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.

Enhancing Clinical Reasoning Development: An AI-Assisted Learning Framework for Unstructured Problem-Solving in Medical Education

hang du
Hunan Normal University
Yang Tan
Hunan Normal University
Jiaxing Zhang
Hunan Normal University
Xuanning Wang
Hunan Normal University
Yufei Li
Hunan Normal University
Xiaofang Hu
Hunan Normal University
Wei Chen
Hunan Normal University
Huaying Wu
Hunan Normal University
Lin Li
lilin@hunnu.edu.cn

Hunan Normal University

Research Article

Keywords: Artificial intelligence, Clinical reasoning, Medical education, Unstructured problems, Systems thinking, Large language models, Randomized controlled trial

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10872685/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

Clinical reasoning for unstructured problems—those lacking clearly defined initial conditions, standard protocols, or singular correct solutions—remains a persistent challenge in medical education. While artificial intelligence (AI) tools have shown promise in structured knowledge tasks, their integration into frameworks designed for unstructured clinical reasoning remains underexplored.

Methods

A systems thinking assessment framework was adapted for medical education and validated through expert review and inter-rater reliability testing (ICC = 0.801, 95% CI: 0.725–0.866). Fifty-five medical students were randomly assigned to three groups: an AI-assisted learning group (AIL, n = 18) using ChatGPT 4.0 as a guided reasoning aid, a conventional learning group (CL, n = 18) using textbooks and non-AI online resources, and a manual group (MG, n = 19) without any supplementary tools. All participants responded to the same unstructured clinical case comprising seven tasks (Q1–Q7), each scored on a 0–3 scale by five blinded raters. As an exploratory benchmark, ChatGPT 4.0 responses to the identical case were independently scored using the same framework to characterize the AI model's performance profile and limitations.

Results

The AIL group achieved the highest mean total score (12.11 ± 1.78), significantly outperforming the CL group (8.94 ± 1.08, $p < 0.0001$) and the MG (9.97 ± 2.37, $p < 0.01$). Significant advantages were observed in Q2 (information needs), Q5 (scenario development), Q6 (challenges), and Q7 (unintended consequences). Subgroup analysis revealed no significant interaction between learning tools and baseline academic performance ($F = 0.213$, $p = 0.930$, $\eta^2 = 0.018$), suggesting consistent benefits across GPA levels. The ChatGPT 4.0 benchmark produced uniformly identical scores (total score 16.5 across all trials), revealing characteristic limitations in response variability and template dependency.

Conclusions

An AI-assisted learning framework grounded in a validated systems thinking assessment tool can enhance medical students' clinical reasoning for unstructured problems, with effects that appear independent of baseline academic performance. The ChatGPT benchmark highlights both the capabilities and the rigidities of current AI models, underscoring the need for hybrid human–AI collaboration rather than AI substitution in clinical reasoning education.

Trial registration:

Chinese Clinical Trial Registry ChiCTR2600133444, 2026-09-28, Retrospectively registered

1. Introduction

Developing clinical reasoning skills for unstructured problems remains a persistent challenge in medical education (1–4). Unlike well-defined problems with established diagnostic pathways, unstructured problems—including complex multisystem cases, atypical presentations, and scenarios involving ethical conflicts—require learners to integrate incomplete information, weigh competing hypotheses, and adapt to evolving clinical contexts (5).

Systems thinking has been recognized as a core competency for navigating such complexity, enabling clinicians to synthesize information across biological, psychological, and social dimensions of patient care (6, 7). However, traditional pedagogical tools—textbooks, lecture-based instruction, and structured case discussions—often assume well-bounded problems with linear reasoning pathways. These approaches may inadequately prepare students for the ambiguity and integrative demands of real-world clinical practice (8, 9).

Recent advances in large language models (LLMs) have opened new possibilities for medical education. AI tools can support knowledge retrieval, hypothesis generation, and structured analysis at scale (10, 11). For instance, Learn Guide, a ChatGPT-based self-directed learning tool, has demonstrated efficacy in enhancing critical thinking and metacognitive awareness among medical trainees (12). Similarly, LLMs have shown utility in unstructured textual knowledge extraction and clinical decision support (13, 14). However, these applications have predominantly been evaluated in structured task contexts, and their integration with frameworks specifically designed for unstructured clinical reasoning remains in its infancy (15).

To address this gap, the present study had two objectives: (1) to develop and validate a systems thinking assessment framework tailored for unstructured clinical problem-solving in medical education, and (2) to evaluate, through a randomized controlled design, whether an AI-assisted learning approach using this framework could enhance medical students' clinical reasoning performance compared with conventional and tool-free learning methods. As a supplementary analysis, we characterized ChatGPT 4.0's own performance on the same assessment to delineate the capabilities and limitations of current AI models as potential educational tools.

2. Methods

2.1 Study design

This study employed a three-arm, parallel-group, randomized controlled design conducted at Hunan Normal University between September 2024 and December 2024. This study was an educational

intervention involving the comparison of routine teaching methods. No patients were involved, no biological samples were collected, and no clinical procedures were performed. All participants were healthy medical students who provided written informed consent. This study was approved by the Biomedical Research Ethics Committee of Hunan Normal University, China (Approval No. 20251051). All methods were carried out in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants. All participant data were anonymized prior to analysis. Patients and the public were not involved in the design, conduct, or reporting of this trial. No changes were made to the trial protocol after commencement.

