İstanbul Med J 2026; 27(2): 97-102
DOI: 10.4274/imj.galenos.2026.04742

Performance Evaluation of Large Language Models in Emergency Medicine Specialty Examination Questions: A Cross-Sectional Study

Şebnem Zeynep Eke Kurt[1], Suphi Bahadırlı[2]

[1]University of Health Sciences Türkiye, Taksim Training and Research Hospital, Clinic of Emergency Medicine, İstanbul, Türkiye
[2]Medipol Mega University Hospital, Clinic of Emergency Medicine, İstanbul, Türkiye

ABSTRACT

Introduction: Large language models (LLMs) have recently shown strong potential in medical education, yet their performance compared with human learners in specialty-level examinations remains unclear. This study aimed to evaluate the performance of LLMs compared to human groups on a 50-question emergency medicine test from the Turkish Medical Specialty Examination.

Methods: A cross-sectional study was conducted at İstanbul Medipol University in 2024, involving 40 medical students and postgraduates and six LLMs (ChatGPT 4o, Claude Sonnet 3.5, Gemini Advanced, ChatGPT 4.0 Mini, Gemini Flash, Claude Haiku). Participants completed a 50-question test. Correct answers were analyzed using Welch's one-way analysis of variance (ANOVA), Levene's test for homogeneity of variances, and Games-Howell post-hoc tests.

Results: Claude Sonnet 3.5 achieved the highest mean correct answers (46.4±0.548), followed by ChatGPT 4o (44.6±1.14) and Gemini Advanced (43.6±1.67). Postgraduates with 5+ years of experience scored 43.5±3.03, while fifth-year medical students scored the lowest (29.1±3.73). Welch's ANOVA indicated significant group differences [$F(9, 20.8): 31.3$, $p<0.001$]. Post-hoc tests revealed LLMs outperformed most human groups, with Claude Sonnet 3.5 significantly surpassing Claude Haiku (mean difference: 9.6, $p=0.028$).

Conclusion: LLMs demonstrated superior performance compared to most human groups, indicating their potential as educational tools in emergency medicine.

Keywords: Large language models, emergency medicine, medical education

Introduction

Emergency medicine requires quick and precise decision-making and skills, both of which are rigorously evaluated by examinations such as the Turkish Medical Specialty Examination (TUS). The TUS assesses competency through complex, scenario-based questions that demand clinical reasoning and practical knowledge (1). Developments in artificial intelligence (AI), particularly large language models (LLMs), have introduced tools capable of addressing medical queries, thereby increasing interest in their performance on standardized tests (2). When trained on extensive datasets, LLMs produce responses that resemble expert knowledge, suggesting applications in medical education and clinical support (3,4). However, their ability to address specialty-specific, time-sensitive questions in emergency medicine remains largely unexamined (5).

Few studies have compared LLMs with human learners in specialty examinations. Studies on general medical licensing exams, such as the United States Medical Licensing Examination (USMLE), have shown that LLMs, such as ChatGPT, outperform medical students (6,7). Emergency medicine, however, presents unique challenges; it requires rapid synthesis of clinical information under pressure and tests both human and AI capabilities (8). The literature lacks comparisons among LLMs, medical students at different training levels, and experienced postgraduates in this field (9). This gap is significant, as LLMs can enhance training by offering accessible educational resources, but their effectiveness in high-pressure specialties requires validation (10).

Unlike prior studies primarily focused on USMLE-style examinations, this study evaluates LLM performance using the TUS, which reflects a different linguistic and curricular context. By directly comparing LLMs

Address for Correspondence: Şebnem Zeynep Eke Kurt, MD, University of Health Sciences Türkiye, Taksim Training and Research Hospital, Clinic of Emergency Medicine, İstanbul, Türkiye
E-mail: sebnemzeynepeke@hotmail.com **ORCID ID:** orcid.org/0000-0003-0778-8884

Cite this article as: Eke Kurt ŞZ, Bahadırlı S. Performance evaluation of large language models in emergency medicine specialty examination questions: a cross-sectional study. İstanbul Med J. 2026; 27(2): 97-102

Received: 24.11.2025
Accepted: 07.02.2026
Publication Date: 12.05.2026

with emergency medicine trainees and experienced specialists, this study provides specialty-specific evidence that extends existing AI examination literature.

