality levels. A third expert moderator reviewed the results and facilitated discussion to resolve discrepancies. Through this process, a shared understanding of the rubric criteria was developed and agreed upon, providing the foundation for consistent scoring during the main evaluation phase.

After calibration, each rater independently assessed all IDs. The raters were blinded to group assignments to minimize bias. Scores were assigned on a dimension-by-dimension basis, and the final score for each ID was calculated as the sum of individual dimension scores (maximum = 100). In cases of scoring discrepancies larger than three points on any dimension, the two raters discussed the case and reached consensus under the guidance of the third expert. However, such cases were rare.

To assess interrater consistency, the intraclass correlation coefficient was calculated using a two-way random-effects model with absolute agreement. This approach evaluates the reliability of the average ratings across multiple raters, thereby providing an assessment of the consistency of the scoring process.

Following the reliability analysis, independent samples *t* tests were conducted to compare the experimental and control groups across the nine rubric dimensions as well as the total score, in order to examine the impact of GenAI support on ID quality. In addition, an analysis of covariance (ANCOVA) was conducted using the pretest *physics teaching theory* scores as a covariate.

3. Interaction content coding

The data for this study primarily consisted of text-based interaction logs generated by preservice teachers during their exchanges with the GenAI system—that is, their prompts and corresponding instructions. In addition, the final ID texts submitted by the participants were used as Supplemental Material [34]. The purpose of the data analysis was to explore the interaction patterns between preservice teachers and GenAI during the process of ID writing.

To achieve this, a coding framework was developed and adapted based on the studies of Kim *et al.* [5,6], while also taking into account the specific characteristics of ID tasks. As shown in Table II, the framework comprised three main categories: instructional design activities, metacognitive activities, and socioemotional activities.

During the data preparation stage, all interaction texts were collected and organized using Excel. Two independent coders conducted a complete coding of the data, covering every instance of interaction between preservice teachers and GenAI throughout the ID task. Following the

principle of complete coding, each interaction entry was divided into semantic units and assigned the most appropriate code label to accurately represent the participant's behavior and intent.

To ensure the reliability of coding, intercoder agreement was assessed using Cohen's Kappa. The result showed a high level of consistency (Kappa = 0.749, SE = 0.029, $p < 0.001$, $N = 271$), indicating substantial agreement between coders. Any remaining discrepancies were discussed and resolved through consensus, ensuring the validity of the final coding scheme.

4. Thematic interview analysis

After completing the ID task, students in the experimental group were first asked to evaluate the usefulness of both the printed materials and DeepSeek during the ID process, using a five-point Likert scale. The researchers then conducted statistical analyses of the questionnaire data. On the following day, a pool of representative cases was first identified based on the characteristics and quality of the instructional design outcomes, from which ten students were randomly selected for semistructured interviews. The interview recordings were transcribed verbatim. Thematic analysis of the interview data was conducted following Braun and Clarke's six-step framework [35]. This study adopted a research question-driven deductive thematic analysis, as both the interview questions and research objectives had been established prior to data collection. The detailed steps were as follows:

Familiarization with data: Two researchers repeatedly read the interview transcripts of the ten pre-service teachers and made preliminary annotations.

Generating initial codes: The researchers independently identified and labeled statements in the text that were relevant to the research questions, thereby forming initial codes.

Searching for themes: The two researchers compared and discussed the codes, grouping similar ones together to form preliminary themes.

Reviewing themes and codes: The research team conducted multiple rounds of discussion and revision to refine the themes and subcodes, ensuring conceptual clarity and close alignment with the research questions.

Defining final themes: All codes were aggregated into overarching themes.

Presenting results: Each theme was described narratively and supported by direct quotes from the participants, with the frequency of each subcode's occurrence among the ten interviewees also reported.

IV. RESULTS

Data collection was completed with some invalid responses being excluded. After data cleaning, the final

TABLE II. Coding scheme for AI-user interaction process.

Category	Subcategory	Code	Definition	Example
Instructional design activity	Problem representation	PR1	Understanding the instructional task or problem based on prior experience or subject knowledge.	What preconceptions do students have about this topic?
		PR2	Focusing on the core of the task, clarifying boundaries, key requirements, and critical points.	I need to design the lesson according to the process shown in the picture.
	Solution generation	SG	Proposing preliminary ideas for instructional steps, strategies, or activity designs.	I plan to use a short video as the introduction.
	Interaction with AI	IA1	Requesting AI to provide an initial plan or example.	Please generate an ID for me.
		IA2	Directly adopting AI-generated output without modification.	I'll just use this one.
		IA3	Modifying or refining prompts to improve AI output.	Please help me improve the key points of instruction.
		IA4	Critiquing or questioning AI-generated content.	Your learning objectives do not align with the core competencies.
		IA5	Asking AI to evaluate one's own ID or ideas.	Please evaluate my ID.
Metacognitive activity	Planning how to solve a problem	PH	Setting goals, conditions, and steps, and making adjustments as needed.	This lesson is for first-year students and should focus on core competencies.
	Monitoring and regulations during problem-solving process	MP	Checking and adjusting the ID process to ensure logical consistency.	Students have not yet studied Newton's first and second laws before this part.
	Evaluation	E	Providing an overall evaluation of the ID or AI output.	Your introduction is good, but the experiment design should not be like this.
Socioemotional activity	Positive emotion	PE	Expressing positive emotions when interacting with AI.	This idea is so helpful—it saved me a lot of time!
	Negative emotion	NE	Expressing negative emotions when interacting with AI.	You do not understand what I mean at all, stupid!

valid sample included 100 IDs from the experimental group and 107 IDs from the control group.

