Comparison of ChatGPT and Online Medical Databases for Information Retrieval Among Medical Students: Quasi-Experimental Study

Isarapong Rattana, Yingyong Chinthammitr, Atthakorn Wutthimanop, Naphatsawan Phongsutham, Parissada Anugoolgarn

Submitted to: JMIR Medical Education
on: September 08, 2025

Table of Contents

Comparison of ChatGPT and Online Medical Databases for Information Retrieval Among Medical Students: Quasi-Experimental Study

Isarapong Rattana[1] MD; Yingyong Chinthammitr[2] MD; Atthakorn Wutthimanop[3] MD; Naphatsawan Phongsutham[4] MD; Parissada Anugoolgarn[4] MD

[1]Division of Hematology, Department of Internal Medicine Maharaj Nakhon Si Thammarat Hospital Nakhon Si Thammarat TH
[2]Division of Hematology, Department of Medicine Faculty of Medicine Siriraj Hospital Mahidol University Bangkok TH
[3]Division of Cardiology, Department of Medicine Maharaj Nakhon Si Thammarat Hospital Nakhon Si Thammarat TH
[4]Department of Medicine Maharaj Nakhon Si Thammarat Hospital Nakhon Si Thammarat TH

Corresponding Author:
Isarapong Rattana MD
Division of Hematology, Department of Internal Medicine
Maharaj Nakhon Si Thammarat Hospital
198 Ratchadamnoen Road, Nai Mueang Subdistrict, Mueang District
Nakhon Si Thammarat
TH

Abstract

Background: Medical students increasingly rely on online tools for clinical information retrieval. While traditional platforms like Google, PubMed, and UpToDate are commonly used, the rise of large language models (LLMs) such as ChatGPT offers new potential in this area.

Objective: This study compared ChatGPT with conventional online resources in terms of accuracy, speed, and user satisfaction when retrieving medical information among sixth-year medical students.

Methods: We conducted a quasi-experimental, posttest-only study with 64 sixth-year medical students at a Thai teaching hospital. Students were initially randomly assigned to use either ChatGPT or traditional online medical resources (Google, PubMed, and UpToDate), with final groups of 35 and 29 participants respectively due to protocol deviations. Participants completed a 10-question open-ended test on internal medicine topics. Responses were scored using expert-validated rubrics with multiple key elements per question, totaling 10 points each. We also recorded completion time and assessed satisfaction via a 15-item Likert questionnaire. Between-group comparisons used independent t-tests.

Results: The ChatGPT group achieved significantly higher accuracy scores (mean 63.79, SD 9.71) than the control group (mean 47.22, SD 8.82, P<.001), reflecting a 35.1% improvement. They also completed the task more quickly (mean 58.59 vs 75.76 minutes, P=.003), with a 22.7% time reduction. Satisfaction scores were higher overall in the ChatGPT group (mean 62.06, SD 7.14 vs 56.72, SD 10.23; P=.018), particularly in the search process domain (P<.001). Performance varied by cognitive level, with application-type questions showing the largest advantage (20.63 percentage points), followed by remembering-level questions (16.14 percentage points) and analysis-level questions (8.54 percentage points). No significant associations were found between ChatGPT use and demographics, prior experience, or AI attitudes.

Conclusions: ChatGPT significantly outperformed traditional online resources in accuracy, speed, and user satisfaction for clinical information retrieval among medical students. These findings suggest its potential as a supportive tool in medical education. However, future integration should include training in critical appraisal and responsible use to ensure safe and effective application in learning environments.

(JMIR Preprints 08/09/2025:83670)
DOI: https://doi.org/10.2196/preprints.83670

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Original Paper

Comparison of ChatGPT and Online Medical Databases for Information Retrieval Among Medical Students: Quasi-Experimental Study

Isarapong Rattana[1], MD; Yingyong Chinthammitr[2], MD; Atthakorn Wutthimanop[3], MD; Naphatsawan Phongsutham[4], MD; Parissada Anugoolgarn[5], MD

[1,3,4,5]Department of Internal Medicine, Maharaj Nakhon Si Thammarat Hospital, Thailand

[2]Department of Internal Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Thailand

Corresponding Author

Isarapong Rattana, MD

Department of Internal Medicine, Maharaj Nakhon Si Thammarat Hospital

198 Ratchadamnoen Rd, Nakhon Si Thammarat, 80000, Thailand

Phone: +6695-854-6156

Email: isarapong2130@gmail.com

Abstract

Background: Medical students increasingly rely on online tools for clinical information retrieval. While traditional platforms like Google, PubMed, and UpToDate are commonly used, the rise of large language models (LLMs) such as ChatGPT offers new potential in this area.

Objective: This study compared ChatGPT with conventional online resources in terms of accuracy, speed, and user satisfaction when retrieving medical information among sixth-year medical students.

Methods: We conducted a quasi-experimental, posttest-only study with 64 sixth-year medical students at a Thai teaching hospital. Students were initially randomly assigned to use either ChatGPT or traditional online medical resources (Google, PubMed, and

UpToDate), with final groups of 35 and 29 participants respectively due to protocol deviations. Participants completed a 10-question open-ended test on internal medicine topics. Responses were scored using expert-validated rubrics with multiple key elements per question, totaling 10 points each. We also recorded completion time and assessed satisfaction via a 15-item Likert questionnaire. Between-group comparisons used independent t-tests.

Results: The ChatGPT group achieved significantly higher accuracy scores (mean 63.79, SD 9.71) than the control group (mean 47.22, SD 8.82, P<.001), reflecting a 35.1% improvement. They also completed the task more quickly (mean 58.59 vs 75.76 minutes, P=.003), with a 22.7% time reduction. Satisfaction scores were higher overall in the ChatGPT group (mean 62.06, SD 7.14 vs 56.72, SD 10.23; P=.018), particularly in the search process domain (P<.001). Performance varied by cognitive level, with application-type questions showing the largest advantage (20.63 percentage points), followed by remembering-level questions (16.14 percentage points) and analysis-level questions (8.54 percentage points). No significant associations were found between ChatGPT use and demographics, prior experience, or AI attitudes.

