Exploring the role of human-AI collaboration in solving scientific problems

Dazhen Tong,[1] Bangjian Jin,[2] Yang Tao,[3] Hongmei Ren,[1]
A. Y. M. Atiquil Islam,[4,5,†] and Lei Bao[6,*]

[1]*School of Physics and Electronic Information, Anhui Normal University, Wuhu, Anhui 241002, China*
[2]*Zhuji City Hailiang Senior High School, Shaoxing, Zhejiang 311816, China*
[3]*Faculty of Education, Research Institute of Science Education, Beijing Normal University, Beijing 100875, China*
[4]*Department of Education Information Technology, East China Normal University, Shanghai 200062, China*
[5]*School of Teacher Education, Jiangsu University, Zhenjiang, Jiangsu 212013, China*
[6]*Department of Physics, The Ohio State University, Columbus, Ohio 43210, USA*

(Received 21 October 2024; accepted 22 April 2025; published 8 May 2025)

[This paper is part of the Focused Collection in Artificial Intelligence Tools in Physics Teaching and Physics Education Research.] Recent research in science education has largely focused on using ChatGPT to solve problems and evaluating its accuracy and problem-solving features. However, as artificial intelligence (AI) is becoming an important tool for human development, effective strategies for learning and education using human-AI collaboration (HAI) learning remain underexplored. HAI is emerging as a powerful tool for enhancing educational outcomes, especially with the integration of advanced natural language AI systems such as ChatGPT-4o. This study explores the role of HAI in solving scientific problems, comparing its efficacy with human-human collaboration (HHC) among high school students. We employed two different groups: one using HAI with ChatGPT-4o, and the other using HHC, focusing on their performance in scientific knowledge and scientific thinking tasks. Results showed that while both collaboration methods significantly improved students' performance in solving scientific problems, HHC exhibited a greater effect size compared to HAI. The use of ChatGPT-4o was particularly noted for its ability to provide interactive learning experiences and guide students through problem solving; however, challenges such as inconsistencies in responses and image recognition hindered its effectiveness. Additionally, students tended to use ChatGPT-4o as a tool for obtaining answers rather than engaging in deeper collaborative explorations. This study underscores the importance of developing targeted training for students in effectively collaborating with AI tools. The results can also provide useful information for future exploration on methods to strengthen HHC by integrating AI tools, leveraging the strengths of both approaches to enhance education outcomes.

DOI: 10.1103/PhysRevPhysEducRes.21.010149

I. INTRODUCTION

In recent years, artificial intelligence (AI) has increasingly been integrated into education, becoming a focal point for academic research. Human-AI collaboration (HAI) is a key topic within this context, referring to the synergistic relationship between humans and AI, wherein their respective strengths are combined to achieve more effective educational outcomes [1,2]. With the advancement of conversational AI systems, human-machine collaboration has increasingly manifested as human-AI interaction [3], where such interaction is facilitated through AI-driven conversational feedback. In the education sector, conversational AI systems employ natural language generation technologies to simulate the instructional behaviors of human teachers, engaging with students in a way that enhances the overall effectiveness of the teaching process [4].

For instance, ChatGPT, an advanced conversational AI model developed from large language models, can simulate humanlike dialogues to provide interactive learning experiences. By generating coherent responses based on user input, ChatGPT guides students through problem-solving processes, helping them manage their learning pace and content progression, thereby enhancing learning effectiveness and maintaining motivation [5]. Additionally, ChatGPT adjusts the question difficulty according to the student's progress and

*Contact author: bao.15@osu.edu
†Contact author: atiq@deit.ecnu.edu.cn

performance, but this adjustment depends on how the student interacts with it. When prompted appropriately, ChatGPT provides a setting for nurturing critical thinking skills [6]. Moreover, ChatGPT can also enhance educators' productivity by providing personalized feedback [7,8]. ChatGPT's potential in subjects requiring high-level reasoning and problem solving, such as physics, indicates substantial future applications.

In fact, recent studies have explored ChatGPT's capabilities in solving physics problems, particularly across different versions of the model. For instance, Gregorcic and Pendrill examined ChatGPT-3.5's ability to address basic mechanics problems, finding that the model's responses contained erroneous physical reasoning [9]. Kortemeyer tested ChatGPT-3.5's performance on the Force Concept Inventory (FCI), finding that its score was double that of novice students [10]. In subsequent research, West reported that ChatGPT-4 demonstrated significantly improved performance, achieving 90%–100% accuracy on FCI problems, comparable to expert-level performance [11]. To evaluate ChatGPT's scientific reasoning, we proposed extending its problem-solving tasks to include both conceptual and open-ended questions. We conducted multiple studies on ChatGPT-4's performance in physics problem solving, which included the FCI [12], the Conceptual Survey of Electricity and Magnetism, and primitive physics problems [13], as well as questions from the Chinese National College Entrance Examination [14] and the Chinese Secondary School Entrance Examination [15]. The results of the study indicate that ChatGPT-4's accuracy in solving physics problems exceeds the average performance of middle school and high school students.

Despite its advantages, the application of ChatGPT in physics education faces notable challenges. First, ChatGPT may struggle with complex physics problems or university-level questions, leading to inadequate or problematic explanations [10,16–18]. Second, even ChatGPT-4 occasionally provides inconsistent responses to identical prompts, which can lead to uncertainty in the answer, posing a significant concern in applying it as a tutoring tool for physics problem solving. Based on the scoring results of GPT-4 and GPT-3.5 on physics exams, Yeadon and Halliday suggested that only weaker students might potentially improve their learning by utilizing AI [17]. Moreover, the lack of transparency and explainability in ChatGPT's thought process remains a critical issue, as students may not fully comprehend how AI-generated solutions are derived. This absence of explanatory derivation and the tendency to present conclusions as self-evident truths hinders the depth of students' understanding in physics and limits their ability to critically assess AI-generated answers. Excessive reliance on AI solutions could undermine students' independent thinking abilities, particularly when solving problems that require creative thinking. In such cases, AI tools may inadvertently inhibit students' active thinking [19]. These existing issues underscore the need for more empirical studies to better understand the practical implications of incorporating ChatGPT into educational contexts.

As AI tools mature, their integration into educational settings is inevitable, raising critical questions about how to design learning environments that maximize their benefits for students. A promising approach is to implement AI within well-established pedagogical frameworks, such as collaborative and cooperative learning—a method where students work together to achieve shared goals through social interaction, fostering skills like communication, critical thinking, and teamwork [20,21]. While collaborative learning is widely endorsed for its ability to deepen understanding and promote metacognition, it faces challenges such as free riding, group conflict, and logistical coordination [22]. These limitations create an opportunity to explore AI's potential as a collaborative peer, particularly given recent advancements in conversational AI.

Comparative studies of conversational AI tools, such as ChatGPT-4, Bing, and Bard, have demonstrated ChatGPT-4's superior performance in generating accurate, detailed responses with minimal internal contradictions [23]. This advantage is further supported by Polverini and Gregorcic's evaluation of visual explanation capabilities across eight conversational agents, where OpenAI's ChatGPT-4o achieved the highest overall performance [24]. While these findings highlight ChatGPT-4o's potential as a tool for facilitating human-AI interaction (HAI), empirical research remains limited on its capacity to act as an effective partner in helping students solve complex scientific problems. Specifically, there is a critical need to investigate how students should engage with ChatGPT-4o to enhance their physics problem-solving skills and whether its capabilities align with the demands of collaborative learning. To address these gaps, this study explores the following research questions:

RQ1: To what extent can ChatGPT-4o serve as a peerlike partner in assisting students with scientific problem solving?

RQ2: What typical roles do students assign to ChatGPT-4o during HAI when solving physics problems?

