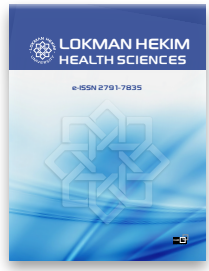
LOKMAN HEKIM HEALTH SCIENCES

DOI: 10.14744/lhhs.2024.94858
Lokman Hekim Health Sci 2026;6(2):00–00



lokmanhekimhs.com

ORIGINAL ARTICLE

The Use of Artificial Intelligence in Medical Education: A Comparative Analysis of Theoretical Exam Performance between ENT Residents and ChatGPT-4o

Tuba Doğan Karataş, Ahmet Aksoy, Adem Bora, Mansur Doğan

Department of Otolaryngology, Sivas Cumhuriyet University Faculty of Medicine, Sivas, Türkiye

Abstract

Introduction: This study assesses the theoretical examination performance of otorhinolaryngology residents and compares their results with those of ChatGPT-4o, an artificial intelligence (AI) language model.

Methods: A 100-item multiple-choice theoretical examination was administered in February 2025 to 17 otolaryngology residents enrolled in an otorhinolaryngology specialty training program. The Department of Otorhinolaryngology at a tertiary care university hospital administered the examination as part of its annual assessment program. The same questions were subsequently presented to ChatGPT-4o, a large language model developed by OpenAI, and its responses were systematically recorded. The numbers of correct answers provided by the residents and ChatGPT-4o were then compared. Each question was assigned a difficulty index based on participant performance and was thematically categorized to enable detailed item-level and domain-specific analyses.

Results: Seventeen otolaryngology residents completed the theoretical examination. The mean examination score among residents was 55.8 out of 100, whereas ChatGPT-4o achieved a score of 64. However, the difference was not statistically significant (p=0.077). Topic-based analysis revealed that ChatGPT-4o performed better on knowledge-based neurotology questions but performed worse on clinically contextual items requiring surgical decision-making. A positive, statistically significant correlation was observed between the duration of residency training and examination performance (r=0.66, p=0.004).

Discussion and Conclusion: ChatGPT-4o demonstrated a performance level comparable to that of human participants in theoretical medical examinations. AI-based educational platforms may serve as supportive tools in the training of medical residents and students.

Keywords: Artificial intelligence; Clinical competence; ChatGPT-4o; Medical education; Otorhinolaryngology

Residency education is a dynamic process aimed at developing residents' knowledge, skills, and attitudes. Artificial intelligence (AI) is playing an increasingly prominent role not only in healthcare services but also in medical education, and its integration offers potential benefits, such as enhancing clinical decision-making systems and supporting diagnostic and therapeutic processes.[1–4] The use of AI in education may help students

Cite this article as: *Doğan Karateş T, Aksoy A, Bora A, Doğan M. The Use of Artificial Intelligence in Medical Education: A Comparative Analysis of Theoretical Exam Performance between ENT Residents and ChatGPT-4o. Lokman Hekim Health Sci 2026;6(2):00–00.*

Correspondence: Tuba Doğan Karataş, M.D. Sivas Cumhuriyet Üniversitesi Tıp Fakültesi, Kulak Burun Boğaz Anabilim Dalı, Sivas, Türkiye
E-mail: tkaratas@cumhuriyet.edu.tr **Submitted:** 22.09.2025 **Revised:** 16.01.2026 **Accepted:** 23.02.2026 **Available Online:** 21.05.2026

enhance their theoretical knowledge and improve decision-making skills.[5,6] Generative AI systems, particularly large language models (LLMs), have emerged as important supportive components in clinical education due to their capacity to provide scenario-based responses.[7]

Clinical reasoning is shaped not only by knowledge but also by affective factors that influence diagnostic accuracy.[8] Moreover, it is a trainable skill; active and collaborative strategies, such as team-based learning, can enhance reasoning performance in medical students.[9]

AI systems have already demonstrated performance comparable to that of humans in high-stakes medical examinations, including physician-level multiple-choice tests[10] and national dental licensing assessments.[11]

Recent studies have shown that LLM-based systems can achieve performance levels comparable to those of humans in medical knowledge examinations and case discussions.[3,7] AI-based educational tools and simulation systems are increasingly used in medical schools.[12] Despite the growing presence of AI in medical education, its comparative performance alongside human trainees in specialty examinations remains underexplored.