2.2 Assessment framework development and validation

2.2.1 Framework adaptation

The Systems Thinking Evaluation Tool developed by Grohs et al. (6), comprising seven assessment items for unstructured problem-solving, was adapted for medical education contexts. The original framework has been successfully applied in healthcare education settings (16). Our adaptation modified four items (Q1, Q2, Q5, Q7) to align with clinical reasoning requirements:

- **Q1 (Problem Identification)**: The original framework's technical protocol and scenario modules were replaced with symptom-history correlation analysis to evaluate clinical information integration and problem definition.
- **Q2 (Information Needs)**: An AI utility scoring component was incorporated into the information source evaluation to quantify modern medical information retrieval proficiency.
- **Q5 (Scenario Development)**: Evidence-based clinical treatment planning replaced generic solution design.
- **Q7 (Unintended Consequences)**: Medical risk assessment—spanning pharmacological, psychological, and socioeconomic domains—replaced generic consequence identification.

The complete adapted framework with scoring criteria is presented in.

2.2.2 Framework validation

Content validity was established through expert review by five medical educators with experienced medical educators with expertise in clinical reasoning and systems thinking, who independently evaluated item relevance, clarity, and clinical appropriateness. Reviewers were five faculty members from the Department of Medical Education. The framework was iteratively revised based on their feedback until consensus.

To assess inter-rater reliability, five raters independently scored all participant responses using the adapted framework (Table 1). Absolute agreement was evaluated using a two-way random-effects intraclass correlation coefficient (ICC) model. The ICC for a single measurement was 0.801 (95% CI: 0.725–0.866, $p < 0.001$), indicating excellent inter-rater reliability. Raters were blinded to group assignment throughout the scoring process.

Table 1
Systems thinking assessment framework for medical education

(adapted from Grohs et al. (6)).

Question	Criteria and Rating Guide	Rating
Q1-Problem Identification: 1. Describe your understanding of the above case and the likely direction of the diagnosis.	Failure to provide a response or identify a problem is not relevant to the case.	0
	Only a single symptom or problem, such as abdominal pain, was identified, without consideration of other related symptoms or medical history.	1
	Multiple symptoms and associated problems were identified but not considered in relation to each other.	2
	Symptoms, history and associated problems were comprehensively identified, and links between them and potential diagnoses were considered.	3
Q2-Information Need: 2. What additional information do you need before you can begin to respond to the above case?	Failure to provide a response or requested information is not relevant to the case.	0
	Identified basic laboratory tests or imaging needs.	1
	Comprehensive screening needs were identified and additional epidemiologic information was considered.	2
	Not only are comprehensive inspection needs identified, but strategies for obtaining additional information are also presented.	3
Q3- Stakeholder Identification: 3. What groups or stakeholders will you involve in analyzing the case issues above? 4. Please briefly describe the process by which you plan to communicate effectively with these stakeholders.	No response is provided or stakeholders are listed without discussion of their roles.	0
	Key stakeholders (e.g., family, physicians) were identified, but their specific roles were not discussed.	1
	All key stakeholders were identified and their roles briefly discussed.	2
	All stakeholders are identified in detail and how to communicate and collaborate effectively is discussed.	3
Q4- Goals: 5. What do you expect a successful program to accomplish?	No response was provided or the target was not relevant to the case.	0
	Short-term goals, such as symptom relief, were identified, but long-term goals were not considered.	1
	Short- and long-term goals were identified, but patient preferences were not considered.	2
	Short- and long-term goals were identified, and patient preferences and cost-effectiveness were considered.	3

Question	Criteria and Rating Guide	Rating
Q5-Scenario Development: 6. What solution would you suggest for the above case	Failure to provide a response or plan is not relevant to the case.	0
	Single treatment options, such as medication, are proposed.	1
	A comprehensive treatment program is proposed, including medications, and lifestyle modifications.	2
	Comprehensive treatment options are presented and the advantages, disadvantages and applicability of different treatment options are considered.	3
Q6-Challenges: 7. What challenges and limitations do you see in implementing the program?	No response is provided or any potential implementation challenges are not identified.	0
	Some potentially simple or short-term challenges were identified, such as resource shortages or patient adherence issues, but long-term or complex challenges were not considered in depth or identified.	1
	Some potentially simple or short-term challenges were identified, such as resource shortages or patient adherence issues, but long-term or complex challenges were not considered in depth or identified.	2
	Potential challenges were comprehensively identified, including technical limitations, resource allocation, staff training, and policy constraints, and the interplay between these challenges and the overall impact on the treatment program were considered.	3
Q7-Unintended Consequences: 8. Please describe any unintended consequences that you believe may result from this program.	Failure to provide a response or identify any potential unintended consequences.	0
	Some potential unintended consequences, such as drug side effects or short-term complications of treatment, were identified, but the broader implications or long-term effects of these consequences were not considered in depth.	1
	Multiple potential unintended consequences were identified and their short-term impacts, such as social isolation or financial burden that may result from treatment, were considered, but long-term consequences were not fully considered.	2
	Multiple potential unintended consequences were comprehensively identified, and the short- and long-term effects of these consequences were considered in depth, including the broader impact on patients' quality of life, family, and society, and how they could be mitigated through preventive measures or adjustments to treatment regimens.	3

2.3 Participants

2.3.1 Inclusion and exclusion criteria

Medical students enrolled in the third-year students at Hunan Normal University were eligible for inclusion. Exclusion criteria included: (1) prior participation in AI-assisted learning interventions; (2) incomplete informed consent; and (3) absence during any study session.

2.3.2 Participant characteristics

Fifty-five medical students (27 male, 28 females; mean age 19–21 years) were enrolled and randomized. Baseline academic performance was assessed using Pre-Clerkship Grade Point Average (GPA), stratified into tertiles for subgroup analysis (High, Intermediate, Low). Baseline demographic and academic characteristics of the three groups are summarized in Table 4. No significant between-group differences were observed at baseline (all $p > 0.05$).