II. METHODS AND DESIGN

A. Participants

The study recruited 120 first-year high school students from two homogenous classes at a top-ranked high school in a southeastern city of China as the participants for this research. First-year high school students were the chosen demographic because students across China complete the study of "mechanical energy and its conservation law" and develop relatively advanced scientific thinking skills by their first year in high school. One class (62 students) served as the control group, where students engaged in human-human collaboration (HHC) to solve scientific problems. The other class (58 students) constituted the experimental group, where HAI was employed, with students working alongside ChatGPT-4o to solve scientific problems.

Currently, there are numerous generative AI models publicly available in China, such as SparkDesk and ERNIE Bot, and most students have experience using these models. However, these students have not had the opportunity to interact with ChatGPT-4o. In everyday learning, when encountering difficulties with homework, these students, like most of their peers in China, often use homework-assistant applications such as Homework Help and Xiaoyuan Search by uploading pictures of problems to obtain answers from a question database. These are important contextual backgrounds that may influence the students' methods of using HAI in learning.

B. Instruments

The scientific problems consisted of scientific knowledge questions and scientific thinking questions. Due to the large number of available knowledge and thinking test sets, it was not feasible to test every set. For practical purposes, we selected our own empirically validated tests on Work and Mechanical Energy (WME) test as the instrument for scientific knowledge questions, and the Inquiry in Scientific Thinking, Analytics, and Reasoning (iSTAR) test as the instrument for scientific thinking questions [25,26]. Each group was required to complete both test sets. After completing the tests, students took the Collaboration Self-Assessment Tool (CSAT) for evaluating their collaborative learning experiences [27].

The WME test consists of 20 multiple-choice questions, categorized into three conceptual groups: work (7 questions), conservation of mechanical energy (8 questions), and the work-energy theorem (5 questions). Most of the questions in the test have been adapted or modified from established instruments, including the Energy Concept Inventory, the Energy Concept Assessment, the Energy and Momentum Concept Survey, as well as several questions from China's national physics college entrance examinations. The validity and reliability of this instrument has been determined in Supplemental Material of the previous study [25].

The iSTAR test is designed to support inquiry-based learning by providing a structured framework for learning objectives and assessment. The framework guides the development and evaluation of instructional approaches aimed at cultivating skills in scientific thinking and reasoning. The original version of iSTAR consists of 35 multiple-choice questions distributed across three skill dimensions: control of variables (9 questions), data analytics (15 questions), and causal decision making (11 questions) [26]. To enhance usability, ongoing efforts aim to divide the full-length iSTAR test into two shorter parallel versions, each containing approximately 20 questions, allowing for random delivery while maintaining equivalent assessment outcomes. The version used in this study is one of the shorter parallel versions, consisting of 22 questions. Its validity and reliability have been confirmed in previous research [28].

The CSAT was originally developed for use by teacher candidates and cooperating teachers. It assesses 11 key collaboration skills: Contribution, Motivation/Participation, Quality of Work, Time Management, Team Support, Preparedness, Problem Solving, Team Dynamics, Interactions with Others, Role Flexibility, and Reflection. Each skill is rated on a 1–4 scale, with detailed descriptors provided for each level [27]. Naturally, the CSAT is not exclusively intended to evaluate collaboration between teacher candidates and cooperating teachers; it is also suitable for use with other education-related groups. In this study, since all materials were provided to participants in advance and no prior preparation was required, the skill of Preparedness was omitted. As a result, the version of the CSAT used in this study assesses only ten key collaboration skills.

C. Experimental design

First, students independently answered the scientific problems using a black pen. Then, the control group discussed the questions with their seatmates, while the experimental group discussed the questions with ChatGPT-4o. Afterward, participants drew on information and insights from the discussions to correct their answers using a red pen. Hereafter, the answers written with a black pen will be referred to as "pretest answers," and those written with a red pen as "post-test answers." Finally, both groups completed the CSAT to assess their collaborative learning experiences. The following content provides further details on the experimental process:

- Students first spent approximately 40 min independently completing the scientific knowledge test (WME) with a black pen.
- They were then given about 20 min to discuss uncertain or difficult questions with their seatmate (control group) or ChatGPT-4o (experimental group) and decide whether to modify their answers using a red pen.
- One to two days later, students spent around 40 min independently completing the scientific thinking test (iSTAR) with a black pen.
- Similarly, they had about 20 min to discuss uncertain or difficult questions and decide whether to modify their answers with a red pen.
- After completing the scientific thinking test, each student spent 10 min completing the CSAT.

In total, each student spent approximately 130 min completing the entire testing process.

D. Data collection

Data were collected by measuring the accuracy rate of the control and experimental groups in solving different types of scientific problems. The CSAT was used to compare the learning experiences of HHC with HAI. Additionally, the text generated in the ChatGPT-4o dialogues was analyzed to summarize the characteristics and strategies of HAI. This analysis included identifying the types of questions posed by students, follow-up prompts for clarification, and the problem-solving approaches suggested by ChatGPT-4o. We also examined the dynamics of the interactions, such as the frequency and depth of iterative questioning by students, the engagement level in response to AI-generated suggestions, and the influence of AI recommendations on students' decision-making processes. By analyzing these interactions, we were able to gain a deeper understanding of both the effective elements that contribute to successful HAI and the challenges that hinder deeper engagement, thereby providing insights into areas for future improvement.

III. ANALYSIS AND RESULTS

A. Analysis of the effectiveness of HAI in solving scientific problems

During preliminary data screening, it was observed that several participants in the HHC group exhibited no peer interaction, while a subset of students in the HAI group engaged in off-topic discussions with ChatGPT-4o unrelated to the test objectives. To ensure data integrity, these individuals were excluded from subsequent analysis, resulting in adjusted final sample sizes of $N = 60$ (HHC) and $N = 55$ (HAI). The accuracy rates in solving scientific knowledge and scientific thinking problems before and after collaborative problem solving are summarized, with paired-samples t tests conducted to examine performance gains across different collaboration methods (HHC and HAI), as presented in Table I.

An independent samples t-test was conducted to compare the pretest results, representing the independent problem-solving performance of participants in the control and experimental groups. The results showed no significant difference between the two groups in the scientific knowledge dimension ($t = 1.254$, $p = 0.213$). However, in the scientific thinking dimension, the control group scored significantly higher than the experimental group ($t = 5.339$, $p < 0.001$). This result confirms that the two classes selected were homogenous in the scientific knowledge dimension but not in the scientific thinking dimension. This finding is consistent with our previous research, which suggested that small differences in the scientific knowledge dimension tend to be amplified in the scientific thinking dimension [28]. The underlying reasons for this phenomenon remain to be explored. However, since our primary focus is on the changes of performances through different collaboration methods, this disparity does not affect the analysis of the results.

The paired-samples t tests results indicate that, while both collaboration methods significantly improved students' success in solving scientific problems, HHC showed a small to moderate level of effect size (0.2–0.5), whereas HAI exhibited a smaller effect size (< 0.2) [29]. In summary, the gain effects in solving scientific problems were greater in HHC compared to HAI.

B. Analysis of the characteristics of HAI in solving scientific problems

1. Statistical analysis of dialogue-starting methods

Although ChatGPT-4o has achieved voice interaction capabilities, this feature is currently available only on the Mac version. For convenience, this study employed the text-based chat interface for HAI in solving scientific problems. In the chat interface, students could either upload screenshots, copy and paste the text of the question, or input their uncertainties and challenges encountered during the problem-solving process in their own words to interact with ChatGPT-4o.

Through the analysis of the dialogue-starting records during HAI, three types of interaction methods were identified: screenshot based, text copying, and question transcription. The proportion of these three methods was

TABLE I. The accuracy rates and paired-samples t test results for scientific knowledge and scientific thinking during the pretest and post-test in HHC and HAI groups.

