Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Effects of AI-Assisted Case Discussion and Virtual Patient Consultation on Medical Students' Learning Outcomes

Hongli Gao

gaohongli@hamu.edu.cn

Henan Medical University

Bing Zhao

Henan Medical University

Xing Hu

Henan Medical University

Chang Liu

Henan Medical University

Dongjun Zhang

Henan Medical University

Hongxing Zhang

Henan Medical University

Research Article

Keywords: AI in education, medical education, virtual patient, case discussion

Posted Date: September 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10644971/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Artificial intelligence (AI) holds broad application prospects in medical education; however, systematic empirical evidence regarding its effectiveness in facilitating case discussions and virtual patient consultation remains scarce. This study recruited undergraduate medical students and conducted experiments to examine the pedagogical effects of AI-assisted case discussion and virtual consultation.

Methods

A total of 207 medical undergraduates participated in this study. Using the self-developed intelligent experimental platform, medical students engaged in AI-assisted case discussion and virtual consultation. Experiment 1 employed a single-factor, three-level between-subjects design to compare the effects of AI-assisted case discussion, peer case discussion, and a learning-only control condition on medical students' knowledge acquisition, intrinsic motivation, and interaction quality. Experiment 2 utilized a single-factor, four-level between-subjects design to examine the effects of virtual patient consultation and the sequencing of AI-assisted case discussion and virtual consultation on medical students' knowledge acquisition, intrinsic motivation, and consultation process quality. The control group participants were the same individuals across both experiments.

Results

Results indicated that the AI-assisted case discussion group achieved significantly higher knowledge scores than the control group. The virtual consultation group, the consultation-then-discussion group, and the discussion-then-consultation group all outperformed the control group in knowledge acquisition. The discussion-then-consultation group demonstrated higher intrinsic motivation and consultation comprehensiveness, yet exhibited lower empathy than the consultation-then-discussion group.

Conclusions

These findings suggest that both AI-assisted case discussion and virtual patient consultation can effectively enhance medical students' knowledge acquisition, while different intervention sequences exert differential effects on intrinsic motivation, consultation comprehensiveness, and empathy. The results provide theoretical foundations and practical guidance for AI-empowered medical education.

1 Background

The rapid development of artificial intelligence (AI) is profoundly transforming various industries, and medical education is no exception. In particular, large language models (LLMs), with their powerful

natural language processing capabilities, are gradually being integrated into medical education, reshaping students' learning experiences and enhancing their knowledge, skills, and diagnostic competencies. Trained on massive datasets, these models can automatically perform a variety of tasks, including clinical case discussion, virtual patient simulation, and personalized learning guidance, thus providing unprecedented opportunities for medical education [1].

Traditional teaching methods (such as classroom lectures, video-based learning, and clinical clerkships in teaching hospitals), although capable of covering foundational knowledge, often lack interactivity and hands-on practice, making it difficult to meet the training needs for clinical communication and diagnostic reasoning skills [2]. To address this, pedagogical approaches such as case-based discussion and role-playing have been widely adopted in medical education, helping students transition from knowledge acquisition to clinical application in simulated environments [3].

Case-based discussion provides opportunities to integrate theoretical knowledge with real clinical scenarios. However, previous implementations of case discussions often fell short of ideal learning outcomes due to delayed teacher feedback, insufficient peer interaction, and low discussion quality [4]. The introduction of AI technology offers new possibilities for case discussion. AI systems, through their natural language understanding and generation capabilities, can provide students with structured guidance, immediate feedback, and targeted prompts, which help stimulate learning motivation and promote deep learning [5, 6]. Despite the potential advantages of AI in facilitating case discussions, empirical research on implementation methods and learning effectiveness remains lacking.

Meanwhile, although standardized patient (SP) training in role-playing scenarios can improve students' communication and diagnostic skills, its implementation costs are high, and it is subject to considerable human variability, making it difficult to conduct frequent, standardized training sessions [7, 8]. To overcome this challenge, researchers have begun exploring AI-assisted virtual standardized patients (VSPs) for medical student education [9, 10]. Through interactive consultations with VSPs, students can practice history-taking, clinical reasoning, diagnostic decision-making, and communication skills in a safe, controlled environment [11, 12, 13]. This approach not only reduces the time and cost required for SP training but also allows students to engage in repeated practice and provides objective assessments of teaching effectiveness [14, 15, 16]. However, most current VSP implementations in teaching remain fragmented and used as add-on tools, failing to integrate organically with course content and instructional workflows [17, 18]. Teaching effectiveness remains inconsistent, and systematic empirical research on how to optimally integrate VSPs into curricula is still lacking.

Experiential learning theory posits that learning is a dynamic process of creating knowledge through the transformation of experience [19]. The learning cycle comprises four stages: concrete experience, reflective observation, abstract conceptualization, and active experimentation. Based on this theory, we can examine two different learning sequences and their effects on learning outcomes. One sequence is "case discussion before virtual consultation", proceeding from abstract to concrete: learners first establish abstract understanding through case discussion, then apply theory to practice in virtual

consultations. The other sequence is "virtual consultation before case discussion," proceeding from concrete to abstract: learners first accumulate experience through direct virtual consultations, then distill theoretical knowledge through case discussion. Investigating these two different learning sequences helps us gain deeper insight into how different learning pathways affect learning outcomes, thereby providing more targeted guidance for educational practice.

To this end, the present study uses depression disorder instruction as an example to investigate the effects of AI-assisted case discussion and virtual patient consultation on medical students' learning outcomes. Study 1 examines the effects of AI-assisted case discussion on medical students' knowledge acquisition and intrinsic motivation. Study 2 explores the effects of AI-assisted virtual consultation and the sequencing of case discussion and consultation on learning outcomes. This research aims to: (1) validate the pedagogical effectiveness of AI-assisted case discussion and virtual consultation in terms of knowledge, motivation, and practical competence; and (2) examine the effects of different instructional activity sequences on learning outcomes. By deeply integrating AI-assisted case discussion and virtual consultation into the instructional workflow, this study explores their combined effects on students' knowledge acquisition, diagnostic competence, and intrinsic motivation. The findings will provide new perspectives and evidence for the effective application of AI in education, helping educators better design learning pathways and curriculum structures to enhance student learning outcomes.

2 Experiment 1: Effects of AI-Assisted Case Discussion on Learning Outcomes

2.1.1 Research Design and Participants

Experiment 1 employed a single-factor, three-level between-subjects design. The experimental conditions included an AI-assisted case discussion group (one participant discussing with an AI), a peer case discussion group (two participants discussing together), and a control group (one participant learning only, no discussion). Dependent variables included knowledge test scores, intrinsic motivation, and the quality of case discussion interactions. Interaction quality was assessed using the interaction analysis model proposed by Gunawardena et al. [20], with two researchers manually coding the textual content.

