Advances in Medical Education and Practice

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Comparative Impact of ChatGPT and Conventional Search Tools on Clinical Reasoning Performance: A Randomized Crossover Study in Preclinical Medical Students

Adisak Nartthanarung[1], Komson Plangsiri[2], Pinkawas Kongmalai[1]

[1]Department of Orthopedics, Faculty of Medicine, Kasetsart University, Bangkok, Thailand; [2]Department of Orthopaedics, Faculty of Medicine, Srinakharinwirot University, Nakhon Nayok, Thailand

Correspondence: Pinkawas Kongmalai, Department of Orthopedics, Faculty of Medicine, Kasetsart University, Bangkok, Thailand, Email pinkawass@hotmail.com

Purpose: To compare the impact of ChatGPT and conventional search strategies on clinical reasoning performance among preclinical medical students.

Patients and Methods: A randomized crossover study was conducted at a single institution during a musculoskeletal system course involving 46 second-year medical students. Participants completed a baseline pre-test followed by two structured intervention phases in which they analyzed standardized clinical cases using either ChatGPT or conventional search tools (Google or PubMed). A 60-minute washout period was implemented before crossover to the alternate modality. Post-tests were administered after each phase. The primary outcome was clinical reasoning performance measured using an eight-point rubric-based scale. Secondary outcomes included self-perceived learning, confidence, and qualitative feedback. Paired t-tests were used for within-subject comparisons, and effect sizes were calculated using Cohen's d.

Results: The mean pre-test score was 3.96 (standard deviation 1.65), increasing to 4.96 (standard deviation 1.71) after the first intervention and to 5.70 (standard deviation 1.50) after crossover. Improvements were statistically significant across all paired comparisons ($p < 0.05$), with a large cumulative effect size (Cohen's $d = 0.97$). Performance improvements were observed across both learning modalities, without evidence that gains were attributable to a single approach. Students reported that ChatGPT facilitated rapid organization of differential diagnoses and management plans, whereas conventional search encouraged more deliberate synthesis and comparison of information sources.

Conclusion: Both artificial intelligence–assisted and conventional search strategies improved short-term clinical reasoning performance within a structured active-learning environment. These findings support a balanced integration of large language models alongside traditional search methods in undergraduate medical education.

Clinical Trial Registration Number: TCTR20260218005.

Keywords: medical education, large language models, information-seeking behavior, crossover study design, active learning, undergraduate curriculum

Introduction

Digital transformation has made online learning and just-in-time information seeking routine in undergraduate medical education. A meta-analysis of online versus offline learning in undergraduate medical education reported that online learning can achieve learning outcomes comparable to or better than offline instruction, supporting its widespread adoption in curricula.[1] In parallel, medical students commonly use mobile devices and general search engines to address knowledge gaps; a survey of preclinical medical students identified Google as the most frequently used search engine for researching unfamiliar terms or concepts.[2]

https://doi.org/10.2147/AMEP.S603679
Received: 19 February 2026
Accepted: 24 March 2026
Published: 30 March 2026

The public release of ChatGPT accelerated attention to artificial intelligence (AI) tools built on large language models (LLMs). GPT-4 can generate coherent, context-sensitive responses, but the technical report emphasizes that the model is not fully reliable and can produce hallucinations, warranting caution when reliability is important. In clinical reasoning contexts, such limitations may affect diagnostic accuracy and management decisions, highlighting the need for empirical evaluation in educational settings. A systematic review of LLMs in medical education describes expanding applications, including tutoring, feedback, assessment support, and exam preparation, while also highlighting variability in study rigor and the need for controlled empirical evaluations.[3] In a crossover study in a surgery clerkship, both ChatGPT and Google search improved postintervention quiz scores, and postintervention performance did not differ significantly between tools, reinforcing that educational effects may depend on how tools are embedded into learning activities rather than on tool novelty alone.[4]

