INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2026.2732482

Taylor & Francis
by informa

Check for updates

From Potential to Reality: Gaps in Realizing Human–AI Complementarity in Educational Tasks

Bo Jiang[a], Jiatong Wang[a], Yijie Lu[b], Jiayi Liu[b] and Yuanhao Jiang[a]

[a]Shanghai Institute of Artificial Intelligence for Education, East China Normal University, Shanghai, China; [b]Department of Education Information Technology, East China Normal University, Shanghai, China

ABSTRACT

Human–artificial intelligence collaboration (HAIC) is emerging as a transformative approach in education, yet empirical evidence regarding its impact on educational task performance remains limited. This study adopts a complementarity perspective and distinguishes three types of HAIC outcomes: inherent complementarity, collaborative complementarity, and collaboration counter-productivity. An experiment was conducted to examine pre-service teachers' task performance across conditions. Specifically, we measured complementarity potential, complementarity rates, and the functional roles of artificial intelligence (AI) participation in HAIC. The results showed that HAIC improved pre-service teachers' immediate task performance, but this performance advantage was not sustained in the post-test. Inherent complementarity was realized substantially more often than collaborative complementarity. By contrast, collaborative complementarity was rarely realized: when both humans and AI initially made incorrect judgments, HAIC seldom produced a correct final judgment. In addition, AI roles in two-way collaboration were highly concentrated in assistive functions, including elaboration, knowledge supplementation, evaluation, feedback provision, and questioning.

HIGHLIGHTS

- We adapted a human–AI complementarity framework to an educational decision-making task.
- We compared complementarity rates across one-way and two-way collaboration modes.
- Inherent complementarity was more readily realized than collaborative complementarity.
- In two-way collaboration, AI roles were concentrated in assistive functions.

KEYWORDS

Human–AI collaboration; complementarity; AI-assisted decision-making; educational annotation

1. Introduction

In many decision-making settings, human judgment is constrained by limited information-processing capacity, finite cognitive resources, and reliance on heuristics, which can hinder optimal decisions (Simon, 1955). Artificial intelligence (AI)-based decision support seeks to mitigate these constraints by providing predictions and recommendations that augment human information processing and judgment (Gonzalez & Heidari, 2025; Hemmer et al., 2024; Steyvers & Kumar, 2024). Yet AI assistance introduces a different challenge: users must assess the reliability of AI outputs, decide when to accept or reject them, and determine how to integrate them with their own judgments (Gonzalez & Heidari, 2025; Steyvers & Kumar, 2024; Vaccaro et al., 2024). Because AI advice is fallible, inaccurate perceptions of system capabilities may lead users to over-rely on incorrect recommendations or reject correct ones, ultimately impairing human–AI team performance (Gonzalez & Heidari, 2025; Vaccaro et al., 2024).

CONTACT Bo Jiang bjiang@deit.ecnu.edu.cn Shanghai Institute of Artificial Intelligence for Education, East China Normal University, No. 3663 North Zhongshan Road, Shanghai 200062, China

© 2026 Taylor & Francis Group, LLC

Against this background, human–AI complementarity has been regarded as a benchmark for evaluating human–AI collaboration (HAIC) performance (Hemmer et al., 2024). Complementary team performance (CTP) is achieved when a human–AI team outperforms both the human and the AI working independently, rather than merely improving human performance (Bansal et al., 2021; Hemmer et al., 2024). AI is often better at processing large-scale data, identifying statistical patterns, and optimizing predefined objectives (Dellermann et al., 2019; Fui-Hoon Nah et al., 2023; Jarrahi et al., 2023; Jiang et al., 2025), whereas humans retain advantages in dealing with uncertainty, contextual information, value judgments, and interpersonal challenges (Harfouche et al., 2023; R. Zhang et al., 2021). These asymmetries provide a basis for complementarity potential. Complementarity potential is converted into realized complementarity effects only when the team successfully exploits these asymmetries to reach a correct or otherwise superior final judgment (Gonzalez & Heidari, 2025; Hemmer et al., 2024). Empirical evidence, however, shows that human–AI teams often outperform humans working alone but fail to outperform the better-performing individual agent, with performance losses particularly common in decision-making tasks (Vaccaro et al., 2024). The central question is therefore no longer simply whether AI improves human performance, but under what conditions complementarity potential can be converted into realized complementarity effects (Steyvers & Kumar, 2024).

Existing studies have begun to support the realization of complementarity effects by redesigning AI role allocation and human–AI interaction modes, thereby restructuring information flow, agency, and decision-making responsibilities between humans and AI (Steyvers & Kumar, 2024). In the more common one-way collaboration mode, AI primarily serves as a provider of recommendations or explanations, while humans are responsible for determining whether to adopt these recommendations and making final decisions (Gomez et al., 2025). This mode relies on humans' ability to accurately recognize the relative strengths of humans and AI. When humans lack metacognitive awareness of their own capabilities or AI capabilities, correct AI recommendations may be rejected, whereas incorrect recommendations may be accepted. In contrast, two-way collaboration allows humans to actively ask questions, request explanations, provide feedback, and prompt AI to revise its outputs, thereby creating more opportunities for exchanging and reorganizing complementary information between both parties (Gomez et al., 2025).

Role allocation and interaction design have therefore received increasing attention as potential conditions for complementarity realization (Steyvers & Kumar, 2024). By determining what information is exchanged, when it is exchanged, and how decision authority is distributed, different interaction modes may shape users' ability to evaluate, question, and revise AI outputs (Gomez et al., 2025). In AI-assisted decision-making, interaction patterns range from relatively one-way advice provision to more interactive, dialogue-based decision-making. Dialogue-based decision-making creates additional opportunities to examine, question, and integrate AI-provided information. However, richer interaction alone does not guarantee better human–AI team performance (Bansal et al., 2021).

Although recent work has begun to distinguish complementarity potential from realized complementarity effects, less is known about how different interaction modes influence whether complementarity potential is realized or lost (Hemmer et al., 2024; Vaccaro et al., 2024). In particular, it remains unclear whether final team judgments arise from retaining an initially correct contribution, generating a new correct judgment through interaction, or overturning a judgment that was initially correct. It is also not well understood which functional roles AI enacts during two-way collaboration (Steyvers & Kumar, 2024). This study addresses these questions in a structured educational decision-making task by comparing individual work, one-way HAIC, and two-way HAIC. It examines complementarity realization through human initial judgments, AI outputs, and final team judgments, and further analyzes AI's functional roles and interaction patterns in two-way collaboration.

RQ1: What are the differences in pre-service teachers' task performance across experimental conditions and task phases?

RQ2: To what extent does complementarity potential exist in HAIC, to what extent is this potential realized as complementarity effects, and how is complementarity realization associated with team performance?

RQ3: What interaction features and functional role structures characterize AI participation in two-way HAIC?

2. Related work

2.1. Human–AI collaboration and interaction modes

The idea of HAIC can be traced back to J. C. R. Licklider's article, Man-Computer Symbiosis, which first introduced the concept of symbiotic computing (Licklider, 1960; Wang et al., 2020). HAIC refers to sustained joint work between humans and artificial intelligence, or autonomous agents, toward a shared goal (Fragiadakis et al., 2025). As generative AI has increasingly been used to assist human decision-making in domains such as medicine (Jussupow et al., 2021), finance (Liu et al., 2023), and education (Koong Lin et al., 2025), AI is no longer viewed only as an external tool. Instead, it is increasingly conceptualized as a teammate. This perspective emphasizes that AI can take on roles such as providing suggestions, conducting analysis, offering prompts, coordinating tasks, and giving feedback in shared tasks (Gonzalez & Malloy, 2026). For example, in educational settings, GenAI-powered teachable agents may be perceived by students as learning companions, facilitators, or collaborative problem-solvers. These perceived roles can influence students' engagement, explanation, and reflection processes (Song et al., 2026). In advertising creativity tasks, AI agents with different anthropomorphic behavioral styles, such as openness, conscientiousness, extraversion, agreeableness, and neuroticism or emotional sensitivity, can affect team collaboration quality and task performance. In particular, when AI is overly compliant or lacks constructive challenge, the quality of creative outputs may decline (Ju & Aral, 2026). In decision-making contexts, AI can be designed either as a tool or as a teammate. However, when AI is positioned as a teammate, smoother dialogue and richer explanations may increase trust without necessarily improving decision quality, and may even induce overreliance (Sultana Samu et al., 2026).

Beyond AI role design, prior studies have also classified HAIC modes according to task agency, decision authority, and interaction structure. From the perspective of HAIC evaluation, Fragiadakis et al. distinguished three modes of HAIC: AI-centric, Human-centric, and symbiotic. These modes describe different structures of task allocation, decision authority, collaborative relationship, and depth of human-AI interaction (Fragiadakis et al., 2025). AI-centric collaboration emphasizes AI-led task processing. Human-centric collaboration emphasizes human leadership with AI assistance. Symbiotic collaboration emphasizes bidirectional interaction, shared decision-making, and continuous feedback between humans and AI. From a broader perspective, Gomez et al. proposed seven HAIC modes: delegation, AI-first, secondary, AI-guided, AI-follow, request-driven, and human-guided. Among these modes, an AI-based decision aid provides recommendations to humans, while humans retain final decision authority (Gomez et al., 2025). One common form of human-AI interaction is therefore AI-assisted decision-making, in which an AI decision-support system provides recommendations and humans make the final decision (Lu et al., 2024).