This study aims to evaluate the performance of LLMs compared with that of medical students and postgraduates on a 50-question emergency medicine test from the TUS, using a cross-sectional design.

Methods

This cross-sectional study was approved by the İstanbul Medipol University Non-Interventional Clinical Research Ethics Committee (decision number: 664, date: 04.07.2024). The research was conducted at a Medipol Mega University Hospital Emergency Medicine Department in 2024. All participants provided informed consent prior to participation in the study. No patient data were collected or used in this research.

The study included 40 human participants: 10 fifth-year medical students, 10 sixth-year medical students, 10 first-year postgraduates, and 10 postgraduates with 5 or more years of experience in emergency medicine. Postgraduate participants consisted of emergency medicine residents (PG1) and board-certified emergency medicine specialists with at least five years of clinical experience (PG5+).

The inclusion criterion was enrollment in or graduation from a medical program, including postgraduates specializing in emergency medicine. No exclusion criteria were applied, as all eligible participants completed the test. Six LLMs "ChatGPT 4o, Claude Sonnet 3.5, Gemini Advanced, ChatGPT 4o Mini, Gemini Flash, and Claude Haiku" were also tested, each of which was evaluated five times to account for response variability. The participants were selected consecutively from the university's medical program and its emergency medicine department to ensure a representative sample.

Large Language Model Configuration and Prompting

The following LLMs were evaluated: ChatGPT-4o and ChatGPT-4o Mini (OpenAI; accessed May 2024), Claude Sonnet 3.5 and Claude Haiku (Anthropic; accessed June 2024), and Gemini Advanced and Gemini Flash (Google DeepMind; accessed June 2024).

All models were accessed through their official web-based interfaces using the default inference settings. Temperature, top-p, and related sampling parameters were not manually adjusted to reflect typical real-world user conditions and enhance reproducibility. A standardized prompt was used for all LLMs. Models were instructed to select the single best answer from the provided multiple-choice options without providing explanations or other commentary.

No chain-of-thought reasoning or clinical justification was explicitly requested. Prompt used for all LLMs: "You are answering a multiple-choice emergency medicine examination question. Select the single best answer (A, B, C, D, or E). Do not provide explanations or additional text."

The primary outcome was the number of correct answers on a 50-question multiple-choice test derived from past TUS examinations that focused on emergency medicine topics. Potential confounders, such as test familiarity or LLM version updates, were minimized by standardizing test conditions and using the latest model versions available in 2024. The test was validated by faculty experts for relevance and difficulty, ensuring content validity. The human participants completed the test under proctored conditions in a controlled environment.

While human participants completed the test under time-limited examination conditions, LLMs were not subject to time constraints.

The 50-item test was derived from TUS questions previously administered between 2018 and 2023. Items covered the core domains of emergency medicine, including trauma, toxicology, cardiology, neurology, infectious diseases, and critical care. Content validity was assessed by two senior emergency medicine faculty members who independently reviewed each item for relevance, clarity, and curriculum alignment. Disagreements were resolved by consensus. No items were modified from their original wording.

Each correct answer was scored as 1 point, with a maximum score of 50. Two researchers independently verified LLM responses, achieving high interrater reliability (kappa=0.95). No specific training was provided for data collection, but the process was standardized to reduce variability.

Selection bias was minimized by including all eligible participants consecutively, whereas information bias was reduced through standardized question presentation and scoring protocols. The variability in the LLM responses was addressed by conducting five test iterations per model and averaging the results. For LLMs, the unit of analysis was the model itself rather than individual runs. Five repeated runs were performed to estimate response variability; model-level mean scores and standard deviations (SDs) were used in all primary analyses.

The sample size was calculated to detect a five-point mean difference in the number of correct answers between groups, with 80% power and an alpha of 0.05, requiring at least 10 participants in each human group and five iterations per LLM, based on prior studies of medical examination performance. The number of correct answers was treated as a continuous variable and summarized via means and SDs.