A. Quantitative analysis of ID quality

This section addresses RQ1 by examining whether there are significant differences in ID quality between preservice teachers with and without DeepSeek support. To ensure the reliability of the scoring process, an interrater consistency analysis was conducted. Table III presents the ICC values for each rubric dimension.

The average-measures ICC values across the rubric dimensions ranged from 0.61 to 0.96, all reaching statistical significance ($p < 0.001$). The overall ICC was 0.91 for the experimental group and 0.90 for the control group, both indicating excellent reliability [36]. Although the dimensions of learner analysis and teaching strategies and methods showed relatively lower

consistency, their ICC values still exceeded the commonly accepted threshold of 0.50, suggesting acceptable reliability. Overall, these findings demonstrate that the scoring process achieved a high level of interrater consistency, providing a solid foundation for the subsequent statistical analyses.

Based on the established reliability, comparisons of pretest and post-test results were conducted to examine the effects of GenAI support on ID quality (see Table IV).

For the pretest, the control group significantly outperformed the experimental group [$t(205) = -4.00$, $p < 0.001$, $d = 0.56$], indicating a moderate effect size [37]. However, this pattern was reversed in the post-test. The experimental group significantly outperformed the control group [$t(205) = 4.46$, $p < 0.001$, $d = 0.62$], reflecting a moderate effect size. The experimental group demonstrated a mean gain of 7.56 points, compared with

TABLE III. Interrater reliability of ID scoring.

Dimension	ICC		<i>p</i>	
	Experimental	Control	Experimental	Control
Teaching content analysis	0.74	0.71	<0.001	<0.001
Learner analysis	0.64	0.73	<0.001	<0.001
Instructional objectives	0.90	0.89	<0.001	<0.001
Key and difficult points	0.70	0.71	<0.001	<0.001
Teaching strategies and methods	0.71	0.61	<0.001	<0.001
Instructional materials	0.90	0.89	<0.001	<0.001
Teaching procedure	0.93	0.80	<0.001	<0.001
Instructional process detailing	0.90	0.96	<0.001	<0.001
Blackboard design	0.94	0.75	<0.001	<0.001
Total	0.91	0.90	<0.001	<0.001

TABLE IV. Independent-samples *t* test results comparing pretest and post-test ID scores between the experimental and control groups.

Test	Group	<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>p</i>	<i>d</i>
Pretest	Experimental	100	76.13	5.93	−4.00	<0.001	0.56
	Control	107	79.67	6.76			
Post-test	Experimental	100	83.69	6.25	4.46	<0.001	0.62
	Control	107	80.02	5.58			

0.35 points in the control group. These findings suggest that, despite initial disadvantages, the experimental group demonstrated greater gains in ID quality after receiving GenAI support.

TABLE V. ANCOVA results with pretest as covariate.

Source	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	Partial η^2
Pretest (cov)	1204	1920.13	74.64	< 0.001	0.27
Group	1204	1384.61	53.83	< 0.001	0.21

To further account for potential baseline differences, an ANCOVA was conducted using the pretest score as a covariate. Table V presents the results.

A significant main effect of group was found [$F(1, 204) = 53.83$, $p < 0.001$, partial $\eta^2 = 0.21$]. This effect corresponds to a large effect size [38], confirming that the experimental group significantly outperformed the control group in ID quality even after controlling for baseline differences.

Independent-samples *t* tests were conducted to compare the experimental and control groups across the dimensions of ID quality (see Table VI).

TABLE VI. Group comparison of ID quality (independent samples *t*-test). Note: * indicates a statistically significant difference between groups ($p < 0.05$).

Dimension	Experimental		Control		<i>t</i>	<i>p</i>	<i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
Teaching content analysis*	8.01	1.08	7.26	0.70	5.97	<0.001	0.83
Learner analysis*	8.66	0.91	7.43	0.88	9.87	<0.001	1.37
Instructional objectives*	8.01	1.62	7.10	1.40	4.33	<0.001	0.60
Key and difficult points	9.06	0.38	8.96	0.63	1.34	0.183	0.19
Teaching strategies and methods	8.78	1.05	8.90	0.79	−0.95	0.346	0.13
Instructional materials*	9.00	1.89	8.52	1.24	2.14	0.034	0.30
Teaching procedure	8.33	2.17	7.94	1.55	1.46	0.146	0.20
Instructional process detailing*	16.52	1.63	15.67	2.05	3.28	0.001	0.46
Blackboard design*	7.35	1.94	8.26	1.09	−4.20	<0.001	0.58
Total*	83.69	6.25	80.02	5.58	4.46	<0.001	0.62