However, HAIC does not necessarily lead to performance gains. Its effects may take different forms. On the one hand, humans and AI may possess distinct forms of intelligence. In decision-making contexts, HAIC may allow both parties to leverage their respective strengths and achieve performance that exceeds either humans or AI working alone. This is often described as complementarity (Bansal et al., 2021; Lu et al., 2024). On the other hand, if human decision-makers fail to make an effective contribution to team performance, AI working alone may produce better outcomes than the human-AI team (Lai & Tan, 2019; Y. Zhang et al., 2020). Therefore, researchers have increasingly shifted attention from whether HAIC is beneficial to the conditions under which humans and AI can realize complementarity. This shift has made complementarity a central issue in HAIC research.

2.2. Complementarity in human–AI collaboration

Information and ability asymmetries between humans and AI provide a potential space for human-AI complementarity, but such differences do not automatically translate into complementarity effects. Information asymmetry includes two aspects: contextual information asymmetry and knowledge

asymmetry. Contextual information asymmetry refers to the fact that humans and AI may access and interpret different types of data in a given context (Hemmer et al., 2024; Holstein & Aleven, 2022). For example, in educational settings, teachers mainly observe students' body language and facial expressions, whereas AI can analyze screen-monitoring activities and interaction data (Harfouche et al., 2023). In housing-price prediction tasks, humans may access visual information such as house photographs, whereas AI processes large-scale numerical market data (Holstein et al., 2023). Knowledge asymmetry refers to AI's advantages in explicit and general knowledge. Because explicit knowledge can be formally represented and stored at scale, AI can use databases and corpora to rapidly retrieve and integrate cross-domain information, exceeding individual humans in knowledge breadth and capacity (Fui-Hoon Nah et al., 2023; Harfouche et al., 2023; Jarrahi et al., 2023).

By contrast, human advantages lie in tacit knowledge and domain expertise. These forms of knowledge are acquired through practice, are difficult to fully formalize, and are internalized as cognitive structures that support transfer, judgment, and intuition (Caldwell et al., 2022; Hemmer et al., 2024; Jarrahi et al., 2023; Louie et al., 2024). Ability asymmetry refers to differences between humans and AI in reasoning and cognitive processes. AI relies on data-driven reasoning, identifying patterns in data and making probabilistic decisions (Holstein et al., 2023). Humans, in contrast, rely on logical reasoning, including induction and deduction (Dellermann et al., 2019), and may show stronger logical reasoning abilities in certain contexts (Harfouche et al., 2023). Humans also have advantages in tasks that require transfer, whereas AI algorithms are often more task-specific (Hendrycks et al., 2020; Torrey & Shavlik, 2010). In HAIC, differences between humans and AI do not automatically lead to realized complementarity effects. Team performance improves only when the team can identify, evaluate, and integrate the relative strengths of both parties. Bansal et al. defined complementary team performance (CTP) as the performance of a human-AI team exceeding that of the best individual member, rather than merely improving human performance (Bansal et al., 2021). Hemmer et al. further distinguished inherent complementarity from collaborative complementarity and operationalized the transformation from potential opportunities to realized complementarity effects using indicators such as complementarity potential, complementarity effect, and complementarity rate (Hemmer et al., 2024). Building on Hemmer et al.'s distinction between inherent and collaborative complementarity, this study further operationalizes negative collaborative outcomes as collaboration counter-productivity (Hemmer et al., 2024). Inherent complementarity is realized when a team successfully identifies and retains an initially correct contribution from either the human or the AI. Collaborative complementarity is realized when interaction produces a correct team judgment that was not present in either party's initial answer. Collaboration counter-productivity occurs when both parties initially provide correct judgments but the team reaches an incorrect final judgment.

2.3. Complementarity mechanisms in human–AI collaboration

From the perspective of complementarity mechanisms, the emergence of complementarity effects in HAIC mainly depends on whether complementarity potential exists between humans and AI and whether this potential can be translated into complementarity effects during collaboration. Complementarity potential is not simply equivalent to higher standalone accuracy of either humans or AI. Rather, it depends on whether the two parties have differentiated and combinable information, representations, and error patterns. The difference between AI and humans should also not be too large. If AI substantially outperforms humans, the performance of HAIC may be lower than that of AI working alone (Steyvers & Kumar, 2024). Whether complementarity potential can be translated into complementarity effects also depends on whether humans can develop appropriate reliance on AI suggestions. Ideally, appropriate reliance in decision-making means that humans accept AI suggestions when AI is correct and reject AI suggestions when AI is incorrect. This process depends on the mental model that humans construct of AI (Q. Zhang et al., 2022). A mental model refers to a simplified internal representation of external reality, which individuals use to predict events and guide behavior (Barnes, 1944). In the context of AI, humans' mental model of AI refers to their cognitive framework regarding AI's capability boundaries, error boundaries, and reliability conditions. This framework helps humans judge when to adopt AI suggestions and when to retain their own judgments (Bansal et al., 2019). If this mental model is biased, humans may develop inappropriate

reliance or miscalibrated trust. Research based on dual-process theory suggests that when humans decide whether to accept algorithmic advice, they tend to rely on heuristic reasoning, or Type 1 processing, as the default cognitive process (Jussupow et al., 2021; R. Zhang et al., 2021). This indicates that human reliance on AI is not entirely based on rational evaluation. Instead, it may be influenced by heuristic cues such as preference for AI, algorithm aversion, and the use of professionalized language (Bauer et al., 2023). Cognitive reflection determines whether individuals can shift from Type 1 intuitive processing to Type 2 deeper cognitive analysis. Therefore, users with different levels of cognitive reflection may understand and use AI in different ways, which may further produce intervention-generated inequalities (Duan et al., 2022; Gajos & Mamykina, 2022).

Furthermore, from the perspective of AI design, Gonzalez and Heidari argued that, in dynamic decision-making contexts, human-AI complementarity depends not only on humans' understanding of AI, but also on whether cognitive AI can model human-like reasoning, memory, and decision-making processes in order to simulate how people perceive, interpret, and respond to complex tasks (Gonzalez & Heidari, 2025). At the same time, cognitive AI needs to provide communication forms that humans can understand and use, including uncertainty quantification, explanations, visualizations, and other communication media such as natural language (Morrison et al., 2024). However, AI explanations do not necessarily promote appropriate reliance. In some cases, explanations may instead intensify overreliance and hinder the translation of complementarity potential into complementarity effects (Morrison et al., 2024). In addition, whether the complementarity potential can be activated is also shaped by the role structure and interaction mode of HAIC. Gonzalez et al. noted that AI roles in human-AI teams exist along a continuum from simple tools to dynamic partners, and that AI contributions vary with different levels of flexibility and agency (Gonzalez et al., 2026).

3. Methods

3.1. Participants

The participants were 97 pre-service teachers from a university in China. As the main task of the experiment was to annotate mathematical problems, participants were recruited from mathematics, education, and other related majors to ensure data quality. Each participant received compensation of 40 yuan per hour. Institutional ethics approval was obtained for this study (Approval No. HR2-0028-2025) from the Academic Ethics Committee of East China Normal University. The manuscript contains no participant-identifiable information. Verbal informed consent for participation in the study was obtained.

3.2. Procedure

Participants were first introduced to the task background: in an AI-assisted context, pre-service teachers were asked to annotate mathematical problems to examine their understanding of which aspects of students' mathematical ability were assessed by each problem. Participants were presented with 20 mathematical problems and asked to label each problem according to the primary cognitive process involved in solving it: computation, recognition, recall, or strategy selection, as shown in Figure 1(a). Next, all participants completed a 20-item pretest to assess their baseline task performance. During the experiment, participants were assigned to three groups, as shown in Figure 1(b): the control group ($n = 25$), experimental group 1 ($n = 35$), and experimental group 2 ($n = 37$). Each group completed the intervention task via a HAIC platform. The platform supported three task-completion modes: personal mode, in which participants completed the annotation task alone; one-way collaboration mode, in which participants received AI-generated annotations and served as the final decision-makers; and two-way collaboration mode, in which participants interacted with AI to jointly produce the final annotation.