Clinical decision support systems offer significant potential, particularly in evaluating patient scenarios and providing personalized feedback.[13] However, ethical and legal considerations should also be taken into account when integrating AI into medical education.[14]

In this context, the present study aims to evaluate the potential role of AI in medical education by comparing the theoretical examination performance of ENT residents with that of ChatGPT-4o, a large language model developed by OpenAI. This study hypothesizes that ChatGPT-4o will achieve performance comparable to that of ENT residents in theoretical examination tasks, particularly in knowledge-based domains.

Materials and Methods

This study was conducted following the examination held on February 12, 2025, at the Department of Otorhinolaryngology, Faculty of Medicine, Sivas University, and was carried out after obtaining approval from the Sivas University Clinical Research Ethics Committee on June 12, 2025 (Approval No: 2025-06/46). The study was conducted in accordance with the principles of the Declaration of Helsinki. Written informed consent was obtained from all participants prior to their inclusion in the study. The study enrolled 17 residents (9 male, 8 female) receiving residency training in otorhinolaryngology. The clinic administers a 100-item multiple-choice examination once a year in February, and the examination curriculum is disclosed 3 months in advance.

All multiple-choice questions were developed by the departmental examination committee in accordance with the institution's standard examination protocol and were peer-reviewed before the annual resident evaluation. Each question consisted of five answer options, with a single best correct answer. In a single session, ChatGPT-4o answered all 100 questions. Each question was presented independently, and no questions were repeated. The level of difficulty of the questions was determined based on the residents' responses.

The examination included questions on ENT diseases, as well as diagnostic and treatment approaches, to evaluate the residents' theoretical knowledge. The numbers of correct and incorrect responses provided by the residents were recorded as part of the study, and their examination performance was subsequently analyzed. In addition, all examination questions were answered once by ChatGPT-4o, a large language model developed by OpenAI; no repeated attempts were allowed. The results were then compared with the performance of the residents. Artificial intelligence–assisted tools were partially used to perform descriptive statistical analyses, calculate question difficulty indices, and assist with data visualization to enhance the clarity of the results.

Because the ethical approval and data collection process were based on the ChatGPT-4o model, which was the latest version available at the time, newer models such as ChatGPT-5.0 were not included in this study.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics for Windows, Version 26.0 (Armonk, NY: IBM Corp.). Descriptive statistics were calculated for the number of correct answers and the examination scores of the participants. The examination performance of otorhinolaryngology residents was compared with that of ChatGPT-4o using a one-sample testing framework. The mean examination score of the residents was compared with the fixed reference score obtained by ChatGPT-4o (64 points). The normality of the residents' score distribution was assessed using the Shapiro–Wilk test and by visually inspecting histograms. As the assumptions for parametric testing were marginally satisfied, a one-sample t-test was used. In addition, a nonparametric one-sample Sign test was performed as a robustness analysis.

The homogeneity of participants' responses to the same questions was evaluated using Cochran's Q test. Intra-group consistency among residents was calculated using Fleiss' kappa coefficient. The artificial intelligence model was not included as a rater in the kappa analysis and was evaluated separately. Intra-group heterogeneity was further examined by calculating question-specific variance distributions. The difficulty index of each question was determined based on the residents' mean scores, and the performance of ChatGPT-4o was evaluated across questions with varying levels of difficulty. These relationships were visualized using regression analysis to explore trends between question difficulty and model performance. Moreover, the relationship between the duration of residency training and examination performance was analyzed using Pearson correlation analysis. A p-value of <0.05 was considered statistically significant.

Results

Seventeen otolaryngology residents completed the theoretical examination. This study compared the performance of 17 residents (mean age=28.1±1.76 years) with that of ChatGPT-4o in a 100-item multiple-choice theoretical examination. The number of correct answers among residents ranged from 18 to 80, with a mean score of 55.8. ChatGPT-4o answered 64 questions correctly, achieving a total examination score of 64. Although ChatGPT-4o's score was higher than that of the residents, the difference was not statistically significant (p=0.077).

While a few residents outperformed ChatGPT-4o, the majority achieved lower scores (out of 100) (Table 1). Cochran's Q test demonstrated statistically significant differences in answer patterns among all participants, including ChatGPT-4o (Q=243.31, p<0.0001). These results indicate that responses to the examination questions were not homogeneous and that human participants and the artificial intelligence model exhibited different performance patterns across questions. The Fleiss' kappa coefficient, calculated to measure intra-group consistency, was 0.137, indicating weak agreement among residents in their responses to the same questions.