Statistical Analysis

Statistical analyses included descriptive statistics to report means and SDs. Normality was confirmed via Shapiro-Wilk tests. Owing to nonhomogeneous variances [Levene's test, F (9, 60): 2.5, p=0.017], Welch's one-way analysis of variance (ANOVA) was used to assess differences in correct answers across groups, followed by Games–Howell post-hoc tests to identify specific group differences. As a sensitivity analysis, results were re-evaluated using model-level mean scores without treating individual LLM runs as independent observations. This approach yielded consistent group-level conclusions. No missing data were observed, so no imputation was needed. Analyses were performed using SPSS version 27. A p value <0.05 was considered significant. The primary outcome was the number of correct answers on the 50-question test.

Results

Participants

Forty human participants (mean age 28.5±4.2 years; 60% male) and six LLMs were included in the study. No exclusions occurred, and no data were missing.

Descriptive Data

Figure 1 shows the mean number of correct answers and the corresponding SDs. Table 1 presents the descriptive performance characteristics of all human groups and LLMs, including measures of central tendency and dispersion, to provide an overall comparison prior to inferential analyses. Claude Sonnet 3.5 scored highest (46.400±0.548), followed by ChatGPT 4o (44.600±1.140), and Gemini Advanced (43.600±1.670). Among human participants, PG5+ physicians scored 43.5±3.03, whereas fifth-year medical students scored the lowest (29.1±3.73).

Main Results

Welch's ANOVA revealed significant group differences [F (9, 20.8): 31.3, p<0.001)] (Table 2).

Games-Howell post-hoc tests (Table 3) indicated that Claude Sonnet 3.5 outperformed both Claude Haiku (mean difference: 9.6; p=0.028) and most human groups (e.g., MS5; mean difference: 17.3; p<0.001).

Table 1. Descriptive performance characteristics across study groups

Group	n (participants/runs)	Mean score	SD	Min–max
MS5	10 participants	29.1	3.73	23–35
MS6	10 participants	33.4	4.86	25–41
PG1	10 participants	38.4	4.21	31–45
PG5+	10 participants	43.5	3.03	38–48
ChatGPT-4o	5 runs	44.6	1.14	43–46
Claude Sonnet 3.5	5 runs	46.4	0.55	46–47
Gemini Advanced	5 runs	43.6	1.67	41–45
ChatGPT-4o Mini	5 runs	38.0	6.80	29–46
Gemini Flash	5 runs	40.3	3.94	35–45
Claude Haiku	5 runs	36.8	4.12	32–43

For human groups, n represents the number of participants. For LLMs, n represents the number of repeated runs performed to estimate response variability. LLMs: Large language models, SD: Standard deviation, MS5: Medical student-phase 5, MS6: Medical student-phase 6, PG1: Postgraduate one year, PG5+: Postgraduate 5+ years, Min-max: Minimum-maximum