The control group completed the annotation task in personal mode. Experimental group 1 completed the task in the one-way collaboration mode. Participants received AI-generated annotations and made final decisions based on their own judgments. Initial human annotations, AI-generated annotations, final human-AI team annotations, and participants' acceptance decisions were recorded. Experimental group 2 completed the task in the two-way collaboration mode by interacting with the AI partner. The initial

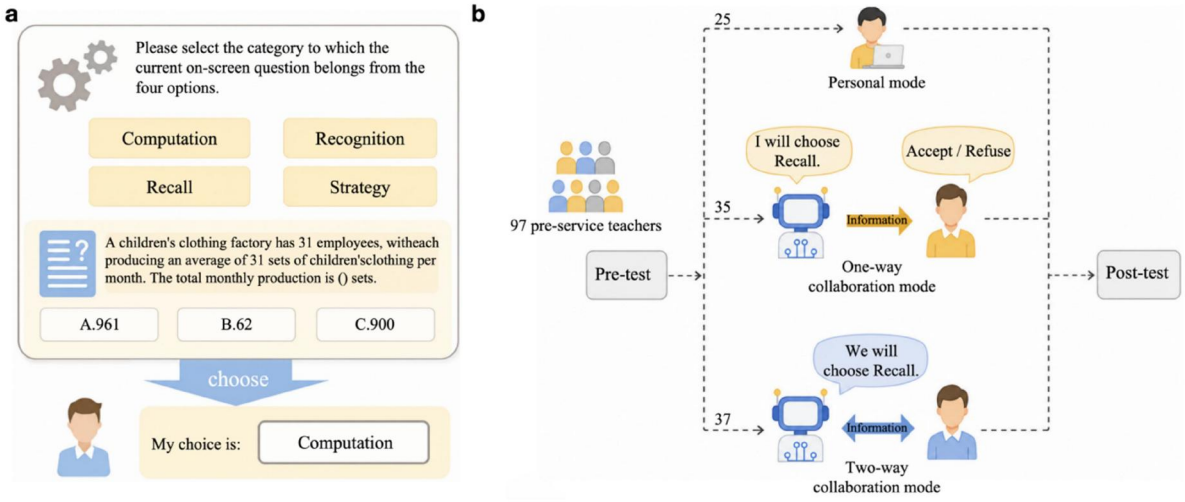


Figure 1. Experimental procedure. (a) An example of the tasks across the pretest, intervention task, and post-test phases. (b) The experimental procedure, including participant allocation and task-completion modes.

human annotations, the final annotations after collaboration, and the human-AI dialogue process were recorded. In addition, AI responses were also recorded for the two experimental groups. Finally, all pre-service teachers completed a 20-item post-test to assess their post-test performance.

4. Measures

4.1. Assessment of complementarity

Complementarity is an important indicator for assessing HAIC in this study. Drawing on Hemmer et al.'s complementarity framework, we operationalized three mutually exclusive item-level collaboration outcome categories: inherent complementarity, collaborative complementarity, and collaboration counter-productivity. For each category, three indicators were calculated: complementarity potential, complementarity effect, and complementarity rate. The concepts and calculation methods of the indicators are as follows:

Inherent complementarity potential refers to the potential space created by preexisting differences between humans and AI in information, knowledge, or ability. In this study, an item is considered to have inherent complementarity potential when the initial human answer and the initial AI answer differ in correctness; that is, one party answers correctly while the other answers incorrectly. For each item, p_{in} is:

$$p_{in} = \begin{cases} 1, & \text{if } \text{correct}_{\text{human}} \neq \text{correct}_{\text{AI}} \\ 0, & \text{if } \text{correct}_{\text{human}} = \text{correct}_{\text{AI}} \end{cases}$$

The total inherent complementarity potential P_{in} is calculated as the sum of the inherent complementarity potential values across all items.

$$P_{in} = \sum_{i=1}^{n} p_{in}$$

The inherent complementarity effect refers to the extent to which human-AI teams actually realize inherent complementarity by selecting the initially correct contribution. In this study, an item is coded as an inherent complementarity effect when it has inherent complementarity potential and the final human-AI team answer is correct. For each item, the inherent complementarity effect e_{in} of the human-AI team is:

$$e_{in} = \begin{cases} 1, & \text{if } p_{in} = 1 \text{ and } \text{correct}_{\text{team}} = 1 \\ 0, & \text{otherwise} \end{cases}$$

The inherent complementarity effect E_{in} in the collaborative process between humans and AI in an experiment is the sum of the inherent complementarity effect values of all items:

$$E_{in} = \sum_{i=1}^{n} e_{in}$$

The inherent complementarity rate R_{in} refers to the proportion of inherent complementarity potential that is realized as an inherent complementarity effect. It is calculated as the ratio of inherent complementarity effects to inherent complementarity potential.

$$R_{in} = \frac{\sum_{i=1}^{n} e_{\text{in}}}{\sum_{i=1}^{n} p_{\text{in}}}$$

Collaborative complementarity refers to performance gains generated through human-AI interaction, such as improved confidence calibration, correction of AI errors, or transferability. Given that the present task was a mathematical problem annotation task, this study focused on an outcome-level operationalization of collaborative complementarity. Specifically, this study examines whether the human-AI team can produce a correct final annotation when neither the human nor the AI initially provides a correct annotation. For each item, the collaborative complementarity potential p_{co} of that item is:

$$p_{co} = \begin{cases} 1, & \text{if } \text{correct}_{\text{human}} = \text{correct}_{\text{AI}} = 0 \\ 0, & \text{otherwise} \end{cases}$$

In the experiment, the collaborative complementarity potential P_{co} between humans and AI in the process of collaboration is the sum of the collaborative complementarity potential values of all items:

$$P_{co} = \sum_{i=1}^{n} p_{co}$$

In this study, an item is coded as showing a collaborative complementarity effect when neither the human nor the AI initially provides a correct annotation, but the final human–AI team annotation is correct. For each item, the collaborative complementarity effect e_{co} of the human-AI team is:

$$e_{co} = \begin{cases} 1, & \text{if } p_{co} = 1 \text{ and } \text{correct}_{\text{team}} = 1 \\ 0, & \text{otherwise} \end{cases}$$

The collaborative complementarity effect E_{co} between humans and AI in the process of collaboration is the sum of the collaborative complementarity effect values of all items in the experiment:

$$E_{co} = \sum_{i=1}^{n} e_{co}$$

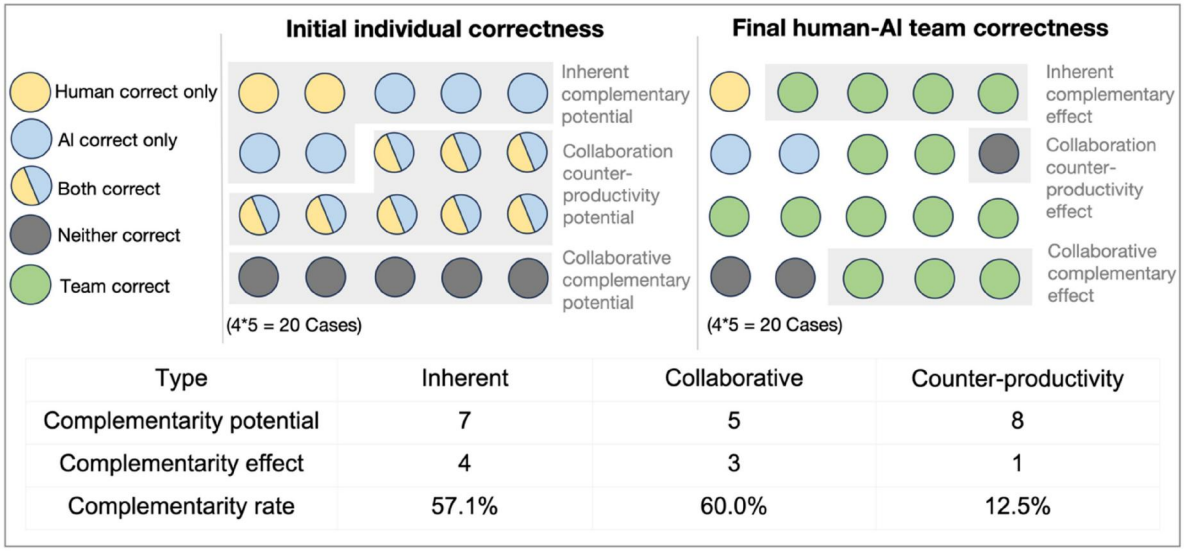
The collaborative complementarity rate R_{co} refers to the proportion of collaborative complementarity potential that is realized as collaborative complementarity effects. It is calculated as the ratio of realized collaborative complementarity effects to collaborative complementarity potential:

$$R_{co} = \frac{\sum_{i=1}^{n} e_{co}}{\sum_{i=1}^{n} p_{co}}$$

Finally, the collaboration counter-productivity potential refers to the risk that collaboration may undermine answers that are initially correct. In this experiment, it is operationalized as the total number of items for which both the human and the AI initially provide correct answers. For each item, the collaboration counter-productivity potential p_{ba} of that item is:

$$p_{ba} = \begin{cases} 1, & \text{if } \text{correct}_{\text{human}} = 1 \text{ and } \text{correct}_{\text{AI}} = 1 \\ 0, & \text{if } \text{correct}_{\text{human}} \neq 1 \text{ or } \text{correct}_{\text{AI}} \neq 1 \end{cases}$$

The collaboration counter-productivity potential P_{ba} is the sum of the collaboration counter-productivity potential values of all items in the experiment:



Type	Inherent	Collaborative	Counter-productivity
Complementarity potential	7	5	8
Complementarity effect	4	3	1
Complementarity rate	57.1%	60.0%	12.5%

Figure 2. Illustration of the calculation method for theoretical complementarity potential and realized complementarity effect in an experimental setting.

$$P_{ba} = \sum_{i=1}^{n} p_{ba}$$

The collaboration counter-productivity effect refers to cases in which HAIC fails to preserve an answer that is initially correct for both the human and the AI. In the experiment, it specifically refers to cases in which the human–AI team ultimately provides an incorrect answer on items that were initially correct for both the human and the AI. For each item, the collaboration counter-productivity effect e_{ba} is:

$$e_{ba} = \begin{cases} 1, & \text{if } p_{ba} = 1 \text{ and } \text{correct}_{\text{team}} = 0 \\ 0, & \text{otherwise} \end{cases}$$

The collaboration counter-productivity effect E_{ba} is the sum of the collaboration counter-productivity effect values of all items in the experiment:

$$E_{ba} = \sum_{i=1}^{n} e_{ba}$$

The collaboration counter-productivity effect rate R_{ba} refers to the proportion of collaboration counter-productivity potential that is realized as a collaboration counter-productivity effect:

$$R_{ba} = \frac{\sum_{i=1}^{n} e_{ba}}{\sum_{i=1}^{n} p_{ba}}$$

Hereafter, the inherent complementarity rate, collaborative complementarity rate, and collaboration counter-productivity rate are abbreviated as ICR, CCR, and CPR, respectively. In the formulas, these rates are denoted as (R_{in}), (R_{co}), and (R_{ba}), respectively. An example of complementarity calculation is shown in Figure 2.