Questions were categorized into 11 thematic titles based on their content. When title-based success rates were evaluated, ChatGPT-4o demonstrated higher performance in the categories "CSF otorrhea," "Hearing Pathophysiology," "Anatomy of the Temporal Bone," and "Vestibular System." In contrast, ChatGPT-4o showed lower performance in areas requiring clinical insight, such as "Surgical Approach." Residents were more successful in certain categories, such as "Temporal Bone Fracture" (Fig. 1; p<0.05). In question-based analyses, the difficulty index of each question was calculated based on the proportion of correct answers provided by residents. Comparison of question difficulty indices demonstrated a statistically significant difference between questions answered correctly and incorrectly by ChatGPT-4o. Questions answered correctly by ChatGPT-4o had significantly higher resident mean scores compared to those answered incorrectly (0.620 vs. 0.468, respectively; Mann–Whitney U test, U=677.5, p<0.001). These findings indicate that ChatGPT-4o more frequently provided correct responses to questions that were relatively more difficult for residents. The distribution of question difficulty according to ChatGPT-4o responses is illustrated in Figure 2.

When the intra-group variance distribution was examined, residents' responses to some questions were highly homogeneous, whereas substantial differences were observed for others. These findings suggest a need for greater standardization of educational materials. For most questions, the variance value was close to 0.25, indicating moderate variability among residents' responses. These data demonstrate that most examination questions

Table 1. Examination Performance of ENT Residents and ChatGPT-4o

Participant	Number of correct answers	Number of incorrect answers	Mean±SD p
Resident 1	71	29	
Resident 2	80	20	
Resident 3	80	20	
Resident 4	57	43	
Resident 5	60	40	
Resident 6	61	39	
Resident 7	80	20	
Resident 8	73	27	
Resident 9	38	62	55.82±16.97
Resident 10	43	57	**p=0.103**
Resident 11	45	55	
Resident 12	18	82	
Resident 13	51	49	
Resident 14	64	36	
Resident 15	33	67	
Resident 16	48	52	
Resident 17	47	53	
ChatGPT-4o	**64**	**36**	

SD: Standard deviation; *: The exam consists of 100 multiple-choice questions.