Conclusions: ChatGPT significantly outperformed traditional online resources in accuracy, speed, and user satisfaction for clinical information retrieval among medical students. These findings suggest its potential as a supportive tool in medical education. However, future integration should include training in critical appraisal and responsible use to ensure safe and effective application in learning environments.

Keywords: Artificial Intelligence; Education, Medical; Educational Technology; Evidence-Based Medicine; Information Storage and Retrieval; Internet; Natural Language Processing; Students, Medical

Introduction

Background

Contemporary medical students and healthcare professionals must develop sophisticated skills for rapidly accessing and evaluating medical information—this has become a fundamental competency [1]. The exponential growth of medical knowledge, with an estimated doubling time of 73 days [2], presents significant

challenges for medical education and clinical practice. Given this information explosion, medical students in their clinical years need strong research competencies to navigate the vast landscape of medical literature effectively [3].

Traditionally, medical students have relied on various online resources for information retrieval. Studies have shown that Google is frequently used by medical students, with one study finding Google usage in 69% of biomedical information-seeking sessions [4]. Although Google offers broad accessibility and ease of use, concerns persist about the quality and reliability of online health information [5]. Professional medical databases such as PubMed provide access to peer-reviewed literature but often require advanced search skills and more time to identify clinically relevant information [6]. Traditional resources like UpToDate have been widely used, but recent studies have shown growing interest in AI-powered tools. Medical students now express positive perceptions about ChatGPT and artificial intelligence in their training [7].

The COVID-19 pandemic accelerated the digital transformation of medical education, increasing reliance on online resources and highlighting the need for efficient digital information literacy skills [8]. This shift emphasized the importance of evidence-based medicine (EBM) training and the development of critical appraisal skills for evaluating online medical information [9].

The Emergence of AI in Medical Education

The introduction of large language models (LLMs), particularly ChatGPT, has created new opportunities and challenges for medical education. ChatGPT, developed by OpenAI, demonstrates remarkable capabilities in generating human-like text responses and processing complex queries [10]. Recent studies have shown that ChatGPT can pass medical licensing examinations, including achieving passing scores on all three steps of the United States Medical Licensing Examination (USMLE) [11]. Furthermore, recent studies have demonstrated ChatGPT's potential in supporting medical education through interactive learning scenarios [12].

However, the integration of AI tools in medical education is not without concerns. Key limitations include potential inaccuracies, reliance on training data that may not reflect current medical knowledge, inherent biases, and the risk of students developing

over-dependence on AI-generated responses [13]. A comparative study in otorhinolaryngology found that UpToDate provided significantly more reliable and useful information than ChatGPT [14], while another study reported that ChatGPT's medical advice achieved only 68.2% on the Patient Education Materials Assessment Tool compared to Google's 89.4% [15].

Study Rationale and Objectives

Although there is clearly growing interest in AI applications in medical education, we have limited research comparing ChatGPT's effectiveness with traditional online medical resources in real-world educational settings, particularly in non-English speaking countries. This study addresses this gap by conducting a systematic comparison in the context of Thai medical education.

The primary objective was to compare the accuracy of medical information retrieved using ChatGPT versus traditional online resources (Google, PubMed, and UpToDate) among sixth-year medical students. Secondary objectives included comparing retrieval time, assessing user satisfaction, and analyzing factors influencing resource selection.

Methods

Study Design and Setting

We conducted a quasi-experimental study using a two-group posttest-only design at the Clinical Medical Education Center, Maharaj Nakhon Si Thammarat Hospital, Thailand. The study was approved by the Institutional Review Board of Maharaj Nakhon Si Thammarat Hospital (IRB No. A032/2568). All participants provided written informed consent, and data were anonymized to protect participant privacy.

Participants

All 64 sixth-year medical students at our institution were invited to participate. Inclusion criteria were: (1) current enrollment as a sixth-

year medical student, (2) voluntary consent to participate, and (3) basic computer and internet literacy skills. Exclusion criteria included inability to complete the entire assessment session or withdrawal of consent during the study.

We calculated our sample size using Cohen's formula for comparing means between two independent groups [16]: $n = 2(Z\alpha/2 + Z\beta)^2\sigma^2/d^2$. Using $\alpha=0.05$ and power=0.80, estimated standard deviation=1.5, and clinically significant difference=1.0, the required sample size was 18 participants per group.

Randomization and Group Assignment

A total of 64 medical students participated in the study and were randomly assigned using simple random sampling to either the experimental group (ChatGPT) or the control group (traditional online medical resources), with an intended allocation of 32 participants per group.

However, due to protocol deviations, 3 participants originally assigned to the control group mistakenly used ChatGPT to complete the task. To maintain the integrity of the analysis and follow intention-to-treat principles, these participants were subsequently reclassified into the experimental group. This deviation occurred due to ambiguous instructions in the initial briefing. To prevent similar issues in future studies, we implemented enhanced participant briefings that include: (1) explicit verbal instructions specifying the assigned information source at the beginning of each session, (2) written reminder cards placed at each workstation listing only the permitted resources for that group, and (3) a demonstration of the assigned tool(s) before commencing the test. Additionally, a research assistant was designated to monitor compliance during the first 5 minutes of the assessment. As a result, the final group sizes were unequal: the experimental group comprised 35 participants and the control group 29 participants. Additionally, two participants in the control group inadvertently used ChatGPT for approximately 40% of their responses. Following intention-to-treat principles, these participants remained in the control group for primary analysis.