From a cognitive load perspective, LLM-generated synthesis may reduce extraneous cognitive load by organizing dispersed information and offering structured explanations, potentially improving efficiency in hypothesis generation and differential diagnosis formulation.[3,5] At the same time, integrating generative AI into medical education raises concerns about inaccurate or fabricated content and uncritical reliance that may promote cognitive offloading, weaken cognitive autonomy, and reduce opportunities for deliberate practice unless paired with faculty guidance and reflective verification strategies.[6,7] These considerations align with scaffolding theory, in which supports are provided to learners and progressively withdrawn as competence develops, and with self-regulated learning frameworks that emphasize meta-cognitive monitoring and active control of learning.[6,8]

Clinical reasoning is widely framed as a core competence for physicians, encompassing information gathering, problem representation, hypothesis generation, differential diagnosis, diagnostic justification, and management planning.[9] Yet definitions and conceptual boundaries vary, complicating alignment across teaching, learning activities, and evaluation.[10] A scoping review of clinical reasoning assessment methods emphasizes selecting complementary assessments that sample different components of this complex construct in ways matched to intended purpose and context.[11] However, despite the increasing use of artificial intelligence tools in medical education, there remains limited empirical evidence directly comparing their impact on clinical reasoning performance with that of conventional search strategies, particularly in preclinical learners within structured case-based learning environments. Given that clinical reasoning is a core competency in medical education and is central to case-based learning, it was selected as the primary outcome of this study. Accordingly, this study compared ChatGPT with conventional search tools during structured case-based learning to evaluate their relative effects on clinical reasoning performance in preclinical medical students.

Materials and Methods

This randomized crossover study was conducted during a scheduled 180-minute instructional session within the musculoskeletal system course for second-year medical students at the Faculty of Medicine, Kasetsart University. The study was implemented within the complete academic cohort of 46 enrolled students.

Participants were randomly assigned in a 1:1 ratio using computer-generated simple random allocation to begin with either ChatGPT or conventional search methods (Google or PubMed). Following a baseline 10-minute pre-test, students completed a 20-minute structured case-based search activity using their assigned modality. A parallel 10-minute post-test (post-test 1) was then administered. A 60-minute washout period involving unrelated academic activities was implemented to minimize potential carryover effects and reduce immediate recall and tool-specific priming before crossover. Participants subsequently switched to the alternate search modality for a second 20-minute session, followed by a second parallel 10-minute post-test (post-test 2). Participants assigned to the ChatGPT group were allowed to use self-directed prompting strategies when interacting with the model. General guidance was provided to encourage engagement with the clinical case, reflecting authentic information-seeking behavior within a structured learning context.

All assessments were designed to evaluate clinical reasoning rather than factual recall. Questions required integration of pathophysiology, prioritization of differential diagnoses, and selection of appropriate investigations and initial management strategies. Assessment design and rubric-based scoring were aligned with established frameworks for clinical reasoning evaluation in medical education.[9,11] Each assessment was scored using a predefined rubric with a maximum score of eight points. The rubric was developed for this study to reflect key domains of clinical reasoning

relevant to the course objectives, including problem representation, generation of differential diagnoses, selection of appropriate investigations, and initial management. Scores were assigned based on the completeness and appropriateness of responses across these domains. The rubric and scoring criteria were reviewed by faculty members involved in the course to ensure content validity, relevance, and clarity.

The primary outcome was objective clinical reasoning performance measured across three time points: baseline, post-test 1, and post-test 2. Secondary outcomes included self-perceived understanding and confidence, collected through structured evaluation forms.

Sample size was calculated for a paired *t*-test comparing post-test scores between learning modalities. Assuming a two-sided α of 0.05, 80% power, an anticipated mean difference of 1.0 point, and a standard deviation of paired differences of 1.5, the minimum required sample size was 18 participants. Inclusion of the full cohort of 46 students exceeded this requirement and increased statistical precision.