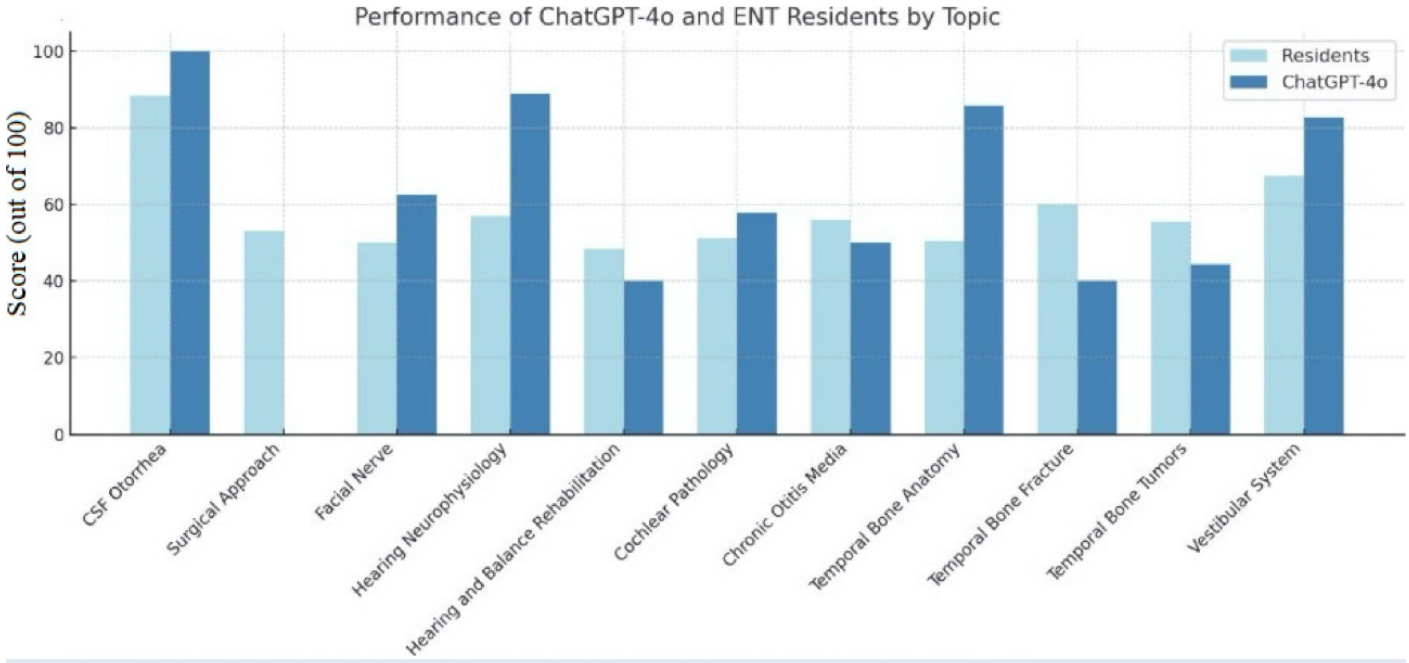


Figure 1. Distribution of examination scores of ENT residents and ChatGPT-4o by topic.

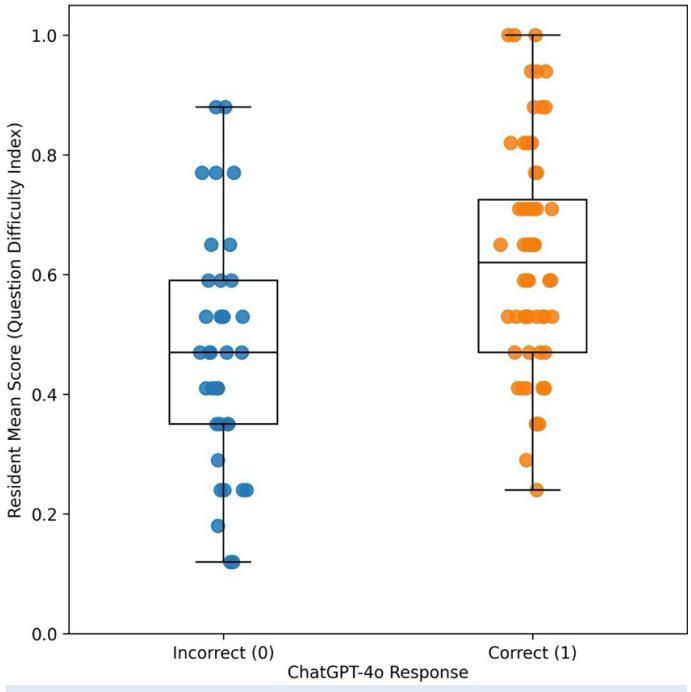


Figure 2. Distribution of resident mean scores by ChatGPT-4o response (0/1). Boxplots indicate median and IQR; each dot is one question (n=100). Mann-Whitney U test.

effectively differentiated among residents' responses and therefore highlight the strength of the examination as an assessment tool in medical education (Fig. 3).

When the relationship between residency year and examination performance was evaluated, a positive correlation was observed. Specifically, the number of

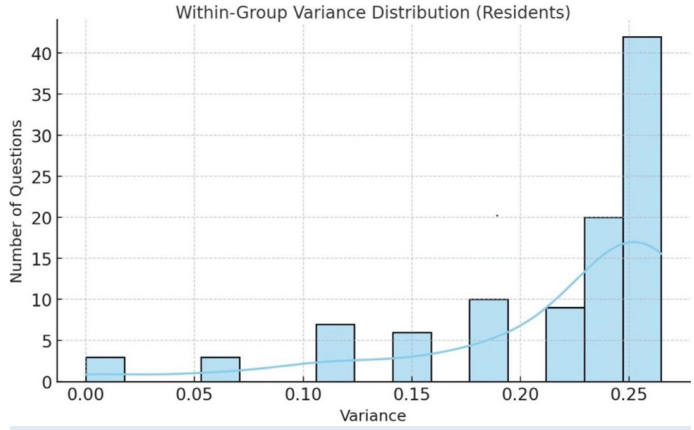


Figure 3. Intra-group variance distribution.

correct answers tended to increase with longer duration of residency training (Fig. 4; Pearson r=0.66, p=0.004).

