

Comparing teacher-centered and student-centered agents based on prompt engineering: Effects on learning performance, cognitive load, flow experience, and empathy perception in physics learning

Yimin Wang^{1,3}, Xiaotong Chen^{1,2}, Yushan Xiong^{1,2},
Shaorui Xu^{1,4,*}, Qiuye Li^{1,2,†}, and Shaona Zhou^{1,2,‡}

¹Guangdong Provincial Key Laboratory of Quantum Engineering and Quantum Materials,
National Demonstration Center for Experimental Physics Education, School of Physics,
South China Normal University, Guangzhou, People's Republic of China

²Guangdong Basic Research Center of Excellence for Structure and Fundamental Interactions of Matter,
Key Laboratory of Atomic and Subatomic Structure and Quantum Control (Ministry of Education),
South China Normal University, Guangzhou, People's Republic of China

³Guangdong University of Education, Guangzhou, People's Republic of China

⁴School of Electronics and Communication, Guangdong Mechanical & Electrical Polytechnic,
Guangzhou, Guangdong, People's Republic of China



(Received 30 October 2025; accepted 1 July 2026; published 9 September 2026)

Large language models (LLMs) often face challenges in generating accurate and coherent explanations when solving physics questions. Prompt engineering serves as an effective strategy for human-artificial intelligence communication, guiding LLMs to generate precise and contextually relevant responses with well-designed instructions. The study utilized prompt engineering on the Coze platform to construct two types of instructional agents: a teacher-centered agent and a student-centered agent. They were designed to respond to students' questions in accordance with their respective roles and communication styles. Specifically, the teacher-centered agent tended to provide authoritative, professionally grounded, and content-focused responses, whereas the student-centered agent was designed to be more attentive to students' prior knowledge and learning needs, offering adaptive and supportive guidance. The study compared the effects of these two agents on students' learning performance, cognitive load, flow experience, and empathy perception. The results revealed differential effects between the two prompt-engineered agent conditions. Compared with the teacher-centered agent, students interacting with the student-centered agent condition showed higher learning performance, lower extraneous cognitive load, stronger germane cognitive engagement, greater flow experience, and higher empathy perception.

DOI: [10.1103/9t5b-twsb](https://doi.org/10.1103/9t5b-twsb)

I. INTRODUCTION

Large language models (LLMs), which exhibit advanced capabilities in natural language processing and content generation, are becoming increasingly prevalent as educational tools [1] for their implementation in domains such as intelligent tutoring [2], personalized learning [3], and the development of instructional materials [4]. However, answering physics questions remains challenging for LLMs, as these tasks require not only conceptual

understanding but also evidence-based reasoning to support scientific explanations. Previous studies have found that LLMs frequently exhibit conceptual confusion, logical errors, and contradictory answers [5] when addressing physics questions in areas such as mechanics, electromagnetism, and thermodynamics [6,7]. Such logical errors and inconsistencies raise reliability concerns that may limit the use of LLMs in physics education.

Prompt engineering is an important tool for guiding LLM outputs. By providing carefully designed instructions, it can help LLMs generate responses that are more accurate and better aligned with users' intentions. Existing research has shown that prompt engineering can reduce hallucinations in LLMs and enhance their performance in interactions with learners [8–10]. Different prompts are also capable of shaping distinct roles and conversational styles. In physics educational contexts, LLMs have been shown to provide learning support to students by embodying various instructional roles [11,12]. However, role-playing prompts are

*Contact author: jayxee@hotmail.co.uk

†Contact author: 2019021855@m.scnu.edu.cn

‡Contact author: zhou.shaona@m.scnu.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.

implemented mainly as brief single-sentence instructions, asking the LLM to answer questions as a physics teacher, without a more explicit design of the role's characteristics, interaction patterns, and pedagogical intentions. More recent studies have also begun to explore more elaborate and systematically designed pedagogical prompting approaches in physics education [13,14]. Nevertheless, the effectiveness of different instructional roles still requires further validation.

This study uses prompt engineering to construct two distinct instructional agents: a teacher-centered agent that provides authoritative, structured, content-focused explanations, and a student-centered agent that is designed to be more attentive to students' prior knowledge and learning needs, thereby offering adaptive and supportive guidance. We will utilize the Coze platform, a tool for developing LLM agents without coding, to create agents capable of directly invoking LLMs for instructional interaction. The purpose of this study is to compare two distinct instructional agents (teacher-centered vs student-centered) and examine their effects on students' learning performance, cognitive load, flow experience, and empathy perception in a physics learning environment.

II. LITERATURE REVIEW

A. Applications of LLMs in physics education

In physics education, LLMs are widely applied to physics problem-solving, automated grading feedback, and the development of instructional tasks. For example, researchers have explored the potential of ChatGPT in physics learning. Kortemeyer [15] and West *et al.* [16] evaluated the performance of ChatGPT 3.5 and GPT-4 on the force concept inventory (FCI). Their findings showed that as model capabilities advanced, ChatGPT's accuracy in the performance of FCI improved from novice level to expert level. Liang's research emphasized the crucial role of structured physics prompts in facilitating LLMs' comprehension of physics questions [17]. He considerably improved the accuracy of the generated responses by using specific prompts to direct ChatGPT in identifying physics variables and connecting them to basic formulas. In the field of automated grading feedback, prompts can also guide LLMs toward humanlike evaluation standards [18]. Wan and Chen utilized few-shot learning methods to instruct ChatGPT in grading students' written responses, which not only improved the model's performance but also enabled the generation of personalized feedback [19]. Similarly, in the area of instructional tasks generation, Kuchemann showed that carefully designed prompts can enhance the capacity of LLMs to produce assignments that more closely align with instructional goals and disciplinary requirements [20]. While LLMs hold considerable promise in physics education, their effectiveness is highly dependent on input control. The application of prompt

engineering can substantially improve LLM performance on specific tasks.

B. The strategies of prompt engineering in LLMs

With the rapid advancement of artificial intelligence (AI), prompt engineering, which involves the design and optimization of prompts that provide specific instructions to LLMs [21], has emerged as a crucial field focused on developing effective strategies to efficiently leverage their potential [22].

Chain-of-thought (CoT) is an effective prompting strategy primarily used for complex tasks that require multistep reasoning [23]. The core mechanism of improving accuracy is to guide LLMs within a series of intermediate reasoning steps before generating the final answers. Existing research has shown that LLMs with CoT prompting significantly outperform those with general prompts [24]. For instance, Kojima *et al.* found that the accuracy of an LLM on multistep multiplication tasks increased from 17.7% to 78.7% by simply adding the phrase "Let's think step by step" to the prompts [25]. With the advancement of large language models, newer reasoning models have become capable of performing internal multistep reasoning before generating a final response. OpenAI's o1 series models, for example, are described in the official documentation as being particularly effective at path planning and strategy development for complex tasks, allocating additional internal reasoning before generating the answer [26]. As a result, explicit prompts such as "think step by step" are not always necessary to elicit reasoning itself. In educational contexts, when researchers expect reasoning models to generate multistep solutions in a pedagogically structured and stepwise manner, prompt design remains essential for specifying task requirements, interaction constraints, and expected output formats. For example, Cohn *et al.* designed a chain-of-thought prompting approach that provided the assessment question and rubric and required the model to produce evidence-based, step-by-step explanations when evaluating students' responses [27]. Among recent reasoning models, DeepSeek R1 has demonstrated strong performance on complex tasks such as code generation and mathematical problem-solving [28]. Unlike o1-style models, which generally do not expose their internal reasoning directly to users, DeepSeek R1 presents an explicit reasoning trace alongside its final response. This design may offer teachers greater visibility into the model's presented reasoning process, which could be beneficial for instructional explanation.

Role-playing is another strategy that has been proven to be effective in prompting the desired response [29]. It has been demonstrated that prompts have the capacity to direct LLMs to assume specific roles within predefined domains. This facilitates the generation of more professional, accurate, and contextually relevant responses [30]. For example, Zheng used prompt engineering to develop a

chemistry assistant [31] that guided ChatGPT to extract experimental conditions for metal-organic frameworks from literature. In physics education, Polverini and Gregorcic demonstrated that LLMs can be prompted to act as physics teachers by specifying the domain, prescribing how to act, and providing background information [30]. Their designed prompt stated, “A Nascar racer won the race in the last lap by grinding against the outer fence of the racetrack. Why did this trick work? Explain it like a physics teacher would from the perspective of forces in circular motion.” This study illustrates how prompt design can encourage LLMs to draw on domain-specific physics knowledge and provide structured explanations aligned with a physics teacher’s perspective.