4.2. Assessment of AI functional roles

Collaborative behavior is an important indicator of interactions among individuals during the collaborative process. Siemon argues that AI-based team roles should be designed by specifying the behaviors, skills, abilities, rights, responsibilities, and task-related expectations that AI teammates are expected to fulfill in human-AI teams (Siemon, 2022). Meanwhile, related research suggests that effective human-AI

Table 1. Definitions of collaborative behavior roles used for annotation in the human-AI complementarity framework.

Label	Concept description
Questioner	Pose a question or raise a doubt.
Feedback provider	Provide feedback on the other person's ideas or suggestions, including agreement or disagreement.
Elaborator	Express, explain, or elaborate on one's own ideas or viewpoints.
Innovator	Propose a new idea or a new answer.
Knowledge supplementer	Enumerate or provide some knowledge, such as definitions, characteristics, concepts, or theorems.
Emotional expresser	Express or convey one's own emotions, or comfort others, providing emotional support or value.
Allocator	Demonstrate behaviors that reflect the role of a team leader, such as allocating and assigning tasks.
Observer	Provide additional information that one has observed.
Decision-maker	Provide a decision.
Evaluator	Evaluate the answers provided by oneself, the other person, and the team.
Task completer	Complete the task assigned by the other person.

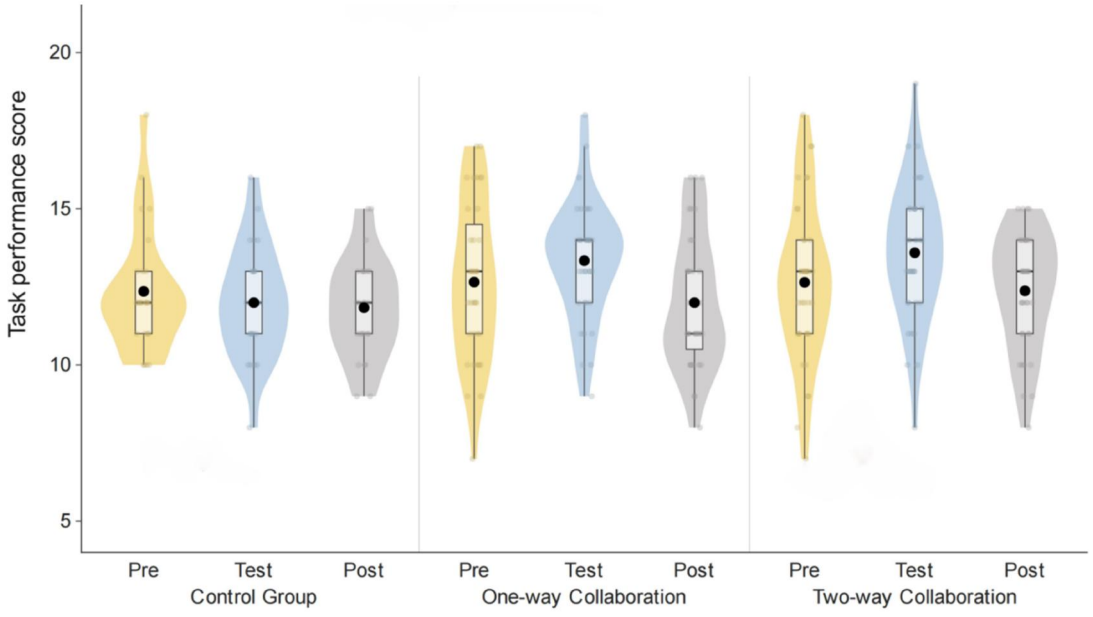


Figure 3. Task performance across pretest, intervention task, and post-test phases by experimental condition.

complementarity is not merely a matter of AI providing recommendations or humans accepting AI suggestions; rather, it requires humans and AI to flexibly adjust their roles, communication patterns, and coordination strategies according to task demands (Gonzalez & Heidari, 2025). Therefore, the AI team-role coding scheme constructed in this study aims to characterize the functional behaviors actually enacted by AI in human-AI dialogue. Specifically, this study drew on the four AI teammate roles proposed by Siemon and the group role framework proposed by Benne and Sheats, and reorganized and refined the relevant roles according to the task context of the present study(see Table 1) (Benne & Sheats, 1948). For example, Benne and Sheats' information seeker and opinion seeker both refer to seeking information, opinions, or judgments from others; therefore, they were merged into questioner in the human-AI dialogue annotation (Benne & Sheats, 1948). Encourager, harmonizer, and compromiser all serve to maintain interaction, alleviate tension, and support collaboration; therefore, they were merged into Emotional expresser. In addition, Siemon's doer emphasizes "pushes for concrete actions so that no time is wasted and can separate the important from the unimportant." In the present study, decision-maker and task completer represent two distinct observable behaviors; therefore, the relevant functions of doer were divided into these two categories (Siemon, 2022).

5. Results

5.1. Task performance in human–AI collaboration

To answer RQ1, we compared pre-service teachers' task performance across experimental conditions and task phases. As shown in Figure 3, the control group showed relatively stable score distributions

Table 2. Results of the 3 × 3 mixed-design ANOVA for task performance.

Source	DF1	DF2	*F*	*p*	η_p^2
Group	2	94	0.382	0.684	0.008
Time	1.866	175.442	15.529	<0.001	0.142
Time × group	3.733	175.442	4.187	0.004	0.082

across the three phases, whereas both HAIC groups showed higher scores during the intervention task and lower scores in the post-test.

Before comparing task performance during the intervention task, we first examined whether the three groups differed in baseline performance. A one-way ANOVA showed no significant difference in pretest performance among the one-way collaboration, two-way collaboration, and control groups ($F(2, 94) = 1.302$, $p = 0.277$, $\eta_p^2 = 0.027$). We calculated pre-to-intervention gain scores as the difference between intervention-task performance and pretest performance. A one-way ANOVA was conducted on the gain scores. Because the homogeneity of variance assumption was not met and group sizes were unequal, Dunnett's T3 was used for post-hoc comparisons. The results showed that both the one-way and two-way collaboration groups demonstrated significantly greater pre-to-intervention gains than the control group. Specifically, the gain score of the one-way collaboration group was significantly higher than that of the control group (M_diff = 2.074, $p < 0.001$), and the gain score of the two-way collaboration group was also significantly higher than that of the control group (M_diff = 2.279, $p < 0.001$). No significant difference was found between the one-way collaboration group and the two-way collaboration group (M_diff = 0.205, $p = 0.917$). Subsequently, a 3 (Group: control group, one-way collaboration group, and two-way collaboration group) × 3 (Time: pretest, intervention task, and post-test) mixed-design ANOVA was conducted(see Table 2). Mauchly's test of sphericity was significant ($p = 0.030$), indicating that the assumption of sphericity was violated; therefore, Greenhouse-Geisser-corrected results were reported. The results showed a significant main effect of Time, $F(1.866, 175.442) = 15.529$, $p < 0.001$, indicating that participants' overall performance changed significantly across the three task phases. The main effect of Group was not significant, $F(2, 94) = 0.382$, $p = 0.684$, indicating that the overall average performance of the three groups across the three phases did not differ significantly. The interaction effect between Time and Group was significant, $F(3.733, 175.442) = 4.187$, $p = 0.004$, indicating that the three groups showed different performance trajectories across the pretest, intervention task, and post-test phases. Further comparisons were conducted to examine performance changes from the intervention task to the post-test. The results showed that the mean post-test scores of all three groups were lower than those in the intervention task. The decline in the control group was small and not significant (M_diff = 0.160, $p = 0.983$). In contrast, the one-way collaboration group showed a significant decline from the intervention task to the post-test (M_diff = 1.343, $p = 0.004$); the two-way collaboration group also showed a significant decline (M_diff = 1.162, $p = 0.013$). These results indicate that HAIC was associated with an immediate performance advantage during the intervention task, but this advantage weakened in the post-test, particularly in both HAIC groups.

5.2. The complementarity in human–AI collaboration

To answer RQ2, we examined complementarity potential, complementarity rates, and their associations with task performance in HAIC. As shown in Figure 4(a), this experiment indicates a high theoretical potential for complementarity between pre-service teachers and AI in HAIC. If the team had always selected the correct initial judgment whenever either the human or the AI was correct, its oracle performance based on inherent complementarity could have reached 16.5.