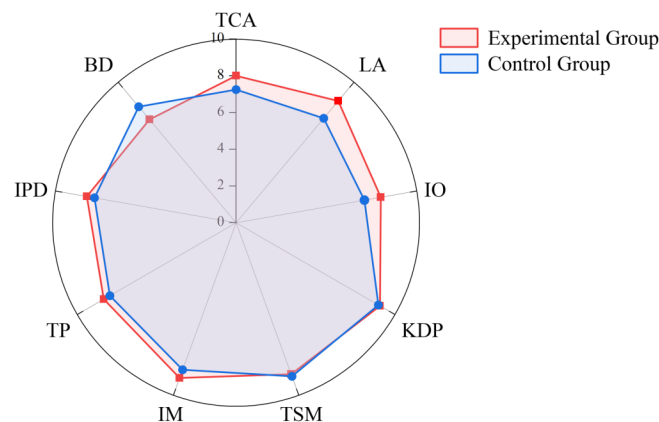


FIG. 2. Radar chart comparing ID quality across nine dimensions between the experimental and control groups. Note: TCA, teaching content analysis; LA, learner analysis; IO, instructional objectives; KDP, key and difficult points; TSM, teaching strategies and methods; IM, instructional materials; TP, teaching procedure; IPD, instructional process detailing; BD, blackboard design.

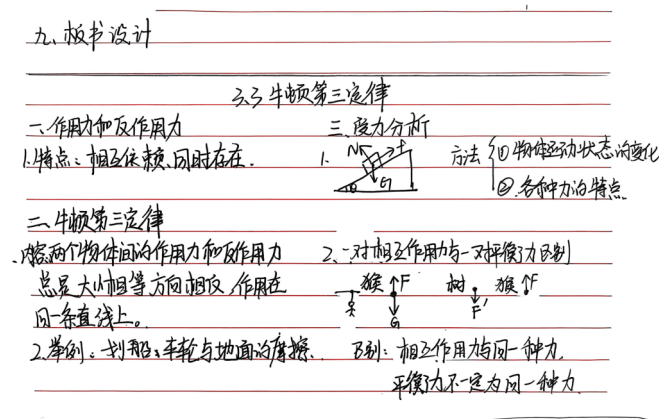


FIG. 3. Example of blackboard design from the control group.

The comparison of dimensions showed that the experimental group scored significantly higher than the control group in teaching content analysis, learner analysis, instructional objectives, instructional materials, instructional process detailing, and total score, indicating statistically significant differences; whereas no significant differences were found between the two groups in key and difficult points, teaching procedure, and teaching strategies and methods. By contrast, the control group scored significantly higher than the experimental group in blackboard design ($t = -4.20$, $p < 0.001$).

To provide a visual comparison across all rubric dimensions, Fig. 2 presents a radar chart contrasting the experimental and control groups. Since the instructional process detailing dimension had a maximum score of 20, its values were standardized to a ten-point scale to ensure comparability with the other dimensions.

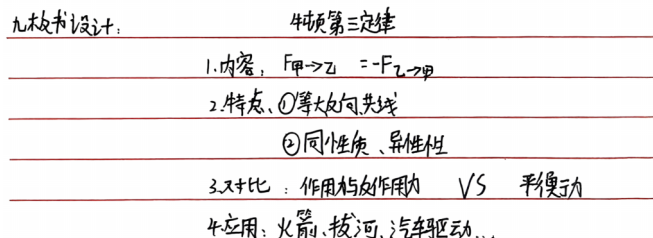


FIG. 4. Example of blackboard design from the experimental group.

During the evaluation process, it was also observed that the blackboard designs produced by the control group were generally more structured, detailed, and pedagogically coherent, while those of the experimental group appeared comparatively cursory and less carefully arranged. Representative blackboard comparisons are presented in Figs. 3 and 4. Overall, these results suggest that AI-assisted preservice teachers performed better on analytical and structural dimensions of ID, while visual presentation remains a potential area for further development.

B. Analysis of DeepSeek interaction processes

To answer RQ2, we analyzed the interaction patterns that emerged during DeepSeek-assisted instructional design, using both descriptive log statistics and ENA. To further explore the cognitive network differences among learners with different performance levels in the AI-assisted ID process, participants in the experimental group were divided according to their ID scores. First, the pretest scores were used as a covariate, and a regression analysis was conducted on the post-test ID scores to eliminate the influence of initial learning differences. The regression residuals—representing the deviation of post-test performance after controlling for pretest effects—were used as an indicator of performance. Participants were then ranked based on the residual scores and according to the quartile split method, the top quartile (top 25%, $N = 25$) was classified as the high group, and the bottom quartile (bottom 25%, $N = 25$) as the low group.

This residual-based grouping approach more accurately reflects learners' "actual improvement level" after the intervention and removes confounding factors such as pretest performance and individual baseline differences. It therefore allows for a more valid comparison of cognitive structure differences between learners of different levels when collaborating with AI.

Subsequently, the interaction records of both groups during the ID task were coded, and ENA was used to compare their cognitive co-occurrence structures. ENA quantifies the co-occurrence relationships among different codes and reveals between-group differences in cognitive structures through network centroids and edge weights.