Study Instruments

ChatGPT Implementation

The experimental group used GPT-4o mini (free OpenAI version) accessed through the ChatGPT application or the web-based interface at chat.openai.com. Students were instructed to use natural language queries in Thai or English according to their preference. Example prompts included: "Please explain the diagnostic criteria and treatment options for [specific condition]" or "What are the contraindications and side effects of [specific medication]?" No specific prompt engineering training was provided to reflect real-world usage patterns.

Clinical Knowledge Test

A standardized test comprising 10 open-ended questions covering major internal medicine topics was developed and validated by five experts (four internists and one experienced nurse practitioner) achieving an Item-Objective Congruence (IOC) index ≥0.6 for all items (see Multimedia appendix 4, page 1-4). Questions were distributed across cardiovascular diseases, respiratory medicine, gastroenterology, endocrinology, nephrology, infectious diseases, neurology, rheumatology, and hematology (see Multimedia appendix 1, page 2-12). Each question was scored using a detailed rubric with multiple key elements totaling 10 points per question.

The questions were analyzed using Bloom's taxonomy: 67.7% (n=23) assessed remembering, 23.5% (n=8) assessed application, and 8.8% (n=3) assessed analysis skills.

User Satisfaction Questionnaire

A 15-item questionnaire using a 5-point Likert scale (1=least satisfied to 5=most satisfied) assessed three domains: satisfaction with the search process (5 items), satisfaction with answer quality (5 items), and satisfaction with practical application (5 items) (see Multimedia appendix 1, page 12). All items achieved IOC ≥0.6, confirming content validity (see Multimedia appendix 4, page 5-6).

Data Collection Procedures

Following informed consent, participants completed a demographic questionnaire including age, gender, GPA, internet usage patterns, experience with various information sources, attitudes toward AI, and access to digital resources validity (see Multimedia appendix 1,

page 1). The clinical knowledge test was administered in a controlled lecture room with participants seated at appropriate distances. Each participant had 150 minutes to complete all 10 questions using their assigned information sources. Time stamps were recorded for each question. Post-test satisfaction questionnaires were completed within 30 minutes.

Scoring and Assessment

A single trained assessor evaluated answer sheets using standardized scoring rubrics. To ensure scoring consistency and minimize bias, the assessor was blinded to group assignments and information sources used by participants. The assessor underwent comprehensive training on the scoring system and criteria before evaluation began. Each question was scored according to predefined key elements with explicit scoring criteria developed through expert consensus. The detailed scoring rubrics provided clear guidelines for awarding partial credit, ensuring systematic and objective assessment across all participants question (see Multimedia appendix 2, page 1-6).

Statistical Analysis

We analyzed data using IBM SPSS Statistics version 20.0 (IBM Corp., Armonk, NY, USA). Descriptive statistics included frequencies, percentages, means, and standard deviations. Between-group comparisons utilized independent t-tests for continuous variables and chi-square tests for categorical variables. All p-values are reported as exact values when ≥.001 and as P<.001 when smaller. Pearson correlation coefficients examined relationships between variables. 95% confidence intervals were calculated for all mean differences. Statistical significance was set at P<.05. Following intention-to-treat principles, all analyses were conducted based on the final group assignments, including the contaminated cases, which provides a more conservative estimate of the intervention effect.

Ethical Considerations

For studies involving human subjects research, researchers must provide a statement summarizing the ethics board review assessment and outcome (approval, waiver, or exemption), including a case or application number if available. This study was approved

by the Institutional Review Board of Maharaj Nakhon Si Thammarat Hospital (IRB No. A032/2568). All participants provided written informed consent before participation. Privacy and confidentiality were maintained through data anonymization and secure storage. No participant compensation was provided. The study posed minimal risk to participants and was conducted in accordance with the Declaration of Helsinki.

Results

Participant Characteristics

Table 1 presents the baseline characteristics of study participants. The mean age was 23.80 (SD 1.35) years, with 39 (60.9%) female participants. The mean GPA was 3.03 (SD 0.35). No significant differences in baseline characteristics were observed between groups (all P>.05). No factors significantly predicted ChatGPT use, including demographics, prior experience with information sources, or attitudes toward AI (all P>.05).

Table 1. Baseline characteristics of study participants (N=64)

Characteristic	ChatGPT Group (n=35)	Control Group (n=29)	Total (N=64)	P value
Age (years), mean (SD)	23.80 (0.90)	23.79 (1.76)	23.80 (1.35)	.984
Gender, n (%)				.231
- Male	16 (45.7)	9 (31.0)	25 (39.1)	
- Female	19 (54.3)	20 (69.0)	39 (60.9)	
GPA, mean (SD)	2.99 (0.40)	3.08 (0.29)	3.03 (0.35)	.267
Has laptop/computer, n (%)	33 (94.3)	29 (100.0)	62 (96.9)	.497
Has high-speed internet, n (%)	35 (100.0)	28 (96.6)	63 (98.4)	.453

Information Source Usage Patterns

Google was the most frequently used information source, with 84.4% using it daily, followed by ChatGPT (34.4% daily users). PubMed and UpToDate showed primarily weekly to monthly usage patterns (Table 2).

Table 2. Frequency of information source usage among participants

Information Source	Daily n (%)	Weekly n (%)	Monthly n (%)	Yearly n (%)	Never n (%)
ChatGPT	22 (34.4)	19 (29.7)	5 (7.8)	6 (9.4)	12 (18.8)
Google	54 (84.4)	7 (10.9)	3 (4.7)	0 (0.0)	0 (0.0)
PubMed	0 (0.0)	21 (32.8)	27 (42.2)	15 (23.4)	1 (1.6)
UpToDate	1 (1.6)	26 (40.6)	23 (35.9)	14 (21.9)	0 (0.0)

Primary Outcome:

Accuracy of Information Retrieval

ChatGPT users demonstrated significantly higher accuracy scores (mean 63.79, SD 9.71) compared to the control group (mean 47.22, SD 8.82, P<.001), corresponding to a 35.1% improvement (mean difference 16.57, 95% CI: 11.87-21.27) (Table 3, Figure 1). When analyzed by individual questions, the ChatGPT group outperformed the control group on all 10 questions (all P<.05), with the largest difference observed in hematology questions, where the mean difference was 2.41 points (95% CI: 1.47-3.35, P<.001), (see Multimedia appendix 3).