Statistical analyses were performed using paired t-tests to compare within-subject score differences across phases. Effect sizes were calculated using Cohen d for paired samples to quantify magnitude beyond statistical significance, consistent with recommendations for reporting educational intervention outcomes.[12] Statistical significance was defined as $p < 0.05$.

The study was approved by the Kasetsart University Research Ethics Committee under expedited review (KUREC-HSR68/072; COA No. COA69/02; approval date 14 January 2026; and was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants prior to enrollment.

Results

All 46 second-year medical students were assessed for eligibility, randomized into two intervention sequences, and completed the baseline pre-test, both intervention phases, and all post-test assessments. No participants were excluded from analysis. Participant progression through the study is presented in Figure 1.

Descriptive statistics and paired comparisons of assessment scores across the three time points are summarized in Table 1. The mean pre-test score was 3.96 (SD 1.65), which increased to 4.96 (SD 1.71) after the first intervention phase and further to 5.70 (SD 1.50) following crossover and completion of the second intervention phase.

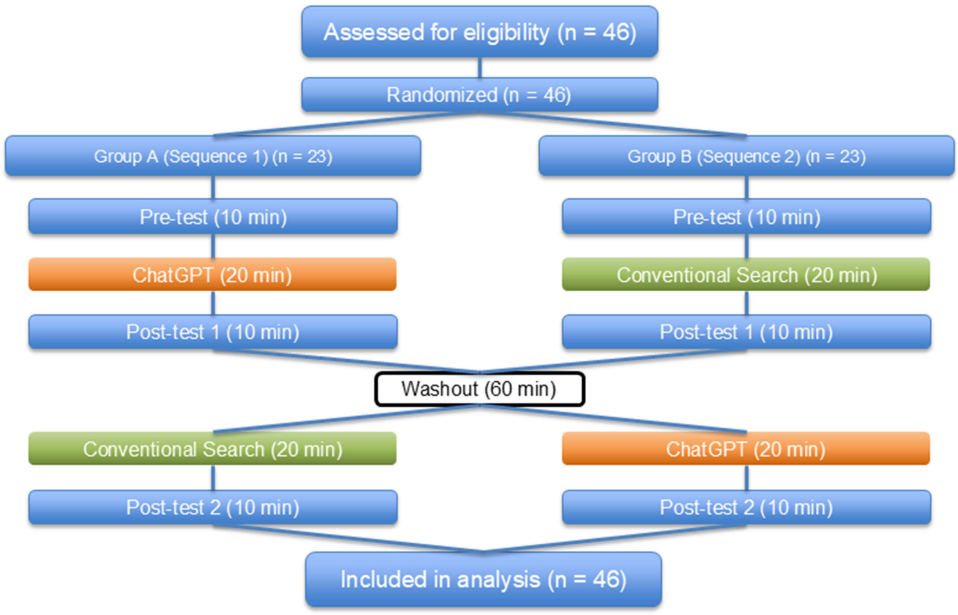

Figure 1 CONSORT flow diagram and schematic representation of the randomized crossover study design. Forty-six second-year medical students were randomized into two intervention sequences. Group A received ChatGPT followed by conventional search, while Group B received conventional search followed by ChatGPT, separated by a 60-minute washout period. Each phase included a 20-minute learning activity followed by a 10-minute post-test. All participants completed both intervention phases and were included in the final analysis.

https://doi.org/10.2147/AMEP.S603679

Table 1 Changes in Clinical Reasoning Performance Across Study Phases and results of Paired Comparisons

Measure	Mean ± SD	Comparison	t (df=45)	p-value	Cohen's d	Effect Size
Pre-test	3.96 ± 1.65	Pre vs Post-test 1	−3.79	<0.05	0.56	Moderate
Post-test 1	4.96 ± 1.71	Post-test 1 vs Post-test 2	−2.96	<0.05	0.44	Small–Moderate
Post-test 2	5.70 ± 1.50	Pre vs Post-test 2	−6.57	<0.05	0.97	Large

Notes: Mean scores increased progressively across study phases. Paired comparisons demonstrate statistically significant improvements with corresponding effect sizes.