C. AI agent based on prompt engineering

With well-designed prompts and development frameworks, LLMs can be shaped into a variety of agents. These agents can assume a variety of roles during the learning process, such as experts, teachers, and partners. For instance, Lyu *et al.* developed six AI assistants with distinct identities through advanced prompt engineering [32], demonstrating that role-playing had a substantial impact on students’ conversational styles and interaction patterns with the large language models. Different prompting strategies can significantly influence the role orientation and interaction patterns of LLMs. For instance, Xu *et al.* configured ChatGPT as an expert agent by explicitly specifying the expert’s identity, skill, and research domain within the prompt [33]. This guided the LLM to respond from the viewpoint of a distinguished expert, thus improving the accuracy of the responses. Conversely, some researchers have attempted to design LLMs as learners with difficulties, which requires them to continuously ask questions and prompt learners to respond. This enhances learners’ comprehension of knowledge during the process of teaching AI. For instance, the teachable agent ALTER-Math, designed by Xing *et al.*, facilitates students’ comprehension of mathematical concepts by offering explanations, clarifications, and corrections to the AI [34]. Schroeder proposes that agents should facilitate learning by using well-designed instructional strategies in educational research [35]. LLMs ought to adopt teaching-oriented prompts rather than providing answers directly, regardless of whether they are expert agents or teachable agents [36]. For instance, Ding *et al.* developed an agent that guides learners in problem-solving through a pedagogical framework consisting of four stages: review, heuristic, rectify, and summarize [37]. Kestin *et al.* developed an AI tutor through targeted, content-rich prompt engineering, integrating generative AI with pedagogical practices to promote learning in undergraduate science education [38]. Tufino and Gregorcic demonstrated a similar approach by building a Socratic physics problem-solving tutor using Google’s Gemini Gems and specifying its pedagogical

behavior via detailed instructions that define the tutor’s role, persona, and interaction rules [13]. Such teaching-oriented prompt designs are consistent with interactive classroom practices and have been advocated in physics education research to better leverage the educational value of LLMs [39].

Some LLM platforms provide tools for creating and customizing LLM-based agents, enabling researchers to configure agent roles, response styles, and interaction rules (e.g., Google’s Gemini Gems). Meanwhile, other development tools support connecting to different LLMs via API calls, allowing for more flexible workflow orchestration and extensibility. Overall, these practices can be broadly grouped into two technical categories. The first category involves platforms such as AutoGen [40] and LangChain [41], which provide robust capabilities for designing multiagent collaborative tasks but require substantial technical expertise. For instance, Ni *et al.* employed AutoGen to generate seven distinct agent roles, including scientists, engineers, executors, and administrators, that collaboratively solved a two-dimensional copper plate elasticity problem through division of labor and mutual verification [42]. The other category includes platforms such as Coze [43] and Chery Studio [44], which are more suitable for teachers without technical backgrounds [45]. For example, Fan *et al.* utilized Coze to develop a corporate agent that facilitated students’ reflective learning and effectively enhanced their metacognitive abilities [46].

D. Impact of LLMs on students’ learning processes

1. Cognitive load

Cognitive load theory was proposed by Sweller *et al.* It posits that the human working memory has a limited capacity for processing information at any given time. When the cognitive demands of learning tasks exceed this capacity, both learning efficiency and problem-solving performance decline significantly [47,48]. The theory categorizes cognitive load into three distinct types: (i) intrinsic cognitive load, (ii) extraneous cognitive load, and (iii) germane cognitive load. Intrinsic cognitive load is determined by the inherent complexity of the instructional material and the learner’s prior knowledge, reflecting the task’s intrinsic difficulty [49]. The phenomenon of extraneous cognitive load, which is caused by the presentation of information, can be mitigated by the provision of clear and well-explained instructions [50]. Germane cognitive load is defined as the effective cognitive effort learners invest in constructing meaningful knowledge representations and forming mental schemas [51]. This makes germane cognitive load the most valuable type of load for learning. Sweller emphasizes that effective instructional design should aim to reduce extraneous cognitive load while actively supporting and enhancing germane cognitive load through appropriate scaffolding and guidance, thereby optimizing overall learning outcomes.

Recent research has demonstrated the dual effect of LLMs on students' cognitive load in education contexts. On the one hand, LLMs can provide personalized instructional support by acting as virtual tutors and assisting students in answering questions [52]. Such support may help reduce learners' cognitive burden during learning [53]. For instance, ChatGPT can offer immediate feedback and customized learning assistance in domains such as language learning and writing support. This enables students to access information more efficiently and reduce the cognitive load associated with information processing and retrieval [54]. When students use ChatGPT to check complex mathematical problems, perceived access to reliable help may reduce uncertainty-related anxiety and encourage them to engage with more challenging tasks, knowing that assistance is available when needed [55]. In physics courses, AI-generated supplementary materials have been found to present conceptual knowledge in a more intuitive manner, helping students construct schemas, improve self-efficacy, and effectively reduce both intrinsic cognitive load and extraneous cognitive load [56].

On the other hand, LLMs also carry the risk of increasing cognitive load. Due to ChatGPT's tendency toward hallucination [57], it may generate inaccurate or fabricated content. As a result, learners may need to spend additional cognitive resources verifying information, identifying errors, and judging the reliability of the generated responses. This may increase both intrinsic cognitive load and extraneous cognitive load.

For instance, in the context of English as a foreign language (EFL) writing instruction, students with limited experience using AI tools were found to experience substantial cognitive load during their interactions with ChatGPT [58].

2. Flow experience

Flow theory, proposed by Csikszentmihalyi, refers to a psychological state characterized by deep immersion, sustained concentration, and intrinsic enjoyment, which occurs when individuals engage with challenges that are well aligned with their skill levels [59]. Within educational contexts, flow has emerged as a critical principle for examining students' learning experiences [60]. It has been demonstrated that students who are experiencing flow typically exhibit heightened concentration, enhanced creativity, and greater satisfaction [61]. Li *et al.* divided flow experience into three dimensions: enjoyment, engagement, and control [62]. In the context of interactions with LLMs, these dimensions are reflected as follows: enjoyment refers to the sense of pleasure and satisfaction that students experience during interaction with LLMs; engagement denotes the degree of concentration and immersion that students exhibit in the dialogue; and control reflects the sense of autonomy that learners experience when engaging

with LLMs, including asking questions, guiding the conversation, and regulating the pace of interaction.

Empirical evidence suggested that chatbots powered by LLMs possess significant potential to facilitate flow experience. For instance, Lademann found that students who used AI-customized chatbots to generate learning materials reported higher levels of positive emotions, such as enjoyment and satisfaction, alongside a significant increase in learning interest [56]. Similarly, Fahad demonstrated that these chatbots can significantly enhance students' intrinsic motivation by stimulating curiosity and sustained engagement, which improves learning performance [63]. Students generally reported that learning with chatbots was a positive experience, and many expressed a desire to continue incorporating such tools in future courses. Furthermore, Wang *et al.* showed that embedding LLMs into learning environments significantly enhances students' flow experiences, particularly in game-based learning, thereby fostering deeper cognitive and emotional involvement in the learning process [64].

3. Empathy perception

Empathy is defined as the ability to understand another person's thoughts and feelings from their perspective rather than one's own [65]. High-quality interactions between teachers and their students have been attributed to empathy, which is considered an essential quality in educators [66]. When teachers demonstrate understanding and acceptance of students' emotions in the classroom, they can significantly enhance students' engagement and performance [67]. It can be attributed to the empathic abilities exhibited by teachers and educational systems, such as the capacity to convey sympathy in response to students' positive and negative emotional expressions, which builds a sense of reliability and security among students [68]. This ability has been shown to boost students' learning motivation and to facilitate the establishment of an effective classroom management model.

With the advancement of artificial intelligence, researchers have begun to explore the empathetic capabilities of LLMs such as ChatGPT, Claude, and Gemini, aiming to transform these systems from informational tools into learning empathic companions. For instance, Zhi proposed a methodology called GuidedEmpathy [69], which classifies the speaker's context and then provides specific guidance to LLMs for generating more empathetic responses. Meng pointed out that constructing LLM agents with humanlike qualities can significantly enhance users' emotional resonance and engagement [70]. Learners are more likely to build trust when they interact with LLMs that exhibit empathic characteristics. For example, Bai observed that when responding to patients' online inquiries, ChatGPT expresses emotional awareness [71] through phrases such as "This is truly unacceptable." Such emotionally resonant responses are more frequently preferred

compared to those provided by medical professionals, as they are perceived to demonstrate a higher level of empathy. We contend that the perceived empathy of LLMs is a fundamental prerequisite for facilitating high-quality interactions. We believe it is essential to further explore the mechanisms of empathetic expression in LLMs when they are acting as teachers. It is worth exploring whether the optimization of prompt engineering could help LLMs to generate emotionally responsive responses.