Discussion

Our study did not find a statistically significant difference between the mean examination performance of ChatGPT-4o and that of ENT residents. These results suggest that artificial intelligence-based models may serve as potential supplementary tools in medical education.[1,2,4] Moreover, large language models may demonstrate performance levels comparable to those of human participants in certain examinations.[3,7] The literature indicates that the integration of artificial intelligence into medical education may improve students' clinical decision-making skills and support educational processes.[5,6,12,13] When integrated

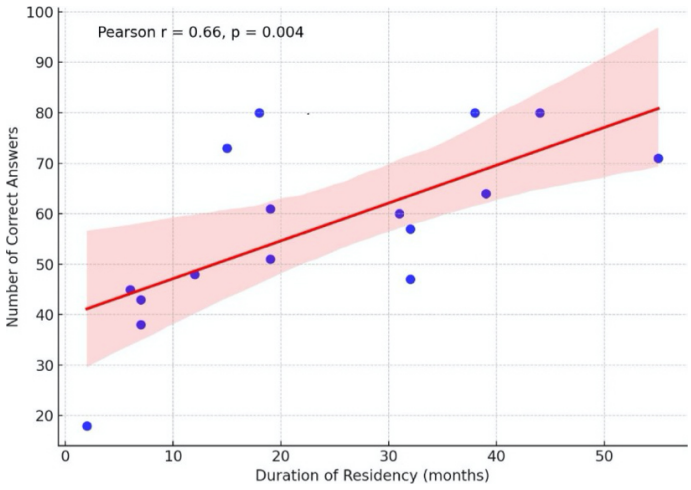


Figure 4. Relation between residency year of ENT residents and examination success (Pearson correlation coefficient (r)=0.66, p=0.004).

into medical education, artificial intelligence can enhance students' clinical decision-making skills, support learning processes, and improve educational efficiency by providing personalized feedback.[5,6,12,13] However, ethical and legal considerations should be taken into account during this integration.[5,6,12,13]

Artificial intelligence-based simulation systems and virtual patients can increase the efficiency of medical education by offering students realistic patient management scenarios.[14,15] Moreover, artificial intelligence-based educational tools can help identify topics in which students experience difficulty and may therefore enable the development of targeted educational programs.[16] However, further studies are needed to determine how artificial intelligence can be used in medical education within appropriate ethical and legal frameworks.[14]

ChatGPT-4o demonstrated higher performance than residents in knowledge-based items, particularly "Cerebrospinal Fluid Otorrhea," "Hearing Pathophysiology," "Anatomy of the Temporal Bone," and "Vestibular System." These findings indicate that artificial intelligence models have a strong capacity to process theoretical knowledge and may outperform humans in selected domains.[3,17]

On the other hand, ChatGPT-4o showed limited success in items requiring clinical decision-making, such as "Surgical Approach." These results suggest that clinical insight, contextual evaluation, and patient-specific variables remain challenging for artificial intelligence models.[18,19] The relatively low performance of both groups in practical areas such as "hearing and balance rehabilitation" and "chronic otitis media" indicates that improvements may be needed not only in education but also in the structure of the examination. The lower performance of residents in the early years of residency training also reflects the effect of the level of supervision. Furthermore, the low performance observed in complex neurotological topics suggests that both human participants and artificial intelligence models may have difficulty addressing rare and anatomically complex subjects.[20]

The results of our study demonstrated that artificial intelligence-based models are comparable to, or may even outperform, residents in theoretical knowledge-based examinations. However, the model's success was limited for some questions requiring clinical and contextual decision-making. These results are consistent with the findings of Tangsrivimol et al.,[20] who reported that, despite ChatGPT's strong capacity to provide information, it may be insufficient in areas such as clinical insight and individualized patient context. Moreover, these authors stated that artificial intelligence can be used as a supplementary tool in medical education, but reliance on incorrect answers may pose serious risks. In this context, large language models such as ChatGPT may be considered educational tools when used in a controlled manner, but they should not be used alone as clinical decision-making tools.

To the best of our knowledge, this study represents one of the earliest investigations to directly compare the theoretical examination performance of otorhinolaryngology residents with that of ChatGPT-4o, thereby underscoring the potential contribution of large language models to postgraduate medical education.