2.3.3 Randomization

Participants were randomly assigned to one of three groups using computer-generated block randomization (block size = 6) to ensure balanced allocation: AI-assisted learning group (AIL, n = 18), conventional learning group (CL, n = 18), and manual group without supplementary tools (MG, n = 19). The random allocation sequence was generated by H.D. using SPSS v27.0 with a computer-generated block randomization procedure (block size = 6). Allocation concealment was ensured using sequentially numbered, opaque, sealed envelopes. A research assistant who was not involved in the study enrolled participants, and H.D. assigned participants to their groups according to the concealed sequence. Participants and intervention providers could not be blinded due to the nature of the educational intervention.

2.4 Intervention

All participants were presented with the same unstructured clinical case: a 45-year-old male patient with persistent abdominal pain, weight loss, and malaise (Fig. 1). Participants were instructed to answer seven open-ended questions (Q1–Q7,Table 1) within a 45-minute time limit. The three groups differed only in the supplementary resources permitted during the response period:

- **AIL group (n = 18)**: Participants were provided access to ChatGPT 4.0 (OpenAI, May 2024 release) as a guided reasoning tool. Standardized prompts were provided to encourage AI use for information retrieval, differential diagnosis generation, and structured analysis, rather than direct answer generation.
- **CL group (n = 18)**: Participants were permitted to use standard textbooks and non-AI online learning resources (e.g., medical databases, clinical guideline websites), but were not allowed to use any AI-based tools.

- **MG group (n = 19)**: Participants responded without access to any supplementary tools or resources, relying solely on prior knowledge and clinical reasoning.
- The intervention was delivered by the research team (H.D., Y.T., and J.Z.). All participants completed the task as instructed. Adherence to the assigned condition was monitored by the research team, and no protocol deviations were observed. No concomitant care was provided during the trial.

2.5 Scoring

Five raters independently evaluated all de-identified participant responses using the assessment rubric (Table 1). Each of the seven tasks (Q1–Q7) was scored on a 0–3 ordinal scale, yielding a maximum possible total score of 21. Raters were blinded to group assignment throughout the evaluation process. To protect participant privacy, all response materials were anonymized prior to scoring. All responses were de-identified and randomly ordered before scoring. Raters were unaware of group allocation and whether a response was generated by ChatGPT or a human participant.

2.6 ChatGPT 4.0 as an AI performance benchmark

As an exploratory supplementary analysis, we submitted the identical clinical case and assessment questions to ChatGPT 4.0 (OpenAI; accessed September,2024) and scored its responses using the same assessment framework and raters. This served not as a clinical gold standard, but as a descriptive benchmark to characterize: (a) the performance profile of a state-of-the-art LLM on an unstructured clinical reasoning task, and (b) AI-specific limitations relevant to educational deployment, such as response homogeneity and template dependency. ChatGPT outputs did not inform the scoring rubric, rater calibration, or the assessment of human participants. The five blinded raters were not informed which responses were AI-generated.

2.7 Statistical analysis

Data normality was assessed using the Shapiro–Wilk test to guide the selection of parametric or non-parametric methods. For normally distributed variables, one-way analysis of variance (ANOVA) was used; for non-normally distributed variables, the Kruskal–Wallis H-test was applied. Normally distributed data are reported as mean ± standard deviation; non-normally distributed data are reported as median [interquartile range]. Effect sizes are reported as η^2 (eta-squared) for ANOVA and as epsilon-squared (ε^2) for Kruskal–Wallis tests.

When omnibus tests indicated significant between-group differences, post-hoc pairwise comparisons were conducted using Dunn's test (for Kruskal–Wallis) with Holm–Bonferroni correction for multiple comparisons, or Tukey's HSD (for ANOVA). A two-way factorial ANOVA was used for subgroup analysis, with learning tool (AIL/CL/MG) and GPA stratification (High/Intermediate/Low) as independent variables. An alpha level of 0.05 (two-tailed) was set for statistical significance. All analyses were performed using SPSS v27.0 (IBM Corp., Armonk, NY, USA) and GraphPad Prism 10 (GraphPad Software, San Diego, CA, USA).

A post-hoc power analysis was conducted using G*Power 3.1 (F test; one-way ANOVA, fixed effects; α = 0.05; three groups; total N = 55; observed Cohen's f = 0.734, derived from $\eta^2 = 0.35$). The achieved power was 0.999, indicating sufficient statistical power to detect the observed effect size with the current sample.

3. Results

3.1 Overall performance and per-task analysis

All 55 randomized participants completed the study and were included in the analysis. No participants were lost or excluded after randomization. No harms or unintended events were reported in any group. Significant between-group differences were observed in total scores and in Q2, Q5, Q6, and Q7 (Table 2). The AIL group achieved the highest mean total score (12.11 ± 1.78), significantly outperforming both the CL group (8.94 ± 1.08, $p < 0.0001$) and the MG (9.97 ± 2.37, $p < 0.01$) (Fig. 2a). No significant differences were observed between groups for Q1 (problem identification), Q3 (stakeholder identification), or Q4 (goals) (Table 2).

Table 2
Intergroup differences in systems thinking competency scores.