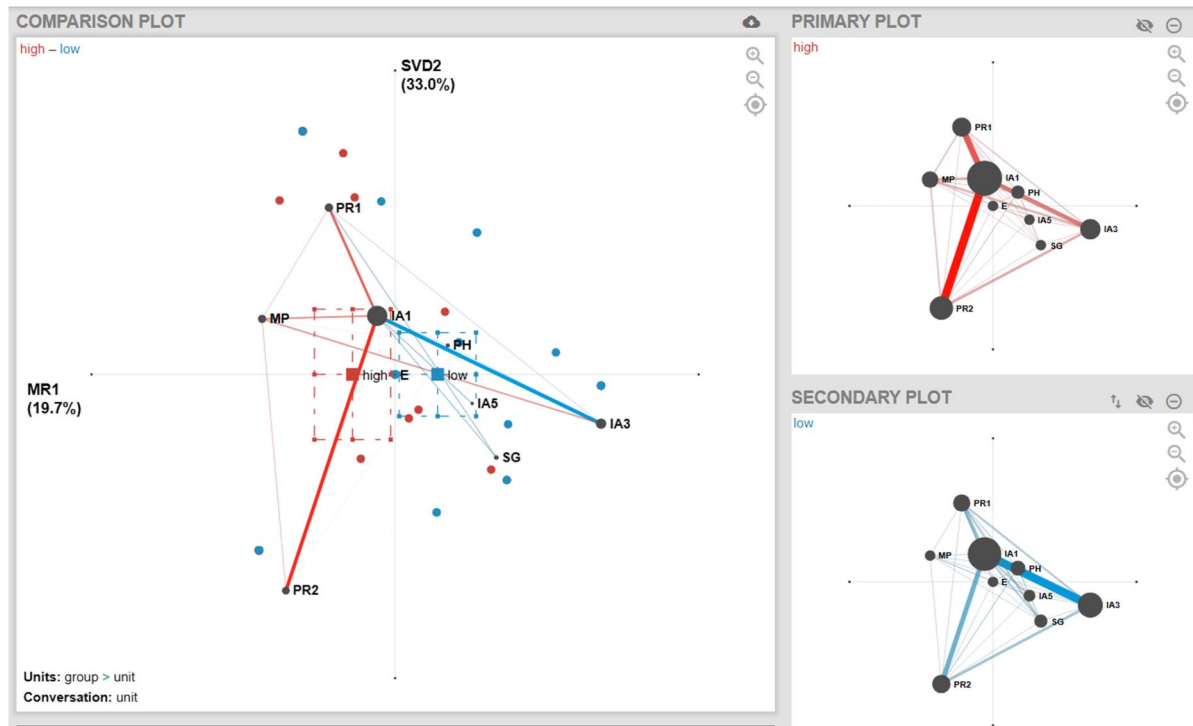


FIG. 5. ENA comparison of cognitive network structures between high- and low-performing learners.

Before conducting the ENA, basic descriptive statistics of AI usage were examined to provide contextual information for the comparison of interaction patterns. Participants in the experimental group issued an average of 2.74 prompts ($SD = 1.65$), with prompt counts ranging from 1 to 11 (median = 3). Further analysis showed that prompt frequency was not meaningfully associated with overall instructional design performance. Specifically, the high group issued an average of 2.96 prompts ($SD = 2.20$), while the low group issued an average of 2.61 prompts ($SD = 1.23$), indicating comparable levels of AI usage across groups. These results suggest that the group differences observed in the subsequent ENA are unlikely to be attributable to differences in the quantity of AI use, but rather reflect differences in interaction patterns and regulation strategies.

To examine the structural differences between the two groups in the ENA space, the centroid coordinates of the high and low groups in the unrotated ENA model were analyzed using the Mann-Whitney U test. The results indicated a significant difference between the two groups on the second SVD dimension ($U = 422.00$, $p = 0.03$, $r = -0.35$). To visualize the structural differences more clearly, a mean rotation was performed on the ENA model so that the horizontal axis aligned with the primary direction of difference between the two group centroids. Figure 5 presents the mean-rotated ENA comparison, which highlights the differences between the high and low groups in problem representation and AI interaction strategies.

As shown in the figure, the high group (upper right) exhibited dense connections among problem representation (PR), planning and monitoring (PH/MP), and AI interaction (IA), forming a tightly integrated PR-IA-MP pathway. In contrast, the low group (lower right) showed connections mainly within AI interaction (IA1–IA3), while the links to problem representation and monitoring were relatively weak. The activity distribution map on the left further demonstrates that the high group's activities were concentrated in problem analysis and regulation dimensions, whereas the low group focused more on AI generation and modification behaviors.

C. Qualitative analysis of experimental group's use of DeepSeek

RQ3 is addressed through questionnaire data and semi-structured interviews, which explore preservice teachers' perceptions of the usefulness and challenges of using DeepSeek in ID. To evaluate the perceived usefulness of the different resources, this study measured the participants' referencing frequency during the ID process. The underlying assumption is that resources perceived as more helpful would be consulted more often. The frequency was assessed using a five-point Likert scale for both DeepSeek and the printed materials, defined as: 0 = never, 1 = seldom, 2 = sometimes, 3 = often, 4 = always.

Results revealed that the students' ratings of the usefulness of the printed materials were mainly concentrated in the 1–2 range (64 out of 85, accounting for 75.3%),

TABLE VII. Themes, subcodes, and frequencies of occurrence from structured interviews ($N = 10$). Note: PT indicates preservice teacher.