The findings of this study indicated that ChatGPT-4o achieved examination scores comparable to those of residents, particularly in domains involving knowledge-based and systematic questions. Lee et al.[17] reported that ChatGPT achieved an examination score of 59.4 in the occupational therapy graduate examination, which was comparable to human performance; these findings are consistent with our results. However, that study emphasized that the model performed poorly on questions requiring clinical decision-making and contextual understanding. Similarly, our study found that ChatGPT demonstrated significantly lower performance in categories such as "Surgical Approach" and "Patient Management." This consistency supports the view that current artificial intelligence models remain limited in terms of clinical experience, contextual insight, and situational understanding.[17]

Yanagita et al.[21] evaluated the performance of ChatGPT in the National Medical Licensing Examination in Japan and reported that the model achieved a score of 72 out of 100

on theoretical knowledge-based multiple-choice questions. This finding is comparable to the score of 64 achieved in our study and supports the view that large language models can demonstrate consistent performance in knowledge-based examinations. However, the same study emphasized that ChatGPT has limited reliability in complex clinical scenarios. These results are consistent with our finding that the model had a low success rate in areas requiring clinical decision-making and contextual interpretation, such as "Surgical Approach." In this context, it can be speculated that ChatGPT may be an effective tool for assessing theoretical knowledge, but it may not reach a sufficient level in examination components that require clinical practice.

A positive correlation was found between residency year and theoretical examination performance (r=0.66, p=0.004). These findings indicate that knowledge gradually improves over time, contributing to examination success. However, one notable observation was that the highest examination scores were achieved by one resident in the 15th month of training and another in the 18th month. These findings suggest that personal factors, such as individual learning pace, prior knowledge, work discipline, and cognitive skills, may play a decisive role in examination performance. Various studies have shown that year of training can be an important factor in examination success and that educational methods and individual learning strategies may also have a substantial influence on performance.[22] Moreover, some studies suggest that training year and experience have a strong effect on examination success.[23,24] This indicates that statistically significant relationships may not apply uniformly to all individuals and that educational evaluation should focus not only on duration but also on the quality of training and individual characteristics. Therefore, the use of qualitative assessment tools in addition to quantitative measures may be recommended when evaluating residents' performance.

Overall, these findings suggest that ChatGPT-4o may serve as a supervised supplementary tool to reinforce theoretical learning and identify knowledge gaps. However, given the risk of incorrect outputs, its use should remain under educator or clinician oversight and should not replace human clinical reasoning. Therefore, in the interest of patient safety and in view of potential medicolegal liability, its outputs should not be used as the sole basis for clinical decision-making.

Limitations

This study has several limitations. First, it was conducted at a single center with a relatively small sample size, which may limit the generalizability of the findings. Second, only the ChatGPT-4o model was evaluated, as the study design and ethics approval were based on this version. Therefore, the results reflect the model's performance during the specific period of data collection. Finally, newer artificial intelligence models, such as ChatGPT-5.0, were not included in the analysis, and their potential effects on reasoning and knowledge performance remain to be explored. Future studies including updated artificial intelligence models would be valuable for evaluating advances in performance and improving the reliability and contemporaneity of the findings.

Conclusion

ChatGPT-4o demonstrated performance on a theoretical multiple-choice examination comparable to that of otorhinolaryngology residents, particularly on knowledge-based questions, while showing limitations in domains requiring clinical decision-making and contextual reasoning. Given the single-center design and small sample size, these findings should be interpreted with caution. Under appropriate supervision, large language models may serve as supplementary tools to reinforce theoretical learning and support formative assessment; however, they should not be used as stand-alone instruments for clinical decision-making. Larger multicenter studies incorporating diverse question formats and updated artificial intelligence models are needed to define safe and effective integration strategies for artificial intelligence-assisted medical education.

Ethics Committee Approval: This study was approved by the Sivas University Clinical Research Ethics Committee (Date: 12.06.2025, Decision no: 2025-06/46). Registration on ClinicalTrials.gov was not performed for this study.

Informed Consent: Written informed consent was obtained from all participants prior to their inclusion in the study.

Conflict of Interest: None declared.

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

Use of AI for Writing Assistance: A statement disclosing the partial use of artificial intelligence–assisted technologies (for statistical analysis) has been added to the manuscript.