Variable	AIL group (n = 18)	CL group (n = 18)	MG (n = 19)	Test statistic	p-value	Effect size
Q1	2.00 [1.88–2.13]	2.00 [2.00–2.00]	2.00 [2.00–2.00]	H = 1.42 (K-W)	0.492	0.011
Q2	1.50 [1.50–2.13]	1.00 [0.50–1.50]	1.00 [0.50–1.50]	H = 21.95 (K-W)	< 0.001***	0.384
Q3	2.50 [2.00-2.50]	2.00 [2.00-2.50]	2.00 [2.00-2.50]	H = 4.18 (K-W)	0.124	0.042
Q4	1.00 [1.00–2.00]	0.75 [0.50–1.13]	1.00 [0.50-2.0]	H = 4.87 (K-W)	0.088	0.055
Q5	1.61 [1.38-2.00]	1.06 [0.50–1.50]	1.21 [0.50-2.00]	H = 8.13 (K-W)	0.017*	0.118
Q6	1.50 [1.00-1.50]	1.00 [1.00–1.00]	1.00 [0.50–1.50]	H = 8.70 (K-W)	0.013*	0.129
Q7	1.50 [1.40-2.00]	1.00 [0.88-1.00]	1.00 [1.00-1.50]	H = 12.28 (K-W)	0.002**	0.198
Total score	12.11 ± 1.78	8.94 ± 1.08	9.97 ± 2.37	F = 14.02 (ANOVA)	< 0.001***	0.350

*Normally distributed data are presented as mean ± SD; non-normally distributed data are presented as median [interquartile range]. Effect sizes: η^2 for ANOVA, ε^2 for Kruskal–Wallis. Significance markers: *p < 0.05, **p < 0.01, ***p < 0.001.*

In Q2 (information needs), the AIL group's median score of 1.50 [1.50–2.13] was significantly higher than both the CL group (1.00 [0.50–1.50], $p < 0.0001$) and the MG (1.00 [0.50–1.50], $p < 0.001$) (Fig. 2b), suggesting that AI assistance facilitates more comprehensive information retrieval in clinical reasoning. In Q5 (scenario development), the AIL group scored 1.61 [1.38–2.00], significantly higher than the CL group (1.06 [0.50–1.50], $p < 0.05$) (Fig. 2c). In Q6 (challenges), the AIL group scored 1.50 [1.00–1.50], exceeding the CL group (1.00 [0.50–1.50], $p < 0.01$) (Fig. 2d). In Q7 (unintended consequences), the AIL group's score of 1.50 [1.40–2.00] was significantly higher than both the CL group (1.00 [0.88–1.00], $p < 0.01$) and the MG (1.00 [1.00–1.50], $p < 0.05$) (Fig. 2e).

3.2 ChatGPT 4.0 benchmark performance

The ChatGPT 4.0 benchmark produced identical scores across all assessment trials (total score: 16.5; Q1: 2.0; Q2: 2.0; Q3: 3.0; Q4: 2.5; Q5: 2.5; Q6: 2.5; Q7: 2.0), with zero variance across repeated submissions. This score ceiling and complete absence of response variability are indicative of algorithmic determinism rather than flexible clinical reasoning. Notably, ChatGPT's scores exceeded all three human groups, suggesting that LLMs can assemble high-scoring responses to structured assessment rubrics through pattern recognition and retrieval—yet the identical outputs across trials also reveal a fundamental limitation: rigid template dependency unsuited to the contextual adaptability expected in clinical practice.

3.3 Subgroup analysis: AI efficacy across academic performance levels

Subgroup analysis stratified by Pre-Clerkship GPA tertiles (High/Intermediate/Low) revealed significant main effects for both learning tool ($F = 18.071$, $p < 0.001$, $\eta^2 = 0.440$) and GPA stratification ($F = 11.441$, $p < 0.001$, $\eta^2 = 0.332$) (Table 3). Critically, the tool-by-GPA interaction was not significant ($F = 0.213$, $p = 0.930$, $\eta^2 = 0.018$), indicating that the benefit of AI-assisted learning was consistent across students of differing baseline academic performance. Detailed stratified scores are provided in Table 3.

Table 3
Tests of between-subjects' effects for subgroup analysis (two-way factorial ANOVA).

Source of effect	F-value	p-value	η^2
Group (AIL/CL/MG)	18.071	< 0.001	0.440
GPA stratification (High/Mid/Low)	11.441	< 0.001	0.332
Tool-by-GPA interaction effect	0.213	0.930	0.018

Table 4
Baseline characteristics of participants by group

Characteristic	AIL group (n = 18)	CL group (n = 18)	MG group (n = 19)	Total (n = 55)	p-value
Age (years), mean ± SD	20.1 ± 0.8	19.9 ± 0.8	20.0 ± 0.7	20.0 ± 0.8	0.74
Age range (years)	19–21	19–21	19–21	19–21	
Gender, n (%)					0.98
Male	9 (50.0)	9 (50.0)	9 (47.4)	27 (49.1)	
Female	9 (50.0)	9 (50.0)	10 (52.6)	28 (50.9)	
Pre-clerkship GPA, n (%)					0.99
High	6 (33.3)	6 (33.3)	7 (36.8)	19 (34.5)	
Intermediate	6 (33.3)	6 (33.3)	6 (31.6)	18 (32.7)	
Low	6 (33.3)	6 (33.3)	6 (31.6)	18 (32.7)	

Data are presented as mean ± SD or n (%). GPA = grade point average; SD = standard deviation. p-values were calculated using one-way ANOVA for age, chi-square test for gender, and chi-square test for GPA strata. No significant between-group differences were observed (all p > 0.05).

4. Discussion

This study evaluated an AI-assisted learning framework for developing clinical reasoning in unstructured problem-solving and found that students using AI as a guided reasoning tool (AIL group) significantly outperformed those using conventional resources or no tools. The effect was consistent across academic performance levels, and the ChatGPT benchmark provided complementary insights into AI's educational strengths and limitations. Below, we discuss these findings across four themes.