		Accuracy rates of scientific knowledge (WME)		Accuracy rates of scientific thinking (iSTAR)	
Test	Group	Mean ($N = 60$)	SD	Mean ($N = 55$)	SD
Pretest	HHC	0.758	0.147	0.732	0.123
	HAI	0.716	0.205	0.545	0.225
Post-test	HHC	0.815	0.137	0.759	0.118
	HAI	0.726	0.210	0.558	0.224

		t	d	p	t	d	p
Paired-samples t tests for pre-post comparisons	HHC	7.572	0.404	0.000	3.102	0.227	0.003
	HAI	2.701	0.053	0.009	2.105	0.040	0.040

TABLE II. Statistical analysis of dialogue-starting methods in HAI for solving scientific problems.

	Screenshot based		Text copying		Question transcription	
Question	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
WME	41	74.5%	11	20.0%	3	5.5%
iSTAR	35	63.6%	13	23.6%	7	12.7%
Total	76	69.1%	24	21.8%	10	9.1%

similar when solving scientific knowledge and scientific thinking problems, as shown in Table II. Specifically, the screenshot method was the most frequently employed, representing approximately 70% of the interactions. This was followed by the text-copying method, which accounted for around 20% of the interactions. The question transcription method was the least utilized, comprising the smallest proportion of interactions.

The results suggest that most students have not yet realized the generative capabilities of artificial intelligence, still treating ChatGPT-4o as a tool for obtaining answers from a database, rather than viewing it as an intelligent partner with whom they can discuss and explore problems. For example, some students, after sending a screenshot or copied text of the question to ChatGPT-4o, received an incorrect answer from the model. Instead of further questioning the response, they chose to change their originally correct answer to the incorrect one provided by ChatGPT-4o.

A small number of students, however, were able to transcribe or follow up on questions with a targeted goal, as demonstrated by the following two students:

Student A (regarding WME Q20): "In a situation where momentum is conserved, is mechanical energy also conserved?"

ChatGPT-4o: Whether mechanical energy is conserved when momentum is conserved depends on the specific situation. Momentum conservation means that the total momentum of a system remains constant in the absence of external forces, while mechanical energy conservation means that the total mechanical energy remains constant in the absence of non-conservative forces (such as friction). Since the conditions for these two conservations are different, we cannot directly conclude that mechanical energy is conserved just because momentum is.

There are two cases where both momentum and mechanical energy conservation can occur simultaneously:

1. *Completely elastic collisions: In this case, both momentum and mechanical energy are conserved, meaning the total momentum and total mechanical energy before and after the collision are equal.*
2. *Systems without non-conservative forces: In systems where only conservative forces (like gravity and spring forces) are at work, both mechanical energy and momentum can be conserved.*

However, in the following case, momentum is conserved but mechanical energy is not:

Inelastic collisions: In such collisions, momentum is still conserved, but part of the mechanical energy is converted into internal energy, sound energy, or other forms of energy, leading to mechanical energy not being conserved.

Therefore, while momentum conservation is an important condition, it does not guarantee that mechanical energy will also be conserved. It is necessary to analyze the specific physical process to determine whether both conservation laws apply".

Student B (regarding iSTAR Q15, appropriate transcription): "To ensure a rose blooms beautifully, it requires sufficient fertilizer, plenty of sunlight, and adequate watering. John and Alex are neighbors and both enjoy roses. In the spring, they bought the same variety of roses from a local nursery and planted them at home. By summer, John's roses bloomed beautifully, while Alex's roses grew only moderately. Based on the above information, which of the following can be confirmed?

a. *John gave his roses sufficient fertilizer.*
b. *Alex did not give his roses enough fertilizer.*
c. *Alex's roses did not get enough sunlight.*
d. *Alex's roses did not get adequate watering.*
e. *Any of 'a', 'b', 'c', or 'd'.*
f. *'a' and any of 'b', 'c', or 'd'.*
g. *'b', 'c', and 'd'.*
h. *All of 'a', 'b', 'c', and 'd'.*
i. *None of the above".*

ChatGPT-4o: "Based on the information provided, John and Alex bought the same variety of roses and planted them at home. By summer, John's roses bloomed beautifully, while Alex's roses grew moderately. From this, we can infer that John took good care of his roses, while Alex may have missed something.

For a rose to bloom beautifully, it requires sufficient fertilizer, plenty of sunlight, and adequate watering. Since Alex's roses grew only moderately, it indicates that at least one of these three conditions was not met. However, we cannot be certain which one. It could be insufficient fertilizer, lack of sunlight, or improper watering—any one or multiple causes.

Analysis of options:

a. *John gave his roses enough fertilizer.—This is possible, but we cannot be certain that this is the only reason.*

b. *Alex did not give his roses enough fertilizer.—This is possible, but we cannot be sure this is the only reason.*
c. *Alex's roses did not get enough sunlight.—This is possible, but we cannot be sure this is the only reason.*
d. *Alex's roses did not get adequate watering.—This is possible, but we cannot be sure this is the only reason.*
e. *Any of 'a', 'b', 'c', or 'd'.—Correct, because any of these could explain why Alex's roses grew moderately.*
f. *'a' and any of 'b', 'c', or 'd'.—Incorrect, because we cannot confirm that 'a' is definitely correct.*
g. *'b', 'c', and 'd'.—Incorrect, because we cannot be certain which or how many of these reasons are true.*
h. *All of 'a', 'b', 'c', and 'd'.—Incorrect, because we cannot confirm that all the conditions were unmet.*
i. *None of the above.—Incorrect, because option 'e' is correct.*

Therefore, the most correct answer is: e. Any of 'a', 'b', 'c', or 'd'".

Student B: "I don't think your explanation is right".

ChatGPT-4o: "I understand your concern. Let's analyze the question and options again carefully. The problem states that John's roses bloomed beautifully, while Alex's roses grew only moderately. We can draw several important conclusions:

To bloom beautifully, roses need enough fertilizer, plenty of sunlight, and adequate watering. John's roses bloomed beautifully, which means they received sufficient fertilizer, sunlight, and watering. Alex's roses grew moderately, indicating that one or more of these three factors were lacking.

Now, let's reanalyze the options:

a. *John gave his roses enough fertilizer.—This can be confirmed, as John's roses bloomed beautifully.*
b. *Alex did not give his roses enough fertilizer.—This cannot be confirmed, as the moderate growth could be due to other reasons.*
c. *Alex's roses did not get enough sunlight.—This cannot be confirmed, as it could be due to other reasons.*
d. *Alex's roses did not get adequate watering.—This cannot be confirmed, as it could be due to other reasons.*
e. *Any of 'a', 'b', 'c', or 'd'.—Incorrect, because option 'a' is confirmed, while 'b', 'c', and 'd' are uncertain.*
f. *'a' and any of 'b', 'c', or 'd'.—Correct, as 'a' is confirmed and any of 'b', 'c', or 'd' is possible.*
g. *'b', 'c', and 'd'.—Incorrect, as 'a' is confirmed, and 'b', 'c', and 'd' are uncertain.*
h. *All of 'a', 'b', 'c', and 'd'.—Incorrect, as 'a' is confirmed, but 'b', 'c', and 'd' cannot all be true.*
i. *None of the above.—Incorrect, as 'f' is the correct answer.*

Therefore, the correct option should be: f. 'a' and any of 'b', 'c', or 'd'".

Student B: "I think your first explanation was better".

ChatGPT-4o: (Repeated the initial answer)

Student B: "I believe option 'i' is correct".

ChatGPT-4o: "Alright, let's analyze this question in detail again.

The problem provides the following information:

John and Alex bought the same variety of roses.

John's roses bloomed beautifully, while Alex's roses grew only moderately.

To bloom beautifully, roses need sufficient fertilizer, plenty of sunlight, and adequate watering. We know that John's roses grew beautifully, while Alex's roses grew moderately. This indicates that John took care of all three factors (fertilizer, sunlight, and watering), while Alex missed at least one. However, we cannot be certain which one or how many factors were lacking.