III. RESEARCH QUESTION

The combined literature confirms that prompt engineering is essential for regulating LLM behavior. It can assume different agent roles to provide learners with guidance in educational contexts. Existing studies have demonstrated that when acting as expert agents based on prompt engineering, LLMs can effectively reduce students' cognitive load and cultivate positive emotional experiences. However, there is limited empirical evidence on other agent roles, such as student-centered agents based on prompt engineering. It is still unclear how different LLM-based agents influence students' learning performance, cognitive load, flow experience, and empathy perception during interactions. To address this gap, this study aims to examine the effects of different instructional agents, designed by prompt engineering, on students' learning performance, cognitive load, flow experience, and empathy perception. Specifically, this study focuses on comparing a teacher-centered agent and a student-centered agent in a physics learning context to clarify how role-playing prompting strategies influence the educational impact of LLMs. The findings are expected to provide empirical insights into the pedagogical potential of LLMs and enhance the design of more effective AI-supported instructional strategies. Based on the above rationale, this study addresses the following research questions:

RQ1: What are the differences in students' learning performance when interacting with the teacher-centered and student-centered agents in physics learning contexts?

RQ2: What are the differences in students' cognitive load, flow experience, and empathy perception when interacting with the teacher-centered and student-centered agents in physics learning contexts?

IV. METHODS

A. Participant

The study recruited 63 high school graduates from a province in southeastern China. All participants had recently completed the National College Entrance Examination and ranked within the top 20.4% of the provincial academic performance. The sample consisted of 20 male and 43 female students, all of whom had completed secondary physics education in accordance

with the national curriculum standards issued by the Ministry of Education (MOE). None of the participants had received any formal instruction related to AI agents or LLMs. They were explicitly informed that their participation was completely voluntary and would not affect their final grades in the course.

B. Materials

According to the compulsory physics textbook under the guidance of the MOE, students have learned that temperature can change the density of substances. The expansion and contraction in temperature of gases is the most significant among ordinary substances. It has been proven that as gases heat up, their density decreases while their volume increases. Conversely, when temperatures are lowered, their density increases while their volume decreases [72]. However, although students have the capacity to comprehend the knowledge, they tend to overlook the specific applications of the convection phenomenon caused by density differences in real situations. Consequently, this study proposed a question: "When two candles of different lengths are lit under a closed glass box, which candle will extinguish first?"

Before the formal experiment began, we distributed this question to 113 high school graduates who were comparable in quality to the formal research subjects. These students' accuracy percentage was determined to be merely 32.86%. This can be ascribed to the high school graduates' insufficient consideration of physics processes occurring in the closed and small glass box. As the candle burns continuously, oxygen is consumed, and carbon dioxide gradually accumulates. The heat generated by combustion causes warm air to rise, leading to an increased concentration of carbon dioxide above the flame, which results in the higher candle extinguishing before the lower one. Therefore, this question was used as one of the pretest questions in the formal research and presented to students as a multiple-choice question with four options. The detailed information can be found in Appendix A.

This study also designed a temperature reduction context to investigate how the direction of air conditioner airflow affects the consistency of temperature distribution in a sealed room [73]. This was also presented as a multiple-choice question with four options.

These two physics questions were administered in both the pretest and the post-test. Furthermore, during the intervention phase, these tasks served as instructional materials for students to engage with their assigned agents. The purpose of the study was to evaluate students' learning performance before and after interaction with the teacher-centered and student-centered agents.

C. Prompt design

This study utilizes the Coze platform to develop a teacher-centered agent and a student-centered agent.

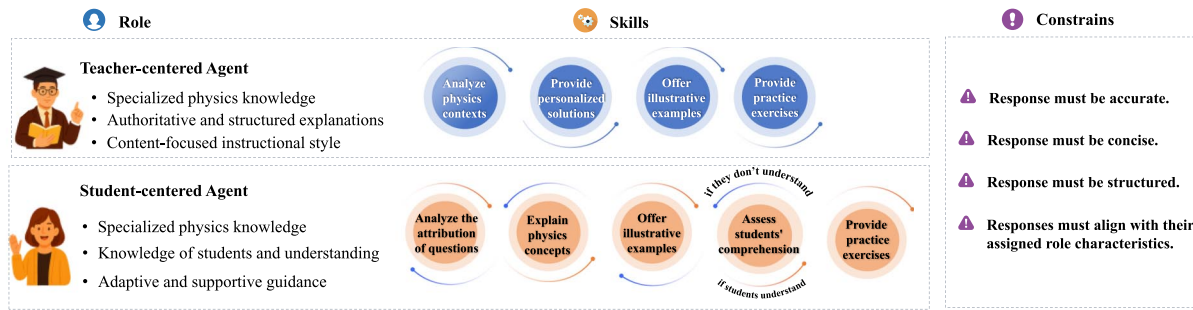


FIG 1. The design framework for prompt engineering.

Large language models are invoked to enable efficient human-machine conversation for conducting experiments. The prompts for both agents were crafted according to their respective roles, skills, and constraints and included chain-of-thought guidance. Although both agents are designed to act as knowledgeable physics teachers, they are guided by different prompt designs and adopt different interaction patterns when responding to students' questions, as illustrated in Fig. 1.

The teacher-centered agent is characterized as an experienced physics teacher with specialized knowledge of physics and extensive teaching expertise. Its capabilities include (1) analyzing physics question contexts, (2) providing personalized solutions, and (3) offering supplementary examples to enhance understanding. To avoid irrelevant responses, the following constraints were imposed: (1) All responses must be accurate and credible. The agent was designed to rely on a bounded knowledge source based on Ministry of Education-approved electronic textbooks. When a question goes beyond this knowledge source, the agent is prompted to explicitly acknowledge that the requested information falls outside its knowledge boundary; (2) physics concepts must be explained from an instructor's perspective; and (3) responses must be concise and structured, excluding extraneous content. These constraints were intended to reduce hallucination, but they could not guarantee the accuracy of the responses.

The student-centered agent is configured as a physics teacher with empathic capabilities. It demonstrates not only professional knowledge of physics but also pedagogical content knowledge (PCK), particularly knowledge of students' understanding (KSU), which is one of the crucial dimensions of PCK. KSU refers to the teacher's ability to diagnose students' comprehension levels and common misconceptions [74]. This capability facilitates the implementation of personalized instructional strategies. The student-centered agent's skills include (1) using the prompt "You have this question because..." to analyze the underlying causes of students' misconceptions; (2) using the sentence "This question involves..." to identify the relevant physics concepts; and (3) presenting a real-world situation that differs from the original

question and explaining it with the statement "This phenomenon is similar to..., also because..." Unlike the teacher-centered agent, the student-centered agent was designed to explain physics concepts from the students' perspective, prioritize interactivity, and maintain a friendly conversation style.

Since the question contexts in the experimental materials required logical reasoning, this study adopted the DeepSeek R1 model, which demonstrates superior reasoning performance in complex problem-solving and multistep analytical tasks, as the foundational support for both agents [75]. In addition to the prompts, we also controlled the temperature parameter. Temperature determines the degree of randomness in LLM text generation: higher values typically yield more diverse and creative outputs. To ensure comparable outputs across the two agents, we set the temperature to 0.3 for both agents to reduce randomness and promote logically coherent responses.

D. Instrument

The two questions in the materials were used as a knowledge test to assess students' conceptual understanding. In addition, a set of questionnaires was administered to measure students' cognitive load, flow experience, and empathy perception during their interactions with the two types of agents.

The cognitive load scale was adapted from the version developed by Chang *et al.* [76], which originally comprised 12 items covering three dimensions: intrinsic cognitive load (4 items), extraneous cognitive load (4 items), and germane cognitive load (4 items). To align with the context of conversations with LLMs, some of the item descriptions were slightly adjusted to focus on conversations with agents, while the original measurement intentions were maintained.

The flow experience scale was adapted from the research of Pearce *et al.* [77]. We used the revised version by Chang *et al.*, further modifying it to suit the context of this study. This scale comprises three dimensions: enjoyment, engagement, and control. Each has 4 items, totaling 12 questions. We also modified some of the item

descriptions to refer to the way of interaction, while maintaining the original measurement intentions.

The empathy measurement tool was based on the Interpersonal Reactivity Index (IRI) developed by Davis [78], which originally contained 28 items covering four dimensions: perspective-taking, fantasy, empathic concern, and personal distress. Among the items in the IRI scales, 2, 5, 7, 9, 11, and 28 are designed to assess students' affective empathy, which is closely aligned with the construct examined in this study. Therefore, these six items were selected to measure students' perceived empathy during their interactions with the LLM agents.