Theme	Category	Direct quotes	Frequency
Conflicts between curriculum standards and textbooks	Inaccurate curriculum content provided	“The content requirements it mentioned were somewhat different from those in the actual curriculum standards.” (PT2)	3
	Outdated version of curriculum standards	“I forgot to check—it turned out to be generated based on the old ‘three-dimensional objectives’ version.” (PT5) “Some parts seem outdated and not aligned with current standards.” (PT6)	4
	Misalignment between generated content and textbooks	“DeepSeek referred to Newton’s Second Law as the second lesson, but in the current textbook, Newton’s Third Law comes first.” (PT10)	5
Language and logical issues	Overly absolute wording	“It said about 90% of students have this misconception—that’s too absolute.” (PT2)	1
	Incomplete expression	“Some sentences are hard to understand and a bit disorganized. The key idea of Newton’s Third Law—equal in magnitude and opposite in direction—wasn’t emphasized.” (PT7)	2
AI-supported insights and assistance	Learner analysis	“In the Learner Analysis part, it considered students’ preconceptions and cognitive characteristics.” (PT7)	3
	Contextual introduction and examples	“The tug-of-war example really engages students and exposes their preconceptions; the magnet interaction example was excellent.” (PT 1) “The introduction cases were quite relatable to everyday life.” (PT3)	4
	Experiment design	“Having two students pull directly makes it more hands-on; using sensors allows real-time data display.” (PT1) “The experiments were simple, easy to observe, and helped students become more involved.” (PT3)	4
	Question chains and analysis	“The question chain generated by DeepSeek matched students’ cognitive structures well.” (PT1)	2
	Teaching procedure	“The Teaching Procedure part was the most helpful.” (PT6)	2
Ways of using AI	Direct adoption	“I basically just copied it.” (PT6)	1
	Modified adoption	“I made some changes because its experiment design was incomplete.” (PT3)	3
	Repeated prompting required	“The main challenge was that I had to repeatedly adjust the prompts before it got accurate.” (PT4)	3
Future intentions and expectations	Willingness to continue use	“I will keep using it.” (PT 5, 6, and 9)	6
	Expectations for improvement	“I hope it can align better with the latest version of the textbook.” (PT3) “I woin’t rely on it too much—it should be combined with the textbook and curriculum standards.” (PT7)	2

whereas their ratings of DeepSeek were primarily concentrated at 3 and above (69 out of 85, accounting for 81.2%). A comparison of means showed that students rated DeepSeek significantly higher ($M = 2.91$) than the printed materials ($M = 2.04$), indicating that they generally

perceived DeepSeek as more helpful in the ID process, see Fig. 6.

For the thematic analysis of the semistructured interviews, five themes were ultimately identified, capturing the complex experiences of all preservice teachers during the

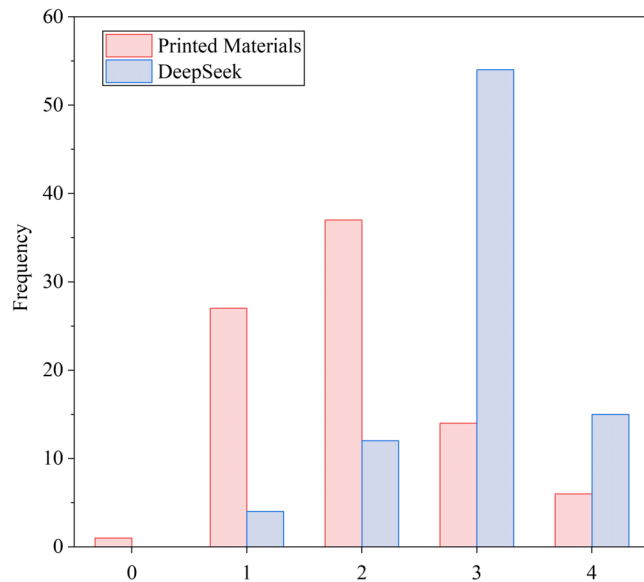


FIG. 6. Bar chart of the experimental group's evaluation of the usefulness of printed materials and DeepSeek.

AI-assisted ID process with DeepSeek. Table VII presents the themes, subcodes, and their frequencies of occurrence.

Results indicated that DeepSeek exhibited clear limitations in maintaining consistency with the current curriculum standards and textbooks. Its generated content primarily relied on the 2001 edition of the national curriculum standards, which led to discrepancies with the latest textbooks and even occasional confusion in instructional logic and sequence. Analysis of interaction logs and final ID drafts showed that 29% of preservice teachers failed to recognize the outdated nature of the instructional objectives and directly adopted the “three-dimensional objectives” generated by DeepSeek.

Interestingly, in the interviews, one preservice teacher who had adopted the “three-dimensional objectives” in his ID was asked about his understanding of the instructional objective structure. He responded, “It should be designed according to the core competencies.” This finding reflects a fundamental characteristic of large language models: Because they are trained on large-scale datasets in which materials related to core competency-based ID are relatively limited, the model is more likely to generate teaching objectives framed around the “three-dimensional objectives.” Therefore, users must possess critical thinking and professional educational judgment rather than relying entirely on AI-generated content.

Notably, some preservice teachers were aware that the “three-dimensional objectives” no longer align with the current curriculum standards. During their ID process, they actively monitored and adjusted the AI output, explicitly prompting DeepSeek to focus on core competency-based objectives to enhance the scientific rigor and contemporary relevance of their designs.