Authorship Contributions: Concept: TDK, AA, AB, MD; Design: TDK, AA, AB, MD; Supervision: TDK; Resource: TDK; Materials: TDK; Data collection and/or processing: TDK, AB; Analysis and/or interpretation: TDK, AA; Literature review: TDK; Writing: TDK, AA; Critical review: AA, MD.

Peer-review: Double blind peer-reviewed.

References

1. Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. JMIR Med Educ 2019;5:e16048. [Crossref]
2. Pucchio A, Eisenhauer EA, Moraes FY. Medical students need artificial intelligence and machine learning training. Nat Biotechnol 2021;39(3)388-9. [Crossref]
3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2(2):e0000198. [Crossref]
4. Joseph G, Bhatti N, Mittal R, Bhatti A. Current application and future prospects of artificial intelligence in healthcare and medical education: a review of literature. Cureus. 2025;17:e77313. [Crossref]
5. Arslan K. Artificial intelligence and applications in education. Western Anatolia Journal of Educational Sciences, 2020;11:71-88. [Article in Turkish]
6. Özer M. Potential benefits and risks of artificial intelligence in education. Bartın University Journal of Faculty of Education 2024;13(2):232-44. [Crossref]
7. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Influence of a large language model on diagnostic reasoning: a randomized clinical vignette study. medRxiv [Preprint]. 2024:2024.03.12.24303785. [Crossref]
8. Merkebu J, Y Soh M, Loncharich M, Hawks MK, Costello JA, Shapiro M, et al. Emotions and clinical reasoning in medical education and clinical practice: a scoping review. Acad Med 2025;100(11):e80-e90. [Crossref]
9. Ishizuka K, Shikino K, Takada N, Sakai Y, Ototake Y, Kobayashi T, et al. Enhancing clinical reasoning skills in medical students through team-based learning: a mixed-methods study. BMC Med Educ 2025;25(1):221. [Crossref]
10. Law AK, So J, Lui CT, Choi YF, Cheung KH, Kei-Ching Hung K, et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. BMC Med Educ 2025;25(1):208. [Crossref]
11. Chan-Chia Lin C, Sun JS, Chang CH, Chang YH, Zwei-Chieng Chang J. Performance of artificial intelligence chatbots in National dental licensing examination. J Dent Sci 2025;204):2307-14. [Crossref]
12. Rincón EHH, Jimenez D, Aguilar LAC, Flórez JMP, Tapia ÁER, Peñuela CLJ. Mapping the use of artificial intelligence in medical education: a scoping review. BMC Med Educ 2025;25(1):526. [Crossref]
13. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2(4):230-43. [Crossref]
14. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018;93(8):1107-9. [Crossref]
15. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2:719-31. [Crossref]
16. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: integrative review. JMIR Med Educ 2019;5(1):e13930. [Crossref]
17. Luo M, Duan Z, Gao J, Sun Y, Chen L, Feng X. Evaluating the role of ChatGPT in rehabilitation medicine: a narrative review. Front Digit Health 2025;7:1618510. [Crossref]
18. Lee SA, Heo S, Park JH. Performance of ChatGPT on the National Korean Occupational Therapy Licensing Examination. Digit Health 2024;10:20552076241236635. [Crossref]
19. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the united states medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312. [Crossref]
20. Tangsrivimol JA, Darzidehkalami E, Virk HUH, Wang Z, Egger J, Wang M, et al. Benefits, limits, and risks of ChatGPT in medicine. Front Artif Intell 2025;8:1518049. [Crossref]
21. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the National Medical Licensing Examination in Japan: Evaluation study. JMIR Form Res 2023;7:e48023. [Crossref]
22. Heltne SF, Hovdenakk S, Kvernenes M, Tenstad O. Study preferences and exam outcomes in medical education: insights from renal physiology. BMC Med Educ 2024;24(1):973. [Crossref]
23. Shaban L, O'Flynn E, Mulwafu W, Borgstein E, Bekele A, Bachheta N, et al. Factors ınfluencing exam performance of surgical trainees in Sub-Saharan Africa: A retrospective analysis of the college of surgeons in East, Central, and Southern Africa membership examination. J Surg Educ 2024;81(3):404-11. [Crossref]
24. Brooks NE, French JC, Sancheti H, Lipman JM. American board of surgery ın-training exam performance predicted by question bank use while unassociated with other learning strategies. J Surg Res 2024;300:191-7. [Crossref]