When analyzed by Bloom's taxonomy, performance varied significantly by cognitive level, with students demonstrating strongest performance on application questions (mean 62.93%, SD 18.46%), followed by remembering questions (mean 58.44%, SD 12.22%) and analysis questions (mean 22.11%, SD 12.61%). The ChatGPT group consistently outperformed the control group across all cognitive domains: remembering questions showed a 16.14 percentage point advantage (65.75% vs 49.61%, P<.001), application questions demonstrated the largest difference of 20.63 percentage points (72.28% vs 51.65%, P<.001), and analysis questions, though challenging for both groups, still favored the ChatGPT group by 8.54 percentage points (25.97% vs 17.44%, P=.006) (see Multimedia appendix 3). These findings indicate that while higher-order thinking tasks proved most difficult for all

participants, ChatGPT users maintained superior performance across all levels of cognitive complexity, with the greatest advantage observed in application-level questions requiring practical knowledge implementation.

Table 3. Comparison of accuracy scores between the ChatGPT and the control groups

Group	n	Mean ± SD	Minimum	Maximum	p-value	Mean Difference (95% CI)
The ChatGPT group	35	63.79 ± 9.71	43.50	77.00	**< 0.001****	16.57 (11.87-21.27)
The Control group	29	47.22 ± 8.82	27.00	63.00		

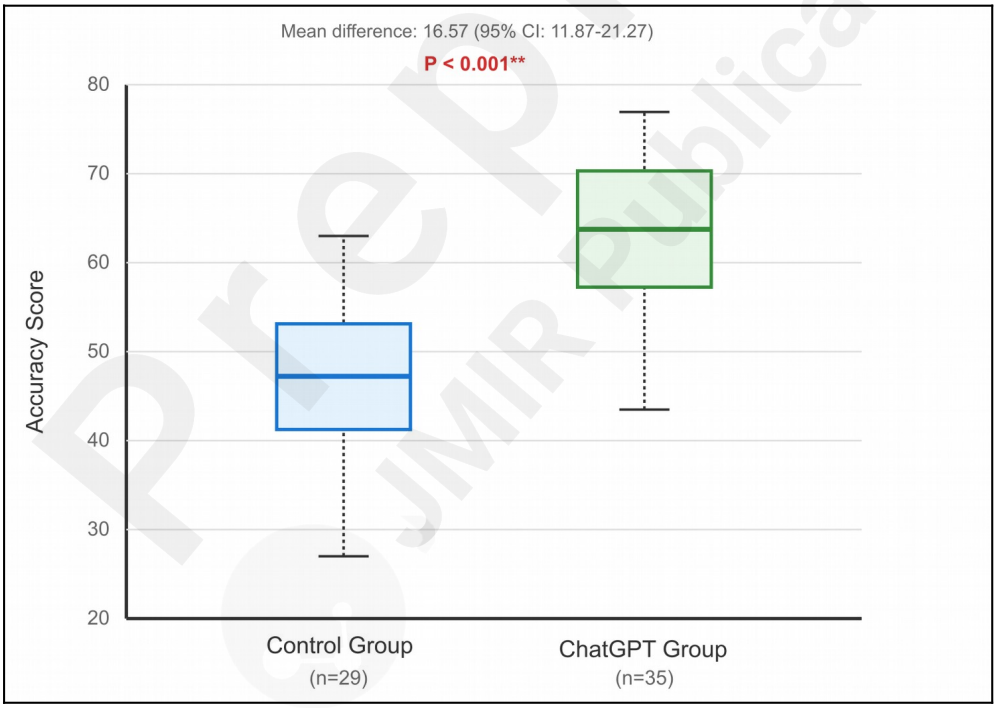


Figure 1. Box plot comparing accuracy scores between the ChatGPT and the control groups. The ChatGPT group (n=35) demonstrated significantly higher accuracy scores on the clinical knowledge test (mean 63.79, SD 9.71) compared to the control group using traditional online resources (n=29; mean 47.22, SD 8.82, P<.001). The plot displays medians, interquartile ranges, and individual data

points for each group.

Secondary Outcomes

Retrieval Time

The ChatGPT group completed the assessment significantly faster (mean 58.59 minutes, SD 19.57) than the control group (mean 75.76 minutes, SD 24.84, P=.003), saving an average of 17.17 minutes (95% CI: 5.98-28.36), representing a 22.7% reduction (Table 4, Figure 2).

Table 4. Comparison of retrieval time between the ChatGPT and the control groups

Group	n	Mean time ± SD (minutes)	Minimum	Maximum	p-value	Mean Difference (95% CI)
The ChatGPT group	34	58.59 ± 19.57	12.00	101.00	**0.003****	-17.17 (-28.36 to -5.98)
The Control group	29	75.76 ± 24.84	19.00	123.00		