At the same time, DeepSeek's language expression displayed issues such as overly absolute wording and incomplete statements, making its output insufficiently precise for pedagogical discourse. However, the interview results also revealed that preservice teachers generally acknowledged DeepSeek's inspirational value in aspects such as contextual introduction, experiment design, and learner analysis, which aligns with previous findings on the “creative support” function of large language models in educational contexts [39,40].

It is worth noting that preservice teachers demonstrated diverse usage patterns: Some admitted to directly copying the generated content, others adopted it after modifications, and several emphasized that repeated prompting was necessary to obtain usable results. This variation highlights the crucial role of users' digital literacy in effectively applying large language models in educational practice. Overall, most preservice teachers expressed a willingness to continue using DeepSeek while simultaneously expecting it to align more closely with the latest textbooks and curriculum standards, and they emphasized that it should not replace printed materials or teachers' professional judgment.

V. DISCUSSION

A. Quantitative findings on AI-assisted instructional design

This study found that GenAI played a significant supportive role in preservice physics teachers' ID. In particular, the experimental group outperformed the control group in dimensions such as teaching content analysis, learner analysis and instructional objectives, indicating that GenAI effectively supported preservice teachers in achieving higher-level performance in the analytical and structural aspects of ID. AI assistance helped learners develop a more systematic understanding of textbook structure, make more accurate judgments about learner characteristics, and set appropriate instructional goals accordingly. This advantage primarily stems from large language models' capacity for knowledge integration and structural generation. For example, DeepSeek, trained on large-scale educational corpora, was able to generate logically coherent and well-structured instructional frameworks and sample cases. These outputs provided structural scaffolding for content analysis and goal setting, partially compensating for preservice teachers' limited practical experience. However, in the dimension of blackboard design, the experimental group performed significantly worse than the control group, indicating that the support provided by GenAI did not yield the expected benefits in this area and may even have impaired students' performance. This pattern may be attributed to the combined effects of a modality mismatch and changes in cognitive resource allocation under AI-assisted conditions. Specifically, DeepSeek primarily provides text-based

support, which aligns well with analytical and verbal aspects of instructional design, but is less compatible with tasks that rely on visual-spatial representation and layout planning, such as blackboard design. In addition, the availability of AI support may have altered how some participants allocated their time and cognitive effort. Under AI-assisted conditions, students in the experimental group may have focused more on text-oriented tasks. As a result, they may have devoted less autonomous effort to visual presentation tasks such as blackboard design within the limited time available. From this perspective, the observed decline in blackboard design performance may reflect a form of cognitive offloading. In this case, reliance on AI support may have reduced some learners' willingness to invest effort in self-generated spatial design, rather than simply indicating a lack of AI support for this dimension.

In addition, no significant differences were found between the experimental and control groups in key and difficult points, teaching procedure, and teaching strategies and methods. These null findings suggest that the contribution of GenAI may be limited in domains that rely more heavily on pedagogical judgment and contextualized instructional decision making.

B. Cognitive processes and learner differences: ENA and interview evidence

Further ENA revealed cognitive differences among learners of varying performance levels in the AI-assisted ID process. The high-performing group's cognitive network was centered on the "PR-IA-MP" structure, reflecting proactive and well-controlled strategic offloading. They tended to engage AI purposefully after clarifying task goals and formed cyclical regulation through planning and monitoring. Such strategic off-loading may reflect a relatively effective mode of cognitive regulation and may help learners allocate more cognitive resources to higher-level integration and optimization. In contrast, the low-performing group exhibited reactive off-loading, frequently querying AI without sufficient task representation and lacking both early analysis and later monitoring. Although this approach may reduce operational burden in the short term, it was associated with insufficient construction of internal representations and less evident metacognitive regulation, making it difficult for learners to form a pathway of continuous improvement [41,42]. This finding is consistent with the results of Kim *et al.* on skill-level differences, suggesting that GenAI does not automatically enhance learning outcomes—its educational value depends on learners' degree of active regulation [6].

The interview results further confirmed this pattern. Most students reported that AI substantially improved efficiency and systematicity in content integration and case provision, but they also noted that the generated content sometimes lacked contextual relevance, disciplinary accuracy, and alignment with curriculum standards, requiring

human judgment and revision. Some students mentioned that while AI-generated suggestions were valuable references, users still needed professional sensitivity and critical thinking to avoid dependency.

C. Implications and limitations

Integrating the quantitative and qualitative findings, it can be concluded that the role of GenAI in teacher education should be regarded as a "cognitive scaffold" rather than a "replacement tool." AI can assist preservice teachers in knowledge integration and structural generation, helping them form more systematic instructional plans in the early stages of design. However, teachers still need adequate disciplinary awareness and professional judgment to evaluate and refine AI outputs. Without corresponding metacognitive and critical thinking skills, AI use may devolve into dependency, weakening learners' capacity for autonomous construction. Therefore, teacher education programs should cultivate AI literacy while strengthening students' metacognitive training in task planning, process monitoring, and outcome evaluation.