All three scales were in the form of a Likert scale, with response options ranging from 1 (strongly disagree) to 5 (strongly agree). For the reverse items in the questionnaire, the scale ranges from 5 (strongly disagree) to 1 (strongly agree).

E. Procedure

This experiment was conducted in three stages, as illustrated in Fig. 2. During the pretest stage, all participants were instructed to complete two questions independently within 10 min. In the intervention phase, students were assigned to one of two agent conditions based on the parity of the last digit of their ID numbers. Students with odd-numbered IDs ($n = 32$) interacted with the teacher-centered agent, whereas students with even-numbered IDs ($n = 31$) interacted with the student-centered agent. During the intervention, students interacted with their assigned agent for approximately 20 min to discuss the two physics questions. Following the session, all participants completed post-test questions and questionnaires on cognitive load, flow experience, and perception of empathy. This stage lasted for about 20 min. Four students did not provide their personal

information in the post-test, so their data were excluded from the analysis. Consequently, 59 students voluntarily completed the questionnaire and shared all conversation information from their interactions with the agents.

F. Data analysis

1. Analysis of knowledge tests

The effects of teacher-centered agents and student-centered agents on students' learning performance were examined by comparing the pretest and post-test mean scores in each condition. Each test included two questions, with 5 points awarded for correct answers and 0 for incorrect ones. The results showed that the pretest ($M = 5.85$, $SD = 3.10$) and post-test ($M = 8.81$, $SD = 2.34$) scores didn't follow a normal distribution (Kolmogorov–Smirnov test, $p < 0.01$). Therefore, the Wilcoxon signed-rank test was used to compare within-condition changes between pre- and post-tests, while the Mann–Whitney U test was used to compare between-condition differences in both tests. Both the Wilcoxon signed-rank test and the Mann–Whitney U test are non-parametric tests; effect size r was reported to indicate the magnitude of the observed differences, with conventional benchmarks of small ($|r| \geq 0.10$), medium ($|r| \geq 0.30$), and large ($|r| \geq 0.50$) effects. The studies compared the effectiveness of teacher-centered and student-centered agents in facilitating students' learning performance.

2. Analysis of three scales

Cronbach's alpha was calculated to evaluate the internal consistency of each scale. Typically, a value greater than 0.7 is regarded as adequate for the internal consistency of the test. In this context, the Cronbach's alpha for the cognitive load scale was 0.857, for the flow experience

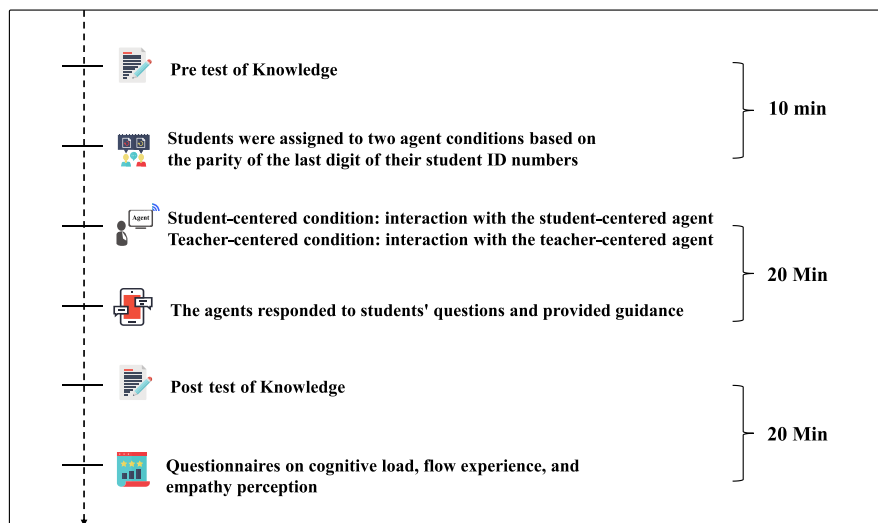


FIG 2. Procedure of the experiment.

scale was 0.893, and for the empathy perception scale was 0.785.

The normality of the three scales of cognitive load, flow experience, and perceived empathy was assessed using the Kolmogorov–Smirnov test. Results showed that the data for cognitive load ($D = 0.091$, $p > 0.05$) and flow experience ($D = 0.085$, $p > 0.05$) were normally distributed. Consequently, independent samples t tests were conducted to compare condition differences on these measures during interactions with the LLM agents, and Cohen’s d was reported as the corresponding effect size measure, with conventional benchmarks of small ($|d| \geq 0.20$), medium ($|d| \geq 0.50$), and large ($|d| \geq 0.80$) effects. Conversely, the questionnaire data for empathy perception were not normally distributed ($D = 0.118$, $p < 0.05$). Therefore, the Mann–Whitney U test was conducted to compare the difference between the two conditions on this dimension, and effect size r was reported.

Because multiple dependent variables were compared across the two agent conditions, the Benjamini–Hochberg procedure was used to control the false discovery rate at 0.05. The adjusted p values are reported in the Results section.

V. RESULTS

A. Learning effects

Figure 3 presents the effects of two agent conditions on students’ learning performance. The Mann–Whitney U test was used to examine between-condition differences, while the Wilcoxon signed-rank test was applied to compare students’ pretest and post-test performance within each condition. The corresponding descriptive statistics and the results of the between- and within-condition comparisons are provided in Appendix B (Tables IV–VI)

In the pretest, the student-centered agent condition achieved a mean score of 5.83 (SD = 3.24), while those in the teacher-centered agent condition obtained a mean score of 5.86 (SD = 3.01). The results of the Mann–Whitney U test indicated no significant difference between the two conditions at pretest ($U = 435$, $Z = 0.000$, $p = 1.000$).

In the post-test, students in the student-centered agent condition obtained a mean score of 9.67 (SD = 1.27), while those in the teacher-centered agent condition scored 7.93 (SD = 2.84). The Wilcoxon signed-rank test showed that students in both the student-centered agent condition ($Z = -4.07$, $p < 0.001$, $r = 0.76$) and the teacher-centered agent condition ($Z = -3.21$, $p < 0.01$, $r = 0.59$) obtained significantly higher post-test scores than pretest scores. Both effect sizes were large, indicating substantial gains in post-test scores relative to pretest scores following student interaction with their assigned agents. In addition, the Mann–Whitney U test showed a significant difference in post-test scores between the two conditions ($U = 298$, $Z = 2.89$, $p < 0.01$, $r = 0.38$), with higher scores in the student-centered agent condition. The corresponding effect size was moderate, indicating a moderate advantage for the student-centered agent condition over the teacher-centered agent condition.

B. Cognitive load

The independent samples t test results for the three dimensions of cognitive load in the teacher-centered agent and student-centered agent conditions are shown in Table I. There was no significant difference between the two conditions in intrinsic cognitive load ($t = -0.57$, $p_{\text{adj}} = 0.58$, $d = -0.15$). However, significant differences were observed in extraneous cognitive load ($t = -2.84$,

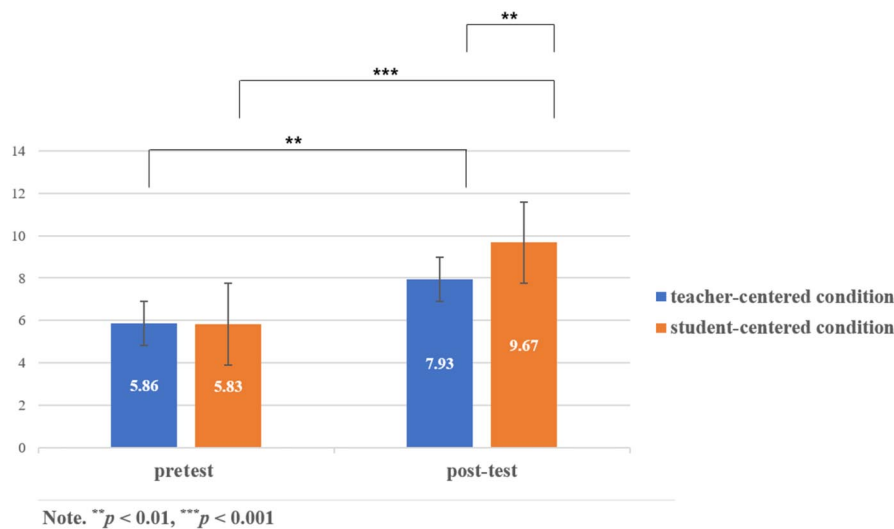


FIG. 3. Between-condition and within-condition differences in pretest and post-test learning performance for the teacher-centered and student-centered agent conditions.

