Cureus
Part of SPRINGER NATURE

Open Access Original Article

ChatGPT-4 Turbo and Meta's LLaMA 3.1: A Relative Analysis of Answering Radiology Text-Based Questions

Review began 11/11/2024
Review ended 11/20/2024
Published 11/24/2024

DOI: 10.7759/cureus.74359

Mohammed Abdul Sami [1], Mohammed Abdul Samad [2], Keyur Parekh [1], Pokhraj P. Suthar [1]

1. Department of Diagnostic Radiology and Nuclear Medicine, Rush University Medical Center, Chicago, USA 2. Department of Osteopathic Medicine, Des Moines University College of Osteopathic Medicine, West Des Moines, USA

Corresponding author: Keyur Parekh, keyur_m_parekh@rush.edu

Abstract

Aims and objectives: This study aimed to compare the accuracy of two AI models - OpenAI's GPT-4 Turbo (San Francisco, CA) and Meta's LLaMA 3.1 (Menlo Park, CA) - when answering a standardized set of pediatric radiology questions. The primary objective was to evaluate the overall accuracy of each model, while the secondary objective was to assess their performance within subsections.

Methods and materials: A total of 79 text-based pediatric radiology questions were selected out of 302 total questions for this comparison. The questions covered seven subsections, including musculoskeletal, chest, and neuroradiology, among others. Image-based questions were excluded to focus on text interpretation and to minimize the sampling bias within each model. Each model was tested independently on the same question set, and the percent accuracy was calculated for both overall performance as well as individual subsections.

Results: GPT-4 Turbo performed at an overall accuracy of 88.6% (70/79 questions), outperforming LLaMA 3.1's 77.2% (61/79). Within subsections, GPT-4 Turbo had higher accuracy in most areas, except for equal accuracy in the neuroradiology section. The subsections with the greatest accuracy for GPT-4 Turbo, in descending order, were chest and cardiac radiology (100%), musculoskeletal system (93.3%), and genitourinary system (92.9%). LLaMA 3.1's highest performance was 86.7% in the musculoskeletal system, while its lowest was 50.0% in chest radiology.

Conclusion: GPT-4 Turbo consistently outperformed LLaMA 3.1 in answering pediatric radiology questions, both overall and within most subsections. These findings suggest that GPT-4 Turbo may offer more accurate responses in specialized medical education, in contrast to LLaMA 3.1's efficient performance, although future research should further evaluate AI models' performance within other fields.

Categories: Healthcare Technology
Keywords: ai in medical education, chatgpt, large language models (llms), llama, pediatric radiology

Introduction

Over the past few years, large language models (LLMs) - artificial intelligence (AI) systems trained on deep learning algorithms - have significantly developed and become increasingly sophisticated. These AI systems, such as OpenAI's GPT-4 Turbo (San Francisco, CA) and Meta's LLaMA 3.1 (Menlo Park, CA), utilize techniques to create human-like text responses [1,2]. Their rapid development has transformed various industries, allowing for tasks like diagnostic problem-solving, language translation, and personalized recommendations, showcasing the potential of AI in the future.

Specifically, LLMs, such as GPT-4 Turbo and LLaMA 3.1, both released in 2023, have so far been the foundation of this potential [1-3]. Each model uses similar deep learning analyses based on extensive datasets, accomplishing a wide variety of tasks from generating simple text to complex reasoning. GPT-4 Turbo, an advanced version of GPT-4.0, GPT-3.5, and GPT-3.0, distinguishes itself with its improved "context window," which means it is more capable of remembering and processing multiple pages of input to generate a more complex output than prior generations [1,4]. Although similar, Meta's LLaMA 3.1 is better known for its efficiency and performance optimization, while GPT-4 Turbo is known for its larger processing capabilities and multimodal task processing [2,4,5].

This study aims to compare these models within a pediatric radiology question set, assessing their capabilities to answer specialized questions accurately. Ultimately, this analysis hopes to delineate each model's strengths and potential areas for improvement within medical education.

Materials And Methods

This study compared the accuracy of ChatGPT-4 Turbo and Meta's LLaMA 3.1 by evaluating their responses to 79 text-based questions from the textbook "Pediatric Imaging: A Core Review" [6]. The primary objective was to assess overall accuracy with a secondary objective to analyze each AI model's performance within specific subsections. From the book's initial pool of 302 questions, 79 were selected based on predetermined criteria that excluded image-based questions. The questions were separated into seven subsections, corresponding with the text's various focuses such as musculoskeletal and chest radiology, with 38 questions being slightly revised to maintain the consistency of the text-only format.

Each LLM's accuracy was measured by calculating the percentage of correct answers for each model across the entire question set as well as within subsections. Each model was primed to await a series of multiple-choice questions and to carefully select what they analyzed to be the correct answer (Figure *1*). Using a binary answering system, questions answered correctly were notated separately as "1" and incorrectly as "0," and the final analysis utilized the sum of the individual notations. Analysis of each performance allowed this study to highlight any differences in accuracy, both overall and within subsection (Figure *2*). This approach allowed for a clear comparison of how GPT-4 Turbo and LLaMA 3.1 handled pediatric radiology questions. This study did not require Institutional Review Board (IRB) approval due to the lack of pertinent patient data, and all data were cross-checked for diagnostic accuracy.

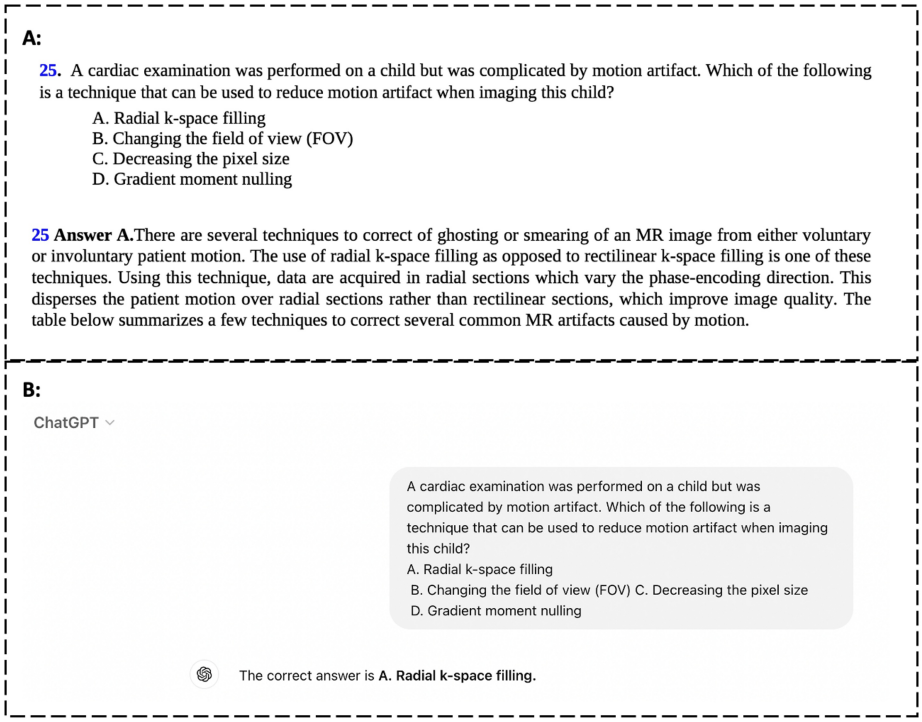

FIGURE 1: Comparison of responses to a sample textbook question.

(A) Part A reflects the original question and answer of one example question as provided in "Pediatric Imaging: A Core Review, 2nd edition," by Blumer et al. [6]. (B) Part B reflects ChatGPT-4 Turbo's response to the identical question after being presented with four answer choices. This methodology was consistently applied to all eligible questions from this textbook.

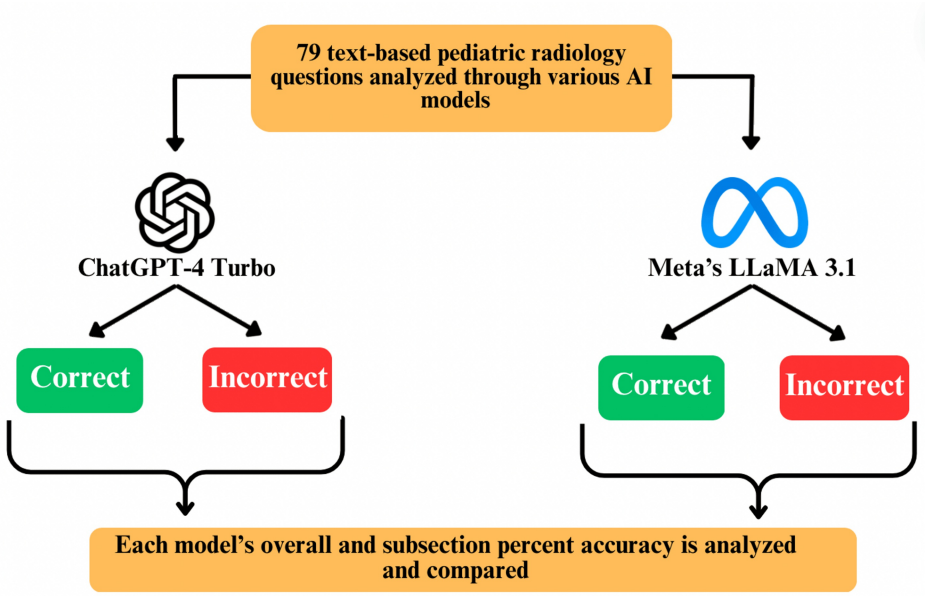

FIGURE 2: Sequence of methodology to compare ChatGPT-4 Turbo and Meta's LLaMA 3.1.

AI: artificial intelligence.

Results

In this analysis, GPT-4 Turbo demonstrated a greater overall accuracy compared to Meta's LLaMA 3.1. GPT-4 Turbo correctly answered 88.6% of the 79 questions (70/79), while LLaMA 3.1 correctly answered 77.2% (61/79). When the results were further divided into subsections, GPT-4 Turbo consistently outperformed LLaMA 3.1 in most areas (Figures *3*, *4*).

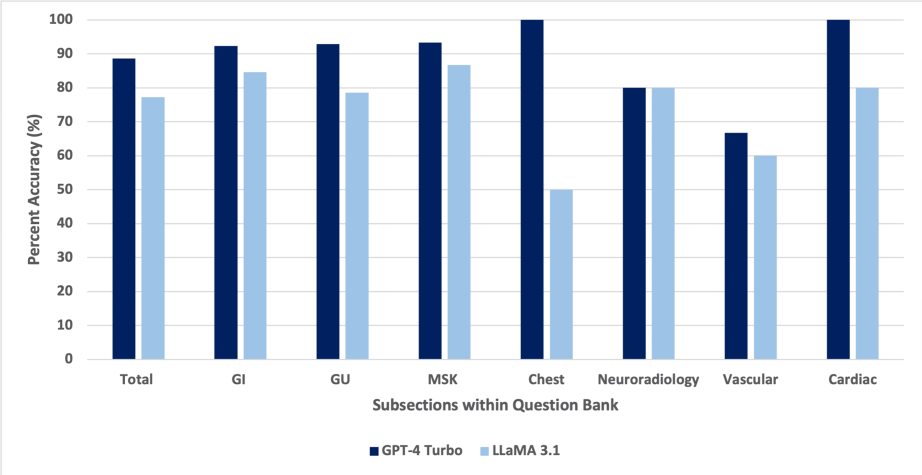

FIGURE 3: Bar graph illustrating differences in performance between ChatGPT-4 Turbo and LLaMA 3.1.

GI: gastrointestinal; GU: genitourinary; MSK: musculoskeletal.

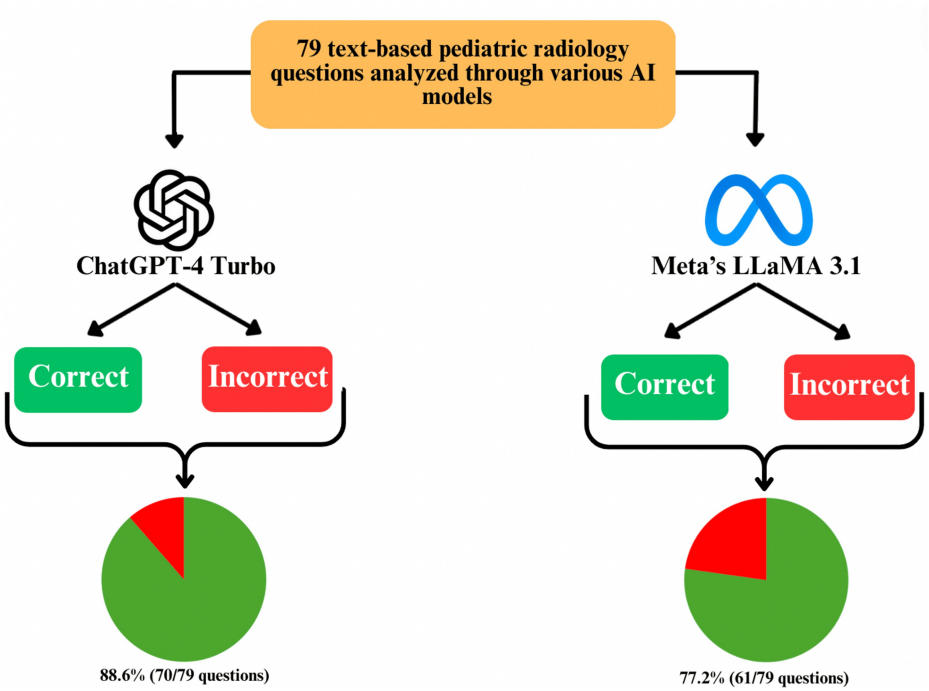

FIGURE 4: Results of comparison and percent accuracies between ChatGPT-4 Turbo and LLaMA 3.1.

AI: artificial intelligence.

GPT-4 Turbo performed at 92.3% (12/13) in the gastrointestinal system, 92.9% (13/14) in the genitourinary system, 93.3% (14/15) in the musculoskeletal system, 100% (2/2) in chest radiology, 80.0% (4/5) in neuroradiology, 66.7% (10/15) in vascular radiology, and 100% (15/15) in cardiac radiology. In comparison, LLaMA 3.1 scored 84.6% (11/13) in the gastrointestinal system, 78.6% (11/14) in the genitourinary system, 86.7% (13/15) in the musculoskeletal system, 50% (1/2) in chest radiology, 80.0% (4/5) in neuroradiology, 60.0% (9/15) in vascular radiology, and 80.0% (12/15) in cardiac radiology (Table *1*).

	Number of questions	GPT4-Turbo's percent accuracy (%)	Number of questions, correct/total	LLaMA 3.1's percent accuracy (%)	Number of questions, correct/total
Overall	79	88.6%	70/79	77.2%	61/79
GI tract	13	92.3%	12/13	84.6%	11/13
GU tract	14	92.3%	13/14	78.6%	11/14
MSK	15	93.3%	14/15	86.7%	13/15
Chest	2	100.0%	2/2	50.0%	1/2
Neuroradiology	5	80.0%	4/5	80.0%	4/5
Vascular	15	66.7%	10/15	60.0%	9/15
Cardiac	15	100.0%	15/15	80.0%	12/15

TABLE 1: Comparative percent accuracies between ChatGPT-4 Turbo and LLaMA 3.1.

GI: gastrointestinal; GU: genitourinary; MSK: musculoskeletal.

In terms of specific accuracy by subsection, GPT-4 Turbo performed best in chest radiology and cardiac radiology (100.0%), followed by the musculoskeletal system (93.3%), gastrointestinal system (92.3%), and genitourinary system (92.9%). The lowest performance for GPT-4 Turbo was in vascular radiology, at 66.7%. LLaMA 3.1's highest accuracy was in the musculoskeletal system (86.7%), and its lowest was in chest

radiology (50.0%).

Discussion

Overall & subsection performance per AI model

GPT-4 Turbo's greater percent accuracy compared to LLaMA 3.1 initially implies its better processing of text-based questions. This performance could likely be attributed to its training on larger, more specialized datasets, which are better equipped for answering specific questions, unlike LLaMA 3.1, which was developed to handle a broader research focus with greater efficiency [1,6]. As stated previously, GPT-4 Turbo's greater context window, which enhances its level of complex problem-solving, may strengthen its ability to handle complex questions [1,2].

Of note, the similar percentage accuracies in smaller categories such as neuroradiology and vascular radiology are likely due to the limited number of text-based questions compared to larger sections. The study's exclusion of image-based questions, which do not mirror a realistic radiology assessment, may have restricted the scope of our research as well.

General strengths and limitations of AI in medicine

AI offers tools and the integration of its ever-growing database that continues to manipulate data more quickly and sometimes more accurately than humans [6,7]. In specific medical fields, such as radiology, these LLMs have the potential to help with quickly interpreting the literature on patient conditions, summarizing patient histories, or creating reports. In fact, prior comparative studies between earlier versions of these LLMs have found significant overall differences in accuracy, which may be a byproduct of each LLM's intended function [8-10]. However, concerns remain about the ability of these models to correctly interpret and analyze images or provide the correct information, especially with reported events in the past of incorrect analyses [7,11]. While the rise of these models has been rapid, their integration in most systems is still in its growing stages, and analyses such as this study continue to mark its ever-growing capabilities and potential for the future.

One interesting note between GPT-4 Turbo and Meta's LLaMA 3.1 is the difference between open-source software and proprietary models [1]. Currently, LLaMA is labeled as open source, which allows developers to openly create or experiment with the model's strengths and limitations. However, GPT-4 Turbo is labeled as a "proprietary" model, which is aimed toward commercial use [1]. Customers are still allowed to use its tools and database, but this limits the allowable experimentation compared to LLaMA [1,4]. This dichotomy between the models may be a talking point in the future if one model becomes more accessible to change, such as in rapidly evolving fields like medicine.

Public perception and trust in AI for healthcare

The ethical consideration of utilizing AI in healthcare fields is complex and significantly intersects with the public perception of AI in general. There is clearly growing excitement for the future of these LLMs within multiple fields, and many view these models with optimism that they may improve the speed of various tasks and reduce human error [12,13]. However, problems arise around the accuracy and acceptability of AI's decision-making, especially when used in urgent situations or given that certain models may only be trained on limited datasets [12-14]. Specifically in medicine, many studies report that patients look for trust in their judgment of AI models and that these models are not currently at that level yet [13-16]. This discrepancy is further noted in this study, where different AI models are seen to have differing performances within the same question set, further emphasizing the lack of standardized training across models [16,17]. Each of these models is trained with specific tasks in mind and while they will continue to improve over decades, their extensive training is yet to be developed in fields where precision is demanded, such as radiology.

Future direction

In the future, studies incorporating image-based questions would be a useful additive to realistically compare each model's true capabilities, especially considering GPT-4 Turbo's improved context windows. Additional models may also be compared such as Microsoft Copilot, Claude 3 Opus, or others with GPT-4 Turbo or LLaMA 3.1, which would better assess the ability of each model within different aspects of medicine and education [17]. Exploring the abilities of these AI models in diagnostic scenarios specifically could offer insight into their practical use in fields like radiology, and following up with longitudinal studies could help observe the evolution of these models' performances over time.

Conclusions

The findings of this study reflect a greater percent accuracy with GPT-4 Turbo compared to LLaMA 3.1 in answering a standardized set of pediatric radiology questions, with GPT-4 Turbo performing at 88.6% correct (70/79) compared to LLaMA 3.1's 77.2% (61/79). These data suggest that GPT-4 Turbo may have an improved dataset for handling specialized text-based tasks. However, the variance across different subsections, such as neuroradiology, was minimal, and statistical significance was not comparably assessed. These findings

highlight the need for cautious interpretation of the data and suggest that both models have unique strengths and limitations that should be considered when integrating AI in various settings.

In conclusion, this study underscores the potential of AI models in medical education, especially in specialized fields such as radiology where accurate interpretations of images are critical. As AI continues to evolve, it offers exciting potential to medicine, radiology, and overall patient outcomes. However, it is vital that these models are utilized ethically and as a supplement to human expertise. Future research should continue to investigate these models to optimize AI while upholding its responsible application.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Mohammed Abdul Sami, Pokhraj P. Suthar, Keyur Parekh, Mohammed Abdul Samad

Acquisition, analysis, or interpretation of data: Mohammed Abdul Sami, Pokhraj P. Suthar, Keyur Parekh, Mohammed Abdul Samad

Drafting of the manuscript: Mohammed Abdul Sami, Pokhraj P. Suthar, Keyur Parekh, Mohammed Abdul Samad

Critical review of the manuscript for important intellectual content: Mohammed Abdul Sami, Pokhraj P. Suthar, Keyur Parekh, Mohammed Abdul Samad

Supervision: Pokhraj P. Suthar, Keyur Parekh

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

To collect data for this article, the authors used ChatGPT 4.0 and Meta's LLaMA 3.1. After using these tools/services, the authors reviewed and edited the content as needed and assume full responsibility for the content of this publication.

References

1. Llama 3 70B vs GPT-4: comparison analysis. (2024). Accessed: September 8, 2024: https://www.vellum.ai/blog/llama-3-70b-vs-gpt-4-comparison-analysis.
2. Suthar PP, Kounsal A, Chhetri L, Saini D, Dua SG: Artificial intelligence (AI) in radiology: a deep dive into ChatGPT 4.0's accuracy with the American Journal of Neuroradiology's (AJNR) "case of the month". Cureus. 2023, 15:e43958. 10.7759/cureus.43958
3. Strong E, DiGiammarino A, Weng Y, Kumar A, Hosamani P, Hom J, Chen JH: Chatbot vs medical student performance on free-response clinical reasoning examinations. JAMA Intern Med. 2023, 183:1028-30. 10.1001/jamainternmed.2023.2909
4. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, Miki Y: ChatGPT's diagnostic performance from patient history and imaging findings on the Diagnosis Please quizzes. Radiology. 2023, 308:e231040. 10.1148/radiol.231040
5. Carlà MM, Gambini G, Baldascino A, et al.: Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. Br J Ophthalmol. 2024, 108:1457-69. 10.1136/bjo-2023-325143
6. Blumer SL, Halabi SS, Biko DM: Pediatric Imaging: A Core Review. 2nd edition . Wolters Kluwer Health, Philadelphia, PA; 2023.
7. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J: Systematic analysis of ChatGPT, Google Search and Llama 2 for clinical decision support tasks. Nat Commun. 2024, 15:2050. 10.1038/s41467-024-46411-8
8. Abdul Sami M, Abdul Samad M, Parekh K, Suthar PP: Comparative accuracy of ChatGPT 4.0 and Google Gemini in answering pediatric radiology text-based questions. Cureus. 2024, 16:e70897. 10.7759/cureus.70897
9. Gupta R, Hamid AM, Jhaveri M, Patel N, Suthar PP: Comparative evaluation of AI models such as ChatGPT 3.5, ChatGPT 4.0, and Google Gemini in neuroradiology diagnostics. Cureus. 2024, 16:e67766.

10.7759/cureus.67766

10. Zhang J, Sun K, Jagadeesh A, et al.: The potential and pitfalls of using a large language model such as ChatGPT, GPT-4, or LLaMA as a clinical assistant. J Am Med Inform Assoc. 2024, 31:1884-91. 10.1093/jamia/ocae184
11. Rydzewski NR, Dinakaran D, Zhao SG, Ruppin E, Turkbey B, Citrin DE, Patel KR: Comparative evaluation of LLMs in clinical oncology. NEJM AI. 2024, 1:10.1056/aioa2300151
12. Bhattarai K, Oh IY, Sierra JM, Tang J, Payne PR, Abrams Z, Lai AM: Leveraging GPT-4 for identifying cancer phenotypes in electronic health records: a performance comparison between GPT-4, GPT-3.5-Turbo, Flan-T5, Llama-3-8B, and spaCy's rule-based and machine learning-based methods. JAMIA Open. 2024, 7:ooae060. 10.1093/jamiaopen/ooae060
13. Sun Z, Ong H, Kennedy P, et al.: Evaluating GPT-4 on impressions generation in radiology reports . Radiology. 2023, 307:e231259. 10.1148/radiol.231259
14. Elkassem AA, Smith AD: Potential use cases for ChatGPT in radiology reporting . AJR Am J Roentgenol. 2023, 221:373-6. 10.2214/AJR.23.29198
15. Mistry NP, Saeed H, Rafique S, Le T, Obaid H, Adams SJ: Large language models as tools to generate Radiology Board-style multiple-choice questions. Acad Radiol. 2024, 31:3872-8. 10.1016/j.acra.2024.06.046
16. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al.: Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ. 2023, 9:e48291. 10.2196/48291
17. Mohammad B, Supti T, Alzubaidi M, Shah H, Alam T, Shah Z, Househ M: The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform. 2023, 305:644-7. 10.3233/SHTI230580