The sample used in this study consisted of physics majors from a single normal university, which may limit the generalizability of the findings across different subjects and institutional contexts. In addition, this study adopted a quasiexperimental design using intact classes rather than random assignment. Although pretest scores were controlled, potential baseline differences between groups cannot be fully ruled out. Moreover, the interview sample was relatively small ($N = 10$), and the qualitative findings should therefore be interpreted as illustrative rather than representative. Finally, the ID task employed in this study was a simulated design task rather than an actual classroom implementation. As a result, the findings primarily reflect cognitive and interactional processes at the design level. Future studies could incorporate teaching practice, classroom behavioral tracking, or multimodal data (e.g., eye-tracking and think-aloud protocols) to further examine AI-assisted ID in authentic educational settings. In addition, more fine-grained research designs could be adopted to better align subjective evaluations with behavioral indicators, thereby enabling examination of the correspondence or divergence between perceived AI use and actual usage patterns.

VI. CONCLUSION

This study employed a quasiexperimental design and ENA to investigate the effects of GenAI on preservice physics teachers' ID performance and cognitive processes. The results showed that AI played a significant role in supporting the systematization and structuring of ID, but its performance was limited in tasks involving visual representation and spatial organization. High-performing learners were able to achieve effective HAI through planned and

reflective engagement, whereas low-performing learners tended to passively rely on AI outputs, which hindered deeper learning and optimization.

Overall, GenAI can serve as an effective tool for promoting analytical thinking and structured ID in teacher education; however, our findings suggest that its educational value is more likely to be realized when learners engage in planned, monitored, and reflective HAI, rather than passively adopting AI-generated outputs. Therefore, future teacher education should not only cultivate AI literacy but also emphasize the development of preservice teachers' proactive use of AI and critical judgment, enabling them to achieve professional growth through HAI. This study provides empirical evidence for understanding the cognitive support mechanisms of GenAI in teacher education and offers practical insights for improving university-based teacher training and ID courses.

ACKNOWLEDGMENTS

The authors would like to acknowledge the help of the Editor and anonymous reviewers. The research is supported by the 2025 Anhui Provincial Philosophy and Social Sciences Planning Project, "Mechanisms of Human-AI Collaboration in Promoting Deep Learning in Secondary School Science Education" (Grant No. AHSKQ2025D69).

Any opinions, findings, and 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 study are not publicly available due to privacy and ethical restrictions, but may be made available from the corresponding author upon reasonable request.

-
- [1] M. G. de Souza, M. Won, D. Treagust, and A. Serrano, Visualising relativity: Assessing high school students' understanding of complex physics concepts through AI-generated images, *Phys. Educ.* **59**, 025018 (2024).
 - [2] Y. Ben-Zion, R. E. Zarzecki, J. Glazer, and N. D. Finkelstein, Leveraging AI for rapid generation of physics simulations in education: Building your own virtual lab, *Phys. Teach.* **63**, 424 (2025).
 - [3] E. Kasneci, S. Krebs, C. Spohn, F. Dorgerloh, A.-S. Mestres, and C. Schulz, ChatGPT for good? On opportunities and challenges of large language models for education, *Learn. Individ. Differ.* **103**, 102274 (2023).
 - [4] M. I. Baig and E. Yadegaridehkordi, ChatGPT in the higher education: A systematic literature review and research challenges, *Int. J. Educ. Res.* **127**, 102411 (2024).
 - [5] J. Kim, S.-S. Lee, R. Detrick, J. Wang, and N. Li, Students-Generative AI interaction patterns and its impact on academic writing, *J. Comput. Higher Educ.* **38**, 504 (2025).
 - [6] J. Kim, Y. Ham, and S.-S. Lee, Differences in student-AI interaction process on a drawing task: Focusing on students' attitude towards AI and the level of drawing skills, *Aust. J. Educ. Technol.* **40**, 19 (2024).
 - [7] D. Tong, B. Jin, Y. Tao, H. Ren, A. Y. M. Atiquil Islam, and L. Bao, Exploring the role of human-AI collaboration in solving scientific problems, *Phys. Rev. Phys. Educ. Res.* **21**, 010149 (2025).
 - [8] S. Halawa, T. C. Lin, and Y. S. Hsu, Exploring instructional design in K-12 STEM education: A systematic literature review, *Int. J. STEM Educ.* **11**, 43 (2024).
 - [9] 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).
 - [10] C. Kulgemeyer, A. Borowski, D. Buschhüter, P. Enkrott, M. Kempin, P. Reinhold, and C. Vogelsang, Professional knowledge affects action-related skills: The development of preservice physics teachers' explaining skills during a field experience, *J. Res. Sci. Teach.* **57**, 1554 (2020).
 - [11] G. Resbiantoro, R. Setiani, and Dwikoranto, A review of misconception in physics: The diagnosis, causes, and remediation, *J. Turk. Sci. Educ.* **19**, 403 (2022), <https://mail.tused.org/index.php/tused/article/view/924>.
 - [12] S. Soeharto and B. Csapó, Evaluating item difficulty patterns for assessing student misconceptions in science across physics, chemistry, and biology concepts, *Heliyon* **7**, e08352 (2021).
 - [13] W. Dick, A history of instructional design and its impact on educational psychology, in *Historical Foundations of Educational Psychology* (Plenum Press, Boston, 1987), p. 183.
 - [14] R. A. Reiser and J. V. Dempsey, *Trends and Issues in Instructional Design*, 2nd ed. (Pearson Education, Upper Saddle River, NJ, 2007), p. 11.
 - [15] L. S. Shulman, Those who understand: Knowledge growth in teaching, *Educ. Res.* **15**, 4 (1986).
 - [16] W. S. Yu, From "basic knowledge and skills" to three-dimensional goals to core literacy: Three stages of curriculum and teaching reform in China over 40 years of reform and opening up, *Curric. Teach. Mater. Method* **39**, 40 (2019).
 - [17] Y. Y. Guo, From three-dimensional curriculum objectives to core literacy in physics, *Phys. Teach.* **39**, 2 (2017), https://kns.cnki.net/kcms2/article/abstract?v=wGRGwcz8A0e7yGb-5Qypp4S1quHM6cxBUyoJ1hUGNwjau5UwoKltOu4zCNrI_yhPs71dKc6n9RqAqK313-hk_2SU7MrpDQ0VjC7OzL3wx6Gw4Q-6FtBSGbs0em9ppzTqFWuZU9