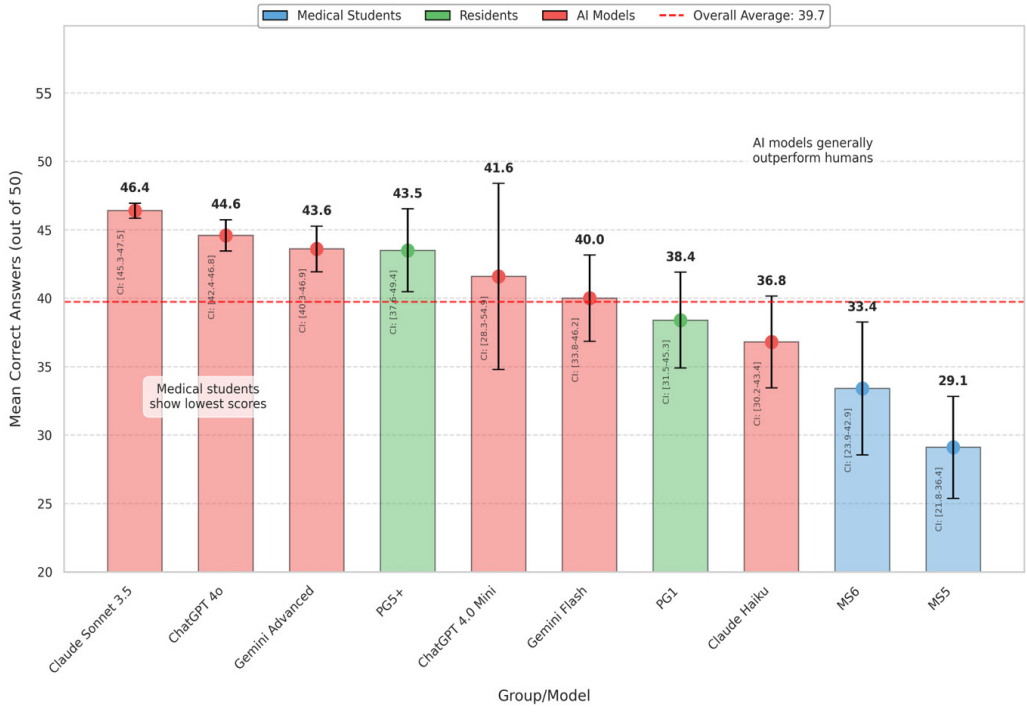

Figure 1. Demographics and analysis of responses to the test. AI: Artificial intelligence, MS5: Medical student-phase 5, MS6: Medical student-phase 6, PG1: Postgraduate one year, PG5+: Postgraduate 5+ years

The difference between Claude Sonnet 3.5 and fifth-year medical students was large (mean difference: 17.3 points; 95% confidence interval: 12.4–22.1; Cohen's d: 3.9). ChatGPT 4o and Gemini Advanced surpassed MS5 and MS6 ($p<0.001$). No significant differences were observed between PG5+ postgraduate physicians and top-performing LLMs (e.g., Claude Sonnet 3.5; mean difference: 2.9, $p=0.21$). While multiple pairwise comparisons were performed, the descriptive summary highlighted clear performance stratification across groups, which guided the interpretation of inferential results presented in Table 3.

Analysis of Performance Trends

The most proficient LLMs, Claude, Sonnet 3.5, ChatGPT 4o, and Gemini Advanced, exhibited noteworthy consistency, as indicated by their minimal SDs (0.548, 1.14, and 1.67, respectively). In contrast, ChatGPT 4o Mini exhibited the highest variability among all the groups (SD: 6.8), indicating potential inconsistencies in its performance. This higher variability may reflect architectural characteristics of lightweight models, which can be more sensitive to question phrasing and may lack domain-specific fine-tuning for emergency medicine content.

Among human groups, the performance gap between experience levels is notable. MS5 students, with the lowest mean score (29.1±3.73), were significantly outperformed by all LLMs and more experienced human groups ($p<0.001$ in most post-hoc comparisons).

This gap highlights the steep learning curve in emergency medicine, where foundational knowledge alone is insufficient without clinical exposure. MS6 students, after an additional year of training, improved to 33.4±4.86, but still lagged behind postgraduates and LLMs. The post-hoc analysis revealed that the difference between MS5 and MS6 was not statistically significant (mean difference: 4.3, $p=0.48$), suggesting that an additional year of medical school training may not substantially enhance performance on specialty-specific exams without targeted emergency medicine experience. Postgraduates with 5+ years of experience (PG5+) achieved a mean score of 43.5±3.03 and closely approached the performance of the top LLMs. The absence of a significant difference between PG5+ and Claude Sonnet 3.5 (mean difference: 2.9, $p=0.21$) or between PG5+ and ChatGPT 4o (mean difference: 1.1, $p=0.985$) underscores the competitive performance of experienced clinicians in this domain. However, their SD (3.03) indicates greater variability than that of the top LLMs, possibly reflecting differences in individual expertise or test-taking strategies among the participants.