4.1 AI-assisted learning enhances information integration and scenario-based reasoning

The AIL group's significant advantages in Q2 (information needs), Q5 (scenario development), Q6 (challenges), and Q7 (unintended consequences) point to AI's particular value in tasks requiring cross-domain information synthesis and anticipatory reasoning—skills central to systems thinking in clinical contexts (6, 16). This pattern is consistent with prior findings that LLMs can support learners in structuring complex information landscapes (17) and that AI-assisted tools enhance task organization and logical coherence in reasoning-intensive domains (18).

A plausible mechanism is that AI reduces the cognitive load associated with information search across disparate knowledge domains, allowing learners to allocate more cognitive resources to higher-order synthesis and evaluation. In traditional learning, students must manually navigate textbooks, guidelines, and literature across basic medicine, clinical medicine, pharmacology, and ethics—a process that is both time-consuming and prone to omissions (19). By providing rapid, cross-domain information retrieval with keyword-aligned relevance, AI may facilitate the construction of more comprehensive and evidence-grounded clinical scenarios, consistent with the framework requirements of Q5 (evidence-based scenario construction).

However, it is critical to interpret these findings within their proper scope. The observed improvements reflect enhanced performance on a structured assessment rubric when AI-assisted information retrieval was available—not necessarily the development of independent, transferable clinical reasoning that persists without AI support. Long-term retention and real-world clinical translation remain untested in the current design.

4.2 Conventional tools underperform in unstructured problem contexts

The CL group's lower performance relative to both the AIL and MG groups across multiple tasks warrants careful interpretation. Two non-mutually-exclusive explanations may account for this pattern. First, conventional tools (textbooks, databases, guidelines) are optimized for structured information retrieval—locating a specific fact, drug interaction, or diagnostic criterion—but may be ill-suited to the dynamic, cross-domain synthesis required for unstructured problems. This interpretation aligns with evidence that traditional resources underperform in non-linear reasoning tasks (20, 21).

Second, a cognitive efficiency explanation is possible: the time and effort required to locate and integrate relevant information from conventional sources may have exceeded the benefit gained, particularly under time-constrained conditions. The MG group, working from consolidated prior knowledge, may have been more efficient than CL participants who divided attention between problem-solving and resource navigation. The AIL group's rapid, AI-mediated access to synthesized information likely conferred the greatest cognitive efficiency advantage.

4.3 Risks and opportunities of AI in clinical reasoning education

The ChatGPT 4.0 benchmark results reveal a dual-edged profile. On one hand, the model achieved the highest scores across all domains, demonstrating that current LLMs can generate responses that score well against structured clinical reasoning rubrics. On the other hand, the identical scores across all trials —a ceiling effect with zero variance—demonstrate a critical limitation: algorithmic determinism manifesting as rigid template dependency, consistent with reports of homogeneous outputs in GPT-3.5-generated medical cases (22) and formulaic response patterns in GPT-4's USMLE performance (4, 23).

Several considerations preclude treating ChatGPT scores as a clinical gold standard. First, LLMs generate text through probabilistic token prediction and may produce plausible-sounding but factually incorrect statements (24). Second, the assessment framework was designed for human clinical reasoning; its applicability to machine-generated text has not been independently validated. Third, the uniformity of ChatGPT's outputs (standard deviation = 0) fundamentally differs from the variability expected in human clinical reasoning, which must adapt to patient-specific contexts.

These findings have direct educational implications. The template-structured outputs of current AI models, while potentially useful as initial reasoning scaffolds for novice learners, must not be conflated with flexible clinical expertise. We propose that AI's optimal role in clinical reasoning education lies not in answer provision but in guided reasoning facilitation—employing stepwise questioning strategies that prompt learners to articulate their logic, identify evidence gaps, and consider alternative explanations (25). Furthermore, a "deconstruct the AI answer" exercise, in which students critically analyze the structure, assumptions, and limitations of AI-generated responses, may cultivate the evaluative judgment essential for clinical practice.

4.4 AI benefits are independent of baseline academic performance

The non-significant interaction between learning tools and GPA stratification ($p = 0.930$, $\eta^2 = 0.018$) suggests that AI-assisted learning benefits students across the academic performance spectrum, consistent with observations that AI tools can reduce performance gaps in reasoning tasks (22, 26). However, this finding should not be interpreted as evidence that AI substitutes for foundational knowledge: the significant main effect of GPA ($p < 0.001$, $\eta^2 = 0.332$) confirms that students with stronger baseline knowledge continue to outperform their peers. Rather, AI appears to provide a complementary scaffolding effect that raises the floor of performance without eliminating the advantage conferred by prior knowledge. This pattern supports a model in which AI serves as a cognitive amplifier—augmenting, but not replacing, the knowledge base that students bring to clinical reasoning tasks.

4.5 Limitations

This study has several limitations that constrain the generalizability of its findings. First, the sample size (N = 55) and single-institution design limit statistical power for detecting moderate effect sizes and preclude direct generalization to other medical education contexts. Multi-center replication with larger cohorts is needed. Second, the cross-sectional, single-session design captured only immediate performance effects; long-term skill retention and transfer to authentic clinical settings were not assessed. Longitudinal studies tracking clinical reasoning development over clerkship rotations would address this gap.

Third, because the AI model version is a rapidly evolving target, findings specific to ChatGPT 4.0 may not generalize to newer or different LLM architectures (27). Continuous evaluation across model iterations is warranted. Fourth, the adapted assessment framework, while demonstrating good inter-rater reliability

(ICC = 0.801), has not been validated against established clinical reasoning measures or real-world clinical performance outcomes—a necessary step before deployment in high-stakes educational assessment.