TABLE I. Results of the independent samples t test on cognitive load between the teacher-centered and student-centered agent conditions.

	The student-centered condition ($N = 29$)	The teacher-centered condition ($N = 30$)			
	$M \pm SD$	$M \pm SD$	t	p_{adj}	Cohen's d
Intrinsic cognitive load	9.47 ± 3.03	9.90 ± 2.82	-0.57	0.58	-0.15
Extraneous cognitive load	8.33 ± 2.83	10.76 ± 3.67	-2.84	0.01*	-0.74
Germane cognitive load	14.27 ± 2.50	12.62 ± 3.31	2.16	0.04*	0.56

Note. * $p < 0.05$.

$p_{\text{adj}} < 0.05$, $d = -0.74$) and germane cognitive load ($t = 2.16$, $p_{\text{adj}} < 0.05$, $d = 0.56$). Figure 4 illustrates the differences in the three dimensions of cognitive load between the teacher-centered and student-centered agent conditions. These effect sizes suggest a relatively substantial difference in extraneous cognitive load and a moderate difference in germane cognitive load between the two agent conditions. Specifically, students in the student-centered agent condition reported statistically lower extraneous cognitive load ($M = 8.33$, $SD = 2.83$) than those in the teacher-centered agent condition ($M = 10.76$, $SD = 3.67$) but a higher germane cognitive load ($M = 14.27$, $SD = 2.50$) than the teacher-centered agent condition ($M = 12.62$, $SD = 3.31$). These findings suggest that, compared with the teacher-centered agent condition, the student-centered agent condition may have reduced students' extraneous cognitive load and facilitated deeper cognitive engagement.

C. Flow experience

From Table II, it is also shown that there were significant differences in enjoyment ($t = 3.88$, $p_{\text{adj}} < 0.001$, $d = 1.00$), engagement ($t = 2.30$, $p_{\text{adj}} < 0.05$, $d = 0.60$), and control ($t = 2.88$, $p_{\text{adj}} < 0.05$, $d = 0.75$) between the teacher-centered and student-centered agent conditions. Figure 5 illustrates the differences in the three dimensions

of flow experience between the teacher-centered and student-centered agent conditions. These effect sizes indicate a large effect for enjoyment, a moderate effect for engagement, and a moderate-to-large effect for control, respectively. In addition, students in the student-centered agent condition showed a higher flow score than the teacher-centered agent condition ($t = 3.54$, $p < 0.01$, $d = 0.92$), with a large effect size. This suggests that the student-centered agent condition was associated with more positive flow-related experiences, which may have supported students' motivation and engagement during learning.

D. Empathy perception

As shown in Table III, a significant difference in empathy perception was observed between the teacher-centered and student-centered agent conditions ($U = 170$, $Z = 4.05$, $p_{\text{adj}} < 0.01$, $r = 0.53$). Figure 6 illustrates the difference in empathy perception between the teacher-centered and student-centered agent conditions. The effect size was large, indicating a substantial difference in empathy perception between the two conditions. Students in the student-centered agent condition reported higher empathy perception ($M = 21.27$, $SD = 2.56$) than those in the teacher-centered agent condition ($M = 18.24$, $SD = 3.70$). These findings suggest that the student-centered agent condition was

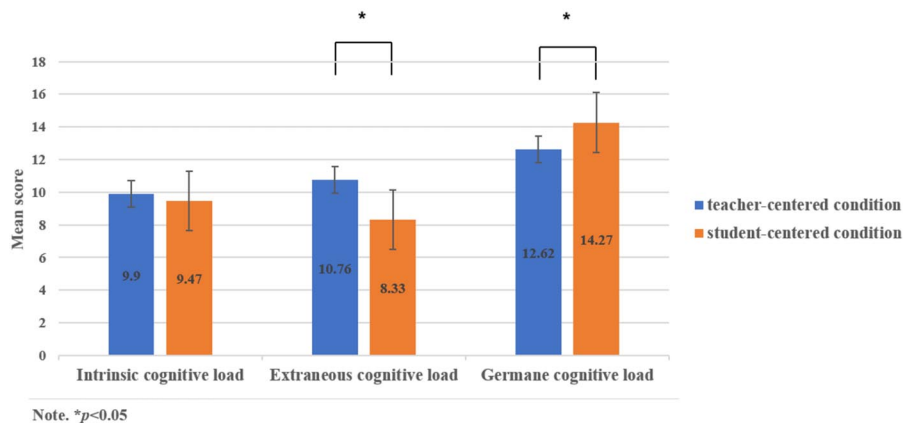


FIG. 4. Differences in cognitive load between the teacher-centered and student-centered agent conditions.

TABLE II. Results of the independent samples t test on flow experience between the teacher-centered and student-centered agent conditions.

	The student-centered condition ($N = 29$)	The teacher-centered condition ($N = 30$)	t	p_{adj}	Cohen's d
	$M \pm SD$	$M \pm SD$			
Flow experience	43.73 ± 5.34	37.97 ± 7.09	3.54	0.003**	0.92
Enjoyment	15.17 ± 2.15	12.66 ± 2.79	3.88	< 0.001***	1.00
Engagement	13.13 ± 2.35	11.76 ± 2.25	2.30	0.03*	0.60
Control	15.43 ± 1.89	13.55 ± 2.99	2.88	0.01*	0.75

Note. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

associated with more positive empathy-related experiences in the physics learning context.

VI. DISCUSSION

The present study utilized prompt engineering to construct teacher-centered and student-centered agents and then compared the impacts of these two types of agents on students' learning performance, cognitive load, flow experience, and empathy perception in a physics learning context.

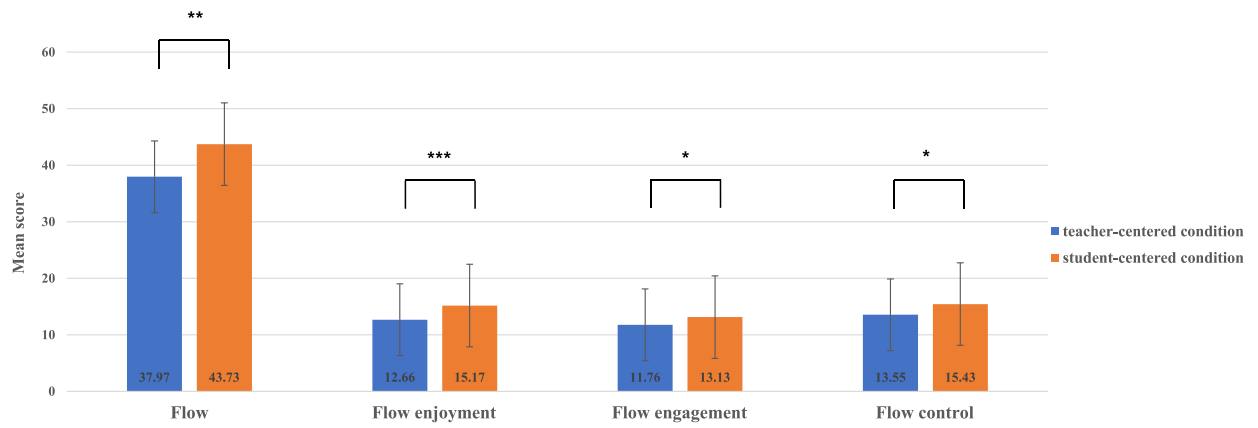
A. Differences between teacher-centered agent and student-centered agent in promoting students' learning performance

Regarding the first question, this study compared students' learning performance before and after interacting with the agents. The results showed that students in both agent conditions obtained higher scores after interacting with their assigned agents than before, suggesting that both teacher-centered and student-centered agents may support students' physics learning in this AI-supported learning context. These findings are consistent with those of Jiang *et al.* [79], who found that Physics-STAR, a physics education tutoring system based on large

language model, significantly improved students' average scores and efficiency on both conceptual and computational problems.

In terms of learning performance, students in the student-centered agent condition performed better than those in the teacher-centered agent condition. As Jiang asserts, the implementation of LLMs in the context of physics education ought to prioritize a student-centered approach [79]. Ding *et al.* also proposed a structured sequence of steps to guide large language models in answering questions: review, guidance, inspiration, correction, and summary [38]. This strategy prioritizes the provision of step-by-step assistance to facilitate students' problem-solving processes. Similarly, the student-centered agent emphasizes the sequential approach to solving physics questions. It first identified the underlying causes of students' misconceptions, then provided specific explanations of essential concepts, and finally utilized similar examples to enhance understanding.