Levene's test result [$F(9, 60)$: 2.5, $p=0.017$] indicates heterogeneity of variances across groups, consistent with the observed differences in SDs. For example, Claude Sonnet 3.5's exceptionally low SD (0.548) contrasts sharply with ChatGPT 4o Mini's high SD (6.8). This heterogeneity justified the use of Welch's ANOVA and Games–Howell post-hoc tests, which are robust to unequal variances.

The post-hoc analysis further reveals that the performance hierarchy among LLMs is not uniform. The significant outperformance of Claude Sonnet 3.5 relative to Claude Haiku (mean difference: 9.6, $p=0.028$) indicates substantial differences in capability even within the same family of models. Similarly, ChatGPT 4o outperforms Claude Haiku with a mean difference of 7.8 ($p=0.052$), although this result is slightly above the conventional significance threshold, indicating a trend rather than a statistically significant difference. A qualitative review suggested that both human participants and LLMs most frequently erred in toxicology and multi-step trauma questions, whereas cardiology and infectious disease questions were associated with higher accuracy across groups.

Table 2. ANOVA and Levene's test results

Test	F	df1	df2	p
One-way ANOVA (Welch's)	31.3	9	20.8	<0.001
Homogeneity of variances test (Levene's)		9	60	0.017

ANOVA: Analysis of variance

Discussion

This study demonstrated that LLMs, particularly Claude Sonnet 3.5, outperformed most human groups who took the 50-question emergency medicine test administered by the TUS. The top-performing LLMs reached scores comparable to or higher than those of postgraduates with more than five years of experience.

Table 3. Post-hoc analysis (Games-Howell)

Group	MS5	MS6	PG1	PG5+	C4o	CS3.5	GA	C4o-m	GF	CH
MS5	-	-4.3 (0.48)	-9.3 (<0.001)	-14.4 (<0.001)	-15.5 (<0.001)	-17.3 (<0.001)	-14.5 (<0.001)	-12.5 (0.122)	-10.9 (0.004)	-7.7 (0.051)
MS6		-	-5.0 (0.273)	-10.1 (0.002)	-11.2 (<0.001)	-13.0 (<0.001)	-10.2 (0.002)	-8.2 (0.442)	-6.6 (0.141)	-3.4 (0.832)
PG1			-	-5.1 (0.062)	-6.2 (0.007)	-8.0 (<0.001)	-5.2 (0.04)	-3.2 (0.98)	-1.6 (0.993)	1.6 (0.994)
PG5+				-	-1.1 (0.985)	-2.9 (0.21)	-0.1 (1.0)	1.9 (0.999)	3.5 (0.595)	6.7 (0.091)
C-4o					-	-1.8 (0.211)	1.0 (0.968)	3.0 (0.98)	4.6 (0.256)	7.8 (0.052)
CS3.5						-	2.8 (0.168)	4.8 (0.817)	6.4 (0.094)	9.6 (0.028)
GA							-	2.0 (0.999)	3.6 (0.509)	6.8 (0.086)
C4o m								-	1.6 (1.0)	4.8 (0.884)
GF									-	3.2 (0.839)
CH										-

The values represent the mean difference (p value). MS5: Medical student-phase 5, MS6: Medical student-phase 6, PG1: Postgraduate one year, PG5+: Postgraduate 5+ years, C-4o: ChatGPT 4o, CS3.5: Claude Sonnet 3.5, GA: Gemini Advanced, C4o m: ChatGPT 4o Mini, GF: Gemini Flash, CH: Claude Haiku

Emergency medicine plays a pivotal role in managing acute, life-threatening conditions, demanding rapid decision-making and extensive clinical knowledge (11). Examinations such as the TUS ensure practitioners are prepared for high-pressure scenarios, highlighting the importance of valid training programs (1,12). The global shortage of emergency physicians highlights the need for innovative solutions to support training, especially in areas with limited access to experienced instructors (13). LLMs offer a potential solution to these gaps by providing accurate answers to complex questions. The performance of these methods in this study suggests that they could improve preparation for standardized tests, such as TUS. The frequent physiological changes during acute medical events highlight the need for accurate and easily accessible knowledge, which LLMs seem to provide effectively (14).