However, only part of this complementarity potential was translated into realized complementarity effects. To further examine the extent to which complementarity potential was converted into complementarity effects, we analyzed the distribution of complementarity rates across collaboration modes. Specifically, three complementarity realization rates were examined: the inherent complementarity rate (ICR), the collaborative complementarity rate (CCR), and the collaboration counter-productivity rate (CPR). ICR measures the proportion of inherent complementarity potential that was translated into

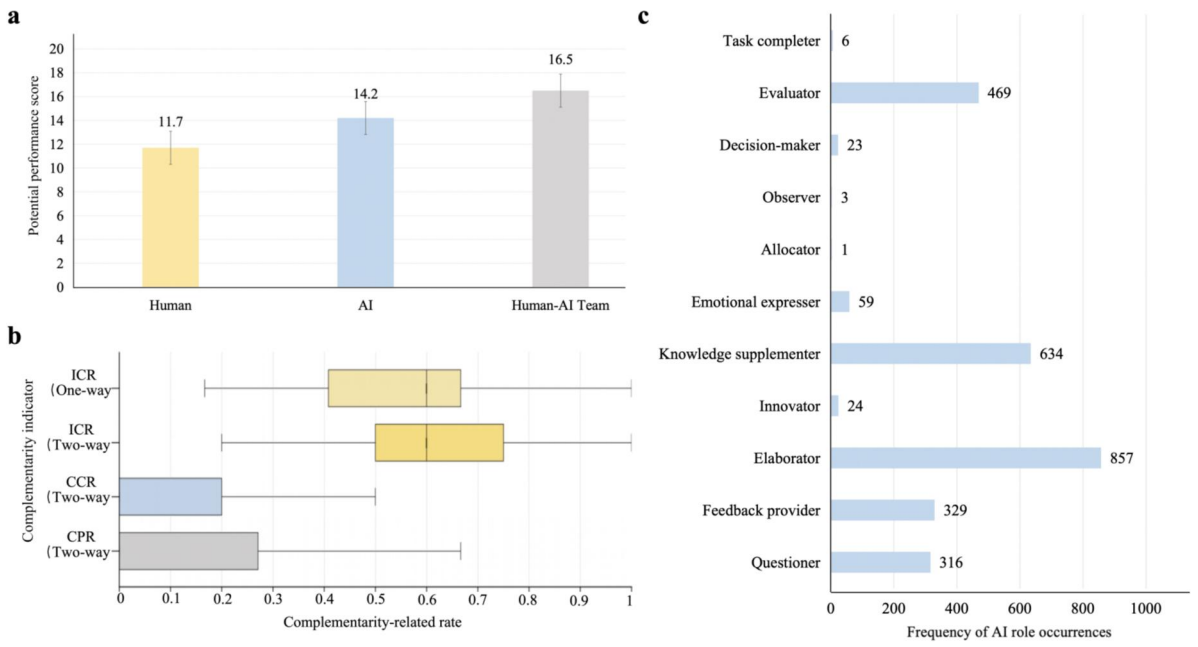


Figure 4. (a) Potential performance of humans alone, AI alone, and human–AI teams. (b) Distributions of ICR, CCR, and CPR across collaboration modes. (c) Frequency distribution of AI role labels in two-way collaboration.

inherent complementarity effects, whereas CCR measures the proportion of collaborative complementarity potential that was translated into collaborative complementarity effects. CPR measures the proportion of collaboration counter-productivity potential that was realized as collaboration counter-productivity effects. The distributions of these realization-related indicators are presented in Figure 4(b). The median inherent complementarity rate reached 0.60 in both one-way and two-way collaboration. The interquartile range was 0.26 in one-way collaboration (Q_1 = 0.41, Q_3 = 0.67) and 0.25 in two-way collaboration (Q_1 = 0.50, Q_3 = 0.75), indicating that a typical human–AI team was able to realize approximately 60% of its available inherent complementarity. In contrast, the median collaborative complementarity rate (CCR) was 0.00 (Q_1 = 0.00, Q_3 = 0.20), indicating that more than half of the teams failed to realize any collaborative complementarity. Counter-productivity rate (CPR) also exhibited a median value of 0.00 (Q_1 = 0.00, Q_3 = 0.27). The average inherent complementarity rate reached 59%, indicating that approximately 41% of the theoretically available inherent complementarity remained unrealized during collaboration. The findings reveal a persistent realization gap between complementarity potential and complementarity effect. Inherent complementarity was more readily realized than collaborative complementarity. While human–AI teams often benefited from differences in their respective strengths, they rarely produced correct solutions when neither the human nor the AI initially possessed the correct answer. Figure 4(b) further shows that counter-productivity was present in a subset of HAIC. This suggests that interaction alone does not guarantee effective complementarity and may occasionally undermine performance when complementary information is not successfully identified or integrated. To further investigate this, we analyzed the correlation between the inherent complementarity rate and intervention-task performance across both the one-way and two-way collaboration groups. In the one-way collaboration group, Shapiro–Wilk tests indicated that the relevant variables were approximately normally distributed ($p > 0.05$). Pearson correlation analysis showed a significant positive correlation between the inherent complementarity rate and intervention-task performance ($r = 0.758$, $p < 0.001$). In the two-way collaboration group, Shapiro–Wilk tests also confirmed the normality of both the inherent complementarity rate ($p = 0.701$) and intervention-task performance ($p = 0.653$). Pearson correlation analysis revealed a significant positive correlation between the inherent complementarity rate and intervention-task performance ($r = 0.57$, $p < 0.001$). These results indicate that, across both one-way and two-way collaboration, intervention-task performance was closely associated with the extent to which human–AI teams successfully realized inherent complementarity.

Table 3. Descriptive statistics of AI role and dialogue indicators in the two-way collaboration condition.

Indicator	Mean	*SD*	Median	Min	Max
Tag	73.54	29.27	70	17	126
Tag type	6.86	1.00	7	5	9
Turn	23.97	9.43	22	6	44

5.3. Functional roles of AI in two-way collaboration

To answer RQ3, we used ChatGPT-4o to conduct multi-label annotation of the interaction data from the two-way collaboration condition based on the AI role coding scheme presented in Table 1. The annotation prompt is provided in Appendix 1. After ChatGPT-4o generated the initial annotations, the researchers manually reviewed all annotation results. Labels that did not follow the coding rules or lacked sufficient semantic support were revised according to the role definitions in Table 1. In total, the two-way collaboration group contained 887 turns and 2721 AI role tags. The frequency distribution of each AI role is shown in Figure 4(c). The top three roles, namely elaborator (857, 31.50%), knowledge supplementer (634, 23.30%), and evaluator (469, 17.24%), together accounted for 72.03% of all tags. When feedback provider (329, 12.09%) and questioner (316, 11.61%) were further included, the top five roles accounted for 95.74% of all tags. This indicates that AI roles were highly concentrated, mainly around elaboration, knowledge supplementation, evaluation, feedback provision, and questioning. In addition, this study extracted interaction indicators including tag, tag type, and turn (see Table 3). Turn refers to the number of interaction rounds in each human-AI dialogue. Tag refers to the total number of AI role labels identified in each human-AI dialogue. Tag type refers to the number of distinct functional AI role categories appearing in each human-AI dialogue, which was used to describe the breadth of AI roles. The two-way collaboration dialogues contained an average of 73.54 AI role tags per participant ($SD = 29.27$, Median = 70), with an average of 23.97 turns ($SD = 9.43$, Median = 22). The mean tag type was 6.86 ($SD = 1.00$, Median = 7), suggesting that each human–AI dialogue typically involved approximately seven functional AI role categories.

6. Discussion

6.1. AI-supported collaboration improved immediate performance, but the advantage was not sustained in the post-test

We compared score differences among the two-way collaboration group, the one-way collaboration group, and the control group across the pretest, intervention task, and post-test phases. Both the one-way collaboration group and the two-way collaboration group showed significantly greater improvement from the pretest to the intervention task than the control group, indicating that HAIC improved participants' immediate performance during the intervention task. The mean post-test scores of all three groups were lower than their scores in the intervention task, and both the one-way and two-way collaboration groups showed significant declines from the intervention task to the post-test. These findings suggest that, in the present task, the advantage of HAIC was mainly reflected in immediate support during the intervention task, whereas evidence for the persistence of this advantage in post-test performance was limited.

This finding is consistent with Akgun and Toker's conclusion that generative AI can enhance learners' immediate task performance, but whether such gains are sustained as later learning outcomes depends on how subsequent AI-supported practice is organized (Akgun & Toker, 2026). One possible explanation is that, in unstructured HAIC, humans may engage in cognitive outsourcing: they tend to directly rely on AI-generated predictions without necessarily engaging in sufficient cognitive participation, explanation generation, or knowledge internalization (Gajos & Mamykina, 2022). From a cognitive psychology perspective, reliance on external support may involve cognitive offloading, whereby external tools reduce the internal information-processing demands of a task (Risko & Gilbert, 2016), in which individuals use physical actions or external tools to reduce the information-processing demands of a task and lower cognitive load (Gilbert et al., 2020). As a result, performance may decline when such cognitive support tools are no longer available (Richmond & Taylor, 2025).