*One participant's time data was missing due to technical error

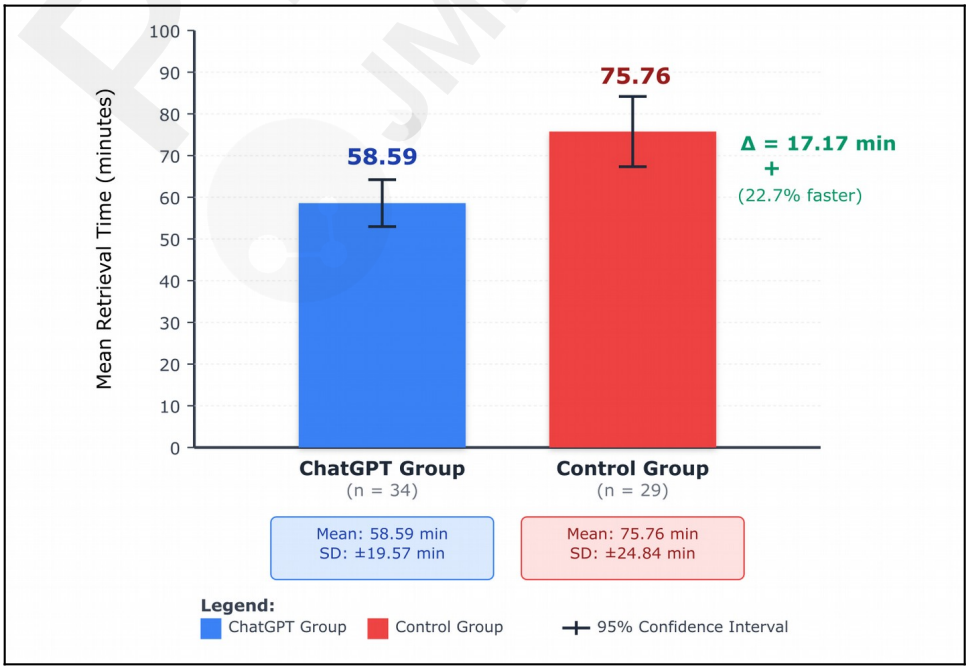


Figure 2. Mean retrieval time for medical information between the ChatGPT group and the control group. Participants in the ChatGPT group (n=34) completed the task in significantly less time (mean 58.59 minutes, SD 19.57) than those in the control group (n=29; mean 75.76 minutes, SD 24.84, P = .003), reflecting a 22.7% reduction. Error bars represent 95% confidence intervals.

User Satisfaction

ChatGPT users demonstrated notably higher overall satisfaction (mean 62.06, SD 7.14) compared to the control group (mean 56.72, SD 10.23, P=.018). The most pronounced difference was in satisfaction with the search process (mean 22.86, SD 1.80 vs mean 18.41, SD 4.07, P<.001) (Table 5).

Table 5. Comparison of satisfaction scores between groups

Satisfaction Domain	The ChatGPT Group Mean (SD)	The Control Group Mean (SD)	P value	Mean Difference (95% CI)
Satisfaction with search process	22.86 (1.80)	18.41 (4.07)	**<.001****	4.45 (2.91-5.99)
Answer quality	18.62 (3.53)	19.00 (4.23)	.697	-0.38 (-2.33 to 1.57)
Practical application	20.49 (3.04)	19.31 (3.66)	.165	1.18 (-0.49 to 2.85)
Overall satisfaction	62.06 (7.14)	56.72 (10.23)	**.018****	5.34 (0.94-9.74)

Correlation Analyses

Accuracy scores correlated positively with search satisfaction (r=0.434, P<.001) but not with retrieval time (r=0.030, P=.816) or GPA (r=-0.036, P=.781). Gender influenced retrieval time, with females taking longer than males (mean 71.95 vs 58.20 minutes, P=.014), (see Multimedia appendix 3).

Discussion

Principal Findings

Our findings demonstrated that ChatGPT outperformed traditional online medical resources in accuracy, speed, and user satisfaction for information retrieval among sixth-year medical students, with a 35.1% improvement in accuracy scores and a 22.7% reduction in retrieval time. These substantial gains highlight both the statistical and practical significance of ChatGPT’s advantages in real-world educational settings, particularly among non-English speaking medical students, and underscore their potential clinical implications for decision-making and professional development.

Through our study implementation, we found that ChatGPT's superior performance can be attributed to several factors. The natural language processing capabilities of ChatGPT allow for more intuitive query formulation. This differs markedly from the structured search syntax that traditional databases require. Additionally, we observed that ChatGPT can synthesize information and provide contextualized responses that reduce the cognitive load associated with evaluating multiple search results. This finding supports cognitive load theory. The theory suggests that reducing extraneous processing demands enhances learning efficiency [17].

The correlation between accuracy and search satisfaction (r=0.434) suggested that students could recognize effective information retrieval. However, the absence of correlation between GPA and performance indicates that information-seeking skills require specific training beyond traditional academic achievement.

Performance Across Cognitive Domains

An unexpected pattern emerged when examining performance through Bloom's taxonomy. The largest performance gap between groups appeared in application-level questions (20.63 percentage points) rather than basic remembering tasks (16.14 percentage points), suggesting ChatGPT's strength lies not merely in information retrieval but in helping students apply knowledge to practical contexts—a crucial skill for clinical practice. However, the uniformly poor performance on analysis-level questions across both groups (25.97% versus 17.44%) highlights a critical limitation: neither ChatGPT nor traditional resources adequately support higher-order analytical thinking. This finding has important implications for medical education, indicating that while AI tools excel at facilitating

knowledge acquisition and application, developing critical reasoning abilities still requires deliberate pedagogical strategies that actively engage students in analytical processes beyond what current information retrieval tools can provide.

Clinical and Educational Implications

Our findings indicate several important implications for medical education and clinical practice:

1. Curriculum Integration: Medical schools should consider incorporating AI literacy training, including appropriate use of ChatGPT and similar tools, while maintaining emphasis on critical thinking and verification skills. The superior performance in application-level questions but poor analytical performance across both groups highlights the need for educational strategies that develop higher-order thinking skills alongside AI tool usage.
2. Clinical Decision Support: The time efficiency gains (22.7% reduction) observed in our study could translate to meaningful benefits in clinical practice, where rapid access to accurate information directly impacts patient care quality and safety.
3. Resource Accessibility: ChatGPT's free availability addresses the access barrier faced by students unable to afford subscription-based resources like UpToDate, potentially democratizing access to medical information.
4. Assessment Methods: The lower performance on analysis-level questions emphasizes the need for developing critical evaluation skills alongside AI adoption.