A total of 120 students from a medical university were recruited. One participant's data was excluded due to an excessively short questionnaire completion time (below three standard deviations), leaving 119 participants aged 15–26 years (M = 18.61, SD = 1.25). All participants had normal or corrected-to-normal vision with no color blindness or color weakness. Participants were randomly assigned to the AI-assisted case discussion group (30 groups: 5 males, 25 females), the peer discussion group (30 groups: 10 males, 50 females), and the control group (29 groups: 4 males, 25 females). Tests confirmed no significant differences in age or gender distribution among the three groups. All participants signed informed consent forms.

2.1.2 Materials

Learning materials: The "Depressive Disorders" segment from the Psychiatry instructional video recorded by Harbin Medical University, available on the Chinese University MOOC platform, was selected. The

video primarily covered the definition, etiology, and pathogenesis of depressive disorders, with a focus on clinical manifestations (depressed mood, psychomotor retardation, somatic symptoms, and other symptoms).

Case for discussion: A case of depressive disorder (see Appendix A) was selected for discussion around the following questions: (1) What symptoms does the client present? (2) What factors might have contributed to these symptoms? (3) What is the probable diagnosis, and on what grounds? (4) What additional questions would you consider necessary to further explore the client's condition? (5) In this discussion, what aspects do you think warrant further exploration, and are there any missing pieces of information?

Knowledge test: The test consists of 9 items, with a total score of 13 points. Items were selected from certified psychotherapist examination questions, covering depressive disorder symptom identification, etiological understanding, and diagnostic reasoning (see Appendix B). The post-test was largely consistent with the pre-test in content, though the phrasing of some items was adjusted.

Intrinsic motivation scale: The intrinsic motivation scale developed by Isen and Reeve [21] was adopted to assess the degree to which students enjoyed their learning experience. The scale consists of 8 items rated on a 6-point Likert scale. The internal consistency coefficient of this scale in the present experiment was $\alpha = 0.944$ (see Appendix C).

2.1.3 Experimental Platform

This study used a self-developed discussion platform, an intelligent dialogue platform based on natural language processing that can interface with multiple models such as ChatGPT for natural language conversations. The model used in this study was ChatGPT-4o. The platform primarily consists of two components: a configuration interface and a discussion interface. In the configuration interface, the prompt presented the depressive disorder case along with the corresponding discussion questions. The AI was instructed to assume the role of an expert and to conduct a case discussion with the participants. The AI was required to offer pedagogical guidance without directly revealing the answers throughout the discussion. Additionally, the AI was programmed to present only one perspective at a time, rather than disclosing all relevant viewpoints simultaneously (see Appendix D).

2.1.4 Procedure

Upon entering the laboratory, participants completed the pre-test questionnaire, watched the learning video, and were then randomly assigned to one of three conditions. In the control condition, participants directly completed the post-test questionnaire, concluding the experiment. In the AI-assisted case discussion condition, participants discussed the depressive disorder case with the AI on the intelligent discussion platform (see Fig. 1). In the peer discussion condition, two human participants discussed on the intelligent discussion platform without AI involvement. Participants received a paper copy of the case materials (including discussion objectives, the case, and questions). The discussion lasted 20 minutes, after which participants proceeded to the post-test phase. The total experiment duration was approximately 35–40 minutes.

2.2 Results

2.2.1 Knowledge Acquisition

An analysis of covariance (ANCOVA) was conducted on post-test knowledge scores across the AI-assisted case discussion group, peer case discussion group, and control group, with pre-test knowledge scores as the covariate. Results showed that pre-test score differences among groups were not significant, $F(1, 115) = 1.37$, $P = 0.244$. For post-test scores, the main effect of group was significant, $F(2, 115) = 4.85$, $P = 0.009$, $\eta_p^2 = 0.078$. Bonferroni post-hoc comparisons revealed that the AI-assisted discussion group ($M = 7.83$, $SD = 1.76$) scored significantly higher than the control group ($M = 6.31$, $SD = 1.37$), $P = 0.007$(see Fig. 2). No significant differences were found between other group pairs.

2.2.2 Intrinsic Motivation

An ANCOVA was conducted on intrinsic motivation scores across the AI-assisted discussion, peer case discussion, and control groups, with pre-test knowledge scores as the covariate. Results showed that pre-test score differences among the three groups were not significant, $F(1, 115) = 0.026$, $P = 0.973$; the main effect on intrinsic motivation scores was not significant, $F(2, 115) = 0.3$, $P = 0.741$.

2.2.3 Interaction Quality of Discussion Processes

Two researchers manually coded the case discussion texts according to the interaction analysis model [20]. Prior to coding, they received training and then jointly coded 33% of the total data. The inter-rater Cohen's Kappa = 0.798, $P < 0.05$, indicating significant agreement between the two coders [22].

Coding results showed that discussions in both groups were primarily concentrated in Phase I (sharing/comparing of information) of the model, as shown in Table 1. Frequency refers to the total number of times a given phase occurred within each group. For the peer case discussion group, the frequency was calculated as the mean of the two participants' scores within each pair.

Because some cell frequencies were below 5, Fisher's exact test was conducted on the total frequencies across phases for both groups. Results indicated that the difference in total frequencies between the two groups was significant, $P < 0.001$, with the peer group exhibiting significantly higher interaction frequencies than the AI group.

Table 1
Interaction analysis of the case discussion processes between two groups

Phase	Code	AI-assisted case discussion (*n* = 30)		peer case discussion (*n* = 30)	
		Frequency	**Percentage**	**Frequency**	**Percentage**
Phase I	[PhI/A]	167	70.46%	339	75.67%
	[PhI/B]	20	8.44%	69.5	15.51%
	[PhI/C]	9	3.8%	3.5	0.78%
	[PhI/D]	36	15.19%	34	7.59%
	[PhI/E]	4	1.69%	2	0.45%
Phase II	[PhII/A]	1	0.42%		

Notes:

Phase I: Sharing/Comparing of Information

PhI/A. A statement of observation or opinion; PhI/B. A statement of agreement from one or more other participants; PhI/C. Corroborating examples provided by one or more participants; PhI/D. Asking and answering questions to clarify details of statements; PhI/E. Definition, description, or identification of a problem.

Phase II: The Discovery and Exploration of Dissonance or Inconsistency Among Ideas, Concepts or Statements

PhII/A. Identifying and stating areas of disagreement

3 Experiment 2: Effects of Virtual Consultation and Instructional Sequencing on Learning Outcomes

3.1 Method

3.1.1 Research Design and Participants

Experiment 2 employed a single-factor, four-level between-subjects design. The experiment included four conditions: (1) virtual consultation condition, in which participants conducted a consultation with a virtual patient; (2) discussion-then-consultation condition, in which participants first engaged in an AI-assisted case discussion and then interviewed the virtual patient; (3) consultation-then-discussion condition, in which participants first interviewed the virtual patient and then engaged in an AI-assisted case discussion; and (4) a learning-only control condition, in which participants completed the same

learning materials without participating in either activity. This control group was identical to the one used in Experiment 1.