6.2. Inherent complementarity was more readily realized than collaborative complementarity

The results showed that there was a potential complementarity space between pre-service teachers and AI. More specifically, the median ICR was 0.60 in both HAIC modes, suggesting that human-AI teams were able to translate approximately 60% of inherent complementarity potential into inherent complementarity effects. In contrast, the overall collaborative complementarity rate was low, indicating that when both the human and the AI initially made incorrect judgments, human-AI teams rarely generated new correct answers through collaboration. The realization of inherent complementarity essentially involves selecting an existing correct answer, whereas the realization of collaborative complementarity requires both parties to re-reason, revise their representations, and generate a new answer when both initial judgments are incorrect. Collaborative complementarity may require more extensive re-reasoning and representation revision than inherent complementarity, because neither party initially possesses the correct answer. Bansal et al. expressed a similar view, arguing that an ideal human-AI team should have minimally overlapping mistakes so that each party has greater opportunities to correct the other's errors (Bansal et al., 2021). In both one-way and two-way collaboration, the inherent complementarity rate was significantly and positively correlated with intervention-task performance. When the initial human and AI judgments differed and one of them was correct, realizing inherent complementarity required the team to retain the initially correct contribution. In one-way collaboration, this would require participants to appropriately accept or reject the AI-generated annotation. In two-way collaboration, dialogue additionally allowed participants to question, verify, and compare the available judgments before reaching a final decision. Related studies support a similar argument: task performance is more likely to depend on whether learners can achieve appropriate reliance and judgment regulation under conditions of ability or information asymmetry, rather than merely on the form of dialogic interaction (Chen et al., 2023). Although AI explanations in the two-way collaboration group may increase the likelihood that humans accept AI suggestions, the present findings show that more AI information or explanatory support does not necessarily lead to better complementarity realization. One possible explanation is that human reliance on AI is not a simple matter of acceptance or rejection, but a process of judgment regulation based on task structure and instance-level information. Learners need not only to understand the available information, action space, and evaluation criteria in order to judge which decision is better, but also to form prior beliefs about AI reliability and update these beliefs after observing specific task information and AI recommendations. Only when learners are able to judge whether AI is more reliable in a specific instance can theoretical complementarity potential be translated into realized complementarity effects (Guo et al., 2025).

6.3. AI participation in two-way collaboration was dominated by assistive roles

In the two-way collaboration condition, dialogues contained an average of 23.97 turns and 6.86 distinct AI role categories per participant. Despite this role breadth, AI participation was highly concentrated in five assistive functions: elaboration, knowledge supplementation, evaluation, feedback provision, and questioning suggesting that AI participation was concentrated in supportive rather than decision-making functions. This role pattern also resonates with Guo et al.'s decision-theoretic framing, in which AI often functions as an informational recommender while the human remains the final decision-maker (Guo et al., 2025). The low frequencies of decision-maker, task completer, allocator, and innovator further indicate that AI did not substantially take on roles related to task planning, final decision-making, or new answer generation. From a cognitive science perspective, people tend to expect AI agents to perform significantly better on average (Kelly et al., 2023). One possible explanation is that, when participants perceive AI as a more stable and more capable information source, they are more likely to assign AI functions related to knowledge supplementation, explanation, evaluation, and feedback, rather than treating AI as a social teammate with whom responsibilities need to be negotiated and jointly distributed. Vaccaro et al. further noted that, in many decision-making tasks, both humans and AI provide complete decisions, with humans typically making the final choice; in such settings, human-AI combinations often produce performance losses rather than synergy (Vaccaro et al., 2024). They argued that human-AI synergy requires humans to be better at some parts of a task and AI to be better at other

parts. Taken together, these findings suggest that although two-way collaboration involved human-AI dialogue, AI was more likely to remain in supportive functions for human decision-making when explicit mechanisms for task allocation, shared control, or role selection were absent. This may explain why, even with sustained dialogue in two-way collaboration, human-AI teams did not necessarily generate new correct annotations when both the human and the AI initially made incorrect judgments. This assistive role profile may have provided limited opportunities for joint answer generation.

7. Conclusion

The findings of this study suggest that differences in initial judgments between pre-service teachers and AI created complementarity potential in HAIC. However, this potential was only partially realized. HAIC improved participants' performance during the intervention task, but the evidence for stable post-test gains was limited. In addition, complementarity was realized mainly through inherent complementarity, that is, by identifying and using an existing correct answer from either the human or the AI. Collaborative complementarity, which required the human-AI team to generate a correct answer when both parties initially answered incorrectly, was rarely realized.

These findings indicate that HAIC should not be understood as an automatically beneficial educational model. Its value depends on whether learners can appropriately evaluate AI outputs, regulate their reliance on AI, and transform human-AI differences into realized complementarity effects. One limitation of this study is that the human-AI collaborative task lasted for a relatively short period, which limited our ability to examine the stability of post-test gains and longer-term changes in participants' annotation performance. Future research should examine more complex and longer-term experimental tasks to obtain clearer evidence regarding the development of pre-service teachers' task performance. In addition, future studies may explore more diverse forms of HAIC, such as multimodal interaction and explicit role allocation mechanisms.

Ethical approval

This study has received ethical review approval (Approval No. HR2-0028-2025) from the Academic Ethics Committee of East China Normal University. All procedures involving human participants were conducted in accordance with institutional and national research ethics guidelines.

Consent to participate

Informed consent was obtained from all participants involved in the study. Participants were informed of their rights to withdraw at any stage without penalty.

Consent for publication

All participants provided consent for anonymized data to be used for academic publication. No identifiable personal information is included in this manuscript.

Author contributions

CRediT: **Bo Jiang**: Conceptualization, Funding acquisition, Project administration, Supervision, Validation, Writing – review & editing; **Jiatong Wang**: Conceptualization, Formal analysis, Methodology, Project administration, Validation, Visualization, Writing – review & editing; **Yijie Lu**: Conceptualization, Data curation, Methodology, Software, Visualization, Writing – original draft; **Jiayi Liu**: Data curation, Formal analysis, Investigation, Visualization, Writing – original draft; **Yuanhao Jiang**: Conceptualization, Funding acquisition, Investigation, Resources, Visualization.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Natural Science Foundation of China under Grant 62477012; the AI for Science Program, Shanghai Municipal Commission of Economy and Informatization, under Grant 2025-GZL-RGZN-BTBX-01014; and the Special Foundation for Interdisciplinary Talent Training in "AI Empowered Psychology / Education" under Grant 2024JCRC-03.

ORCID

Jiatong Wang http://orcid.org/0009-0000-8820-4398

Data availability statement

The de-identified dataset and supporting documentation have been deposited in Figshare under the reserved DOI [10.6084/m9.figshare.33094661]. The dataset is not publicly available due to ethical and privacy considerations related to participant data. Access to the dataset can be requested from the corresponding author upon reasonable request.

Materials and code availability

All experimental materials and analytical scripts are available from the authors upon reasonable request.

References

Akgun, M., & Toker, S. (2026). *Do gains from generative AI-enabled adaptive pretesting persist? Evidence from a retention study*. https://arxiv.org/abs/2606.22328

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*(1), 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445717

Barnes, W. H. F. (1944). The nature of explanation. *Nature*, *153*(3890), 605–605. https://doi.org/10.1038/153605a0

Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*, *34*(4), 1582–1602. https://doi.org/10.1287/isre.2023.1199

Benne, K. D., & Sheats, P. (1948). Functional roles of group members. *Journal of Social Issues*, *4*(2), 41–49. https://doi.org/10.1111/j.1540-4560.1948.tb01783.x

Caldwell, S., Sweetser, P., O'donnell, N., Knight, M. J., Aitchison, M., Gedeon, T., Johnson, D., Brereton, M., Gallagher, M., & Conroy, D. (2022). An agile new research framework for hybrid human-AI teaming: Trust, transparency, and transferability. *ACM Transactions on Interactive Intelligent Systems*, *12*(3), 1–36. https://doi.org/10.1145/3514257

Chen, V., Vera Liao, Q., Wortman Vaughan, J., & Bansal, G. (2023). *Understanding the role of human intuition on reliance in human-AI decision-making with explanations*. https://arxiv.org/abs/2301.07255 https://doi.org/10.1145/3610219

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, *61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

Duan, X., Ho, C.-J., & Yin, M. (2022). The influences of task design on crowdsourced judgement: A case study of recidivism risk evaluation. In *Proceedings of the ACM Web Conference 2022 (WWW '22)* (pp. 1685–1696). Association for Computing Machinery. https://doi.org/10.1145/3485447.3512239

Fragiadakis, G., Diou, C., Kousiouris, G., & Nikolaidou, M. (2025). Evaluating human-AI collaboration: A review and methodological framework. arXiv preprint arXiv:2407.19098. https://arxiv.org/abs/2407.19098

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, *25*(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814

Gajos, K. Z., & Mamykina, L. (2022). Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In G. Jacucci, S. Kaski, C. Conati, S. Stumpf, T. Ruotsalo, & K. Z. Gajos (Eds.), *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22)* (pp. 794–806). Association for Computing Machinery. https://doi.org/10.1145/3490099.3511138