This study found that Claude Sonnet 3.5 outperformed most human groups, which is consistent with recent literature on AI in medical education. Roos et al. (15) reported that LLMs outperformed medical students on MCAT-style questions, suggesting their strength in knowledge-based assessments. A meta-analysis of 32 studies by Waldock et al. (16) confirmed that LLMs were consistently superior in general medical examinations, although specialty-specific research remains limited. A recent study from Japan by Akitomo et al. (17) revealed similar AI performance in dental board exams, supporting the current findings. However, Johri et al. (18) noted that LLMs sometimes struggle with context-specific reasoning in clinical scenarios; this challenge was not evident in this study, likely owing to the structured nature of the test questions. These results suggest that LLMs perform well in standardized settings but may face challenges in less structured environments (19,20).

Although a large number of pairwise comparisons were conducted, the primary aim of this study was not to interpret each contrast in isolation but to identify overarching performance patterns across participant groups and model types.

When the results are examined collectively, three consistent trends emerge. First, a marked performance gap is observed between undergraduate medical students and both LLMs and experienced emergency physicians. Second, performance convergence is evident between senior emergency physicians (PG5+) and top-performing LLMs, suggesting comparable performance on examination-based assessments. Third, substantial variability among LLMs themselves highlights the importance of model architecture and design choices in determining examination performance.

Given their consistency and high accuracy, advanced LLMs have the potential to be reliable tools for emergency medicine training. However, the variability among models, such as ChatGPT 4.0 Mini, suggests that not all LLMs are equally suitable for such applications, necessitating careful model selection for educational purposes (21). For human learners, these data highlight the importance of clinical experience for improving performance, as evidenced by the progression from MS5 to PG5+ (22).

The findings are likely applicable to academic settings with standardized emergency medicine examinations, particularly in urban tertiary institutions. The strong performance of LLMs suggests that they could enhance training by providing scalable educational resources, especially where access to instructors is limited (23,24). Their epidemiological and clinical importance lies in improving training efficiency, which could help increase the supply of qualified emergency medicine specialists.

The results indicate the potential of LLMs as educational tools in emergency medicine, based on their test performance. Their integration into medicine could support exam preparation and knowledge reinforcement. Further studies are needed to validate their clinical utility and address variability in responses.

Study Limitations

The study has several limitations that should be taken into account. The single-center design may limit its applicability to other educational settings. The controlled test environment does not replicate clinical pressures, potentially overestimating LLM performance. The variability in LLM responses across iterations indicates potential inconsistencies, which could affect reliability.

Limited qualitative analysis of response patterns restricts deeper insight into LLM reasoning processes. The absence of time pressure for LLMs may partially lead to an overestimation of their comparative performance. Given the large number of pairwise comparisons, the interpretation focused on contrasts with the greatest educational and clinical relevance rather than an exhaustive discussion of all individual differences.

Conclusion

This study demonstrates that LLMs outperform most human groups and perform comparably to experienced postgraduates on emergency medicine exam questions. These findings suggest that LLMs could be valuable educational tools for enhancing medical training in critical specialties. Integrating LLMs into educational programs may improve exam preparation and knowledge acquisition, particularly in resource-limited settings. Future research should investigate their clinical applications, address response variability, and validate findings across diverse contexts to ensure their effective integration into medical education.

Ethics

Ethics Committee Approval: This cross-sectional study was approved by the appropriate İstanbul Medipol University Non-Interventional Clinical Research Ethics Committee (decision number: 664, date: 04.07.2024).

Informed Consent: All participants provided informed consent prior to participation in the study.