Dependent variables included knowledge test scores, intrinsic motivation, and consultation process quality. Consultation process quality was manually coded by two researchers and comprised diagnostic accuracy, consultation comprehensiveness, communication quality, and empathy, with higher scores indicating better consultation quality. Diagnostic accuracy was classified into three categories: correct diagnosis, incorrect diagnosis, and no clear diagnosis. Consultation comprehensiveness referred to the number of core symptom categories covered during the consultation. Communication quality included participants' proportion of in-depth questions, language fluency, and semantic coherence. Empathy included participants' emotional responses, understanding, and exploration during the consultation.

Ninety students from a medical university were recruited. Two participants were excluded due to incomplete data, leaving 88 participants aged 17–23 years (M = 18.53, SD = 1.11). All participants had normal or corrected-to-normal vision with no color blindness or color weakness. Participants were randomly assigned to the virtual consultation group (30 groups: 4 males, 26 females), the discussion-then-consultation group (28 groups: 6 males, 22 females), and the consultation-then-discussion group (30 groups: 6 males, 24 females). Tests confirmed no significant differences in age or gender among the three groups. All participants signed informed consent forms.

3.1.2 Materials

Learning materials, discussion case, knowledge test, and intrinsic motivation scale: Same as in Experiment 1.

Virtual patient configuration: The virtual patient was named "Zhang Xiaoming", whose clinical characteristics were configured according to the depressive disorder features described in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [23].

3.1.3 Experimental Platform

This study used the self-developed intelligent platform; the setup of AI-assisted case discussion was the same as in experiment 1. In the virtual consultation, the AI assumed the role of a virtual patient and received consultations from participants. The prompt supplied the demographic and clinical background of a patient with depressive disorder for the AI to enact, and required the AI to role-play this patient during a simulated clinical interview with the participants. An illustrative example of the interview dialogue was also provided in the prompt (see Appendix E).

3.1.4 Procedure

Upon entering the laboratory, participants first completed the pre-test questionnaire, then watched the learning video, and were subsequently randomly assigned to the experimental groups. In the virtual consultation group, the AI role-played a patient with depressive disorder, and participants, taking on the role of a psychological counselor, were required to conduct a consultation with this patient (see Fig. 3).

In the discussion-then-consultation group, after watching the video, participants first discussed a depression case with the AI (the case discussion process was the same as in Experiment 1), then conducted a virtual consultation. In the consultation-then-discussion group, participants first conducted a virtual consultation, then engaged in case discussion with the AI. Both case discussion and virtual consultation lasted 20 minutes, after which participants proceeded to the post-test phase. The total experiment duration was approximately 35–50 minutes.

3.2 Results

Two researchers manually coded the validity of virtual patient simulation and consultation process quality. First, both coders independently coded 37.5% of the total data. Inter-rater reliability analysis showed that Cohen's Kappa between the two coders was 0.652, $p < 0.05$, indicating significant agreement [22]. Subsequently, the remaining data was independently coded by the two researchers separately. Finally, all coding results were verified by a third researcher.

3.2.1 Simulation Quality of AI Virtual Patients

The validity of virtual patient simulation was assessed through two indicators—fidelity and prompt adherence—both scored on a 0–2 scale, with higher scores indicating better simulation validity. Fidelity measured the extent to which the virtual patient's performance approximated real patients' behaviors and conversational patterns. Prompt adherence assessed whether the content expressed by the virtual patient remained within the specified constraints, aiming to evaluate whether the ChatGPT-simulated virtual patient maintained consistency in symptom presentation and expression style to ensure comparability of consultation outcomes across different groups. Results showed that AI virtual patient simulation scores for both fidelity and prompt adherence were very high, approaching the maximum (see Table 2).

Table 2
Fidelity and prompt adherence of AI virtual patients

Fidelity (***M* ± *SD***)	**virtual consultation** (***n* = 30**)	**discussion-then-consultation** (***n* = 28**)	**consultation-then-discussion** (***n* = 30**)
	1.95 ± 0.15	**1.96 ± 0.18**	**1.97 ± 0.18**
Prompt Adherence (*M* ± *SD*)	1.95 ± 0.2	1.91 ± 0.27	1.97 ± 0.18

3.2.2 Quality of Virtual Consultation

(1) Diagnostic Accuracy

Based on consultation process records, participants' diagnostic outcomes were categorized as correct diagnosis, incorrect diagnosis, or no clear diagnosis, with frequencies for each category shown in Table 3. Because the data distribution did not meet the assumption of homogeneity of variance, a nonparametric independent-samples Kruskal-Wallis test was conducted on diagnostic outcomes across the three groups. Results indicated no significant differences among the groups (P = 0.088).

Table 3
Diagnostic frequency coding results under different consultation conditions

Group	No Clear Diagnosis	Incorrect Diagnosis	Correct Diagnosis
Virtual Consultation (n = 30)	19	1	10
Discussion-then-Consultation (n = 28)	19	0	9
Consultation-then-Discussion (n = 30)	26	1	3

(2) Consultation Comprehensiveness

The consultation process covered 12 thematic categories: mood, interest, social functioning, energy, sleep, cognitive ability, weight and appetite, somatic symptoms, self-worth, self-harm or suicide, medical history, and screening. Two researchers manually coded the themes covered during each participant's consultation, then calculated the average number of distinct thematic categories addressed per group as the indicator of consultation comprehensiveness.

A one-way ANOVA was conducted on the consultation comprehensiveness indicator with group as the independent variable. Results showed significant differences among the three groups (virtual consultation, discussion-then-consultation, and consultation-then-discussion) in consultation comprehensiveness, $F(2, 85) = 4.21$, $P = 0.018$, $\eta_p^2 = 0.09$. Bonferroni multiple comparisons revealed that the discussion-then-consultation group (M = 5.79, SD = 2.13) demonstrated significantly higher consultation comprehensiveness than the virtual consultation group (M = 4.2, SD = 1.88), P = 0.019 (see Fig. 4).

(3) Communication Quality

Communication quality primarily comprised participants' proportion of in-depth questions, fluency, and coherence during the consultation. Two researchers manually annotated the types of questions asked by participants during the virtual consultation. The in-depth questioning proportion was calculated as the number of in-depth questions divided by the total number of questions. A one-way ANOVA conducted on the in-depth questioning proportion across the three groups showed no significant between-group differences, $F(2, 85) = 2.68$, P = 0.074.

Regarding fluency and coherence, two researchers coded each utterance in the virtual consultation transcripts for fluency and coherence on a 0–2 scale. Group-level mean scores were then derived. Because the data distribution did not satisfy the assumption of homogeneity of variance, independent-samples Kruskal-Wallis tests were conducted. Results showed no significant differences among groups in fluency ($P = 0.907$) or coherence ($P = 0.055$).