Table 2 Summary of Qualitative Feedback on the Two Learning Modalities

Aspect	ChatGPT	Conventional Search
Perceived benefit	Facilitated rapid organization of differential diagnoses and management plans	Promoted comparison across multiple information sources
Learner reflection	Students emphasized the importance of verifying artificial intelligence outputs against reliable references	Students reported that using this approach enhanced their ability to critically evaluate information sources

Paired analyses demonstrated statistically significant improvements across all comparisons, including pre-test to post-test 1 ($t(45) = -3.79$, $p < 0.05$), post-test 1 to post-test 2 ($t(45) = -2.96$, $p < 0.05$), and pre-test to post-test 2 ($t(45) = -6.57$, $p < 0.05$), with corresponding effect sizes ranging from moderate to large. No reduction in performance was observed following crossover.

Qualitative evaluation indicated that students perceived both learning modalities as beneficial. ChatGPT was described as facilitating rapid organization of differential diagnoses and management plans, whereas conventional search required more time but promoted comparison across multiple information sources. Students reported increased confidence when applying knowledge immediately after the instructional session and emphasized the importance of verifying artificial intelligence outputs against reliable references. Several participants noted that experiencing both approaches enhanced their ability to critically evaluate information sources. Qualitative feedback regarding the perceived strengths of each learning modality is summarized in Table 2.

Discussion

This randomized crossover study demonstrates that structured integration of both ChatGPT and conventional search tools significantly enhances short-term clinical reasoning performance among preclinical medical students. The large cumulative effect size observed across study phases suggests meaningful educational impact beyond simple exposure to digital tools, and the improvement across both intervention phases is consistent with reinforcement of reasoning processes rather than a transient novelty effect.

Recent literature has begun to explore the comparative utility of ChatGPT and traditional search tools in medical education. In a crossover study of third-year medical students during a surgery clerkship, ChatGPT was comparable to Google Search in improving quiz performance, with no statistically significant difference in postintervention scores between groups.[4] Although that study provided early evidence that LLM-based tools may function as viable learning aids, its emphasis was feasibility and short-term score comparison within a small sample. In contrast, our study extends prior work by incorporating a larger cohort, reporting standardized effect sizes to convey magnitude beyond statistical significance,[12] and examining cumulative improvement across sequential learning phases. Importantly, our primary outcome targeted clinical reasoning performance assessed through structured rubrics rather than simple knowledge recall, aligning with recommendations to match assessment methods to specific components of the clinical reasoning construct.[9,11] Because crossover studies increase efficiency by using learners as their own controls but can be threatened by period and carryover effects if not addressed explicitly, careful design and transparent reporting remain essential.[13,14]

From a cognitive load perspective, AI-generated synthesis may reduce extraneous load by organizing dispersed information and offering structured explanations, allowing learners to allocate attention to integration and application.[5,15] In parallel, scaffolding theory suggests that supports should be deliberately calibrated and gradually withdrawn to promote independent performance.[8] However, LLM outputs can be fluent yet incorrect; the GPT-4 technical report acknowledges residual reliability limitations, and clinical guidance highlights hallucinations as a practical risk that requires verification against trusted sources.[16] The progressive improvement observed after crossover may therefore reflect transfer of reasoning strategies across tools within a structured activity rather than dependence on a single tool.

Our interpretation of attitudes also differs from some prior reports. While the clerkship study reported reluctance among students to use ChatGPT for learning during clinical rotations,[4] our qualitative findings suggest emerging critical appraisal behavior and metacognitive monitoring. This is educationally important because uncritical reliance on generative AI has been argued to promote cognitive offloading and threaten cognitive autonomy without explicit pedagogical scaffolding and faculty oversight.[7,17] Taken together with recent systematic and scoping reviews emphasizing both educational promise and persistent accuracy, ethics, and governance concerns, our findings support a hybrid approach that integrates LLMs as supervised cognitive supports while maintaining conventional search and appraisal skills.[3,18]