Here's the detailed analysis for each option:

a. *John gave his roses enough fertilizer.—This is possible, but we cannot be sure it's the only reason.*
b. *Alex did not give his roses enough fertilizer.—This is possible, but we cannot be sure it's the only reason for the moderate growth.*
c. *Alex's roses did not get enough sunlight.—This is possible, but we cannot be sure it's the only reason.*
d. *Alex's roses did not get adequate watering.—This is possible, but we cannot be sure it's the only reason.*
e. *Any of 'a', 'b', 'c', or 'd'.—While each option is possible, we cannot confirm any single one, so this option is incorrect.*
f. *'a' and any of 'b', 'c', or 'd'.—Incorrect, because we cannot confirm that 'a' is definitely correct, but 'b', 'c', or 'd' is possible.*
g. *'b', 'c', and 'd'.—Incorrect, because we cannot confirm which or how many of the reasons are correct.*
h. *All of 'a', 'b', 'c', and 'd'.—Incorrect, because we cannot confirm that all conditions were unmet.*
i. *None of the above.—This is correct because none of the options are fully accurate based on the information provided.*

Therefore, the most correct answer is: i. None of the above".

Student B: "Okay, thank you".

ChatGPT-4o: "You're welcome! I'm glad I could help. If you have any more questions, feel free to ask. Happy studying!"

After the test, student B wrote a comment on their test paper: "ChatGPT-4o is like a flip-flopper, it goes whichever way you guide it." This suggests that the student started to question ChatGPT-4o's critical thinking abilities. We encountered a similar issue when using ChatGPT-3.5 to solve FCI questions [30].

2. Analysis of modifications in collaborative problem solving

The AI chatbot and human partner can provide both correct and incorrect answers during a discussion.

The correctness of answers does not definitively lead the other collaborator to modify his or her initial responses toward the correct or incorrect direction. In light of this complex process, the analysis was directed toward the modifications made by students following their collaborations rather than focusing on the accuracy of the information provided by the chatbot or human partners. This approach was considered to provide a more direct indication of collaborative effectiveness, as it revealed whether the interactions had positive or negative impacts on students' reasoning in evaluating their initial answers.

Before engaging in collaborative problem solving, students had independently attempted to solve the problems. Following discussions, whether through HAI or HHC, five potential outcomes were observed: incorrect to correct, incorrect to incorrect, correct to incorrect, correct unchanged, and incorrect unchanged. Since students used black pens for their initial independent problem solving and red pens for modifications made after collaboration, only the first three types of outcomes were reflected in red. Consequently, the statistical analysis of modifications during collaborative problem solving focused exclusively on these three outcomes, as shown in Table III.

Table III illustrates that after collaborative problem solving, the majority of modifications by students were made to correct their initial incorrect answers. However, a subset of students failed to truly resolve their problems, instead changing their answers from incorrect options to other incorrect ones. Additionally, a small number of students, very likely due to a lack of confidence in their original answers, were misled by ChatGPT-4o or their peers during the collaboration and changed their correct answers to incorrect ones.

By comparing the modification outcomes between HAI and HHC, it is evident that HHC has a much higher frequency and proportion of "incorrect to correct" changes ($N = 127,\ \ P = 77.9\%$) compared to HAI $N = 38$, $P = 55.9\%$), with a statistically significant difference ($t = -2.488,\ \ p = 0.008$). Conversely, in the case of "correct to incorrect" changes, the frequency and proportion in HHC ($N = 11$, $P = 6.8\%$) are lower than in HAI ($N = 14,\ \ P = 20.6\%$), though the difference is not statistically significant ($t = 1.049,\ \ p = 0.847$). This suggests that, in terms of collaborative effectiveness, HAI is significantly less effective than HHC. ChatGPT-4o behaves like an overconfident student who does not always have the correct views but is confident enough to provide guidance on problem solving, which is useful in some cases but may also mislead students in others. For example, prior to engaging in HAI, student C correctly selected answer A. However, after interacting with ChatGPT-4o, the student changed the answer to the incorrect option C.

Student C: (sends a screenshot of WME question 12 to ChatGPT-4o), see Fig. 1.

ChatGPT-4o: "To solve this question, we need to analyze the work done by the forces acting on the two blocks.

Given:

Two blocks, A and B, are on a frictionless horizontal surface.

A force, F_0, acts on both blocks: F_0 acts to the right on block A, and $-F_0$ (or F_0 to the left) acts on block B.

Each block moves a distance of d_0.

The work done by a force is given by the formula:

$$W = Fd\cos(\theta),$$

where F is the force, d is the distance, and θ is the angle between the force and the direction of motion.

For block A:

The force F_0 and displacement d_0 are in the same direction.

Thus, $\theta = 0^o$, and $\cos(\theta) = 1$.

The work done on block A is:

$$W_A = F_0 d_0 \cos(0^o) = F_0 d_0.$$

For block B: The force $-F_0$ and displacement d_0 are in opposite directions.

Thus, $\theta = 180^o$, and $\cos(180^o) = -1$.

TABLE III. Statistical analysis of modifications in collaborative problem solving.

Collaboration Style	Question	Incorrect to correct		Incorrect to incorrect		Correct to incorrect	
		Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
HAI	WME	19	63.3%	5	16.7%	6	20.0%
	iSTAR	19	50.0%	11	28.9%	8	21.1%
	Total	38	55.9%	16	23.5%	14	20.6%
HHC	WME	89	78.8%	15	13.3%	9	8.0%
	iSTAR	38	76.0%	10	20.0%	2	4.0%
	Total	127	77.9%	25	15.3%	11	6.8%

12. As shown in the figure, two identical wooden blocks are pushed by two hands on a horizontal frictionless surface with constant forces F_0 in opposite directions. When both wooden blocks are pushed to move with a displacement of d_0 towards each other, how much total work is done by the two forces?

A. $2F_0d_0$
B. F_0d_0
C. 0
D. $-F_0d_0$
E. $-2F_0d_0$

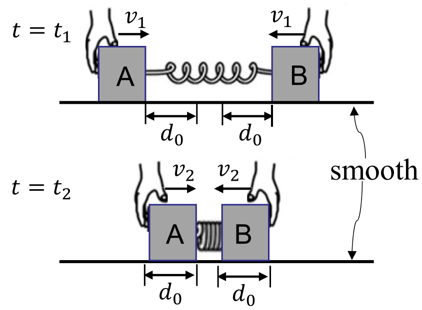


FIG. 1. A screenshot of WME question 12.

The work done on block B is:

$$W_B = F_0 d_0 \cos(180^o) = -F_0 d_0.$$

The total work done by the two forces is the sum of the work done on each block:

$$W_{\text{total}} = W_A + W_B = F_0 d_0 + (-F_0 d_0) = 0.$$

Therefore, the total work done by the two forces is 0, and C is the correct answer".

Student C: (changes his answer from A to C)

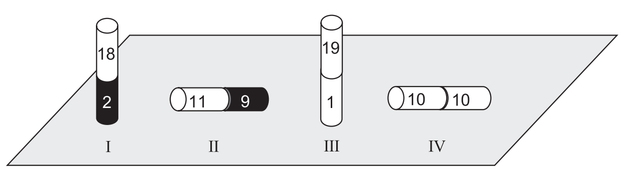
3. Challenges in HAI for problem solving

(1) *Inaccurate image recognition.* The image recognition feature was introduced as part of ChatGPT-4 around October 2023. This represented a significant advancement in ChatGPT's multimodal capabilities, enabling it to not only process text but also understand and generate responses to image-related queries. However, the accuracy of image recognition remains suboptimal.

For example, *Student D:* (sends a screenshot of iSTAR question 2 to ChatGPT-4o), see Fig. 2.

ChatGPT-4o: "This question describes an experimental problem involving transparent vertical glass tubes. Based on the description and the illustration, we can deduce the following steps and analysis:

2. Four glass tubes that each contains twenty fruit flies are sealed shut. Tubes I and II are partially covered with black paper; Tubes III and IV are not covered. These tubes are placed on a clear horizontal glass table with tubes I and III standing up on their ends and tubes II and IV lying flat as shown. For five minutes the tubes are exposed to red light coming from all directions. At the end of the five minutes, the numbers of flies in the covered and uncovered parts of each tube are counted as shown below. Note that a total of 20 fruit flies are still in each tube.