(4) Empathy

Two researchers rated each participant utterance on three empathy dimensions using a 0–2 scale. Total scores were then calculated for each group. One-way ANOVAs were conducted on the three dimensions of empathy across the three groups. Results showed no significant between-group differences on the emotional response and understanding dimensions. On the exploration dimension, due to heterogeneity of variance, a Welch-corrected analysis was employed, revealing significant between-group differences, Welch $F = 6.28$, $P = 0.004$, $\eta_{\mathrm{p}}^{2} = 0.129$. Games-Howell post-hoc tests on the exploration dimension showed that the virtual consultation group ($M = 2.73$, $SD = 2.48$) scored significantly higher than the discussion-then-consultation group ($M = 1.14$, $SD = 1.38$), $P = 0.01$; the consultation-then-discussion group ($M = 2.40$, $SD = 2.22$) also scored significantly higher than the discussion-then-consultation group ($M = 1.14$, $SD = 1.38$), $P = 0.032$ (see Fig. 5).

3.2.3 Knowledge Acquisition

Because pre-test knowledge scores differed significantly among groups, $F(3, 112) = 4.82$, $P = 0.03$, hierarchical regression analysis was used to compare between-group differences in post-test knowledge scores. In the hierarchical regression, pre-test scores served as the control variable, post-test scores as the dependent variable, and experimental group (virtual consultation group, discussion-then-consultation, consultation-then-discussion, and control group) as the independent variable. The control group served as the reference group, with regression coefficients for other experimental groups representing their effects relative to the control group. Results indicated that the virtual consultation group (B = 1.402, β = 0.286, $t = 2.62$, $P = 0.01$), the discussion-then-consultation group (B = 1.41, β = 0.281, $t = 2.60$, $P = 0.011$), and the consultation-then-discussion group (B = 1.517, β = 0.31, $t = 2.84$, $P = 0.005$) all had significantly higher post-test scores than the control group (see Fig. 6).

3.2.4 Intrinsic Motivation

An ANCOVA was conducted on intrinsic motivation scores across the virtual consultation, discussion-then-consultation, consultation-then-discussion, and control groups, with pre-test knowledge scores as the covariate. Results showed that pre-test differences were not significant, $F(1, 112) = 0.12$, $P = 0.73$; intrinsic motivation score differences were significant, $F(3, 112) = 3.41$, $P = 0.02$, $\eta_{\mathrm{p}}^{2} = 0.084$. Bonferroni multiple comparisons revealed that the discussion-then-consultation group ($M = 5.50$, $SD = 0.46$) had significantly higher intrinsic motivation than the virtual consultation group ($M = 5.1$, $SD = 0.59$, $P = 0.034$)

and the control group (*M* = 5.12, *SD* = 0.56, *P* = 0.047). No significant differences were found among other group pairs.

4 Discussion

4.1 Effects of AI-Assisted Case Discussion on Learning Outcomes

This study found that AI-assisted case discussion demonstrated significant advantages in promoting medical students' mastery of depressive disorder knowledge. Compared to the peer discussion group and the no-discussion control group, the AI discussion group achieved higher post-test knowledge scores. The findings suggest that AI can play a positive role in learning environments, particularly when working with first-year students who lack a systematic medical knowledge base. AI can provide clearer reasoning pathways and richer information supplementation, thereby overcoming the inefficient communication caused by weak cognitive foundations in traditional discussions and improving both efficiency and accuracy [24].

Regarding the interaction quality of case discussions, previous researchers using the interaction analysis model found that the vast majority of discussions remained at Phase I, with few advancing to Phases II or III [25]. The results of this study indicate that in both the AI-assisted group and the peer discussion group, interactions primarily remained at the information sharing and comparison stage, with fewer occurrences of higher-level interaction stages such as dissonance, negotiation, and co-construction. These results are consistent with existing research.

Furthermore, although AI-assisted discussion improved knowledge mastery, it did not significantly enhance learners' intrinsic motivation and even underperformed the peer discussion group in interaction frequency. This may be because the prompts for the AI were insufficiently precise, causing the AI to provide overly lengthy responses in answering questions and discussing content. It may also be related to the "instrumental" nature of current AI interactions—while the feedback is accurate, its impact on learner motivation is limited [26].

4.2 Effects of the Sequence of Case Discussion and Consultation on Learning Outcomes

In this study, the AI-simulated patient's fidelity and content consistency both approached maximum scores, indicating effective patient simulation by the AI. Results showed that whether through virtual consultation alone or through a multi-stage learning pathway combining virtual consultation with case discussion, students' mastery of depressive disorder knowledge was significantly enhanced. These findings are consistent with previous research, further demonstrating that AI possesses universal advantages in structured knowledge supplementation [27]. Meanwhile, learners who participated only in

virtual consultation scored significantly higher on the exploration dimension, echoing Chen et al.'s [28] research conclusion that VSPs can strengthen emotional resonance.

Huwendiek et al.'s [17] research indicated that learners preferred the learning sequence of "course learning → 1–2 virtual patients → group discussion → real patients." However, this remained at the level of preference description and did not verify the actual effects of different sequences on motivation and consultation quality. The present study further found that different sequences of AI-assisted case discussion and virtual consultation had comparable effects on knowledge acquisition but differential effects on the consultation process and intrinsic motivation. The case discussion-then-consultation sequence significantly improved intrinsic motivation and consultation comprehensiveness, possibly because learners first established a knowledge framework and then participated in practical tasks more purposefully [29], and initial structured knowledge construction helps enhance the initiative and engagement of subsequent learning. In contrast, learners who first engaged in virtual consultation demonstrated higher empathy, consistent with previous research 30]; when learners exhibited more empathy during consultation, their consultation comprehensiveness was lower.

This study supports Kolb's experiential learning model [19]: different learning pathways (abstract → practice vs. practice → abstract) are suited to different types of learning objectives. When the learning goal is to enhance consultation comprehensiveness, AI-assisted case discussion can be conducted first, followed by virtual consultation. When the learning goal emphasizes empathy during the consultation process, direct virtual consultation can be conducted. Educators can flexibly arrange the sequencing of case discussion and virtual consultation based on their instructional objectives.

In summary, the findings of this study demonstrate the feasibility and effectiveness of AI as an auxiliary pedagogical tool in medical education: in theoretical knowledge learning, AI can provide clearly structured case analysis support; in practical skill development, AI virtual patients can offer standardized, highly repeatable, low-cost clinical communication training to enhance medical students' consultation competencies; in instructional design, the rational sequencing of AI-based teaching activities will further optimize learning outcomes and motivation.

4.3 Limitations and Future Directions

First, when validating the effectiveness of virtual consultation, given the complexity of training standardized patients, we did not compare virtual consultations with standardized patient interviews. Second, this study used a laboratory experimental method, and the duration of AI-assisted case discussion and virtual consultation was relatively short. Consequently, the generalizability of our findings to real-world clinical settings requires further investigation.

5 Conclusions

This study draws the following main conclusions: (1) AI-assisted case discussion and virtual patient consultation can significantly enhance medical students' knowledge mastery; (2) The sequencing of AI-

assisted case discussion and virtual consultation affects consultation process quality and intrinsic motivation.