Fifth, as an educational intervention study rather than a clinical trial, this study was not prospectively registered. Sixth, the ecological validity of the single-case assessment format may not fully capture the complexity and time pressure of real clinical reasoning. Future work should integrate this framework into longitudinal curricula and evaluate its impact on clinical performance metrics.

5. Conclusion

This study demonstrates that an AI-assisted learning framework, grounded in a validated systems thinking assessment tool, can enhance medical students' performance in solving unstructured clinical problems. The intervention's effects—significant improvements in information integration (Q2), evidence-based scenario construction (Q5), challenge identification (Q6), and risk anticipation (Q7)—were consistent across academic performance levels. The ChatGPT 4.0 benchmark simultaneously illustrates the potential and the pitfalls of current AI: strong rubric-aligned performance coexists with rigid template dependency and zero response variability.

These findings argue for a model of human–AI collaboration in clinical reasoning education in which AI serves as a guided reasoning scaffold rather than an answer provider. The integration of AI tools into medical curricula, paired with critical evaluation exercises that teach students to assess AI outputs, represents a promising direction for preparing healthcare professionals to navigate the complex, unstructured problems of clinical practice.

Declarations

Ethics approval and consent to participate

This study was approved by the Biomedical Research Ethics Committee of Hunan Normal University, China (Approval No. 20251051). All methods were carried out in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants. No patients were involved, no biological samples were collected, and no clinical procedures were performed. All participant data were anonymized prior to analysis.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the Educational Reform Projects of the Hunan Provincial Department of Education, China (Grant Nos. HNJG-2023-0242, HNJG-2022-0062, HNJG-2023-1595, HNJG-202401000470). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

Hang Du: Conceptualization, Methodology, Investigation, Writing – original draft, Visualization. **Yang Tan**: Investigation, Data curation, Formal analysis, Validation. **Jiaxing Zhang**: Methodology, Validation, Formal analysis, Writing – review & editing. **Xuanning Wang**: Data curation, Investigation, Project administration. **Yufei Li**: Investigation, Resources, Writing – review & editing. **Xiaofang Hu**: Validation, Writing – review & editing, Supervision. **Wei Chen**: Writing – review & editing, Resources, Visualization. **Huaying Wu**: Project administration, Writing – review & editing, Methodology. **Lin Li** (Corresponding author): Conceptualization, Supervision, Funding acquisition, Writing – review & editing, Project administration.

All authors read and approved the final manuscript and agree to be accountable for all aspects of the work.

Acknowledgements

The authors thank the medical students who participated in this study for their time and engagement.

Reporting guidelines

This study adheres to the CONSORT 2025 guidelines for reporting randomized trials. A completed CONSORT checklist is provided as a supplementary file.

References

1. Darok MC, Ho L, Holstrom-Mercader MS, Lorenz FJ, Freiberg AS, Dellasega CA. How Medical Students Make Meaning of Uncertainty through a Novel Pediatric Psycho-Oncology Rotation. Health Professions Educ. 2024;10(1):7.
2. Kwon YG, Namgung M, Park SH, Kim MK, Yoo HH, Kim CW. Medical students' perceptions of improving physician satisfaction and patient care: a text network analysis approach. BMC Med Educ. 2024;24(1):1549.
3. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: An assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. Med Teach. 2024;46(3):366–72.
4. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.
5. Naamati-Schneider L, Alt D. The Complex and Ever-Changing Environment of Healthcare Systems: Characteristics and Future Challenges. Technology and Competency-Oriented Learning: Effective Methods for Training the Next Cohort of Healthcare Managers. Springer; 2024. pp. 11–23.
6. Grohs JR, Kirk GR, Soledad MM, Knight DB. Assessing systems thinking: A tool to measure complex reasoning through ill-structured problems. Think Skills Creativity. 2018;28:110–30.
7. Kelly M, Power L, Lee A, Boudreault N, Ali M, Hubinette M. The tip of the iceberg: Generalism in undergraduate medical education, a systems thinking analysis. Med Educ. 2024;58(12):1536–44.
8. Díaz B, Nussbaum M. Artificial intelligence for teaching and learning in schools: The need for pedagogical intelligence. Comput Educ. 2024;217:105071.
9. Prince G, Osipov R, Mazzella A, Chelminski P. Linking the humanities with clinical reasoning: proposing an integrative conceptual model for a graduate medical education humanities curriculum. Acad Med. 2022;97(8):1151–7.
10. Fong L, Wynne K, Verhoeven B. Easing the cognitive load of general practitioners: AI design principles for future-ready healthcare. Technovation. 2025;142:103208.
11. Fouladvand S, Pierson E, Jankovic I, Ouyang D, Chen JH, Daneshjou R. Session Introduction: Artificial Intelligence in Clinical Medicine: Generative and Interactive Systems at the Human-Machine Interface. Biocomputing 20242024. pp. 1–7.
12. Shalong W, Zuo Y, Zou B, Liu G, Zhou J, Zheng Y et al. Enhancing self-directed learning with custom GPT AI facilitation among medical students: A randomized controlled trial. Med Teach.1–8.
13. Yuxian J. Bridging the Knowledge-Skill Gap: The Role of Large Language Model and Critical Thinking in Education. Computers & Education. 2025:105357.
14. Xu Y, Shao Y, Dong J, Shi S, Jiang C, Li Q. Advancing Problem-Based Learning with Clinical Reasoning for Improved Differential Diagnosis in Medical Education. ArXiv. 2025;abs/2503.06099.
15. Kim J, Yu S, Detrick R, Li N. Exploring students' perspectives on Generative AI-assisted academic writing. Educ Inform Technol. 2025;30(1):1265–300.