Red light comes from all directions. Table is made of clear glass.

This experiment shows that flies respond to ("respond to" means move towards or away from):

a. red light but not gravity
b. gravity but not red light
c. both red light and gravity
d. neither red light nor gravity
e. cannot be determined

FIG. 2. A screenshot of iSTAR question 2.

Experimental Steps:

Each of the four tubes is sealed with 20 fruit flies.

Tubes I and II have their upper half (including the top) covered with black paper, while tubes III and IV have no blackened areas.

The four tubes are placed vertically in a horizontally divided glass-separated area (tubes I and III are in region 1, and tubes II and IV are in region 2).

The fruit flies are allowed to disperse within the tubes for 5 minutes.

The distribution of fruit flies in each tube is recorded.

Illustration Analysis:

The illustration shows the distribution of fruit flies in the four tubes:

Tube I: 1 fruit fly

Tube II: 11 fruit flies

Tube III: 9 fruit flies

Tube IV: 10 fruit flies

Question: Under what conditions do the fruit flies climb up the tube (i.e., move toward the indicated direction or away)?

a. *When there is light*
b. *When there is no light*
c. *Unknown environmental factors*
d. *Color factors*

Answer Analysis:

From the illustration, it can be seen that:

In the illuminated tubes III and IV, the distribution of fruit flies is relatively even (9 and 10, respectively).

In the partially blackened tubes I and II, the fruit flies are primarily concentrated in the non-blackened areas (1 in tube I and 11 in tube II).

This suggests that fruit flies tend to move toward light or away from dark areas. Therefore, the correct answer is: a. When there is light".

Student D: "The analysis is incorrect; the number of fruit flies in the four tubes doesn't match".

ChatGPT-4o: "Let's analyze the problem and the illustration again:

Experiment Description and Illustration:

Experimental Steps:

Each of the four tubes contains 20 fruit flies and is sealed.

The tops of tubes I and II are blackened, while tubes III and IV have no blackened areas. The four tubes are placed vertically in a horizontally transparent glass-separated area. The distribution of fruit flies in each tube is recorded every 5 minutes.

Illustration:

Tube I: 2 fruit flies

Tube II: 11 fruit flies

Tube III: 9 fruit flies

Tube IV: 10 fruit flies

Question: Under what conditions do the fruit flies climb up the tube (i.e., move toward the indicated direction or away)?

a. *When there is light*
b. *When there is no light*
c. *Unknown environmental factors*
d. *Color factors*

Detailed Analysis:

Tube I (blackened top): 2 fruit flies

Tube II (blackened top): 11 fruit flies

Tube III (unblackened): 9 fruit flies

Tube IV (unblackened): 10 fruit flies

Conclusion: The data indicates that more fruit flies are found in the unblackened (illuminated) tubes III and IV, while fewer are present in the blackened (non-illuminated) tubes I and II. Therefore, it can be concluded that fruit flies tend to climb up the tubes in illuminated conditions.

Correct Answer: a. When there is light".

From the above example, it can be seen that ChatGPT-4o not only had issues extracting the image information in the question but also modified the answer choices. Even after the student pointed out its mistakes, ChatGPT-4o made minor adjustments, yet the response remained incorrect. This issue is prevalent in questions containing images.

Therefore, while ChatGPT has shown a relatively high accuracy rate in solving some scientific problems, this is primarily based on researchers converting image information into text or excluding questions with images from the tests [9,10,13,31]. However, in this study, the methods of human-AI interaction were diverse. The use of screenshots and images in the questions required ChatGPT-4o to extract this information. This extraction process significantly increased the difficulty in analyzing and solving the problems. As a result, the lower accuracy of ChatGPT-4o observed in this study is not inconsistent with the conclusions of previous research.

(2) *Mismatch between problem analysis and answer options.* While ChatGPT-4o exhibits significantly enhanced logical reasoning capabilities compared to earlier versions, it is surprising that, despite being able to correctly analyze the problem and derive the correct answer, it occasionally fails to accurately match the answer with the provided options. For example:

Student E: (sends a screenshot of iSTAR question 11 into ChatGPT-4o), see Fig. 3.

ChatGPT-4o: "Based on the description in the image, let's analyze the logical reasoning behind this question:

Problem Description:

Albert and Jon are playing a new card game. Each card has an integer on one side and is either gray or white on the other side. Jon made a hypothesis: If one side of the card shows an even number, then the other side must be gray.

We need to verify whether Jon's hypothesis is correct and determine which cards need to be flipped. The problem presents four cards showing the following: 3, 8, white, and gray.

11. Albert is playing a new card game with his friend Jon. Each card contains a number on one side and is either white or gray on the other side. After a while, Jon claims a discovery: "If a card has an even number on one side, then it is gray on the other side."
In one case, four cards are presented to Albert: 3, 8, white, and gray as shown below.
Which card(s), when turned over, is(are) useful to determine whether Jon's statement is false?

a. the 3 card only
b. the 8 card only
c. the 3 card and the white card
d. the 3 card and the gray card
e. the 8 card and the white card
f. the 8 card and the gray card
g. all four cards

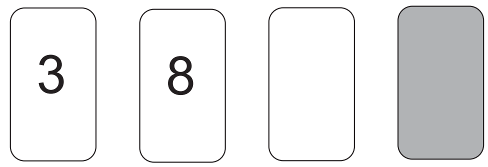


FIG. 3. A screenshot of iSTAR question 11.

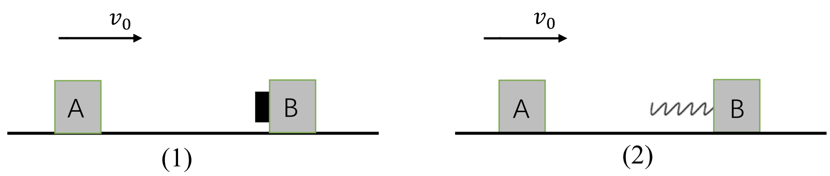
To verify Jon's hypothesis, we need to check the following:
If a card shows an even number, the other side must be gray.
If a card shows white, the other side must not be an even number.
Analysis of the cards provided:
The first card shows 3, which is an odd number, so it does not affect the hypothesis and does not need to be flipped.
The second card shows 8, which is an even number, so we need to check if the other side is gray.
The third card shows white, so we need to check if the other side is an even number (as the white side cannot show an even number).
The fourth card shows gray, which already fits the hypothesis, so it does not need to be flipped.
Thus, we need to flip the second and third cards to verify whether Jon's hypothesis is correct. The correct answer is: c. Flip 8 and the white card".
Student E: "The option recognition is wrong".
ChatGPT-4o: "I apologize for the recognition error. Here is the correct identification and analysis of the options:
Problem Description:
Albert and Jon are playing a new card game. Each card has an integer on one side and is either gray or white on the other side. After a while, Jon made a hypothesis: "If one side of the card shows an even number, then the other side must be gray".
Assuming Albert randomly draws four cards showing the following: 3, 8, white, and gray (as shown below). To verify whether Jon's hypothesis is false, which card(s) should be flipped?
Options:
a. *Flip only 3.*
b. *Flip only 8.*
c. *Flip 3 and the white card.*
d. *Flip 8 and the white card.*
e. *Flip 3 and the gray card.*
f. *Flip 8 and the gray card.*
g. *Flip all four cards.*
h. *None of the above.*
Analysis:
Flipping the card showing 3 is unnecessary because it is an odd number and does not impact the verification of the hypothesis.
Flipping the card showing 8 is necessary because it is an even number, and we need to check if the other side is gray.
Flipping the white card is necessary because we need to verify if the opposite side shows an even number.
Flipping the gray card is unnecessary because the gray side already fits the hypothesis and does not require verification.
The correct answer is: d. Flip 8 and the white card".
Student E: "Please recognize the problem again".
ChatGPT-4o: (Although ChatGPT-4o provided the correct reasoning process, it still selected an answer that did not match the given options in the original question.)