In addition, large language models may generate hallucinations, producing responses that appear plausible but contain inaccuracies. In musculoskeletal case-based learning, such outputs may affect diagnostic reasoning or management planning if not critically appraised.[19–21] Furthermore, variability in prompting strategies can influence the quality and accuracy of responses, potentially affecting learning outcomes and limiting reproducibility. These considerations highlight the importance of verifying AI-generated information against reliable sources and providing appropriate guidance when integrating artificial intelligence into medical education. Artificial intelligence tools may therefore be incorporated into medical curricula as supportive learning resources within structured educational frameworks, with emphasis on maintaining critical appraisal and verification skills.

Despite these contributions, several strengths and limitations warrant consideration. Strengths of this study include its randomized crossover design, which improves internal validity by allowing each participant to serve as their own control, the complete participation rate within an authentic classroom setting, and the use of structured rubric-based assessment aligned with clinical reasoning theory. In addition, reporting standardized effect sizes enhances interpretability beyond statistical significance alone. However, limitations include the single-institution context and relatively small sample size, which may limit the generalizability of the findings; the short-term outcome assessment without evaluation of long-term retention or transfer to clinical environments; and potential residual period effects inherent to crossover designs, including carryover effects, which cannot be fully excluded despite implementation of a 60-minute washout interval to reduce immediate recall and tool-specific priming. Furthermore, the quality of large language model outputs is dependent on prompt formulation and model version, which may affect reproducibility over time. Future studies should investigate longitudinal retention, application in clinical-year learners, and structured artificial intelligence literacy curricula designed to optimize safe and effective integration.

Conclusion

Both artificial intelligence–assisted and conventional search strategies improved clinical reasoning performance, supporting their complementary roles within a structured learning environment rather than the superiority of one approach. The integration of artificial intelligence into medical education should be accompanied by appropriate faculty oversight and emphasis on verification of information to mitigate potential risks, including inaccurate or hallucinated outputs.

Data Sharing Statement

The datasets generated and analyzed during the current study are not publicly available due to institutional data protection policies involving student academic performance data. De-identified data may be made available from the corresponding author upon reasonable request and with approval from the institutional review board.

Ethics Approval and Informed Consent

The study was approved by the Kasetsart University Research Ethics Committee (KUREC), Faculty of Medicine, Kasetsart University, Bangkok, Thailand (study code KUREC-HSR68/072; COA No. COA69/02; approval date January 14, 2026; review method: expedited). The study was conducted in accordance with the Declaration of Helsinki and applicable international guidelines for human research protection. Written informed consent was obtained from all participants prior to enrollment. Participation was voluntary, and students were assured that their academic evaluation would not be affected by participation or non-participation.

Consent for Publication

Not applicable. This study did not include identifiable images, videos, recordings, or personal data requiring individual consent for publication.

Acknowledgments

The authors thank the second-year medical students who participated in this study and the faculty members who supported the implementation of the structured learning session. No professional writing assistance was used in the preparation of this manuscript.

Author Contributions

PK made the primary contribution to this work, including the conception and design of the study, execution of the intervention, data collection, data analysis, interpretation of the data, and drafting of the manuscript. AN and KP contributed to the study design, provided academic supervision, and critically reviewed and revised the manuscript for important intellectual content. All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This research received no external funding.

Disclosure

The authors declare that they have no financial or non-financial competing interests related to this work.