16. Zhang P, Li H, Tan J. AI-assisted learning for systems thinking in resolving ill-structured problems among healthcare students. Med Educ. 2024;58(11):1413–4.
17. Corbita DP, Mendoza M, Cordova M, Balmas R, Vijayan PM, Madronero J. Evaluating the Reliability and Impact of AI-Assisted Tools on Argumentative Writing Competency Among EFL Omani Learners. In: Mansour N, Bujosa Vadell LM, editors. Finance and Law in the Metaverse World: Regulation and Financial Innovation in the Virtual World. Cham: Springer Nature Switzerland; 2024. pp. 509–18.
18. Bakkum MJ, Hartjes MG, Piët JD, Donker EM, Likic R, Sanz E, et al. Using artificial intelligence to create diverse and inclusive medical case vignettes for education. Br J Clin Pharmacol. 2024;90(3):640–8.
19. Zhao Y-m, Liu S-s, Wang J. Application of data-driven blended online-offline teaching in medicinal chemistry for pharmacy students: a randomized comparison. BMC Med Educ. 2024;24(1):738.
20. Longo UG, De Salvatore S, Valente F, Villa Corta M, Violante B, Samuelsson K. Artificial intelligence in total and unicompartmental knee arthroplasty. BMC Musculoskelet Disord. 2024;25(1):571.
21. Brynjolfsson E, Li D, Raymond L. Generative AI at Work. SSRN Electron J. 2023.
22. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ. 2024;17(5):926–31.
23. Luke WANV, Lee SC, BK H, Chen HWA, Lee ZX. Is ChatGPT 'ready' to be a learning tool for medical undergraduates and will it perform equally in different subjects? Comparative study of ChatGPT performance in tutorial and case-based learning questions in physiology and biochemistry. Med Teach. 2024;46(11):1441–7.
24. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature. 2024;630(8017):625–30.
25. La N, Rattanapitoon SK, Aeksanti T, Rattanapitoon NK. Expanding the Role of Generative AI in Paediatric Intensive Care Education: Beyond Factual Knowledge Toward Integrated Clinical Reasoning. J Paediatr Child Health. 2026;62(1):148–9.
26. Shao H, Wang R. Generative AI an academic equalizer? The differential impact of AI-assisted learning on self-efficacy and intrinsic motivation among university students. Front Psychol. 2026;Volume 17–2026.
27. Bubeck S, Chandrasekaran V, Eldan R, Gehrke JA, Horvitz E, Kamar E et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. ArXiv. 2023;abs/2303.12712.