From the above example, it is evident that ChatGPT-4o provides logically consistent answers, but after arriving at the correct answer, it incorrectly matches the options. Even after the student's correction, it stubbornly switched options d and e without acknowledging the mistake. This issue is also present in the HAI in WME.

Student F: (sends a screenshot of WME question 20 into ChatGPT-4o), see Fig. 4.

ChatGPT-4o: "To solve this problem, we need to analyze the conservation of mechanical energy in two scenarios:
Scenario (1):
Given: Block A moves to the right with an initial velocity v_0. Block B is initially at rest. The collision between A and B is completely inelastic due to the sticky puck.
Analysis:
Before the collision, the total mechanical energy is the kinetic energy of block A, $\frac{1}{2}mv^2$.

20. Two blocks, A and B, have the same mass and are on a frictionless horizontal plane. Block A moves to the right with an initial velocity v_0, and block B is initially at rest.
(1) A massless sticky puck is attached to the left side of block B. A and B then collide and stick together while moving to the right. Block A and B (including the puck) are treated as a system.
(2) A massless spring is attached to the left side of block B. A and B then collide. Consider the process from A is in contact with the spring until the spring is compressed the most. Here, blocks A and B and the spring are treated as a system.



Which of the following statement about whether mechanical energy is conserved in these two situations is true?
A. Mechanical energy is not conserved in (1), but it is conserved in (2).
B. Mechanical energy is conserved in (1), but it is not conserved in (2).
C. Mechanical energy is conserved in both (1) and (2).
D. Mechanical energy is not conserved in both (1) and (2).

FIG. 4. A screenshot of WME question 20.

During the collision, since it is inelastic, part of the kinetic energy is converted into internal energy (heat, deformation, etc.).
After the collision, blocks A and B move together at a common velocity.
Conclusion:
Mechanical energy is not conserved because part of the kinetic energy is transformed into other forms of energy during the inelastic collision.
Scenario (2):
Given: Block A moves to the right with an initial velocity v_0. Block B is initially at rest and connected to a spring. The collision is elastic and involves the spring.
Analysis:
Before the collision, the total mechanical energy is the kinetic energy of block A.
During the collision, mechanical energy is transferred between kinetic energy and elastic potential energy (as the spring compresses).
At the point of maximum spring compression, the kinetic energy of the blocks is momentarily zero, and all mechanical energy is stored as potential energy in the spring. As the spring releases, energy is again converted back into kinetic energy.
Conclusion:
Mechanical energy is conserved throughout the process because the system includes elastic potential energy, and no energy is lost.
Therefore, the correct answer is: B. (1) Mechanical energy is not conserved during the process; (2) Mechanical energy is conserved during the process".

Once again, ChatGPT-4o arrived at the correct answer: mechanical energy is not conserved in (1), mechanical energy is conserved in (2), yet incorrectly selected option B and tampered with the content of this option.

The mismatch between problem analysis and option may seem minor, but this phenomenon is common in HAI and rarely occurs in HHC. Such mismatches or unauthorized modifications to the answer options reduce students' trust in ChatGPT-4o and increase the cognitive load in HAI. Consequently, students' experience in using AI and the effectiveness of their collaborations may be negatively affected.

(3) *Consistency issues in ChatGPT-4o's responses.* In the most basic version of the text-generation process, the same prompt will consistently produce the same output—the one deemed most likely based on the statistics of the training data and the given prompt. However, randomness can be introduced by adjusting a parameter known as "temperature." In the ChatGPT, the temperature is automatically set to a value above zero, allowing the chatbot to generate varied responses even when presented with the same prompt [32]. Some researchers, in pursuit of a more accurate result, have gone to great lengths to conduct up to 60 tests and then average the correct response rate [33]. The same issue is present in this study, where different students, despite interacting with ChatGPT-4o using the same method for the same question, received different responses.

For example, regarding WME question 12, two students interacted with ChatGPT-4o via screenshot and received completely different answers. One response, as seen in the previous reply to student C, was:

$$W_{\text{total}} = W_A + W_B = F_0 d_0 + (-F_0 d_0) = 0.$$

Thus, the total work done by the two forces is 0, and C is the correct answer.

In contrast, ChatGPT-4's other response was: *"To solve this problem, we need to calculate the work done by the two forces as they move the two blocks a distance of d_0.*

Given:

Blocks A and B are each pushed by a constant force F_0 in opposite directions.

Each block moves a distance of d_0.

The surface is smooth, with no friction.

Let's break down the work done on each block:

For block A: Force F_0 acts to the right, and block A moves a distance d_0 to the right. The work done on block A is $W_A = F_0 d_0$.

For block B: Force F_0 acts to the left, and block B moves a distance d_0 to the left. The work done on block B is $W_B = F_0 d_0$.

Since both forces do work in the direction of the blocks' motion, the total work done by the two forces is:

$$W_{\text{total}} = W_A + W_B = F_0 d_0 + F_0 d_0 = 2F_0 d_0.$$

Therefore, the correct answer is: A. $2F_0 d_0$."

Another example: In iSTAR question 15, two students used the method of copying the question text to interact with ChatGPT-4o.

15. In order to produce good blossoms, roses must have plenty of fertilizer, full sunlight, and proper watering. John and Alex are neighbors who both like roses. During the springtime, they each planted several rose bushes of the same kind bought from a local nursery. When summer came, the roses at John's house were flowering beautifully, while the roses at Alex's house were struggling. Based on this result, which of the following statements must be true?

a. *John applied plenty of fertilizer to his roses.*
b. *Alex didn't apply enough fertilizer to his roses.*
c. *Alex's roses didn't receive enough sunlight.*
d. *Alex's roses didn't receive proper watering.*
e. *"a" and any one of "b", "c", and "d"*
f. *"b", "c", and "d"*
g. *All of "a", "b", "c", and "d"*
h. *None of "a", "b", "c", and "d"*

ChatGPT-4o Response 1:

"Based on the information provided, John's roses bloomed beautifully, while Alex's roses grew only moderately. From this, we can infer that John's roses received sufficient fertilizer, adequate sunlight, and proper watering, while Alex's roses lacked at least one key factor. However, based on this information alone, we cannot determine which specific factor caused Alex's roses to grow poorly. Therefore, the most reasonable answer is: e. Any of 'a', 'b', 'c', or 'd'".

ChatGPT-4o Response 2:

"To answer this question, we need to carefully analyze the given information and the meaning of each option.

Given:

Roses need sufficient fertilizer, plenty of sunlight, and adequate watering to bloom well.

John's roses bloomed beautifully.

Alex's roses grew only moderately.

Analysis:

John's roses bloomed beautifully, which indicates that he provided all the necessary conditions: sufficient fertilizer, adequate sunlight, and proper watering.

Alex's roses grew moderately, suggesting that Alex might have failed to meet one or more of the requirements for fertilizer, sunlight, or watering.

Option Analysis:

a. *John gave his roses enough fertilizer.*
   This can be confirmed, as roses need enough fertilizer to bloom well.
b. *Alex didn't give his roses enough fertilizer.*
   This isn't necessarily true, as Alex's roses may have grown moderately due to insufficient fertilizer, sunlight, or watering.
c. *Alex's roses didn't get enough sunlight.*
   This isn't necessarily true for the same reasons as above.
d. *Alex's roses didn't get adequate watering.*
   This also isn't necessarily true for the same reasons.
e. *Any of 'a', 'b', 'c', or 'd'.*
   This is incorrect because 'a' can be confirmed, whereas 'b', 'c', and 'd' are uncertain.
f. *'a' and any of 'b', 'c', or 'd'.*
   This is also incorrect because 'b', 'c', and 'd' are uncertain.
g. *'b', 'c', and 'd'.*
   This is incorrect because all of 'b', 'c', and 'd' are uncertain.
h. *'a', 'b', 'c', and 'd'.*
   This is incorrect because 'b', 'c', and 'd' are uncertain.
i. *None of the above.*
   This is also incorrect because 'a' can be confirmed.