This study contributes empirical evidence to the field of AI-enhanced medical education by demonstrating that AI-assisted case discussion and virtual patient consultation significantly improve medical students' knowledge acquisition. The identification of a sequencing effect—whereby the order of the two interventions affects consultation quality and intrinsic motivation—provides novel insights for optimizing instructional design. These insights move beyond the simplistic "whether it works" question to address the more nuanced "how and when it works best" question, offering practical guidance for educators in structuring AI-assisted learning activities.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Committee of Henan Medical University (HYYLL- 20260543). Written informed consent was obtained from all participants before participation.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Medical Education Research Project of Medical Education Branch of the Chinese Medical Association and National Center for Health Professions Education Development “Competency-Oriented AI Virtual Patient Training and Assessment for Depression” (2025B95), Higher Education Teaching Reform Research and Practice Project of Henan Medical University “Research and Practice on AI-Empowered Competency-Based Graded Training and Assessment in Psychiatry” (2026-HYJG-025). The funders had no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript.

Author Contribution

HG conceived the study, secured funding, supervised the work, and wrote the original draft. BZ developed the methodology, curated data, performed investigation, and co-wrote the original draft. XH contributed

to methodology, software, data curation, and original draft writing. CL assisted with writing, methodology, and data curation. DZ reviewed and edited the manuscript and supervised the research. HZ acquired funding, reviewed and edited the manuscript, administered the project, and provided supervision. All authors read and approved the final manuscript.

Data Availability

The datasets generated or analysed during the current study are available in the figshare repository, https://doi.org/10.6084/m9.figshare.33078098

References

1. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions[J]. JMIR Med Educ. 2023;9(1):e48291. https://mededu.jmir.org/2023/1/e48291.
2. Gutiérrez-Maldonado J, Alsina-Jurnet I, Rangel-Gómez MV, et al. Virtual Intelligent Agents to Train Abilities of Diagnosis in Psychology and Psychiatry[M]//Tsihrintzis. In: Virvou GA, Howlett M RJ, editors. et al. New Directions in Intelligent Interactive Multimedia. Studies in Computational Intelligence. Volume 142. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. pp. 497–505. https://doi.org/10.1007/978-3-540-68127-4_51.
3. Zhang X, Zeng D, Wang X, et al. Analysis of virtual standardized patients for assessing clinical fundamental skills of medical students: a prospective study[J]. BMC Med Educ. 2024;24(1):981. https://doi.org/10.1186/s12909-024-05982-2.
4. Ertmer PA, Koehler AA. Online case-based discussions: examining coverage of the afforded problem space[J]. Education Tech Research Dev. 2014;62(5):617–36. https://doi.org/10.1007/s11423-014-9350-9.
5. Lu J, Zhang L, Ye L. Construction and application of dentalvirtual reality simulation teaching system in China: Necessity and strategies[J]. Eur J Dent Educ. 2025;29(3):570–5. https://doi.org/10.1111/eje.12952.
6. Grunhut J, Marques O, Wyatt AT. Needs, challenges, and applications of artificial intelligence in medical education curriculum[J]. JMIR Med Educ. 2022;8(2):e35587. https://doi.org/10.2196/35587.
7. Zhang S, Soreide KK, Kelling SE, et al. Quality assurance processes for standardized patient programs[J]. Currents Pharm Teach Learn. 2018;10(4):523–8. https://doi.org/10.1016/j.cptl.2017.12.014.
8. Gillette C, Stanton RB, Rockich-Winston N, et al. Cost-effectiveness of using standardized patients to assess student-pharmacist communication skills[J]. Am J Pharm Educ. 2017;81(10):6120. https://doi.org/10.5688/ajpe6120.

9. Laverde NGrévisseC, ,Jaramillo S, et al. Integrating large language model-based agents into a virtual patient chatbot for clinical anamnesis training[J]. Comput Struct Biotechnol J. 2025;272481–2491. https://doi.org/10.1016/j.csbj.2025.05.025.
10. Wang G, Li Y, Qin J, Zhao W, Wu J, Xu Y. Application and prospect of virtual standardized patient in psychiatry teaching[J]. J Neurosci Mental Health. 2025;25(1):68–72. https://doi.org/10.3969/j.issn.1009-6574.2025.01.010.
11. Sezer B, Sezer TA, Teker GT, et al. Developing a virtual patient: design, usability, and learning effect in communication skills training[J]. BMC Med Educ. 2023;23:891. https://doi.org/10.1186/s12909-023-04860-7.
12. Zhang X, Zeng D, Wang X, et al. Analysis of virtual standardized patients for assessing clinical fundamental skills of medical students: a prospective study. BMC Med Educ. 2024;24:981. https://doi.org/10.1186/s12909-024-05982-2.
13. Cook AD. Creating virtual patients using large language models: scalable, global, and low cost. [J] Med teacher. 2024;47(1):1–3. https://doi.org/10.1080/0142159x.2024.2376879.
14. Kelly S, Smyth E, Murphy P, et al. A scoping review: virtual patients for communication skills in medical undergraduates[J]. BMC Med Educ. 2022;22(1):429. https://doi.org/10.1186/s12909-022-03474-9.
15. Quail M, Brundage SB, Spitalnick J, et al. Student self-reported communication skills, knowledge and confidence across standardised patient, virtual and traditional clinical learning environments[J]. BMC Med Educ. 2016;16:1–12. https://doi.org/10.1186/s12909-016-0577-5.
16. Maicher KR, Stiff A, Scholl M, et al. Artificial intelligence in virtual standardized patients: combining natural language understanding and rule based dialogue management to improve conversational fidelity[J]. Med Teach. 2023;45(3):279–85. https://doi.org/10.1080/0142159x.2022.2130216.
17. Huwendiek S, Duncker C, Reichert F, et al. Learner preferences regarding integrating, sequencing and aligning virtual patients with other activities in the undergraduate medical curriculum: a focus group study[J]. Med Teach. 2013;35(11):920–9. https://doi.org/10.3109/0142159x.2013.826790.
18. Lee J, Kim H, Kim KH, et al. Effective virtual patient simulators for medical communication training: a systematic review[J]. Med Educ. 2020;54(9):786–95. https://doi.org/10.1111/medu.14152.
19. Kolb DA, Boyatzis RE, Mainemelis C. Experiential Learning Theory: Previous Research and New Directions[M]//Perspectives on Thinking, Learning, and Cognitive Styles. Routledge; 2001. https://doi.org/10.4324/9781410605986-9.
20. Gunawardena CN, Lowe CA, Anderson T. Analysis of a Global Online Debate and the Development of an Interaction Analysis Model for Examining Social Construction of Knowledge in Computer Conferencing[J]. J Educational Comput Res. 1997;17(4):397–431. https://doi.org/10.2190/7mqv-x9uj-c7q3-nrag.
21. Isen AM, Reeve J. The influence of positive affect on intrinsic and extrinsic motivation: facilitating enjoyment of play, responsible work behavior, and self-control[J]. Motivation Emot. 2005;29(4):295–323. https://doi.org/10.1007/s11031-006-9019-8.