Generalizability and International Context

While this study was conducted in Thailand, the findings likely have broader applicability. The use of ChatGPT in both Thai and English languages demonstrates its versatility across linguistic contexts. However, cultural factors affecting information-seeking behaviors and attitudes toward AI may vary internationally. Future multi-national studies should explore these variations to better understand global applicability.

Comparison with Prior Work

Our findings align with recent studies demonstrating ChatGPT's

capabilities in medical knowledge assessment [11,12] while extending the evidence to real-world educational applications. The 63.79% accuracy rate in our study is lower than ChatGPT's reported USMLE performance, likely reflecting the open-ended nature of our assessment compared to multiple-choice examinations.

Interestingly, our results differ from Caglar et al [14], who found UpToDate superior to ChatGPT in otorhinolaryngology. One possible explanation for this discrepancy is that our focus on general internal medicine topics and the predominance of remembering-level questions (67.7%) in our assessment, where ChatGPT's broad knowledge base provides advantages.

Limitations

While our study provides valuable insights, several limitations may affect the generalizability of our findings. First, the quasi-experimental design and single-institution setting may limit generalizability to other medical education contexts. Second, the assessment's focus on factual recall (67.7% remembering-level questions) may not fully capture complex clinical reasoning skills, potentially favoring ChatGPT's information retrieval strengths. Third, protocol deviations occurred in two ways: three participants were reclassified from control to ChatGPT group after using ChatGPT exclusively, and two control participants partially used ChatGPT (approximately 40% of responses) but remained in the control group for analysis. This contamination may have led to underestimation of the true effect size. Specifically, the contamination in the control group likely resulted in a conservative bias, potentially underestimating the true effect size of ChatGPT's superiority. The actual performance difference between pure ChatGPT use and traditional resources may be larger than our reported 35.1% improvement in accuracy. This contamination effect, while strengthening the validity of our positive findings, means that the true magnitude of ChatGPT's advantage in medical information retrieval may be even more substantial than demonstrated in this study. Fourth, we used ChatGPT's free version, and results may differ with paid versions or other AI models. Finally, the rapid evolution of AI technology means our findings represent a temporal snapshot that may quickly become outdated.

Future Research Directions

Future studies should investigate: (1) long-term retention of information retrieved using ChatGPT versus traditional sources, (2) integration strategies that combine AI tools with evidence-based databases, and (3) development of AI literacy curricula for medical education. (4) assessment methods that better evaluate higher-order thinking skills when using AI tools.

Conclusions

ChatGPT appeared to demonstrate superior performance compared to traditional online medical resources for medical information retrieval among sixth-year medical students, achieving higher accuracy scores, faster retrieval times, and greater user satisfaction with particular advantages in application-level questions, though limitations remain in supporting higher-order analytical thinking. These findings point to significant potential for ChatGPT as a supplementary tool in medical education. However, based on our experience, its integration must be thoughtfully implemented with appropriate guidelines, emphasis on critical evaluation skills, and recognition of current limitations. As AI technology continues to evolve, we will need to prepare future physicians who can effectively leverage these tools while maintaining the critical thinking and clinical reasoning skills essential for quality patient care.

Acknowledgments

We thank all participating medical students and the Clinical Medical Education Center staff at Maharaj Nakhon Si Thammarat Hospital for their support. We acknowledge the use of ChatGPT-4o mini (OpenAI) for grammar checking of the final manuscript only. No generative AI tools were used in the conception, data analysis, interpretation, or drafting of the manuscript.

Conflicts of Interest

None declared.

Funding

This research did not receive any funding. The researcher is solely responsible for all research expenses.

Data Availability

The deidentified dataset supporting the conclusions of this article is available from the corresponding author upon reasonable request, subject to institutional review board approval and completion of a data sharing agreement.

Authors' Contributions

According to CRediT taxonomy:

- Conceptualization: Isarapong Rattana
- Data curation: Isarapong Rattana
- Formal analysis: Yingyong Chinthammitr
- Investigation: Isarapong Rattana, Atthakorn Wutthimanop, Naphatsawan Phongsutham, Parissada Anugoolgarn
- Methodology: Isarapong Rattana, Yingyong Chinthammitr
- Project administration: Isarapong Rattana
- Resources: Isarapong Rattana
- Software: Yingyong Chinthammitr, Isarapong Rattana
- Supervision: Yingyong Chinthammitr
- Validation: Yingyong Chinthammitr
- Visualization: Isarapong Rattana, Yingyong Chinthammitr
- Writing - original draft: Isarapong Rattana
- Writing - review & editing: All authors

Abbreviations

AI: artificial intelligence EBM: evidence-based medicine GPA: grade point average IOC: Item-Objective Congruence IRB: institutional review board LLM: large language model SD: standard deviation SPSS: Statistical Package for the Social Sciences USMLE: United States Medical Licensing Examination

References

1. Loda T, Masters K, Zipfel S, Herrmann-Werner A. Can Medical Students Evaluate Medical Websites?: A mixed-methods study from Oman. Sultan Qaboos Univ Med J. 2022 Aug;22(3):362-369. doi:10.18295/squmj.8.2021.114 PMID:36160142
2. Densen P. Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc. 2011;122:48-58. PMID:21686208 PMCID:PMC3116346
3. Davies K. The information-seeking behaviour of doctors: a review of the evidence. Health Info Libr J. 2007 Jun;24(2):78-94.