Furthermore, at the end of each interaction round, the agent would ask students, "Have you comprehended the point of knowledge?" or "Have you understood this phenomenon?" This instructional strategy may enhance students' engagement, often referred to as their sense of



Note. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

FIG. 5. Differences in flow experience between the teacher-centered and student-centered agent conditions.

TABLE III. Results of the Mann–Whitney U test between the teacher-centered and student-centered agent conditions in empathy perception.

	The student-centered condition ($N = 29$)	The teacher-centered condition ($N = 30$)	U	Z	p_{adj}	r
	$M \pm SD$	$M \pm SD$				
Empathy perception	21.27 ± 2.56	18.24 ± 3.70	170	4.06	0.003**	0.53

Note. ** $p < 0.01$.

participation. Within the context of human-machine interaction, the sense of participation is defined as the user's subjective experience of taking an active role in task completion, as opposed to remaining a passive recipient of information [80]. The student-centered agent has been shown to effectively enhance students' sense of participation in solving physics questions by actively asking questions and awaiting feedback. This enabled students to remain immersed in the conversation context and promote the comprehension and implementation of acquired knowledge.

B. The differences between teacher-centered and student-centered agents in terms of cognitive load, flow experience, and empathy

The second research question examined differences between teacher-centered agent and student-centered agent conditions in students' cognitive load, flow experience, and empathy perception. In terms of cognitive load, there was no difference in intrinsic cognitive load between the two conditions, as the materials selected in this study exhibited equivalent complexity for all participants. In contrast, students in the student-centered agent condition reported lower extraneous cognitive load than those in the teacher-centered condition. Sweller has indicated that extraneous cognitive load is typically the result of deficiencies in instructional design [81]. It has a

substantial negative effect on learning outcomes. Students were able to concentrate on information processing as the student-centered agent adopted a step-by-step conversational strategy with continuous feedback, compared to the teacher-centered agent's responses. Wang's research confirmed that students are able to focus on the process of knowledge construction and enhance learning outcomes when the system provides immediate and personalized feedback [64]. The progressive and interactive design of the student-centered agent in responding to questions efficiently reduces students' extraneous cognitive load. Additionally, students in the student-centered condition also reported higher germane cognitive load. According to Sweller, learners allocate more cognitive resources to germane cognitive load when the instructional design effectively reduces extraneous cognitive load [82]. This allows them to process intrinsic complexity more efficiently, which in turn facilitates concept construction and knowledge integration. The findings of this study indicated that the higher level of germane cognitive load observed in the student-centered agent condition suggests that its instructional structure and presentation effectively guided learners to allocate their cognitive resources toward solving physics questions, thereby improving their understanding of physics concepts.

The results indicated that students in the student-centered agent condition reported higher flow experience than those in the teacher-centered condition, with significant differences in enjoyment, engagement, and control. This difference may be attributed to the student-centered agent's interactive and feedback strategy, which keeps students engaged with the conversation and facilitates their entry into the flow state. It suggests that an interaction style emphasizing encouragement and timely feedback may be associated with more positive learning experiences. According to Csikszentmihalyi, flow experience was a high level of immersion, concentration, and enjoyment experienced when facing challenging but manageable tasks [59]. A previous study indicated that when the learning task aligns with learners' abilities, instructional designs that support autonomy, competence, and task relevance can effectively promote the process of entering the flow experience [83]. Both of the questions selected in this study were particularly challenging, especially for

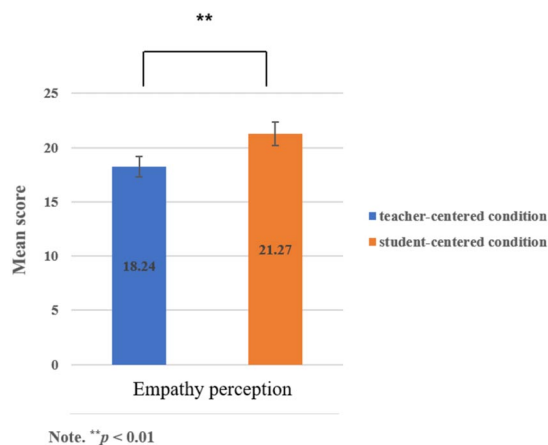


FIG. 6. Differences in empathy perception between the teacher-centered and student-centered agent conditions.

students who had no prior experience using large language models to solve physics questions. Dialogue-based learning represents an innovative and exploratory instructional method. The student-centered agent takes the student's perspective as the core and adopts encouraging language and timely feedback strategies to guide students to gradually solve questions. The design enhances students' sense of control and satisfaction after solving physics questions, thereby sustaining their engagement in the flow state.

In terms of empathy perception, students in the student-centered agent condition also reported higher empathy perception than those in the teacher-centered condition. Unlike the teacher-centered agent, the student-centered agent adopted a student-oriented perspective that takes into account learners' prior knowledge and potential cognitive difficulties. These responses generated by LLMs allow students to perceive the agent's empathy. Ortega-Ochoa *et al.* have noted that agents designed with empathetic abilities can enhance learners' emotional experiences, while encouraging interactions may help boost students' confidence and engagement [84]. In this study, the student-centered agent was programmed to deliver encouraging feedback at the end of each dialogue turn, such as "Congratulations! You got it right!" or "You are so great!" These expressions contributed to a stronger affective connection between students and the agent, thereby increasing their identification with it. Previous research also supported this finding, suggesting that empathetic agents are more capable of regulating users' emotions and alleviating boredom. The student-centered agent received unanimous positive feedback from students. A substantial proportion of students also reported in the questionnaire that they would recommend the student-centered agent to their peers, further confirming its positive impact on empathy perception.

VII. CONCLUSIONS AND FUTURE WORK

This study used the Coze platform to develop two types of LLM agents based on different prompting strategies: the teacher-centered agent and the student-centered agent. These agents were designed with distinct role orientations and interaction styles to investigate how different prompting strategies affect students' learning outcomes, cognitive load, flow experience, and empathy perception in physics learning. The results showed that

students in both agent conditions performed better on the post-test than on the pretest, suggesting that both types of agents may have supported students' understanding of physics concepts and learning performance. Compared with the teacher-centered agent condition, the student-centered agent condition was associated with lower extraneous cognitive load, higher germane cognitive load, and higher levels of flow experience and empathy perception. These results suggest that the design of the student-centered agent, which emphasizes encouragement, adaptability, and student-centered interaction, may have been better aligned with students' cognitive characteristics and emotional needs in an AI-supported physics learning context.

This study has several limitations that should be noted, particularly in terms of sample representativeness and the scope of experimental design. First, the participants were limited to students at a specific academic level. This limited the generalizability of the findings to a broader student population. Second, only two physics questions were used for students to engage in dialogue with the agents, which may have constrained the depth and frequency of the interactions. Future studies will address these limitations by increasing the sample coverage and integrating a wider variety of physics questions to obtain a more comprehensive understanding of LLM-assisted physics learning.

ACKNOWLEDGMENTS

The authors would like to acknowledge the valuable suggestions of the anonymous reviewers. This research was supported by the 2024 Humanities and Social Sciences Planning Project funded by the Ministry of Education under (Grant No. 24YJA880105), Guangdong Provincial Undergraduate University Teaching Quality and Teaching Reform Project [2024] No. 9, Item 611, the Guangdong "14th Five-Year" Planning Project of Philosophy and Social Sciences under (Grant No. GD23XJY63), Guangdong Provincial Young Innovative Talents Project for Higher Education Institutions (Grant No. 2024WQNCX109).

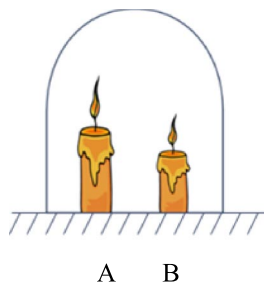
DATA AVAILABILITY

There are no publicly available research data or software supporting this manuscript. Requests for further information or data should be sent to the authors.

APPENDIX A: KNOWLEDGE TEST

This appendix presents two multiple-choice questions used to assess students' conceptual understanding of convection caused by temperature differences in the pre- and post-tests of the experiment.

1. As shown in the following figure, two candles, A and B, are placed at different vertical positions. After lighting them and covering them with a transparent glass cup of appropriate height, which of the following phenomena will occur? ()



- A. Candle A goes out first.
- B. Candle B goes out first.
- C. Candle A and Candle B go out simultaneously.
- D. It cannot be determined.

2. As illustrated in the figure below, the air-conditioning outlets in the two rooms are oriented diagonally downward and diagonally upward, respectively. Assuming identical air-conditioning power and room dimensions, which room will exhibit a more uniform temperature distribution when the cooling mode is activated? ()



Room 1



Room 2

- A. The temperature distribution in Room 1 is more uniform.
- B. The temperature distribution in Room 2 is more uniform.
- C. The temperature distribution in Room 1 and Room 2 is the same.
- D. It cannot be determined.