22. Landis JR, Koch GG. The measurement of observer agreement for categorical data[J]. Biometrics. 1977;159–74. https://doi.org/10.2307/2529310.
23. Regier DA, Narrow WE, Kuhl EA, et al. The Conceptual Development of DSM-V[J]. Am J Psychiatry. 2009;166(6):645–50. https://doi.org/10.1176/appi.ajp.2009.09020279.
24. Sun L, Yin C, Xu Q, et al. Artificial intelligence for healthcare and medical education: a systematic review[J]. Am J Translational Res. 2023;15(7):4820–8. https://europepmc.org/article/PMC/PMC10408516.
25. Gunawardena CN, Flor NV, Sánchez DM. Knowledge Co-Construction in Online Learning: Applying Social Learning Analytic Methods and Artificial Intelligence. 1st ed. Routledge; 2005. https://doi.org/10.4324/9781003324461.
26. Chen CH, Chang CL. Effectiveness of AI-assisted game-based learning on science learning outcomes, intrinsic motivation, cognitive load, and learning behavior[J]. Educ Inform Technol. 2024;29(14):18621–42. https://doi.org/10.1007/s10639-024-12553-x.
27. Xu L, Xu Q, ,Liu C et al. Virtual Standardized Patients for Improving Clinical Thinking Ability Training in Residents: Randomized Controlled Trial[J]. JMIR Med Educ,2025;11e73196. https://doi.org/10.2196/73196
28. Chen S, Wu M, Zhu KQ et al. LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation[J]. arXiv preprint arXiv:2305.13614, 2023.
29. Posel N, Mcgee JB, Fleiszer DM. Twelve tips to support the development of clinical reasoning skills using virtual patient cases[J]. Med Teach. 2015;37(9):813–8. https://doi.org/10.3109/0142159x.2014.993951.
30. Kameg K, Howard VM, Clochesy J, et al. The Impact of High Fidelity Human Simulation on Self-Efficacy of Communication Skills[J]. Issues Ment Health Nurs. 2010;31(5):315–23. https://doi.org/10.3109/01612840903420331.