Gilbert, S. J., Bird, A., Carpenter, J. M., Fleming, S. M., Sachdeva, C., & Tsai, P.-C. (2020). Optimal use of reminders: Metacognition, effort, and cognitive offloading. *Journal of Experimental Psychology. General*, *149*(3), 501–517. https://doi.org/10.1037/xge0000652

Gomez, C., Cho, S. M., Ke, S., Huang, C.-M., & Unberath, M. (2025). Human-AI collaboration is not very collaborative yet: A taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science*, *6*, 1521066. https://doi.org/10.3389/fcomp.2024.1521066

Gonzalez, C., Donahue, K., Goldstein, D. G., Heidari, H., Jalali, M. S., Schelble, B., Singh, A., & Woolley, A. W. (2026). Toward a science of human-AI teaming for decision making: A complementarity framework. *PNAS Nexus*, *5*(3), pgag030. https://doi.org/10.1093/pnasnexus/pgag030

Gonzalez, C., & Heidari, H. (2025). A cognitive approach to human–AI complementarity in dynamic decision-making. *Nature Reviews Psychology*, *4*(12), 808–822. https://doi.org/10.1038/s44159-025-00499-x

Gonzalez, C., & Malloy, T. (2026). Toward complementary intelligence: Integrating cognitive and machine AI. *Current Directions in Psychological Science*, *35*(3), 122–130. https://doi.org/10.1177/09637214251407571

Guo, Z., Wu, Y., Hartline, J., & Hullman, J. (2025). *A decision theoretic framework for measuring AI reliance*. https://arxiv.org/abs/2401.15356

Harfouche, A., Quinio, B., Saba, M., & Saba, P. B. (2023). The recursive theory of knowledge augmentation: Integrating human intuition and knowledge in artificial intelligence to augment organizational knowledge. *Information Systems Frontiers*, *25*(1), 55–70. https://doi.org/10.1007/s10796-022-10352-8

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2024). *Complementarity in human-AI collaboration: Concept, sources, and evidence*. https://arxiv.org/abs/2404.00029

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2020). *Aligning AI with shared human values*. arXiv preprint arXiv:2008.02275. arXiv. https://arxiv.org/abs/2008.02275

Holstein, K., & Aleven, V. (2022). Designing for human-AI complementarity in K-12 education. *AI Magazine*, *43*(2), 239–248. https://doi.org/10.1002/aaai.12058

Holstein, K., De-Arteaga, M., Tumati, L., & Cheng, Y. (2023). Toward supporting perceptual complementarity in human-AI collaboration via reflection on unobservables. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–20. https://doi.org/10.1145/3579628

Jarrahi, M. H., Askay, D., Eshraghi, A., & Smith, P. (2023). Artificial intelligence and knowledge management: A partnership between human and AI. *Business Horizons*, *66*(1), 87–99. https://doi.org/10.1016/j.bushor.2022.03.002

Jiang, Y.-H., Chen, Z.-W., Zhao, C., Tang, K., Duan, J., & Zhou, Y. (2025). Explainable learning outcomes prediction: Information fusion based on grades time-series and student behaviors. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)* (pp. 1–11). Association for Computing Machinery. https://doi.org/10.1145/3706599.3721212

Ju, H., Aral, S. (2026). *Personality pairing improves human-AI collaboration*. https://arxiv.org/abs/2511.13979

Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research*, *32*(3), 713–735. https://doi.org/10.1287/isre.2020.0980

Kelly, M., Kumar, A., Smyth, P., & Steyvers, M. (2023). Capturing humans' mental models of AI: An item response theory approach. In L. Nannini, A. Balayn, & A. L. Smith (Eds.), *Proceedings of the 2023 ACM Conference on Fairness Accountability and Transparency (FAccT '23)* (pp. 1723–1734). Association for Computing Machinery. https://doi.org/10.1145/3593013.3594111

Koong Lin, H. C., Tsai, M. C., Wang, T. H., & Lu, W. Y. (2025). Application of WSQ (watch-summary-question) flipped teaching in affective conversational robots: impacts on learning emotion, self-directed learning, and learning effectiveness of senior high school students. *International Journal of Human–Computer Interaction*, *41*(7), 4319–4336. https://doi.org/10.1080/10447318.2024.2351708

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In d. boyd & J. H. Morgenstern (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT '19)* (pp. 29–38). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287590

Licklider, J. C. R. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, *HFE-1*(1), 4–11. https://doi.org/10.1109/THFE2.1960.4503259

Liu, C.-W., Yang, M., & Wen, M.-H. (2023). Judge me on my losers: Do robo-advisors outperform human investors during the COVID-19 financial market crash? *Production and Operations Management*, *32*(10), 3174–3192. https://doi.org/10.1111/poms.14029

Louie, R., Nandi, A., Fang, W., Chang, C., Brunskill, E., & Yang, D. (2024). Roleplay-doh: Enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles. arXiv preprint arXiv: 2407.00870. https://arxiv.org/abs/2407.00870

Lu, Z., Wang, D., & Yin, M. (2024). Does more advice help? The effects of second opinions in AI-assisted decision making. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–31. https://doi.org/10.1145/3653708

Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., & Perer, A. (2024). The impact of imperfect XAI on human-AI decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–39. https://doi.org/10.1145/3641022

Richmond, L. L., & Taylor, R. G. (2025). The benefits and potential costs of cognitive offloading for retrospective information. *Nature Reviews Psychology*, *4*(5), 312–321. https://doi.org/10.1038/s44159-025-00432-2

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Siemon, D. (2022). Elaborating team roles for artificial intelligence-based teammates in human-AI collaboration. *Group Decision and Negotiation*, *31*(5), 871–912. https://doi.org/10.1007/s10726-022-09792-z

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852

Song, Y., Kim, J., Liu, Z., Li, C., & Xing, W. (2026). Students' perceived roles, opportunities, and challenges of a generative AI-powered teachable agent: A case of middle school math class. *Journal of Research on Technology in Education*, *58*(3), 535–554. https://doi.org/10.1080/15391523.2024.2447727

Steyvers, M., & Kumar, A. (2024). Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science*, *19*(5), 722–734. https://doi.org/10.1177/1745691623118118

Sultana Samu, M. S., Khan, N., Toufique Elahi, K., Binte Rahman, T., Rakibul Islam, M., Sadeque, F. (2026). *AI as teammate or tool? A review of human-AI interaction in decision support.* https://arxiv.org/abs/2602.15865

Torrey, L., & Shavlik, J. (2010). Transfer learning. In E. S. Olivas, J. D. M. Guerrero, M. Martinez-Sober, J. R. Magdalena-Benedito, & A. J. Serrano López (Eds.), *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques* (pp. 242–264). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-60566-766-9.ch011

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, *8*(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., & Wang, Q. (2020). From human-human collaboration to human-AI collaboration: Designing AI systems that can work together with people. In R. Bernhaupt, F. F. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjørn, S. Zhao, B. P. Samson, & R. Kocielnik (Eds.), *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–6). Association for Computing Machinery. https://doi.org/10.1145/3334480.3381069

Zhang, Q., Lee, M. L., & Carter, S. (2022). You complete me: Human-AI teams and complementary expertise. In S. D. J. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. M. Drucker, J. R. Williamson, & K. Yatani (Eds.), *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)* (pp. 1–28). Association for Computing Machinery. https://doi.org/10.1145/3491102.3517791

Zhang, R., McNeese, N. J., Freeman, G., & Musick, G. (2021). An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW3), 1–25. https://doi.org/10.1145/3432945

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In M. Hildebrandt & C. Castillo (Eds.), *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT '20)* (pp. 295–305). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372852

About the authors

Bo Jiang is a Professor at East China Normal University and Deputy Dean of the Shanghai Institute of Intelligent Education. His research focuses on educational agents, explainable learner modeling, and AI literacy education. He has served on editorial boards of international SCI/SSCI journals and chaired GCCCE 2023 and ICCE 2025.

Jiatong Wang is a PhD student at the Shanghai Institute of Artificial Intelligence for Education, East China Normal University, China. Her research focuses on human–AI collaboration, artificial intelligence in education, and learning analytics, with particular interests in human–AI collaborative problem solving.

Yijie Lu is an MS student in the Department of Educational Information Technology at East China Normal University, China. Her research focuses on human–AI collaboration and synergy in educational contexts, with particular interests in collaborative patterns and the synergistic interplay between humans and AI systems in learning environments.

Jiayi Liu is an MS student in the Department of Educational Information Technology at East China Normal University, China. Her research focuses on multi-agent-based educational systems and human–computer interaction.

Yuanhao Jiang is currently a PhD candidate at the Shanghai Institute of Artificial Intelligence for Education, East China Normal University. He is currently visiting the National Institute of Education, Nanyang Technological University. His research interests include AI for Education, large language models, and human–computer interaction.