Conclusion: The only thing that can be confirmed is that 'John gave his roses enough fertilizer', so the correct answer is a. John gave his roses enough fertilizer".

It can be observed that for the same dialogue content and format, ChatGPT-4o provides different responses. This inconsistency can significantly reduce students' trust in ChatGPT-4o and may even cause students who had originally a correct problem-solving approach to become confused during the interaction with the chatbot.

4. Student self-assessment analysis in collaborative problem solving

The original version of the CSAT contains 11 self-assessment questions, each rated on a scale of 1–4, with a maximum score of 44 points. Ofstedal and Dahlberg

TABLE IV. Self-assessment score statistics and comparison in collaborative problem solving. The numbers in parentheses in the table represent the standard deviation of the means.

Group	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	Total
HHC	3.20	2.67	2.83	3.10	3.00	3.27	2.87	3.13	2.95	2.95	29.97
($N = 60$)	(0.78)	(1.10)	(0.83)	(0.71)	(0.78)	(0.61)	(0.79)	(0.70)	(0.95)	(0.91)	(5.63)
HAI	2.73	2.45	3.00	2.78	2.75	2.96	2.75	3.09	2.89	2.78	28.18
($N = 55$)	(1.03)	(1.12)	(0.86)	(0.90)	(0.95)	(0.86)	(0.91)	(0.87)	(1.05)	(0.99)	(7.39)
t	2.799	1.025	−1.059	2.124	1.577	2.198	0.765	0.290	0.318	0.947	1.464
p	0.006	0.308	0.292	0.036	0.117	0.030	0.446	0.772	0.751	0.346	0.146

provided a scoring guide: 11–25 (collaboration skills are emerging), 26–34 (collaboration skills are developing), and 35–44 (collaboration skills are established) [27]. Since the CSAT used in this study only contains 10 questions, the scoring guide was adjusted proportionally: 10–23 (collaboration skills are emerging), 24–31 (collaboration skills are developing), 32–40 (collaboration skills are established). The average scores for each group on individual CSAT questions and the overall score were calculated, followed by an independent samples t-test. The results are shown in Table IV.

Overall, there is no significant difference in the self-assessment scores between the HHC group and the HAI group, and both groups fall within the "collaboration skills are developing" stage. Looking at the subdimensions of the self-assessment, the HHC group scored higher than the HAI group on nine out of the total ten dimensions. However, significant differences were found only in C1 Contribution, C4 Time Management, and C6 Problem Solving. This suggests that, compared to HAI, students in HHC are more willing to engage in timely and open-minded communication with their peers. It is worth noting that in the C3 Quality of Work dimension, the HAI group scored slightly higher than the HHC group, although the difference is not significant. One could speculate that in HAI, students tend to be less reliant on team members and focus more on the quality of their own work, showing a stronger sense of self-monitoring.

IV. CONCLUSION

Although many participants in this study had limited experience with AI technologies, they were familiar with using non-AI online tools for learning purposes. This study was the first time for the HAI group of students to use an AI chatbot. As a result, some hesitancy was observed during its use, mainly reflected in their reluctance to share their own opinions with ChatGPT-4o, their skepticism toward its generated responses, and, in some cases, tentative attempts to correct ChatGPT-4o's analysis and or answers. Additionally, the depth of HAI was influenced by the method of interaction. The use of a text-based dialogue box led most students to opt for convenience, primarily using screenshots or copying the question text to interact with ChatGPT-4o, with few engaging in deeper discussions about specific points of confusion. This significantly affected the frequency and depth of dialogue in HAI problem solving. This could be an explanation for why the effect size for HAI on scientific problem solving, $d_{(\mathrm{HAI})} = 0.053$, is much smaller than the effect size for HHC, $d_{(\mathrm{HHC})} = 0.378$. Furthermore, this observation aligns with the theory of collaborative learning, which suggests that richer and more interactive dialogues often lead to deeper engagement and better problem-solving outcomes [34]. This finding addresses the first research question of this study: ChatGPT-4o, at the current stage, does not provide as much support in collaborative learning for solving scientific problems as a peerlike partner would.

By analyzing the dialogues in HAI, it was found that approximately 70% of students used screenshots, and around 20% used copied text to start the interaction with ChatGPT-4o. After ChatGPT-4o provided its analysis and responses, a few students would continue to ask followup questions. Most of the participants either chose to trust ChatGPT-4o and modify their answers or remained confident in their original answers and ignored ChatGPT-4o's responses. This kind of interaction, which requires little transformation, limited the collaboration to a simple question-and-answer format, lacking deeper exchanges of ideas and critical thinking. As a result, in HAI, most students did not effectively engage with ChatGPT-4o as a learning partner but instead treated it more like a database for obtaining answers. This might be the reason for the lower effectiveness of HAI.

The rapidly iterating generative AI tools have gained significant attention in industries such as education, emotional intelligence, healthcare, information technology, psychology, and social media marketing and have made important impacts on various business sectors and services [35]. However, the generative AI technology is still in its early stages of development, with certain immaturities that can potentially hinder the establishment of a close collaborative relationship. This may be due to challenges such as inaccurate image recognition, mismatches between analysis results and answer options, and inconsistent answers to the same question, which may present barriers to effective communication between students and AI. During collaboration, students are required not only to reflect on their own problem-solving strategies but also to constantly monitor ChatGPT-4o for these types of errors and correct them when necessary. If a student attempts to correct ChatGPT-4o and it fails to rectify the mistake promptly, the student may lose patience. On the other hand, if ChatGPT-4o uncritically agrees with the student's corrections, it can reinforce wrong answers, leading the student further down the incorrect path, hindering learning. Moreover, when solving scientific problems with definitive answers, ChatGPT-4o's inconsistency in providing correct solutions can severely undermine students' trust in the tool.

In light of the challenges associated with the integration of generative AI into educational contexts, it is evident that comprehensive training for both students and educators is imperative to optimize its potential. For students, structured training is essential to cultivate a critical and reflective approach to engaging with AI tools. Such training should enable learners to identify and rectify errors in AI-generated outputs, as well as critically assess the validity and applicability of AI-derived solutions. This skill set is crucial for fostering critical thinking ability and a deep understanding of the scientific content knowledge. For educators, professional development training must address the unique characteristics of generative AI, including its inherent variability and occasional inconsistency. Training should focus on equipping teachers with the pedagogical strategies necessary to harness these features effectively, transforming potential limitations into opportunities for deeper learning. Specifically, educators must be adept at managing the uncertainty associated with AI outputs and leveraging this variability as a controlled instructional tool to stimulate critical thinking and promote higher-order cognitive engagement among students.

By providing both students and teachers with the required competencies to engage with generative AI in a deliberate and informed manner, educational stakeholders can mitigate potential risks and maximize the technology's capacity to enhance learning outcomes. To achieve this goal, it is important to conduct further empirical research to explore best practices for AI integration in educational settings, ensuring that its implementation aligns with pedagogical goals and contributes meaningfully to the learning process.

This study has several limitations that should be considered when interpreting the results. First, the scope of the HAI was restricted to ChatGPT-4o's text-based dialogue interface. Other generative AI models and interaction modes (e.g., voice based or multimodal) were not explored, which limits the generalizability of the findings. Additionally, the participants' prior exposure to AI, particularly within the context of Chinese education, may have influenced their interactions and perceptions of ChatGPT-4o, potentially affecting the results. A broader investigation with different AI tools and interaction modes would provide a more comprehensive understanding of AI's role in collaborative problem solving. Second, the assessment of scientific knowledge in this study focused solely on work and mechanical energy, limiting the scope of the findings to these topics. Future research should extend to other core ideas, such as electromagnetism and thermodynamics, to explore how AI can support learning across a broader range of scientific concepts. Expanding the content coverage would offer deeper insights into the potential of HAI in various areas of science education. Finally, since this study primarily focused on the characteristics of HAI, it has somewhat neglected to explore the features and advantages of HHC. As a result, some of the valuable aspects of HHC were not fully investigated. Moving forward, future research should aim to investigate these differences in more detail, particularly in areas such as shared understanding, emotional intelligence, and dynamic feedback, to better understand how these factors might provide advantages in HHC compared to HAI, and how the benefits of such factors may be adapted into the HAI context to improve its efficacy.