Figures

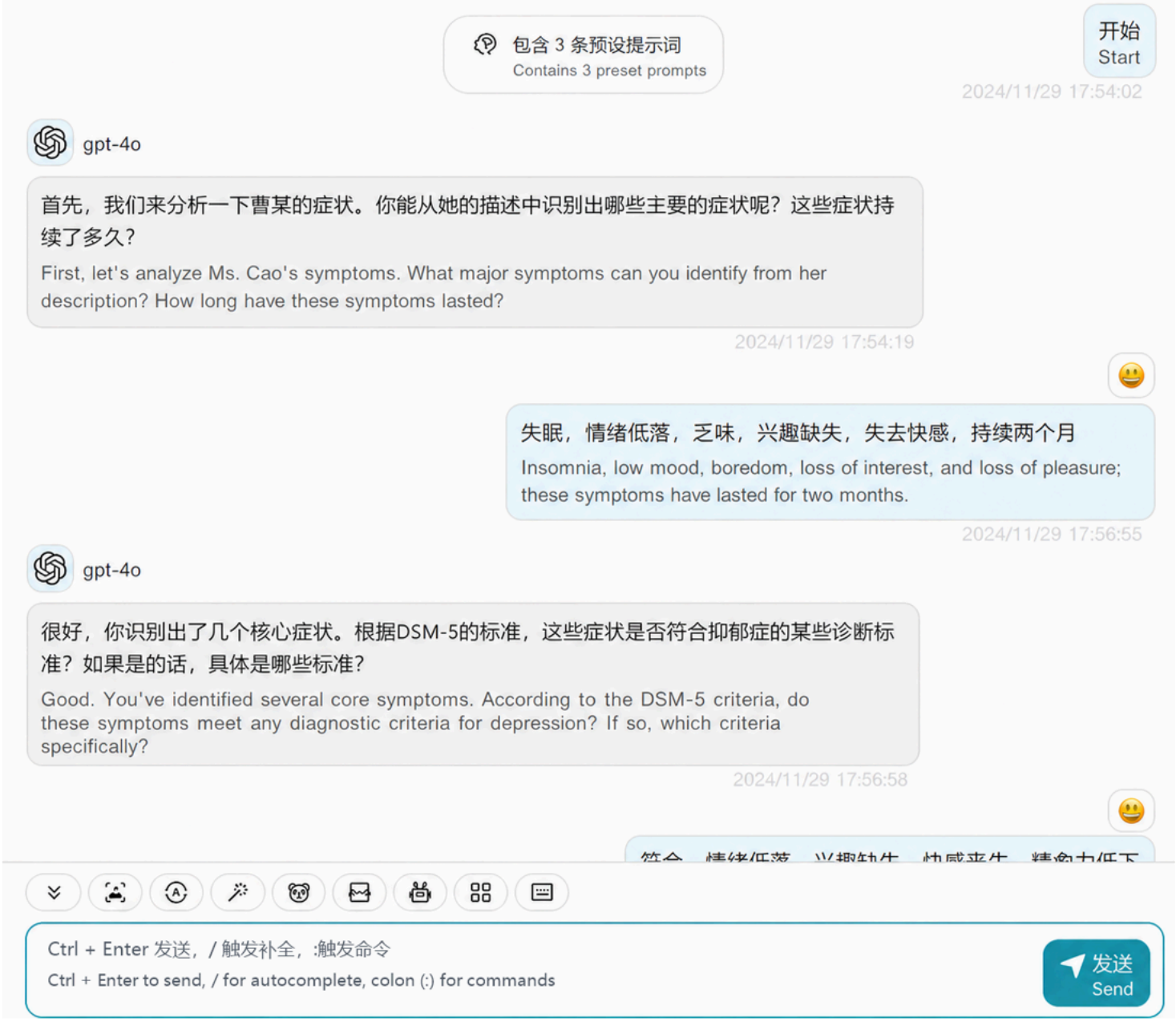

Figure 1

The intelligent experimental platform interface for AI-assisted case discussion

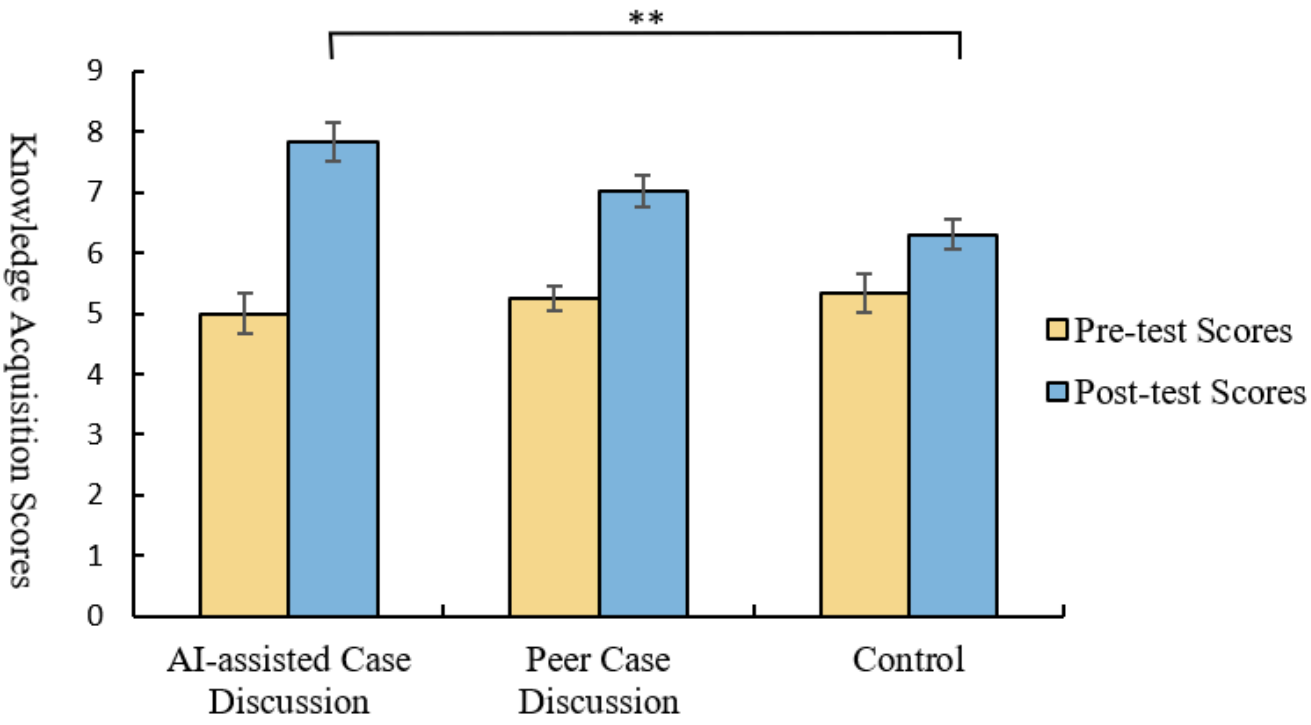

Figure 2

Pre- and post-test knowledge score comparisons across groups

Note: * indicates $P < 0.05$, ** indicates $P < 0.01$, same below.