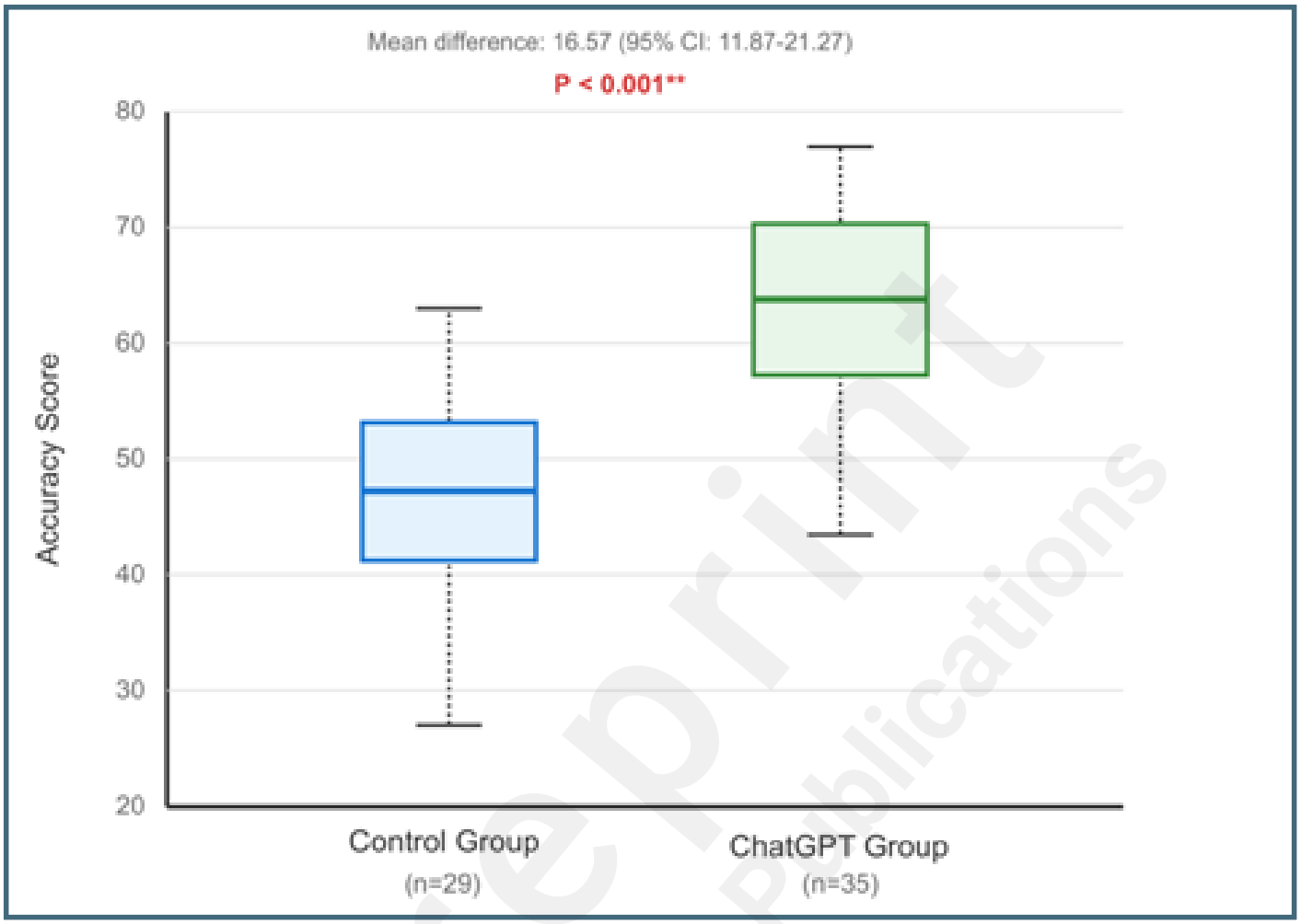
doi:10.1111/j.1471-1842.2007.00713.x PMID:17584211
4. Judd T, Kennedy G. Expediency-based practice? Medical students' reliance on Google and Wikipedia for biomedical inquiries. Br J Educ Technol. 2011;42(2):351-360. doi:10.1111/j.1467-8535.2009.01019.x
5. Daraz L, Morrow AS, Ponce OJ, Beuschel B, Farah MH, Katabi A, et al. Can patients trust online health information? A meta-narrative systematic review addressing the quality of health information on the internet. Am J Med Qual. 2018 Sep/Oct;33(5):487-492. doi:10.1177/1062860617751639 PMID:29345143
6. Thiele RH, Poiro NC, Scalzo DC, Nemergut EC. Speed, accuracy, and confidence in Google, Ovid, PubMed, and UpToDate: results of a randomised trial. Postgrad Med J. 2010 Aug;86(1018):459-465. doi:10.1136/pgmj.2010.098053 PMID:20709766
7. Alkhaaldi SMI, Kassab CH, Dimassi Z, et al. Medical Student Experiences and Perceptions of ChatGPT and Artificial Intelligence: Cross-Sectional Study. JMIR Med Educ. 2023;9:e51302. doi:10.2196/51302 PMID:38133911
8. Schäfer C, Schröder T, Beißbarth T. Information needs and information retrieval behaviour of medical students during the COVID-19 pandemic. BMC Med Educ. 2021 Aug 12;21(1):425. doi:10.1186/s12909-021-02839-w PMID:34384421
9. Fourtassi M, Abda N, Bentata Y, Hajjioui A. Medical education in Morocco: Current situation and future challenges. Med Teach. 2020 Sep;42(9):973-979. doi:10.1080/0142159X.2020.1779921 PMID:32608301
10. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. J Educ Eval Health Prof. 2024 Mar 15;21:6. doi:10.3352/jeehp.2024.21.6 PMID:38486402
11. 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 Feb 9;2(2):e0000198. doi:10.1371/journal.pdig.0000198 PMID:36812645
12. Thomae AV, Witt CM, Barth J. Integration of ChatGPT Into a Course for Medical Students: Explorative Study on Teaching Scenarios, Students' Perception, and Applications. JMIR Med Educ. 2024 Aug 22;10:e50545. doi:10.2196/50545 PMID:39177012

13. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on medical licensing examinations? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023 Feb 8;9:e45312. doi:10.2196/45312 PMID:36753318
14. Caglar OK, Allahverdiyev I, Agayarov OY, Demir D, Almuradova E. ChatGPT vs UpToDate: comparative study of usefulness and reliability of Chatbot in common clinical presentations of otorhinolaryngology-head and neck surgery. Eur Arch Otorhinolaryngol. 2024 Apr;281(4):2145-2151. doi:10.1007/s00405-023-08423-w PMID:38036900
15. Ayoub F, Swartz JI, Reyes JA, Francis PA, Das D. Evaluation of medical advice provided by ChatGPT. J Med Internet Res. 2023 Jul;25(7):e48328. doi:10.2196/48328 PMID:37462124
16. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
17. Sweller J, van Merrienboer JJG, Paas FGWC. Cognitive architecture and instructional design. Educ Psychol Rev. 1998;10:251-296. doi:10.1023/A:1022193728205
18. OpenAI. ChatGPT (GPT-4o mini) [Computer software]. 2024. URL: https://chat.openai.com (accessed 2025-02-15)

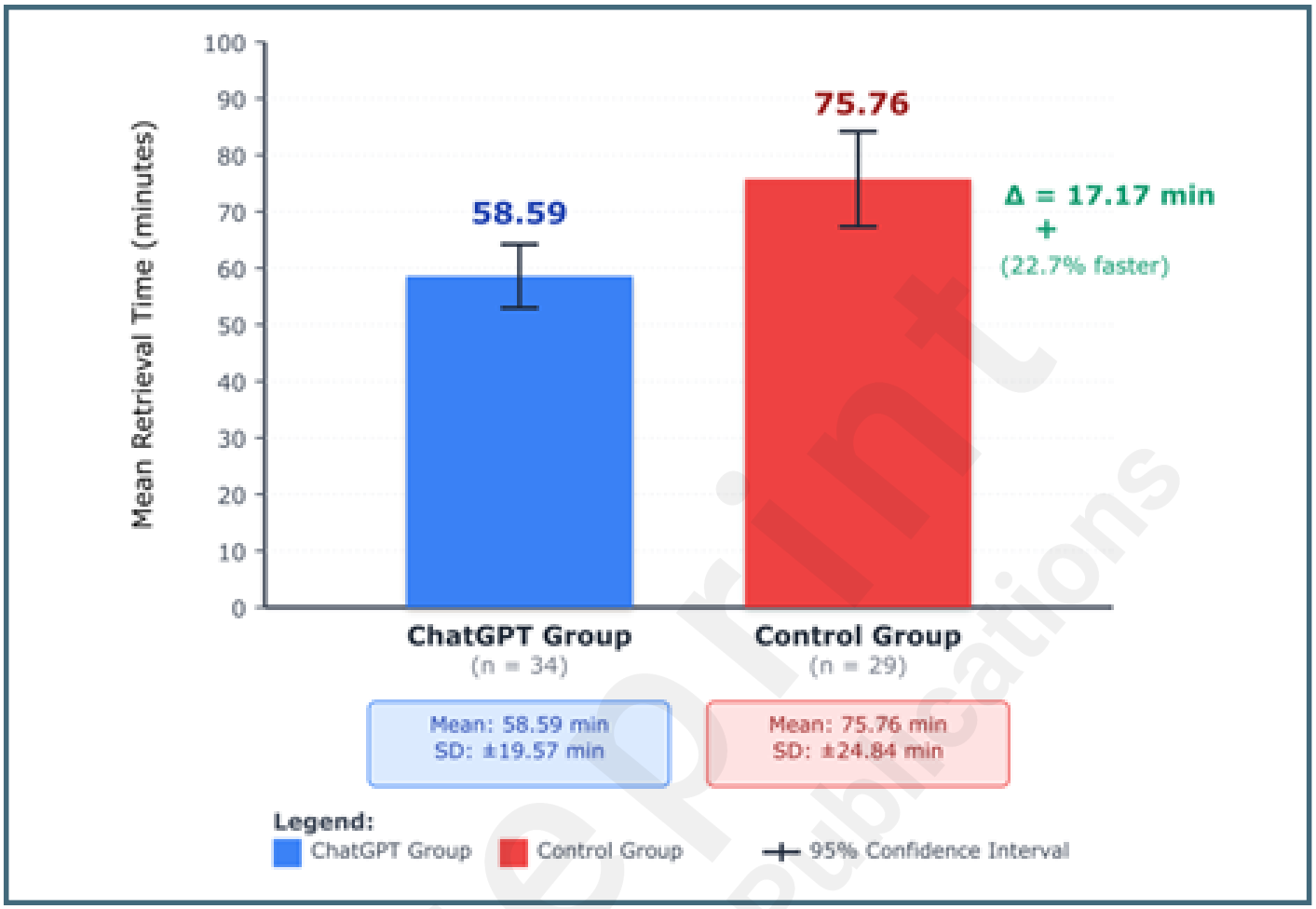
Supplementary Files

Figures

Box plot comparing accuracy scores between the ChatGPT and the control groups.



Mean retrieval time for medical information between the ChatGPT group and the control group.



Multimedia Appendixes

Questionnaires and test materials.
URL: http://asset.jmir.pub/assets/b281cdd1f123aadb918e42c3199fdbcb.docx

Scoring rubrics for clinical knowledge test questions.
URL: http://asset.jmir.pub/assets/be1c0de8c4cd222088ac59c8c61db768.docx

Comprehensive performance analysis.
URL: http://asset.jmir.pub/assets/9872e99e0f0aefba213a749575efc8ca.docx

Content validity assessment form for essay test items.
URL: http://asset.jmir.pub/assets/1db475a49305f2cf4b477c9fbcd95d59.docx

Image license and permission statement.
URL: http://asset.jmir.pub/assets/fb2c658fa28352125fa4336d344ff486.docx

TOC/Feature image for homepages

Comparative performance of ChatGPT versus traditional online medical resources (Google, PubMed, UpToDate) showing 35.1% higher accuracy and 22.7% faster retrieval time among 64 medical students.