Figures

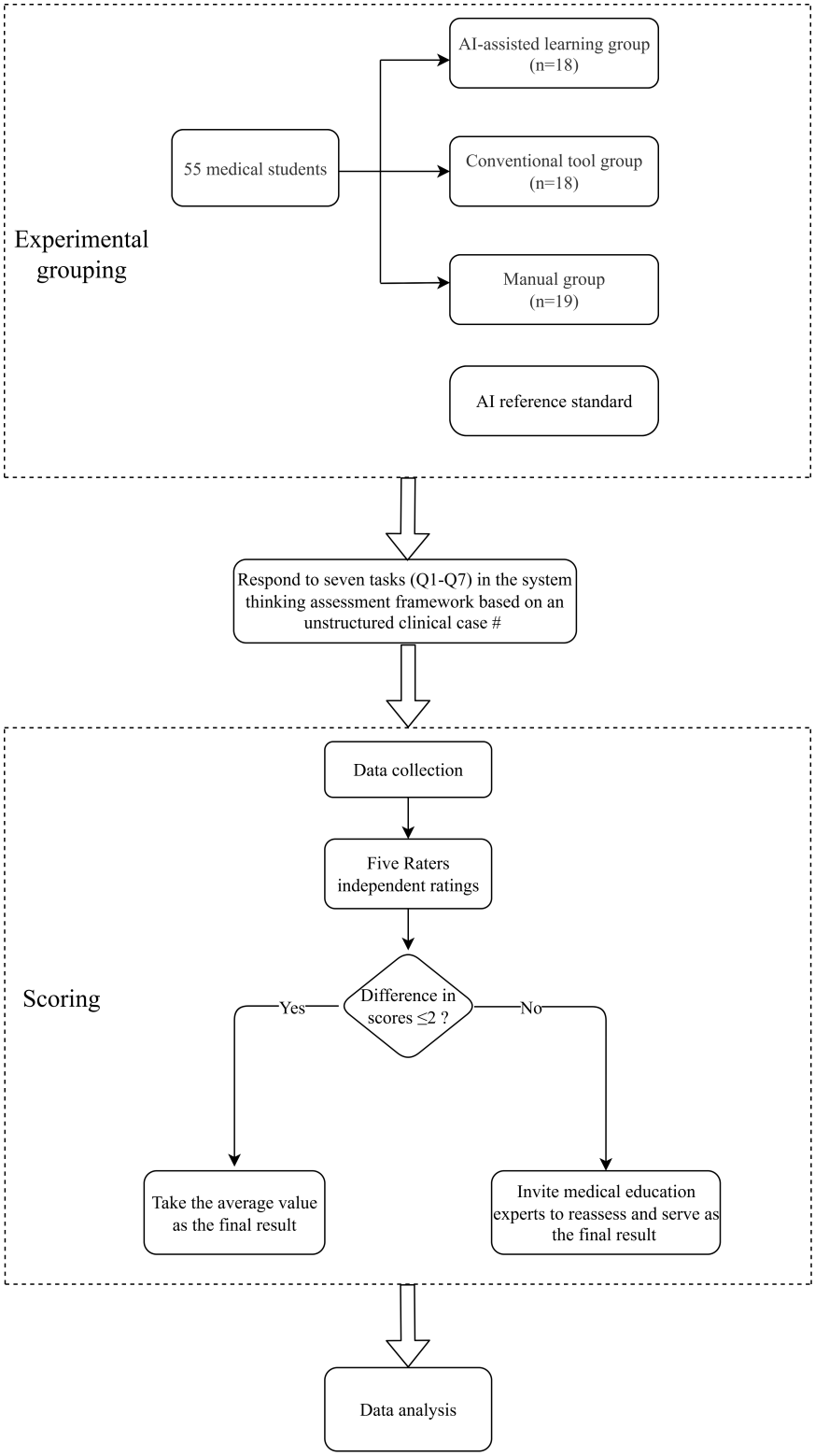


Figure 1

Study design flowchart.

A case of a 45-year-old male patient with persistent abdominal pain, weight loss, and malaise was presented to all three groups.

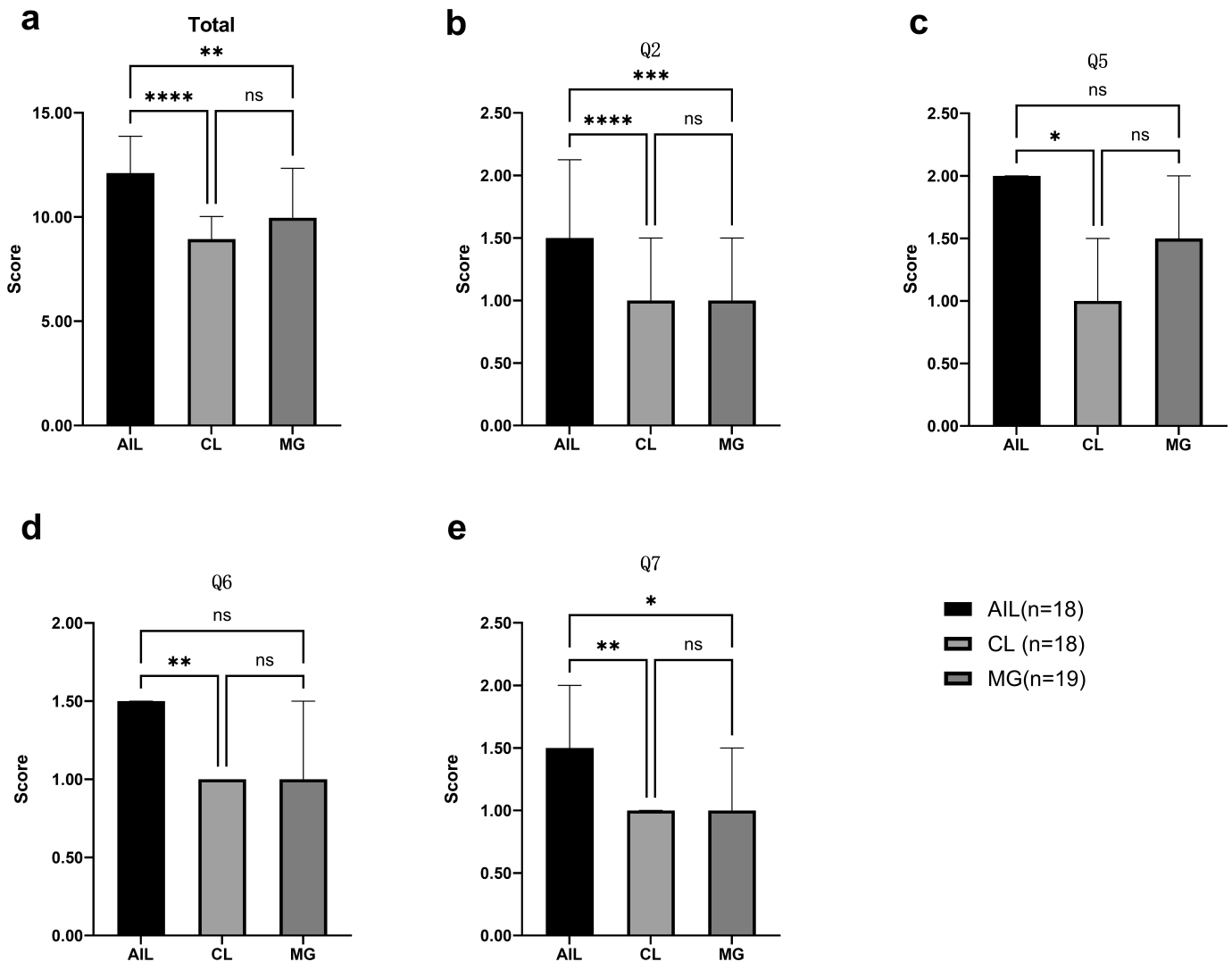


Figure 2

Intergroup differences in systems thinking competency scores.

*(a) Total scores (one-way ANOVA). (b) Q2 scores (Kruskal–Wallis test). (c) Q5 scores (Kruskal–Wallis test). (d) Q6 scores (Kruskal–Wallis test). (e) Q7 scores (Kruskal–Wallis test). Data in (a) are presented as mean ± SD; data in (b–e) are presented as median with interquartile range. Significance markers: *p<0.05, **p<0.01, ***p<0.001. All post-hoc comparisons used Dunn's test with Holm–Bonferroni correction.*

Supplementary Files

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

- SupplementaryMaterials.docx
- CONSORT2025editablechecklist.docx