In conclusion, the suboptimal gain and self-assessment experiences in HAI for solving scientific problems can be attributed to both students' perceptions of ChatGPT-4o's role and the limitations of ChatGPT-4o itself. However, the results do not suggest that HAI for scientific problem solving is unfeasible, but rather, the outcomes of this study call for improvements to enhance both the effectiveness and experience of collaboration. First, students need to develop a deep understanding of generative AI, including its strengths and weaknesses. This will help students accurately position AI in collaboration, leveraging its strengths while mitigating its weaknesses, allowing them to use it as a true learning partner. Second, training in HAI strategies is essential. Demonstrating successful examples of AI-assisted problem-solving or effective methods of interacting with AI before collaboration may likely enable students to engage in deeper questioning and discussion, thereby enriching the collaboration. Finally, developers of generative AI technologies must continue to improve the current technical limitations of tools such as ChatGPT-4o. With more research and development on the way, the HAI will provide promising applications in shaping education.

ACKNOWLEDGMENTS

The authors would like to acknowledge the help of the editor and anonymous reviewers. The research is supported in part by the Anhui Provincial Philosophy and Social Science Planning Project, "Research on Promoting the Cooperation and Sharing of High-Quality Digital Resources in Basic Education between Anhui and Shanghai, Jiangsu, Zhejiang" (Project Approval No. AHSKZ2022D21).

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

[1] P. Dillenbourg and A. John, Self, intelligent tutoring systems: Second international conference, in *People Power: A Human-Computer Collaborative Learning System* (ITS'92, Montrøal, Canada, 1992).
[2] P. Dillenbourg and A. John, Self, computer supported collaborative learning, in *Designing Human-Computer Collaborative Learning* (Springer, Berlin, Heidelberg, 1995), pp. 245–264.
[3] D. Wang, E. Churchill, P. Maes, X. Fan, B. Shneiderman, Y. Shi, and Q. Wang, From human-human collaboration to human-AI collaboration: Designing AI systems that can work together with people, in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 10.1145/3334480.3381069.
[4] K. Guo, J. Wang, and S. K. W. Chu, Using chatbots to scaffold EFL students' argumentative writing, Assess. Writ. **54**, 100666 (2022).
[5] Y. Liang *et al.*, Exploring the potential of using ChatGPT in physics education, Smart Learn. Environ. **10**, 52 (2023).
[6] P. Bitzenbauer, ChatGPT in physics education: A pilot study on easy-to-implement activities, Contemp. Educ. Technol. **15**, ep430 (2023).
[7] M. Halaweh, ChatGPT in education: Strategies for responsible implementation, Contemp. Educ. Technol. **15**, 421 (2023).
[8] E. Kasneci *et al.*, ChatGPT for good? On opportunities and challenges of large language models for education, Learn. Individ. Diff. **103**, 102274 (2023).
[9] B. Gregorcic and A.-M. Pendrill, ChatGPT and the frustrated Socrates, Phys. Educ. **58**, 035021 (2023).
[10] G. Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course?, Phys. Rev. Phys. Educ. Res. **19**, 010132 (2023).
[11] C. G. West, Advances in apparent conceptual physics reasoning in GPT-4, advances in apparent conceptual physics reasoning in GPT-4, arXiv:2303.17012.
[12] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[13] D. Tong *et al.*, Investigating ChatGPT-4's performance in solving physics problems and its potential implications for education, Asia Pacific Educ. Rev. **25**, 1379 (2024).
[14] Y. Zeng and D. Tong, A study on ChatGPT-4's ability to solve scientific problems: A case study of the Chinese National College Entrance Examination Physics Test, Middle Sch. Phys. **42**, 22 (2024).
[15] D. Liu and D. Tong, A study on ChatGPT-4's ability to solve scientific problems: A case study of the Chinese Secondary School Entrance Examination from Beijing, Shanghai, and Guangzhou, Middle Sch. Phys. **42**, 15 (2024).
[16] 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).
[17] W. Yeadon and D. P. Halliday, Exploring Durham University physics exams with large language models, arXiv:2306.15609.
[18] W. Yeadon and T. Hardy, The impact of AI in physics education: A comprehensive review from GCSE to university levels, arXiv:2309.05163.
[19] J. Chang and X. Lu, The study on students' participation in personalized learning under the background of artificial intelligence, in *Proceedings of the 2019 10th International Conference on Information Technology in Medicine and Education (ITME)* (IEEE, New York, 2019).
[20] D. W. Johnson and R. T. Johnson, *Learning Together and Alone: Cooperative, Competitive, and Individualistic Learning*, 5th ed. (Allyn and Bacon, Boston, MA, 1999).
[21] R. E. Slavin, *Co-operative learning. What makes group-work work?* edited by H. Dumont, D. Istance, and F. Benavides in The Nature of Learning, Using Research to Inspire Practice (OECD Publishing, Paris, France, 2010).
[22] L. Springer, M. Stanne, and S. Donovan, Effects of small-group learning on undergraduates in science, mathematics, engineering and technology: A meta-analysis, Rev. Educ. Res. **69**, 21 (1999).
[23] R. P. dos Santos, Enhancing physics learning with ChatGPT, bing chat, and bard as agents-to-think-with: A comparative case study, arXiv:2306.00724.
[24] G. Polverini and B. Gregorcic, Evaluating vision-capable chatbots in interpreting kinematics graphs: A comparative study of free and subscription-based models, arXiv:2406.14685.
[25] D. Tong, J. Liu, Y. Sun, Q. Liu, X. Zhang, S. Pan, and L. Bao, Assessment of student knowledge integration in learning work and mechanical energy, Phys. Rev. Phys. Educ. Res. **19**, 010127 (2023).
[26] L. Bao, K. Koenig, Y. Xiao, J. Fritchman, S. Zhou, and C. Chen, Theoretical model and quantitative assessment of scientific thinking and reasoning, Phys. Rev. Phys. Educ. Res. **18**, 010115 (2022).
[27] K. Ofstedal and K. Dahlberg, Collaboration in student teaching: Introducing the collaboration self-assessment tool, J. Early Child. Teach. Educ. **30**, 37 (2009).
[28] D. Tong, Research on deep learning in physics with a focus on knowledge integration: Taking mechanical energy and its conservation law as an example, Ph.D. thesis, East China Normal University, Shanghai, China, 2024.
[29] J. Cohen, A power primer, Psychol. Bull. **112**, 155 (1992).
[30] D. Tong and H. Ren, A study on the performance of ChatGPT-3.5 in solving physics problems, Middle Sch. Phys. **41**, 11 (2023).
[31] C. G. West, AI and the FCI: Can ChatGPT project an understanding of introductory physics? arXiv:2303.01067.
[32] C. Kwan Lo, What is the impact of ChatGPT on education? A rapid review of the literature, Educ. Sci. **13**, 410 (2023).
[33] G. Polverini and B. Gregorcic, Performance of ChatGPT on the test of understanding graphs in kinematics, Phys. Rev. Phys. Educ. Res. **20**, 010109 (2024).
[34] R. M. Gillies, Promoting academically productive student dialogue during collaborative learning, Int. J. Educ. Res. **97**, 200 (2019).
[35] A. K. Kar, P. S. Varsha, and S. Rajan, Unravelling the impact of generative artificial intelligence (GAI) in industrial applications: A review of scientific and grey literature, Glob. J. Flex. Syst. Manage. **24**, 659 (2023).