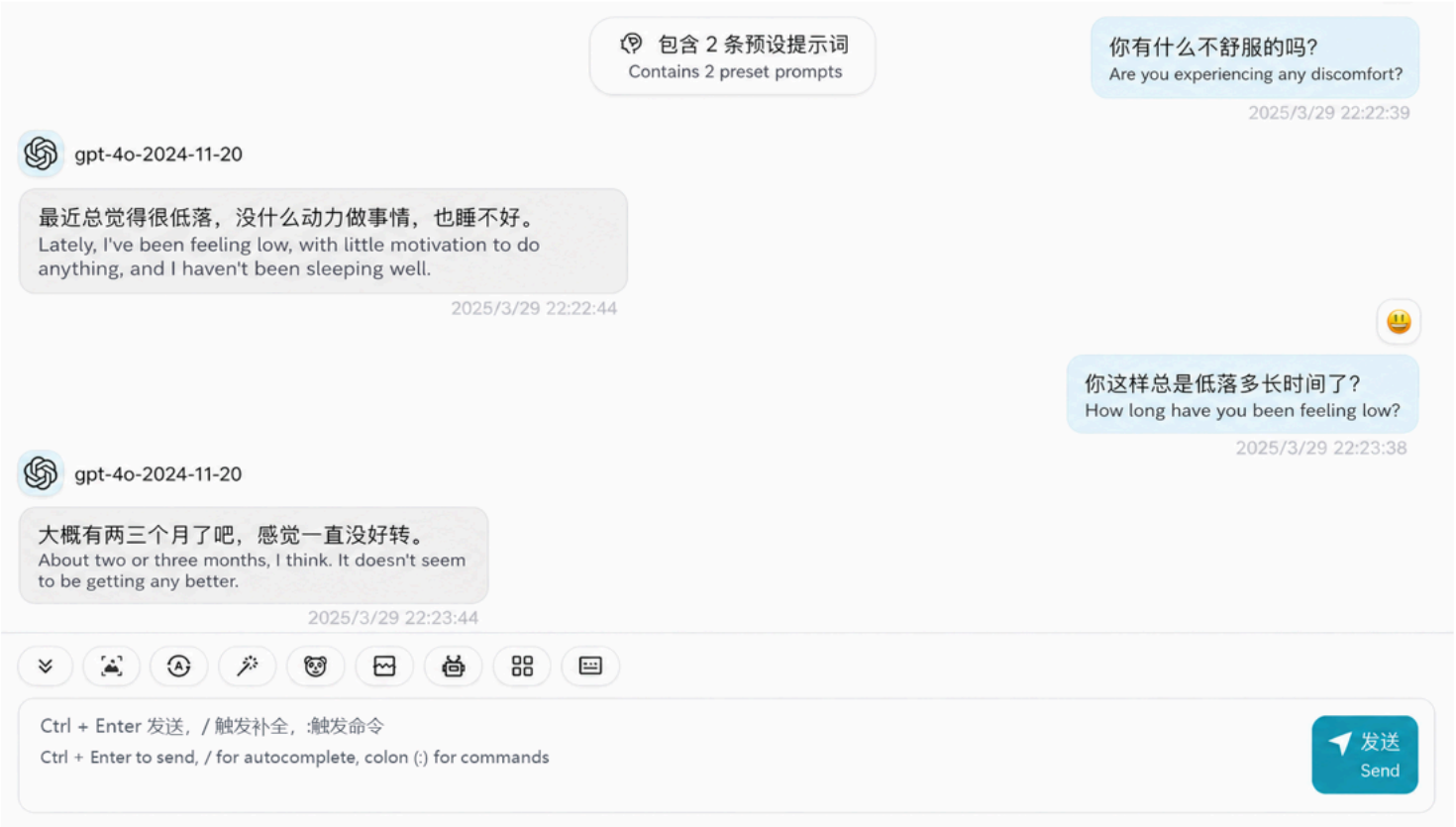

Figure 3

The intelligent experimental platform interface for virtual consultation

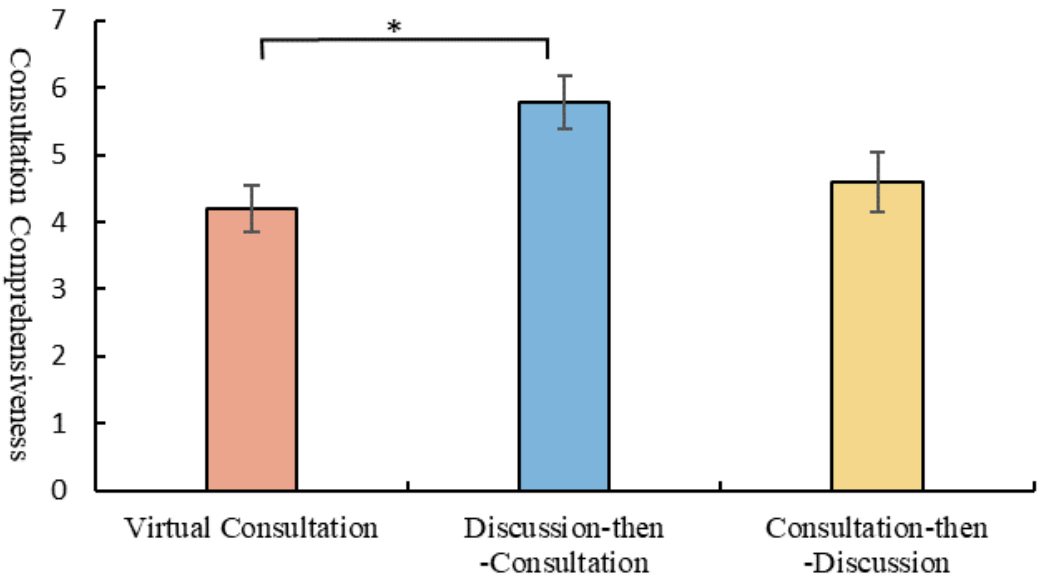

Figure 4

Comparison of consultation comprehensiveness differences across consultation conditions

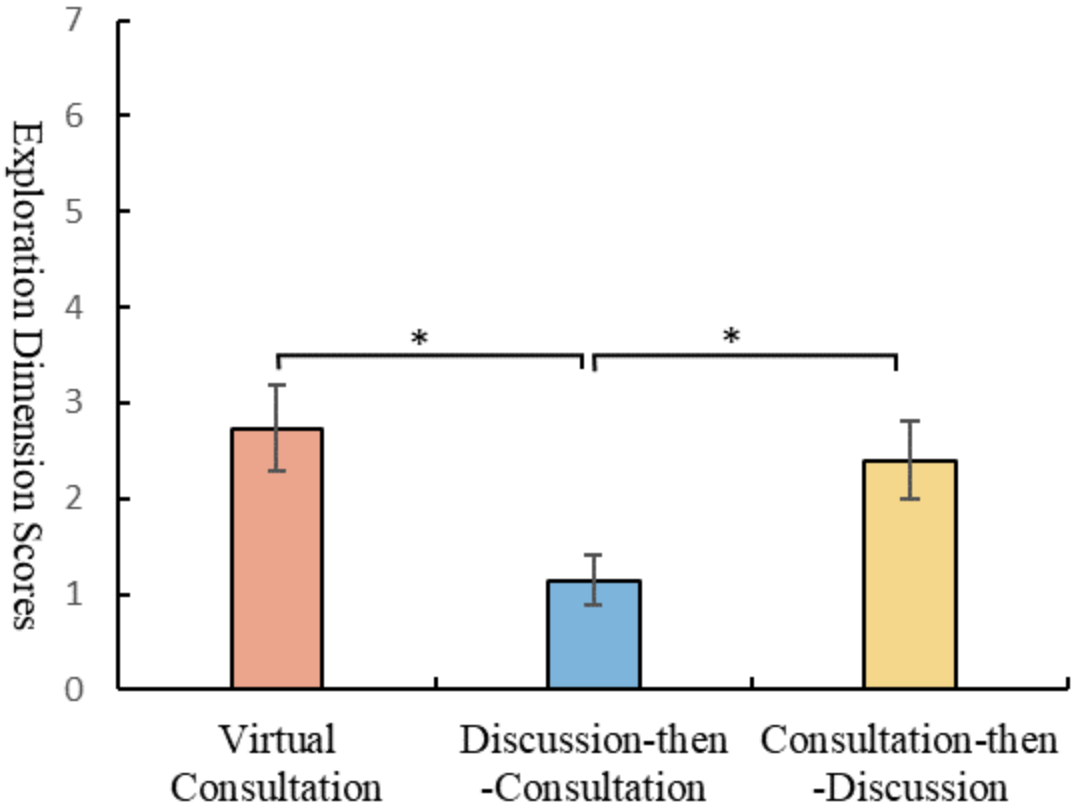

Figure 5

Comparison of exploration dimension differences across consultation conditions

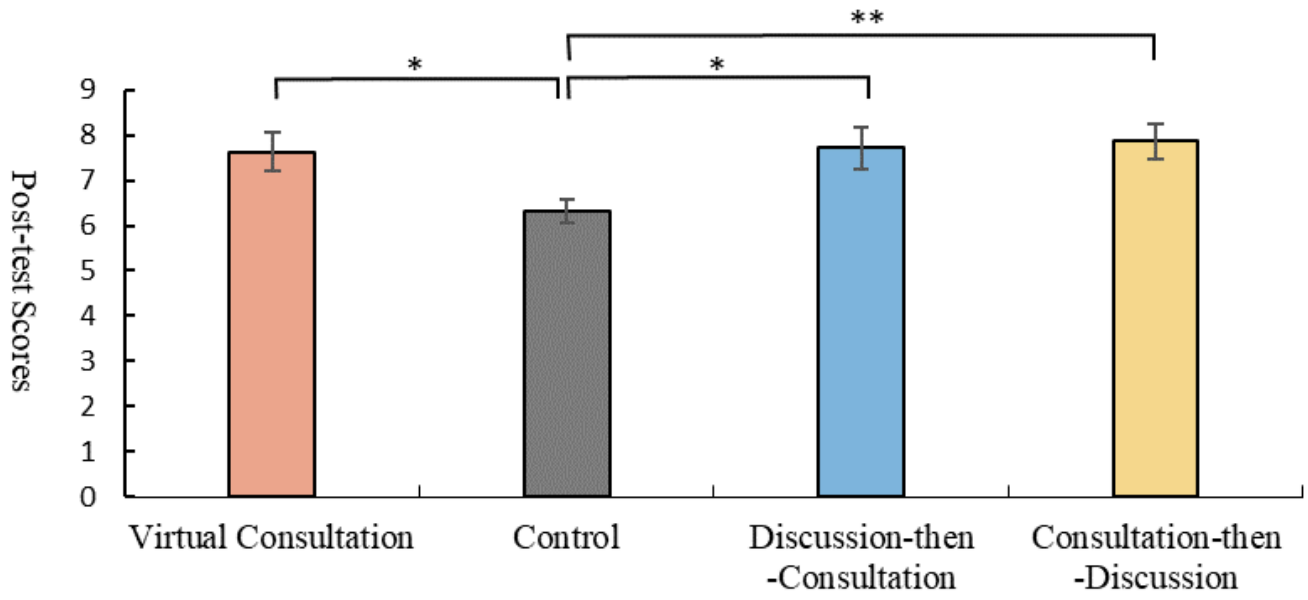

Figure 6

Post-test knowledge score differences between consultation conditions and the control group

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx