***ORIGINAL ARTICLE***

# Performance of Large Language Models on Iran's Medical Informatics Graduate Entrance Exams

Zeinab Ghaffari, Masoumeh Khedri, Ali Mohammad Hadianfard*

## ABSTRACT

*Large language models (LLMs) with advanced natural language processing are increasingly used in medical education. This study evaluated and compared the accuracy of four LLMs (GPT-4O, O3-mini, Gemini, and Copilot) in answering questions from Iran's master's and doctoral entrance exams in medical informatics. Multiple-choice questions from the 2024 exams, 116 for master's and 96 for doctoral programs, were submitted to the free versions of the models using uniform prompts. Responses were compared with the official answer key to*

Received 31/01/2026
Accepted for publication 13/05/2026
Published 21/06/2026

* *Correspondence to*: *Ali Mohammad Hadianfard, Department of Health Information Technology, School of Allied Medical Sciences, Ahvaz Jundishapur University of Medical Sciences, Golastan Blvd., Ahvaz, Iran Postal code: 61357-15794 Tel: +98 61 3311 2551 Email: dr.ali.hadianfard@gmail.com*

**About the authors**:

**Zeinab Ghaffari**; PhD student in Medical Informatics, Department of Medical Informatics, School of Allied Medical Sciences, Urmia University of Medical Sciences, Urmia, Iran.

**Masoumeh Khedri**; PhD student in E-Learning in Medical Sciences, Department of E-Learning in Medical Sciences, Smart University of Medical Sciences, Tehran, Iran.

Department of Health Information Technology, School of Allied Medical Sciences, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran.

**Ali Mohammad Hadianfard**; BSc, MSc, PhD in Medical Informatics, Associate Professor, Department of Health Information Technology, School of Allied Medical Sciences, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran.

*measure accuracy, uncertainty, and error rates. Statistical analyses included Chi-Square tests and logistic regression. At the master's level, O3-mini performed best, while Copilot was weakest, though the differences were not significant (p = 0.2088). At the doctoral level, GPT-4O and O3-mini outperformed Gemini, which had 79.44% accuracy and a 21.55% error rate, with significant differences (p < 0.001). Model performance varied across specialized subjects. These results indicate that LLM performance depends on model type, educational level, and the nature of the content, providing a foundation for more accurate assessments and targeted AI applications in specialized education.*

**Keywords**: Educational Measurement, Large Language Models, Generative Artificial Intelligence, Medical Informatics, Distance Education

## INTRODUCTION

Large Language Models (LLMs) are trained on vast amounts of data. They can process and understand natural language to perform a wide range of tasks, including text generation, question answering, logical reasoning, machine translation, summarization, and multimodal support (1,2).

Some chatbots that are based on large language models include ChatGPT, Gemini, and Copilot. The first version of the ChatGPT chatbot was released by the company OpenAI on 30 November 2022 (3). The company subsequently released newer models with more advanced capabilities, including GPT-4o (13 May 2024), O1, O1-mini, and O3-mini (31 January 2025) (4). O3-mini offers exceptional STEM capabilities—with strong performance in science, mathematics, and programming—while maintaining the low cost and reduced latency of O1-mini (5). Microsoft also introduced its first model, called Bing, which was powered by the GPT-4 artificial intelligence model (OpenAI). This version was known as an intelligent chatbot called Bing Chat. Eventually, the company introduced its new model, Copilot, on 16 March 2023 (6). In addition, Google released its chatbot, Bard, in March 2023 (7). On 6 December 2023, Sundar Pichai, the CEO of Google and Alphabet, introduced Gemini, the most capable model ever developed by the company (8). These chatbots are used in various fields, including healthcare, education, and web development. The opportunities offered by large language models in medical education and academic research include improved access to information, curriculum development, educational methodologies, study plans, personalized educational materials, assessment and evaluation, assistance with medical writing, medical research, and literature review, and program monitoring and evaluation. However, the use of LLMs in medical education also raises concerns such as plagiarism, misinformation, overreliance on artificial intelligence, reduced critical thinking, algorithmic bias, privacy concerns, and copyright-related issues (9,10).

Given the opportunities and concerns mentioned, numerous studies have been conducted to evaluate the performance of chatbots powered by large language models. In one study, ChatGPT's performance was assessed on the United States Medical Licensing Examination (USMLE), and the results showed that ChatGPT's accuracy is improving. Given its ability to generate novel insights, this artificial intelligence can be helpful for education in medical training environments (11). The performance of GPT-4o and GPT-4

on the Arabic GAT exam demonstrates the high potential of these chatbots to support students in Arabic education and their applicability to non-English languages. However, they are not yet reliable as independent preparation tools for students (12). Large language models (GPT-4, Claude 3 Oplus, Gemini 1.0 Pro) demonstrated strong performance across all subjects in the Thailand Medical Licensing Examination and show high potential in medical education, particularly in accurately interpreting and responding to a wide range of exam questions (13). Microsoft Copilot, ChatGPT-4, and Google Gemini performed well on the Italian Healthcare Entrance Exam, and their narrative coherence was generally logical. However, it is generally recommended to use this new technology cautiously, as a supplement to learning rather than as a primary source (14). In the continuing education exam of the American Academy of Periodontology (AAP), ChatGPT-4o demonstrated strong capability in answering AAP questions in terms of accuracy and reliability, while Google Gemini and ChatGPT-3.5 showed weaker performance. These findings emphasize the potential of large language models as educational tools in periodontics and oral implantology. However, the limitations of these models, such as their inability to process image-based requests effectively and their tendency to generate inconsistent responses to identical prompts, should be taken into account (15). A study showed that computer science students, especially experienced students (96% use LLMs), use large language models to generate ideas and quickly access information, but they have concerns about accuracy (80%) and the impact on academic integrity (16). Large language models, when combined with human supervision and continuous refinement, can serve as virtual learning assistants, significantly improving accessibility and student engagement in computer science courses (17). A comprehensive review of the applications of large language models in smart education (LLM4Edu) showed that these models have significant potential to transform education, including providing personalized learning, adaptive feedback, access to diverse resources, and the creation of educational content, ultimately enhancing the quality and effectiveness of teaching and learning (18). Given the growing role of large language models in medical education and academic research, one area that could benefit in particular from this technology is medical informatics.

Medical informatics is continuously evolving and is influenced by technological advances, changes in healthcare delivery, and the increasing volume of healthcare data. This field encompasses a wide range of topics, including health data management, clinical informatics, telemedicine, bioinformatics, and more. The primary goal of medical informatics is to improve patient care, enhance the efficiency of healthcare systems, and facilitate medical research through the effective use of information technology and data science (19).

In Iran, the first steps toward academic training in medical informatics were taken in the late 1990s. The first official master's and doctoral programs in medical informatics in Iran began in 2008 and 2009 (20). The aforementioned exams are held annually in Iran for candidates of the master's and doctoral entrance exams in medical informatics. Given the nature of this field, professors and students can benefit from the chatbots mentioned for educational purposes. Therefore, evaluating the accuracy and reliability of these chatbots' responses is of significant importance.

**Objectives**

This study was conducted to evaluate the accuracy and reliability of responses provided by large language models to multiple-choice questions from the master's and doctoral entrance exams in the field of medical informatics in Iran.

## METHODS

This comparative study examined the performance of large-language-model-based platforms, including GPT-4o, O3-mini, Gemini (Flash 2.0), and Copilot. The multiple-choice questions used in this study were extracted from the master's and doctoral entrance exams in the field of medical informatics in Iran, which were held in the year 1403 (2024–2025). The master's entrance exam in medical informatics consisted of 160 questions, covering various areas including programming principles and data structures, biostatistics and mathematics, health information management, health informatics, medical terminology and concepts, and general language. The doctoral entrance exam in medical informatics consisted of 130 questions, covering areas such as medical informatics, health information management and technology, and both specialized and general language. This study focused on evaluating the ability of large language models to answer questions in the specialized field of medical informatics; therefore, questions related to the general language course were not included. Questions containing images were excluded from the total due to the models' limitations in image analysis. Ultimately, 116 questions from the master's entrance exam were included in the study, and out of 130 questions from the doctoral entrance exam, 100 remained. Additionally, 4 questions were removed by the doctoral exam organizing body (Sanjesh Organization), resulting in a final total of 96 questions included in the study.

The questions, without any modifications and categorized according to the domains specified in the master's and doctoral entrance exams in medical informatics, were presented to GPT-4o, O3-mini, Gemini (Flash 2.0), and Copilot on 2 March 2025, using a single prompt in Persian: "Select the correct answer from the options A, B, C, D."

In addition, because access to advanced versions of these chatbots was limited, the free versions were used. To ensure uniform conditions, each model's default settings were applied, and responses were recorded without intervention. The responses generated by the large language models were evaluated based on answer accuracy, the rate of uncertain responses, and the error rate.

To evaluate the models, their responses were compared with the official answer key published by the exam-organizing body (Sanjesh Organization). For each model, the percentages of correct, incorrect, and uncertain responses were calculated both overall and separately for each specialized subject area.

Statistical analyses were conducted using the Spyder environment and the Python programming language. Given the nominal and binary nature of the data (correct/incorrect), the Chi-Square test was employed to compare differences in accuracy proportions among the various Large Language Models (LLMs). In instances where the Chi-Square test yielded significant results, post-hoc pairwise comparisons with Bonferroni correction were performed to identify specific differences between models while controlling for Type I error. Furthermore, to evaluate the simultaneous impact of 'LLM type' and 'subject matter' on the probability of error, a Logistic Regression model was utilized. Model fit indices, including Pseudo R-squared and regression coefficients (β),

were reported to assess the predictive power and the strength of associations within the model. Since the analyses were conducted using non-parametric tests (Chi-square) and logistic regression models, the assumption of normality was not required and therefore was not assessed.

**Ethical Statement**

As this study did not involve human participants or human data, compliance with research ethics guidelines and ethics committee approval were not required.

## RESULTS

At the master's level, O3-mini demonstrated the best performance, with an accuracy of 80.17%, while Copilot showed the lowest, at 69.83%. Overall, a small proportion of indeterminate responses was observed for both O3-mini and Copilot (approximately 0.86%). The error rate for Copilot was higher (29.31%) than that of the other models, whereas O3-mini exhibited the lowest error rate (17.97%) among all models (Table 1).

**TABLE I.** OVERALL PERFORMANCE OF LARGE LANGUAGE MODELS IN THE MEDICAL INFORMATICS MASTER'S ENTRANCE EXAMINATION

| *Metric (%)* | *GPT-4o* | *O3-mini* | *Gemini* | *Copilot* |
|---|---|---|---|---|
| *Accuracy* | 71.55 | 80.17 | 77.59 | 69.83 |
| *uncertain* | 0.00 | 0.86 | 0.00 | 0.86 |
| *Error* | 28.45 | 17.97 | 22.41 | 29.31 |

Subject-wise analysis at the master's level revealed that the O3-mini model achieved absolute superiority in 'Mathematics and Biostatistics' (94.74%) and 'Medical Terminology' (95%) (Table 3). However, the Chi-Square statistical test indicated that the overall difference in accuracy among the four models at this level was not statistically significant ($\chi^2 = 4.55$, $p = 0.207$).

At the doctoral level, both GPT-4o and O3-mini ranked first with an accuracy of 80.21%, whereas the Gemini model demonstrated the weakest performance at 44.79%. Correspondingly, the error rate for Gemini was significantly higher (55.21%) than that of the other models (Table 2).

**TABLE III.** OVERALL PERFORMANCE OF LARGE LANGUAGE MODELS IN THE MEDICAL INFORMATICS DOCTORAL ENTRANCE EXAMINATION

| *Metric (%)* | *GPT-4o* | *O3-mini* | *Gemini* | *Copilot* |
|---|---|---|---|---|
| *Accuracy* | 80.21 | 80.21 | 79.44 | 78.12 |
| *uncertain* | 0.00 | 0.00 | 0.00 | 0.00 |
| *Error* | 19.79 | 19.79 | 55.21 | 21.88 |

At the doctoral level, the Chi-Square test indicated a highly significant difference in performance among the models ($\chi$^2 = 42.15, $p < 0.001$). Regarding specialized doctoral subjects, the highest performance in 'Medical Informatics' was achieved by O3-mini (80%), while GPT-4o demonstrated the superior accuracy in 'Health Information Management and Technology' (84.78%).

At the master's level, variations in response accuracy were observed across different subjects; O3-mini outperformed other models in 'Programming and Data Structures' (64.71%), 'Mathematics and Biostatistics' (94.74%), and 'Health Informatics' (86.97%).

Furthermore, in 'Health Information Management' and 'Medical Terminology', O3-mini exhibited performance comparable to that of the Gemini model (Table 3).

**TABLE IIIII.** PERFORMANCE (ACCURACY) OF LARGE LANGUAGE MODELS ACROSS MASTER 'S-LEVEL SPECIALIZED COURSES IN MEDICAL INFORMATICS

| *Course/Subject Area (%)* | *GPT-4o* | *O3-mini* | *Gemini* | *Copilot* |
|---|---|---|---|---|
| *Programming and Data Structures* | 52.94 | 64.71 | 52.94 | 52.94 |
| *Mathematics and Biostatistics* | 60.00 | 94.74 | 75.00 | 55.00 |
| *Health Information Management* | 90.00 | 83.33 | 83.33 | 80.00 |
| *Health Informatics* | 65.52 | 86.97 | 75.86 | 72.41 |
| *Medical Terminology* | 80.00 | 95.00 | 95.00 | 80.00 |

At the doctoral level, the best performance in Medical Informatics was observed for O3-mini (80%), while Gemini exhibited the lowest performance (18%) (Table 4). In Health Information Management, GPT-4o achieved the highest accuracy (84.78%), whereas Gemini showed the lowest performance (73.91%).

**TABLE IVV.** PERFORMANCE (ACCURACY) OF LARGE LANGUAGE MODELS ACROSS DOCTORAL-LEVEL SPECIALIZED COURSES IN MEDICAL INFORMATICS

| *Course/Subject Area (%)* | *GPT-4o* | *O3-mini* | *Gemini* | *Copilot* |
|---|---|---|---|---|
| *Medical Informatics* | 76.00 | 80.00 | 18.00 | 74.00 |
| *Health Information Management & Technology* | 84.78 | 80.43 | 73.91 | 82.61 |

The Chi-square test indicated that the differences in model performance at the master's level were not statistically significant ($p = 0.207$); however, at the doctoral level, the differences were highly significant ($\chi^2 = 42.15$, $p < 0.001$). Given the significant Chi-square result at the doctoral level, post hoc pairwise comparisons with Bonferroni correction were conducted. The results demonstrated that the performance of the Gemini model was significantly poorer than that of GPT-4o ($p < 0.001$), O3-mini ($p < 0.001$), and Copilot ($p < 0.001$). However, no statistically significant differences were observed among GPT-4o, O3-mini, and Copilot in pairwise comparisons ($p > 0.05$).

At the master's level (with model fit of Pseudo $R^2 = 0.059$), the "type of specialized course" emerged as the primary predictor of error. Specifically, the courses "Health Information Management" ($\beta = -1.42$, 95% CI: [−2.11, −0.72]) and "Medical Terminology" ($\beta = -1.69$, 95% CI: [−2.52, −0.87]) were associated with negative and highly significant coefficients ($p < 0.001$), substantially reducing the odds of error occurrence, indicating greater model proficiency in these domains.

In contrast, at the doctoral level (with model fit of Pseudo $R^2 = 0.124$), the "type of language model" played a determining role, with the Gemini model identified as the primary factor increasing the likelihood of error, as evidenced by a positive and significant coefficient ($\beta = 1.57$, 95% CI: [0.92, 2.22], $p < 0.001$). Additionally, at this level, the course "Health Information Management and Technology" ($\beta = -1.04$, $p < 0.001$) was associated with fewer errors compared to "Medical Informatics.

## DISCUSSION

This study compared the performance of four large language models, GPT-4o, O3-mini, Gemini (Flash 2.0), and Copilot, in answering multiple-choice questions from the master's and doctoral entrance exams in Medical Informatics in Iran in 1403 (2024-2025).

Although previous studies have evaluated the performance of large language models (LLMs) on standardized tests across various countries and languages (13–15, 21–30), this study's focus on Iranian graduate entrance exams in Persian represents a distinctive contribution.

The findings of this study indicated that, overall, large language models demonstrated relatively high accuracy in predicting correct answers. These results are consistent with a study that reported notable performance by ChatGPT on medical licensing exams, confirming its potential as an assistive tool for answering medical questions (31). Furthermore, another study reported that high internal consistency and low internal contradiction in ChatGPT reflect sound clinical reasoning and high-quality explanations, supporting the model's relative capability in medical education (11).

A more detailed analysis revealed significant differences among the models. O3-mini demonstrated highly satisfactory performance at both the master's and doctoral levels. At the doctoral level, the performance of this model was comparable to GPT-4o. In contrast, the weakest performance at the master's level was observed for Copilot, while at the doctoral level, it was observed for Gemini. This pattern aligns with previous studies reporting the superiority of advanced GPT-family models (15, 21) and suggests that these models may have applicability in medical domains under human supervision.

A more detailed examination of chatbot performance revealed that the accuracy of large language models varied across different subjects, including Mathematics, Statistics, and Medical Informatics. In other words, there were significant differences in response quality across courses. These findings are consistent with previous studies (23, 24) that report variations in accuracy across topics. Such fluctuations in accuracy may be attributed to differences in the volume and diversity of training data available for each scientific domain, highlighting the importance of optimizing models for specialized content.

The results indicated that, at the master's level, no significant differences were observed in the performance of large language models. However, at the doctoral level, significant differences emerged, primarily due to Gemini's poor performance. These findings are consistent with previous studies in the field, which have shown that the performance of language models on complex, analytical tests is directly influenced by the type of exam, question complexity, cognitive level, and subject domain (13, 15, 21, 23)

Furthermore, regression analysis in this study indicated that the error rate is simultaneously influenced by both the type of model and the subject area, underscoring the importance of the specialized context in assessing model capabilities. Therefore, it is recommended that the evaluation of large language model performance incorporate rigorous statistical methods alongside qualitative assessments of responses by human experts to gauge the models' capabilities across different cognitive levels accurately.

Differences in the accuracy of large language models can be attributed to variations in neural network architectures, the quality, diversity, and volume of training data, as well as the techniques employed in model design and training, all of which play a significant role in optimizing the models' final performance. However, due to the black-box nature

of large language models, it is not possible to determine the definitive reasons for these performance differences.

This study faced several limitations, including:

a) The exclusion of questions containing images due to the inability of language models to process visual content;

b) The use of free versions of the models due to the lack of unrestricted access to the advanced versions of the studied platforms; and

c) The use of questions from a single year's entrance exam because of the unavailability of official answer keys for previous years, as well as variations in question difficulty and design style across different years.

d) The unbalanced distribution of questions across different subject areas—inherent in the official budgeting of the National Organization for Educational Testing

This resulted in small cell counts within certain subgroup analyses, potentially reducing the statistical power required to detect subtle differences between the models.

## CONCLUSION

This study assessed the performance of four advanced large language models in medical education. Results showed that error rates depend not only on the model type but also significantly on educational level and question content. These findings align with global trends and highlight less-explored factors, such as educational level and specialized courses, that influence model performance. While many international studies focus on model performance in specific, often English-language exams, this research emphasizes that educational level and subject context were also key to effectiveness. Consistent with other research, combining statistical and qualitative approaches and designing context-aware assessments can improve the use of large language models in higher and specialized education. Although the results were promising in terms of accuracy, there is room for further improvement. Differences in the comprehensiveness of each model also underscore the need for ongoing AI development to enhance reliability and trust. Lastly, as AI continues to evolve rapidly, newer, more optimized models are constantly being created and released. Consequently, the performance of the models evaluated here may be limited compared to future versions.

### Declaration of the Use of Artificial Intelligence Tools

This study evaluated several artificial intelligence models; however, no generative AI tools were used in the writing or editing of the manuscript. All analyses were conducted by the authors using statistical software.

### Contributorship Statement

Conceptualization and study design: Z.G., A.H., Data collection, analysis, and interpretation: Z.G., M.K., Drafting of the manuscript: Z.G., M.K., Critical review of the manuscript for important intellectual content: A.H., Statistical analysis: Z.G., M.K., Supervision of the study: A.H., All authors reviewed, commented on, and approved the final manuscript, as well as taking responsibility for its content.

**Funding Statement**

This research did not receive any specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**Declaration Of Conflicting Interests**

The authors declare no conflicts of interest regarding the research, authorship, and publication of this article.

**Data Availability Statements**

No human data was used for the research described in the article. All the analyzed data are presented in the article in tables.

**REFERENCES**

1. Shao M, Basit A, Karri R, Shafique M. Survey of Different Large Language Model Architectures: Trends, Benchmarks, and Challenges. IEEE Access. 2024;12:188664-706. https://doi.org/10.1109/ACCESS.2024.3482107
2. Yang J, Jin H, Tang R, Han X, Feng Q, Jiang H, et al. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. ACM Trans Knowl Discov Data. 2024;18(6):Article 160. https://doi.org/10.1145/3649506
3. OpenAI. Introducing ChatGPT: OpenAI; 2022 [cited 2/26/2025]. Available from: https://openai.com/index/chatgpt/
4. OpenAI. Pioneering research on the path to AGI [cited 4/6/2025]. Available from: https://openai.com/research/
5. OpenAI. OpenAI o3-mini 2025 [cited 4/6/2025]. Available from: https://openai.com/index/openai-o3-mini/
6. Microsoft. Introducing Microsoft 365 Copilot – your copilot for work 2023 [cited 2/26/2025]. Available from: https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/
7. D. Hassabis SP. Introducing Gemini: our largest and most capable AI model 2023 [cited 4/6/2025]. Available from: https://blog.google/technology/ai/google-gemini-ai/#sundar-note
8. S P. Introducing Gemini: our largest and most capable AI model 2023 [cited 4/6/2025]. Available from: https://blog.google/technology/ai/google-gemini-ai/#sundar-note
9. 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. https://doi.org/https://doi.org/10.2196/48291
10. Benítez TM, Xu Y, Boudreau JD, Kow AWC, Bello F, Van Phuoc L, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. Journal of the American Medical Informatics Association. 2024;31(3):776-783. https://doi.org/10.1093/jamia/ocad252
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 Digital Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
12. Alahmadi MD, Alharbi, M., Tayeb, A., & Alshangiti, M. Evaluating Large Language Models' Proficiency in Answering Arabic GAT Exam Questions. Engineering, Technology & Applied Science Research. 2024;14(6):17774-80. https://doi.org/10.48084/etasr.8481
13. Saowaprut P, Wabina RS, Yang J, Siriwat L. Evaluation of Large Language Models in Thailand's National Medical Licensing Examination. medRxiv. 2024:2024.12. 20.24319441. https://doi.org/10.1101/2024.12.20.24319441
14. Rossettini G, Rodeghiero L, Corradi F, Cook C, Pillastrini P, Turolla A, et al. Comparative accuracy of ChatGPT-4, Microsoft Copilot, and Google Gemini in the Italian entrance test for healthcare sciences degrees: a cross-sectional study. BMC Medical Education. 2024;24(1):694. https://doi.org/10.1186/s12909-024-05630-9

15. Sabri H, Saleh MH, Hazrati P, Merchant K, Misch J, Kumar PS, et al. Performance of three artificial intelligence (AI)-based large language models in standardized testing; implications for AI-assisted dental education. Journal of periodontal research. 2025;60(2):121-33. https://doi.org/10.1111/jre.13323
16. Weber JL, Martinez Neda B, Carbajal Juarez K, Wong-Ma J, Gago-Masague S, Ziv H, editors. Measuring cs student attitudes toward large language models. Proceedings of the 55th ACM Technical Symposium on Computer Science Education V 2; 2024. https://doi.org/10.1145/3626253.3635604
17. Liu M, M'hiri F, editors. Beyond traditional teaching: Large language models as simulated teaching assistants in computer science. Proceedings of the 55th ACM Technical Symposium on Computer Science Education V 1; 2024. https://doi.org/10.1145/3626252.3630789
18. Gan W, Qi Z, Wu J, Lin JC-W, editors. Large language models in education: Vision and opportunities. 2023 IEEE international conference on big data (BigData); 2023: IEEE. https://doi.org/10.1109/BigData59044.2023.10386291
19. Gomathy C, Dhanush M, Shyam BSK. A Study on Medical Informatics. International Journal of Scientific Research in Engineering and Management. 2023;07(11). http://doi.org/10.55041/IJSREM
20. Sarafi Nejad A, Fatehi F. Medical Informatics in Iran and the Emergence of Clinical Informatics. Iranian Journal of Medical Sciences. 2022;47(6):503-4. https://doi.org/10.30476/ijms.2022.48773
21. Colluoglu B, Dikici S. Transforming neurosurgical practice with large language models: comparative performance of ChatGPT-omni and Gemini in complex case management. World Neurosurgery. 2025:124103. https://doi.org/10.23736/s0390-5616.25.06447-1
22. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? Patterns. 2024;5(3). https://doi.org/10.48550/arXiv.2207.08143
23. Giannos P, Delardas O. Performance of ChatGPT on UK standardized admission tests: insights from the BMAT, TMUA, LNAT, and TSA examinations. JMIR Medical Education. 2023;9(1): e47737. https://doi.org/10.2196/47737
24. Guigue PA, Meyer R, Thivolle-Lioux G, Brezinov Y, Levin G. Performance of ChatGPT in French language Parcours d'Accès Spécifique Santé test and in OBGYN. International Journal of Gynecology & Obstetrics. 2024;164(3):959-63. https://doi.org/10.1002/ijgo.15083
25. Saowaprut P, Rodis Wabina RS, Yang J, Siriwat L. Evaluation of Large Language Models in Thailand's National Medical Licensing Examination. medRxiv. 2024:2024.12. 20.24319441. https://doi.org/10.1101/2024.12.20.24319441
26. Sharma P, Thapa K, Thapa D, Dhakal P, Upadhaya MD, Adhikari S, et al. Performance of ChatGPT on USMLE: Unlocking the potential of large language models for AI-assisted medical education. arXiv preprint arXiv:230700112. 2023. https://doi.org/10.1371/journal.pdig.0000198
27. 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 digital health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
28. Laupichler MC, Rother JF, Kadow ICG, Ahmadi S, Raupach T. Large language models in medical education: comparing ChatGPT-to human-generated exam questions. Academic Medicine. 2024;99(5):508-12. https://doi.org/10.1097/acm.0000000000005626
29. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. Correction: 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. 2024;10: e57594. https://doi.org/10.2196/45312
30. Beaulieu-Jones BR, Berrigan MT, Shah S, Marwaha JS, Lai S-L, Brat GA. Evaluating capabilities of large language models: performance of GPT-4 on surgical knowledge assessments. Surgery. 2024;175(4):936-42. https://doi.org/10.1016/j.surg.2023.12.014
31. Ebrahimian M, Behnam B, Ghayebi N, Sobhrakhshankhah E. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. BMJ Health & Care Informatics. 2023;30(1):e100815. https://doi.org/10.1136/bmjhci-2023-100815