Footnotes

Authorship Contributions: Surgical and Medical Practices - S.B.; Concept - Ş.Z.E.K.; Design - S.B.; Data Collection or Processing - S.B.; Analysis or Interpretation - S.B.; Literature Search - Ş.Z.E.K.; Writing - Ş.Z.E.K.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Koçak M, Oğuz AK, Akçalı Z. The role of artificial intelligence in medical education: an evaluation of large language models (LLMs) on the Turkish Medical Specialty Training Entrance Exam. BMC Med Educ. 2025; 25: 609.
2. 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: e0000198.
3. Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. BMC Med Educ. 2025; 25: 129.
4. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023; 6: 120.
5. Preiksaitis C, Ashenburg N, Bunney G, Chu A, Kabeer R, Riley F, et al. The role of large language models in transforming emergency medicine: scoping review. JMIR Med Inform. 2024; 12: e53787.
6. Knoedler L, Alfertshofer M, Knoedler S, Hoch CC, Funk PF, Cotofana S, et al. Pure wisdom or potemkin villages? A comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE step 3 style questions: quantitative analysis. JMIR Med Educ. 2024; 10: e51148.
7. Morjaria L, Burns L, Bracken K, Ngo QN, Lee M, Levinson AJ, et al. Examining the threat of ChatGPT to the validity of short answer assessments in an undergraduate medical program. J Med Educ Curric Dev. 2023; 10: 23821205231204178.
8. Meral G, Ateş S, Günay S, Öztürk A, Kuşdoğan M. Comparative analysis of ChatGPT, Gemini and emergency medicine specialist in ESI triage assessment. Am J Emerg Med. 2024; 81: 146-50.
9. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ. 2024; 58: 1276-85.
10. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023; 620: 172-80.
11. Vella KM, Hall AK, van Merrienboer JJG, Hopman WM, Szulewski A. An exploratory investigation of the measurement of cognitive load on shift: application of cognitive load theory in emergency medicine. AEM Educ Train. 2021; 5: e10634.
12. Lombardi CV, Chidiac NT, Record BC, Laukka JJ. USMLE Step 1 and Step 2 CK as indicators of resident performance. BMC Med Educ. 2023; 23: 543.
13. Ayoola AS, Acker PC, Kalanzi J, Strehlow MC, Becker JU, Newberry JA. A qualitative study of an undergraduate online emergency medicine education program at a teaching Hospital in Kampala, Uganda. BMC Med Educ. 2022; 22: 84.
14. Williams CYK, Zack T, Miao BY, Sushil M, Wang M, Kornblith AE, et al. Use of a large language model to assess clinical acuity of adults in the emergency department. JAMA Netw Open. 2024; 7: e248895.
15. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Med Educ. 2023; 9: e46482.
16. Waldock WJ, Zhang J, Guni A, Nabeel A, Darzi A, Ashrafian H. The accuracy and capability of artificial intelligence solutions in health care examinations and certificates: systematic review and meta-analysis. J Med Internet Res. 2024; 26: e56532.
17. Akitomo T, Hamada M, Tsuge Y, Kaneki A, Ogawa M, Nishimura T, et al. Artificial intelligence's performance on the Japanese National Dental Examination. Cureus. 2024; 16: e73103.
18. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med. 2025; 31: 77-86.
19. Sonoda Y, Kurokawa R, Hagiwara A, Asari Y, Fukushima T, Kanzawa J, et al. Structured clinical reasoning prompt enhances LLM's diagnostic capabilities in diagnosis please quiz cases. Jpn J Radiol. 2025; 43: 586-92.
20. Watanabe T, Baba A, Fukuda T, Watanabe K, Woo J, Ojiri H. Role of visual information in multimodal large language model performance: an evaluation using the Japanese nuclear medicine board examination. Ann Nucl Med. 2025; 39: 217-24.
21. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024; 7: e2440969.
22. Davis D, Galbraith R; American College of Chest Physicians Health and Science Policy Committee. Continuing medical education effect on practice performance: effectiveness of continuing medical education: American College of Chest Physicians Evidence-Based Educational Guidelines. Chest. 2009; 135: 42S-8S.
23. Atsukawa N, Tatekawa H, Oura T, Matsushita S, Horiuchi D, Takita H, et al. Evaluation of radiology residents' reporting skills using large language models: an observational study. Jpn J Radiol. 2025; 43: 1204-12.
24. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ. 2023; 9:e48291.