Appendix

Experimental tasks

Based on the experimental objectives, we designed a collaborative problem-solving task in which pre-service teachers completed mathematical problem annotation tasks with or without AI support. The computer-assisted decision-making scenario is a relatively mature experimental scenario in the field of human-AI interaction. In this scenario, we designed a mathematical problem label annotation task. In the task, participants were shown 20 mathematical problems and asked to choose the most appropriate one from four literacy labels (operation, identification, recall, strategy selection). We chose the task for the following reasons: 1) AI-assisted decision-making is currently a common form of human-AI collaboration, with a sufficient theoretical basis. Choosing this task can make the results more universal. 2) The task of annotating mathematical problems is a well-structured problem that facilitates the calculation of accuracy and complementarity. 3) In the field of AI-empowered digital education, large-scale semi-automatic labeling is an educational problem needs to be addressed. The platform supported pre-service teachers' annotation tasks through two collaboration modes: one-way collaboration and two-way collaboration.

Experimental platform

Participants completed all experimental tasks on the experimental platform that we implemented, including the pretest, intervention task, and post-test. Figure 5 provides an overview of the experimental platform and user interfaces. Specifically, Figure 5(a) shows the main interface, Figure 5(b) shows the personal mode used by the control group, Figure 5(c) shows the one-way collaboration mode used by Experimental Group 1, and Figure 5(d) shows the two-way collaboration mode used by Experimental Group 2.

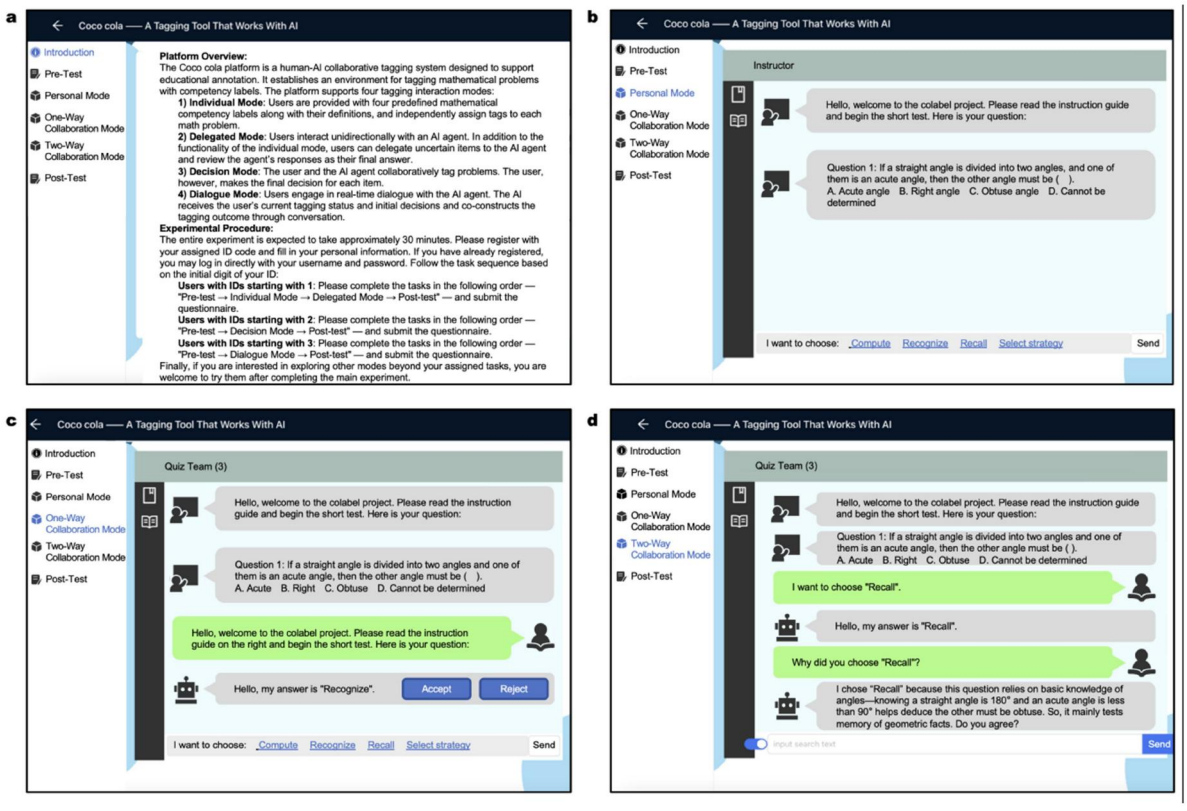


Figure 5. Overview of the experimental platform and user interfaces. Subfigure (a) shows the main interface, which provides participants with an overview of the experimental procedure and operational instructions. Subfigure (b) displays the interface for the control group, where pre-service teachers independently select tags for mathematical problems without AI assistance. Subfigure (c) illustrates the interface for Experimental Group 1: after pre-service teachers select initial tags, the AI immediately presents its own suggestion, and pre-service teachers decide whether to accept or reject the AI's input to produce the final answer. Subfigure (d) demonstrates the interface for Experimental group 2: following the initial tagging, participants are redirected to a chat-based interface that supports bidirectional discussion with the AI before finalizing their response.

We designed a collaborative mathematical problem annotation platform, CocoCola. The basic function of the platform is to provide participants with mathematical problems and require them to choose the literacy labels covered by the problems. Specifically, users will complete the pretest, intervention task, and post-test in sequence according to the instructions on the main page.

The platform mainly includes three annotation modes: personal mode (Figure 5(b)), one-way collaboration mode (Figure 5(c)), and two-way collaboration mode (Figure 5(d)). The personal mode was used by the control group. Participants are informed of the definitions of four labels and independently choose the literacy labels for all mathematical problems. The one-way collaboration mode (Figure 5(c)) was used by experimental group 1.In the one-way collaboration mode, the AI partner presented its annotation to participants through the interface, and the pre-service teachers take the role of decision-makers to decide whose answer to use when completing the task. However, participants could not communicate with the AI in this mode; therefore, the AI partner did not have access to participants' reasoning processes.

In the two-way collaboration mode (Figure 5(d)), which was used by Experimental Group 2, pre-service teachers could communicate with the AI agent through a chat box.

The AI agent will obtain information such as the participant's current response and initial answer, and work together with the user to obtain the label of the mathematical problem. Pre-service teachers' individual information, answer records between users and AI, communication records, and other data are collected by the platform. The AI agents on this platform all use the ChatGPT-4o interface, and the prompts are as follows.

Prompts for math problem annotation task

Scenario: You are involved in the task of assigning competency labels to math problems.

Requirement: Please read the mathematical questions asked by the user and clarify that 1) recall refers to the definition of recall (the difference between recall and identification is that recall definitions often provide definitions and explanations for nouns, and then ask pre-service teachers to answer what they are), terminology, unit rate, operation laws, geometric properties (such as triangle interior angles and 180 degrees), and operation formulas (such as the formula for rectangular area $S = ab$; distance = velocity $\times$ time). 2) Identification refers to the identification of numbers, expressions, quantities, and common geometric shapes (the difference from recall is judged based on the properties of the shapes, such as the fact that all faces of a rectangular prism are rectangles), and the identification of equivalent relationships between different presentation forms in mathematics (such as the equality of sizes between fractions, decimals, and percentages; simple geometric shapes presented from different angles). 3) Operations refer to the $+$, $-$, $\times$, and $\div$ operations, estimation, and simple algebraic operations performed on natural numbers, integers, fractions, decimals, and their combinations. 4) The strategy selection refers to determining effective and appropriate computational strategies and tools to solve conventional mathematical problems. The strategy selection does not require complex strategy selection, nor does it require the problem to contain multiple solutions. As long as the problem does not directly require pre-service teachers to calculate, but rather allows pre-service teachers to find a solution, the difference between the strategy selection and computation is that the solution method is relatively vague in the question stem. Choose from these four tags, and you must answer the tag you have chosen (two or four words).

Case: The adjacent sides of a parallelogram surrounded by a wire are 12 centimeters and 6 centimeters, respectively. The circumference of this parallelogram is x centimeters. Your reply is: operation. Please read the instructions carefully! If you make a mistake, you will be punished! The following are the questions that require labels:

Prompts for two-way collaboration mode

Scenario: You are a pre-service teacher majoring in mathematics education and are currently collaborating with me to complete a task: assigning mathematical literacy labels to mathematical problems.

Requirement: In this process, we will discuss together the selection of matching competency labels for each math problem, and there are optional competency labels and definitions in the Reference Information below. In my conversation, I will inform you of your own answer and mine. Please always remember this information to maintain consistency in the conversation!! First, you must answer the questions I have raised or complete the tasks I have assigned to you. Second, you can refer to the dialogue mode in the Reference Information below to have a conversation, but be sure to keep the conversation context appropriate and brief.

Supplementary Materials: 1) Available tags: Memory, Memory Definition, Terminology, Unit Rate, Laws of Operations, Geometric Properties, and Operation Formulas, etc.; Identify, recognize numbers, expressions, quantities, and common geometric shapes. Identify equivalent relationships between different forms of presentation in mathematics; Operations refer to the $+$, $-$, x, $\div$ operations, estimation, and simple algebraic operations performed on natural numbers, integers, fractions, decimals, and their combinations; Select strategies, determine effective and appropriate computational strategies and tools to solve conventional mathematical problems. 2) In addition to answering the questions I have raised, you can also ask the other person about their labeling strategy; Explain why you chose this answer; Agree or praise my ideas; Provide feedback on my explanation (agree or disagree); Provide me with some knowledge; Propose other possibilities beyond my own and my answers; Express whether you have confidence in your answer, compare the explanations and understanding of two possible answers, etc.