References

1. Pei L, Wu H. Does online learning work better than offline learning in undergraduate medical education? A systematic review and meta-analysis. *Med Educ Online*. 2019;24(1):1666538. doi:10.1080/10872981.2019.1666538
2. Singh K, Sarkar S, Gaur U, et al. Smartphones and educational apps use among medical students of a smart university campus. original research. *Fronti Commun*. 2021;6. doi:10.3389/fcomm.2021.649102
3. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Med Educ*. 2024;58(11):1276–1285. doi:10.1111/medu.15402
4. Araji T, Brooks AD. Evaluating the role of ChatGPT as a study aid in medical education in surgery. *J Surg Educ*. 2024;81(5):753–757. doi:10.1016/j.jsurg.2024.01.014
5. Sweller J. Cognitive load during problem solving: effects on learning. *Cognitive Sci*. 1988;12(2):257–285. doi:10.1207/s15516709cog1202_4
6. Zimmerman BJ. Becoming a self-regulated learner: an overview. *Theory Into Pract*. 2002;41(2):64–70. doi:10.1207/s15430421tip4102_2
7. Izquierdo-Condoy JS, Arias-Intriago M, Tello-De-la-Torre A, Busch F, Ortiz-Prado E. Generative artificial intelligence in medical education: enhancing critical thinking or undermining cognitive autonomy? *J Med Internet Res*. 2025;27:e76340. doi:10.2196/76340
8. Wood D, Bruner JS, Ross G. The role of tutoring in problem solving. *J Child Psychol Psychiatry*. 1976;17(2):89–100. doi:10.1111/j.1469-7610.1976.tb00381.x
9. Connor DM, Durning SJ, Rencic JJ. Clinical reasoning as a core competency. *Acad Med*. 2020;95(8):1166–1171. doi:10.1097/acm.0000000000003027
10. Young M, Thomas A, Lubarsky S, et al. Drawing Boundaries: the Difficulty in Defining Clinical Reasoning. *Acad Med*. 2018;93(7):990–995. doi:10.1097/acm.0000000000002142
11. Daniel M, Rencic J, Durning SJ, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. *Acad Med*. 2019;94(6):902–912. doi:10.1097/acm.0000000000002618

12. Sullivan GM, Feinn R. Using effect size—or why the P value is not enough. *J Grad Med Educ*. 2012;4(3):279–282. doi:10.4300/jgme-d-12-00156.1
13. Lim C-Y, In J. Considerations for crossover design in clinical study. *Korean J Anesthesiol*. 2021;74(4):293–299. doi:10.4097/kja.21165
14. Li T, Yu T, Hawkins BS, Dickersin K, Manzoli L. Design, analysis, and reporting of crossover trials for inclusion in a meta-analysis. *PLoS One*. 2015;10(8):e0133023. doi:10.1371/journal.pone.0133023
15. van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. *Med Educ*. 2010;44 (1):85–93. doi:10.1111/j.1365-2923.2009.03498.x
16. Roustan D, Bastardot F. The clinicians' guide to large language models:a general perspective with a focus on hallucinations. *Interactive J Med Res*. 2025;14:e59823. doi:10.2196/59823
17. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ*. 2023;9:e48291. doi:10.2196/48291
18. Aster A, Laupichler MC, Rockwell-Kollmann T, Masala G, Bala E, Raupach T. ChatGPT and other large language models in medical education — scoping literature review. *Med Sci Educ*. 2024;35(1):555–567. doi:10.1007/s40670-024-02206-6
19. Saglam S, Uludag V, Karaduman ZO, Arıcan M, Yücel MO, Dalaslan RE. Comparative evaluation of artificial intelligence models GPT-4 and GPT-3.5 in clinical decision-making in sports surgery and physiotherapy: a cross-sectional study. *BMC Med Inf Decis Making*. 2025;25(1):163. doi:10.1186/s12911-025-02996-8
20. Sánchez-Rosenberg G, Magnéli M, Barle N, et al. ChatGPT-4 generates orthopedic discharge documents faster than humans maintaining comparable quality: a pilot study of 6 cases. *Acta Orthopaedica*. 2024;95:152–156. doi:10.2340/17453674.2024.40182
21. Zamora T, Salas P, Zuñiga S, Botello E, Andia ME. Generative artificial intelligence, large language models and ChatGPT in musculoskeletal oncology: current applications and future potential. *J Clinl Orthopaedics Trauma*. 2025;69:103161. doi:10.1016/j.jcot.2025.103161