APPENDIX B: DESCRIPTIVE STATISTICS AND INFERENCE RESULTS FOR THE KNOWLEDGE TEST

This appendix presents the descriptive statistics and inferential results for the knowledge test, including the pretest and post-test performance of the two conditions.

TABLE IV. Descriptive statistics for pretest and post-test learning performance in the two conditions.

Condition	<i>N</i>	Pretest		Post-test	
		<i>M</i>	SD	<i>M</i>	SD
The student-centered condition	29	5.83	3.24	9.67	1.27
The teacher-centered condition	30	5.86	3.01	7.93	2.84

TABLE V. Mann–Whitney *U* test results for between-condition comparisons on pretest and post-test learning performance.

Learning performance				
The student-centered condition vs the teacher-centered condition				
	<i>U</i>	<i>Z</i>	<i>p</i>	<i>r</i>
Pretest	435	0.000	1.000	...
Post-test	298	2.89	0.004**	0.38

Note. ** $p < 0.01$.

TABLE VI. Wilcoxon signed-rank test results for the pretest and post-test comparison within the student-centered and teacher-centered conditions.

Learning performance				
Pretest vs post-test				
	<i>N</i>	<i>Z</i>	<i>p</i>	<i>r</i>
The student-centered condition	29	−4.07	< 0.001***	0.76
The teacher-centered condition	30	−3.21	0.001**	0.59

Note. ** $p < 0.01$, *** $p < 0.001$.

-
- [1] A. Extnance, ChatGPT has entered the classroom: How large language models could transform education, *Nature (London)* **623**, 474 (2023).
 - [2] Y. Zhu, C. Zhu, T. Wu, S. Wang, Y. Zhou, J. Chen, F. Wu, and Y. Li, Impact of assignment completion assisted by large language model-based chatbot on middle school students' learning, *Educ. Inf. Technol.* **30**, 2429 (2025).
 - [3] A. Coban, D. Dzsotjan, S. Küchemann, J. Durst, J. Kuhn, and C. Hoyer, AI support meets AR visualization for Alice and Bob: Personalized learning based on individual ChatGPT feedback in an AR quantum cryptography experiment for physics lab courses, *EPJ Quantum Technol.* **12**, 15 (2025).
 - [4] R. Malik, D. Abdi, R. Wang, and D. Demiszky, Scaffolding middle school mathematics curricula with large language models, *Br. J. Educ. Technol.* **56**, 999 (2025).
 - [5] B. Gregorcic and A.-M. Pendrill, ChatGPT and the frustrated Socrates, *Phys. Educ.* **58**, 035021 (2023).
 - [6] G. Polverini, J. Melin, E. Önerud, and B. Gregorcic, Performance of ChatGPT on tasks involving physics visual

- representations: The case of the brief electricity and magnetism assessment, *Phys. Rev. Phys. Educ. Res.* **21**, 010154 (2025).
- [7] R. Loubet, P. Zittlau, L. Vollmer, M. Hoffmann, S. Fellenz, F. Jirasek, H. Leitte, and H. Hasse, Using large language models for solving textbook-style thermodynamic problems, *Comput. Chem. Eng.* **204**, 109333 (2026).
- [8] H. Liu, H. Yin, Z. Luo, and X. Wang, Integrating chemistry knowledge in large language models via prompt engineering, *Synth. Syst. Biotechnol.* **10**, 23 (2025).
- [9] N. Xu and X. Ma, DecoPrompt: Decoding prompts reduces hallucinations when large language models meet false premises, [arXiv:2411.07457](https://arxiv.org/abs/2411.07457).
- [10] J. Cheng, T. Su, J. Yuan, G. He, J. Liu, X. Tao, J. Xie, and H. Li, Chain-of-thought prompting obscures hallucination cues in large language models: An empirical evaluation, [arXiv:2506.17088](https://arxiv.org/abs/2506.17088).
- [11] 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).
- [12] L. Gong, J. Chen, and F. Wu, Is ChatGPT a competent teacher? Systematic evaluation of large language models on the competency model, *IEEE Trans. Learn. Technol.* **18**, 530 (2025).
- [13] E. Tufino and B. Gregorcic, Creating a customisable Socratic AI physics tutor, *Phys. Educ.* **60**, 065037 (2025).
- [14] A. Sirnoorkar and N. S. Rebello, Feedback that clicks: Introductory physics students' valued features in AI feedback generated from self-crafted and engineered prompts, [arXiv:2509.08516](https://arxiv.org/abs/2509.08516).
- [15] G. Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course?, *Phys. Rev. Phys. Educ. Res.* **19**, 010132 (2023).
- [16] C. G. West, Advances in apparent conceptual physics reasoning in GPT-4, [arXiv:2303.17012](https://arxiv.org/abs/2303.17012).
- [17] Y. Liang, D. Zou, H. Xie, and F. L. Wang, Exploring the potential of using ChatGPT in physics education, *Smart Learn. Environ.* **10**, 52 (2023).
- [18] Z. Chen and T. Wan, Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy, *Phys. Rev. Phys. Educ. Res.* **21**, 010126 (2025).
- [19] 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).
- [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] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with ChatGPT, [arXiv:2302.11382](https://arxiv.org/abs/2302.11382).
- [22] L. Giray, Prompt engineering with ChatGPT: A guide for academic writers, *Ann. Biomed. Eng.* **51**, 2629 (2023).
- [23] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, *Adv. Neural Inf. Process. Syst.* **35**, 24824 (2022).
- [24] J. Wu, J. Li, Z. Ge, M. Xu, L. Lin, and R. Zhang, Effectiveness of generative AI in automated written corrective feedback with prompting, *J. Educ. Comput. Res.* **63**, 1493 (2025).
- [25] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, Large language models are zero-shot reasoners, [arXiv:2205.11916](https://arxiv.org/abs/2205.11916).
- [26] OpenAI, Reasoning models, https://developers.openai.com/api/docs/guides/reasoning?utm_source=chatgpt.com.
- [27] C. Cohn, N. Hutchins, T. Le, and G. Biswas, A chain-of-thought prompting approach with LLMs for evaluating students, *Formative Assess. Responses Sci., AAAI* **38**, 23182 (2024).
- [28] Z. Lai, G. Dai, Y. Lu, Y. Xu, J. Li, and X. Zhu, DeepSyn: Operable chemical retrosynthetic design with the DeepSeek R1 model, *J. Phys. Chem. Lett.* **16**, 4962 (2025).
- [29] Y. Walter, Embracing the future of artificial intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education, *Int. J. Educ. Technol. High. Educ.* **21**, 15 (2024).
- [30] G. Polverini and B. Gregorcic, How understanding large language models can inform the use of ChatGPT in physics education, *Eur. J. Phys.* **45**, 025701 (2024).
- [31] Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes, and O. M. Yaghi, ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis, *J. Am. Chem. Soc.* **145**, 18048 (2023).
- [32] B. Lyu, C. Li, H. Li, H. Oh, Y. Song, W. Zhu, and W. Xing, The role of teachable agents' personality traits on student-AI interactions and math learning, *Comput. Educ.* **234**, 105314 (2025).
- [33] B. Xu, A. Yang, J. Lin, Q. Wang, C. Zhou, Y. Zhang, and Z. Mao, ExpertPrompting: Instructing large language models to be distinguished experts, [arXiv:2305.14688](https://arxiv.org/abs/2305.14688).
- [34] W. Xing, Y. Song, C. Li, Z. Liu, W. Zhu, and H. Oh, Development of a generative AI -powered teachable agent for middle school mathematics learning: A design-based research study, *Br. J. Educ. Technol.* **56**, 2043 (2025).
- [35] N. L. Schroeder, W. L. Romine, and S. D. Craig, Measuring pedagogical agent persona and the influence of agent persona on learning, *Comput. Educ.* **109**, 176 (2017).
- [36] R. Xiao, X. Hou, R. Ye, M. Kazemitabaar, N. Diana, M. Liut, and J. Stamper, Improving student-AI interaction through pedagogical prompting: An example in computer science education, [arXiv:2506.19107](https://arxiv.org/abs/2506.19107).
- [37] Y. Ding, H. Hu, J. Zhou, Q. Chen, B. Jiang, and L. He, Boosting large language models with Socratic method for conversational mathematics teaching, [arXiv:2407.17349](https://arxiv.org/abs/2407.17349).
- [38] G. Kestin, K. Miller, A. Klales, T. Milbourne, and G. Ponti, AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting, *Sci. Rep.* **15**, 17458 (2025).
- [39] S. Küchemann, S. Steinert, J. Kuhn, K. Avila, and S. Ruzika, Large language models—Valuable tools that require a sensitive integration into teaching and learning physics, *Phys. Teach.* **62**, 400 (2024).