- YDvUG2kO91IToMMiwwL8a3EMUJ0AVxhngudMATfwUra3JqtW==&uniplatform=NZKPT&language=CH.
- [18] N. Li, J. Zhang, and Z. Tang, Analysis and implications of the current situation of middle school physics teaching design from the perspective of core literacy: Based on an analysis of physics teacher articles from 2016 to 2020, *Phys. Teach.* **42**, 63 (2021), https://kns.cnki.net/kcms2/article/abstract?v=wGRGwcz8A0cvwhnr2JUPj-i_lrgIqC5Jar6lXABZ6AQr5LKF_bK7ugGoSmzNLc1yvUDArAAUFaQlexJO_msYt2o6D9A1MEeFrtGkIpZ7W1Rvm9Z7Wbybu-DQy6l2ubJ5RWm3ALinTtyeEr4Uw2rvfll5tdGVp3-BKcBWv7HybkrAKf5Ak6RUQ==&uniplatform=NZKPT&language=CHS.
- [19] S. Kapon and A. Merzel, Content-specific pedagogical knowledge, practices, and beliefs underlying the design of physics lessons: A case study, *Phys. Rev. Phys. Educ. Res.* **15**, 010125 (2019).
- [20] 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).
- [21] 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).
- [22] T. Wang 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).
- [23] G. Kortemeyer, Toward AI grading of student problem solutions in introductory physics: A feasibility study, *Phys. Rev. Phys. Educ. Res.* **19**, 020163 (2023).
- [24] K. E. Avila, S. Steinert, S. Ruzika, J. Kuhn, and S. Küchemann, Using ChatGPT for teaching physics, *Phys. Teach.* **62**, 536 (2024).
- [25] D. T. T. Mai, C. V. Da, and N. V. Hanh, The use of ChatGPT in teaching and learning: A systematic review through SWOT analysis approach, *Front. Educ.* **9**, 1328769 (2024).
- [26] P. Tschisgale, P. Wulff, and M. Kubsch, Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory, *Phys. Rev. Phys. Educ. Res.* **19**, 020123 (2023).
- [27] T. Jiang, Z. Sun, S. Fu, and Y. Lv, Human-AI interaction research agenda: A user-centered perspective, *Data Inf. Manage.* **8**, 100078 (2024).
- [28] J. Kim, S. Yu, S. S. Lee, and R. Detrick, Students' prompt patterns and its effects in AI-assisted academic writing: Focusing on students' level of AI literacy, *J. Res. Technol. Educ.* **58**, 638 (2025).
- [29] A. R. Malik, Y. Pratiwi, K. Andajani, I. W. Numertayasa, S. Suharti, A. Darwis, and Marzuki, Exploring artificial intelligence in academic essay: Higher education students perspective, *Int. J. Educ. Res. Open.* **5**, 100296 (2023).
- [30] A. Nguyen, Y. Hong, B. Dang, and X. Huang Human-AI collaboration patterns in AI-assisted academic writing, *Stud. Higher Educ.* **49**, 847 (2024).
- [31] C. Zhai, S. Wibowo, and L. D. Li, The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review, *Smart Learn. Environ.* **11**, 28 (2024).
- [32] 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).
- [33] W. Dick, L. Carey, and J. O. Carey, *The Systematic Design of Instruction*, 7th ed. (Pearson, Upper Saddle river, NJ, 2008).
- [34] See Supplemental Material at <https://link.aps.org/supplemental/10.1103/bxql-gzfk> for details.
- [35] V. Braun and V. Clarke, Using thematic analysis in psychology, *Qual. Res. Psychol.* **3**, 77 (2006).
- [36] 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).
- [37] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Lawrence Erlbaum Associates, Hillsdale, NJ, 1988).
- [38] J. T. E. Richardson, Eta squared and partial eta squared as measures of effect size in educational research, *Educ. Res. Rev.* **6**, 135 (2011).
- [39] S. Zha, Y. Qiao, Q. Hu, Z. Li, J. Gong, and Y. Xu, Designing child-centric AI learning environments: Insights from LLM-enhanced creative project-based learning, *arXiv:2403.16159*.
- [40] G. Franceschelli and M. Musolesi, On the creativity of large language models, *AI Soc.* **40**, 3785 (2025).
- [41] E. F. Risko and S. J. Gilbert, Cognitive offloading, *Trends Cogn. Sci.* **20**, 676 (2016).
- [42] S. J. Gilbert, Strategic offloading of delayed intentions into the external environment, *Q. J. Exp. Psychol.* **68**, 971 (2015).