- [40] Significant Gravitas, AutoGPT, <https://github.com/Significant-gravitas/AutoGPT>.
- [41] LangChain, <https://www.langchain.com/agents>.
- [42] B. Ni and M. J. Buehler, MechAgents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge, *Extreme Mech. Lett.* **67**, 102131 (2024).
- [43] Coze, <https://www.coze.cn/docs/guides/welcome>.
- [44] Cherry Studio, <https://www.cherry-ai.com/>.
- [45] M. Liu, M. Yang, Z.-M. Wu, and J. Liao, Development, application status and future prospect of large language model agents for education, *Mod. Educ. Technol.* **34**, 5 (2024).
- [46] F. Fan, S. Wang, M. D. M. Hemuti, X. Nie, and L. T. Yang, Enhancing students' metacognition with innovative IA-based metacognitive reflective learning tool, *IEEE Trans. Learn. Technol.* **18**, 699 (2025).
- [47] J. Sweller, Cognitive load during problem solving: Effects on learning, *Cogn. Sci.* **12**, 257 (1988).
- [48] P. Chandler and J. Sweller, Cognitive load theory and the format of instruction, *Cogn. Instr.* **8**, 293 (1991).
- [49] J. Sweller, Cognitive load theory, learning difficulty, and instructional design, *Learn. Instr.* **4**, 295 (1994).
- [50] J. Sweller, J. J. G. van Merriënboer, and F. G. W. C. Paas, Cognitive architecture and instructional design, *Educ. Psychol. Rev.* **10**, 251 (1998).
- [51] Q. Liu, S. Yu, W. Chen, Q. Wang, and S. Xu, The effects of an augmented reality based magnetic experimental tool on students' knowledge improvement and cognitive load, *J. Comput. Assist. Learn.* **37**, 645 (2020).
- [52] D. Baidoo-Anu, and L. Owusu Ansah, Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning, *J. AI* **7**, 52 (2023).
- [53] L. Yan, R. Martinez-Maldonado, Y. Jin, V. Echeverria, M. Milesi, J. Fan, L. Zhao, R. Alfredo, X. Li, and D. Gašević, The effects of generative AI agents and scaffolding on enhancing students' comprehension of visual learning analytics, *Comput. Educ.* **234**, 105322 (2025).
- [54] W. K. Monib, A. Qazi, and M. M. Mahmud, Exploring learners' experiences and perceptions of ChatGPT as a learning tool in higher education, *Educ. Inf. Technol.* **30**, 917 (2025).
- [55] L. P. Patac, A. V. Patac Jr. Jr., Using ChatGPT for academic support: Managing cognitive load and enhancing learning efficiency—A phenomenological approach, *Soc. Sci. Humanit. Open* **11**, 101301 (2025).
- [56] 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).
- [57] Z. Ji *et al.*, Survey of hallucination in natural language generation, *ACM Comput. Surv.* **55**, 1 (2023).
- [58] D. J. Woo, D. Wang, K. Guo, and H. Susanto, Teaching EFL students to write with ChatGPT: Students' motivation to learn, cognitive load, and satisfaction with the learning process, *Educ. Inf. Technol.* **29**, 24963 (2024).
- [59] M. Csikszentmihalyi, Flow: The psychology of optimal experience (1991).
- [60] C.-C. Wang and M.-C. Hsu, An exploratory study using inexpensive electroencephalography (EEG) to understand flow experience in computer-based instruction, *Inf. Manage.* **51**, 912 (2014).
- [61] U. Konradt, R. Filip, and S. Hoffmann, Flow experience and positive affect during hypermedia learning, *Br. J. Educ. Technol.* **34**, 309 (2003).
- [62] Q. Li, S. Xu, Y. Xiong, W. He, and S. Zhou, Comparing the perception of emergency remote teaching experience between physics and nonphysics students, *Phys. Rev. Phys. Educ. Res.* **19**, 020131 (2023).
- [63] F. S. Al-Hafdi and S. M. AlNajdi, The effectiveness of using chatbot-based environment on learning process, students' performances and perceptions: A mixed exploratory study, *Educ. Inf. Technol.* **29**, 20633 (2024).
- [64] M. Wang, D. Zhang, J. Zhu, and H. Gu, Effects of incorporating a large language model-based adaptive mechanism into contextual games on students' academic performance, flow experience, cognitive load and behavioral patterns, *J. Educ. Comput. Res.* **63**, 662 (2025).
- [65] S. van Hoof, *The Handbook of Virtue Ethics* (Acumen, Durham, 2014).
- [66] S. Stojiljković, G. Djigić, and B. Zlatković, Empathy and teachers' roles, *Procedia - Soc. Behav. Sci.* **69**, 960 (2012).
- [67] Z. Zhang, Toward the role of teacher empathy in students' engagement in English language classes, *Front. Psychol.* **13**, 880935 (2022).
- [68] S. Meyers, K. Rowell, M. Wells, and B. C. Smith, Teacher empathy: A model of empathy for teaching for student success, *Coll. Teach.* **67**, 160 (2019).
- [69] L. Zhi, GuidedEmpathy: Guiding large language models for empathetic response generation with situational awareness, in *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence* (ACM, Beijing, China, 2024), pp. 354–360. [10.1145/3709026.3709029](https://doi.org/10.1145/3709026.3709029)
- [70] M. T. Shih, M. Y. Hsu, and S. C. Lee, Empathy-GPT: Leveraging large language models to enhance emotional empathy and user engagement in embodied conversational agents, in *The 37th Annual ACM Symposium on User Interface Software and Technology* (ACM, Pittsburgh PA, USA, 2024), pp. 1–3. [10.1145/3672539](https://doi.org/10.1145/3672539)
- [71] X. Bai, G. Chen, T. He, C. Zhou, and C. Guo, A holistic comparative study of large language models as emotional support dialogue systems, *Cogn. Comput.* **17**, 71 (2025).
- [72] *Physics: Grade 8*, edited by Q.-C. Peng and S.-B. Huang (People's Education Press, Beijing, 2016), Vol. 1, p. 147.
- [73] H. Liu, Z.-W. Chen, J.-J. Li, and Z.-K. Zhai, Thermal comfort research of air conditioner outlet locations, *Home Appliance Sci. Technol.* **2**, 48 (2022).
- [74] S. Park and Y.-C. Chen, Mapping out the integration of the components of pedagogical content knowledge (PCK): Examples from high school biology classrooms, *J. Res. Sci. Teach.* **49**, 922 (2012).
- [75] Z. Deng, W. Ma, Q.-L. Han, W. Zhou, X. Zhu, S. Wen, and Y. Xiang, Exploring DeepSeek: A survey on advances, applications, challenges and future directions, *IEEECAA J. Autom. Sin.* **12**, 872 (2025).
- [76] C.-C. Chang, C. Liang, P.-N. Chou, and G.-Y. Lin, Is game-based learning better in flow experience and various

- types of cognitive load than non-game-based learning? Perspective from multimedia and media richness, *Comput. Hum. Behav.* **71**, 218 (2017).
- [77] J. Pearce, M. Ainley, and S. Howard, The ebb and flow of online learning, *Comput. Hum. Behav.* **21**, 745 (2005).
- [78] M. H. Davis, A multidimensional approach to individual differences in empathy, *JSAS Catalog Sel. Doc. Psychol.* **10**, 85 (1980).
- [79] Z. Jiang and M. Jiang, Beyond answers: Large language model-powered tutoring system in physics education for deep learning and precise understanding, [arXiv:2406.10934](https://arxiv.org/abs/2406.10934).
- [80] C. Heeter, Being there: The subjective experience of presence, in *Presence: Teleoperators and Virtual Environments* (MIT Press, Cambridge, MA, 1992), pp. 262-271.
- [81] J. Sweller, Cognitive load theory and educational technology, *Educ. Technol. Res. Dev.* **68**, 1 (2020).
- [82] J. Sweller, J. J. G. van Merriënboer, and F. Paas, Cognitive architecture and instructional design: 20 years later, *Educ. Psychol. Rev.* **31**, 261 (2019).
- [83] J. Hamari, D. J. Shernoff, E. Rowe, B. Coller, J. Asbell-Clarke, and T. Edwards, Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning, *Comput. Hum. Behav.* **54**, 170 (2016).
- [84] E. Ortega-Ochoa, M. Arguedas, and T. Daradoumis, Empathic pedagogical conversational agents: A systematic literature review, *Br. J. Educ. Technol.* **55**, 886 (2024).