Cureus
Part of SPRINGER NATURE

Open Access Original Article

Comparative Accuracy of ChatGPT 4.0 and Google Gemini in Answering Pediatric Radiology Text-Based Questions

Review began 09/02/2024
Review ended 10/01/2024
Published 10/05/2024

DOI: 10.7759/cureus.70897

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 Diagnostic Radiology, 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 evaluates the accuracy of two AI language models, ChatGPT 4.0 and Google Gemini (as of August 2024), in answering a set of 79 text-based pediatric radiology questions from “Pediatric Imaging: A Core Review.” Accurate interpretation of text and images is critical in radiology, making AI tools valuable in medical education.

Methods: The study involved 79 questions selected from a pediatric radiology question set, focusing solely on text-based questions. ChatGPT 4.0 and Google Gemini answered these questions, and their responses were evaluated using a binary scoring system. Statistical analyses, including chi-square tests and relative risk (RR) calculations, were performed to compare the overall and subsection accuracy of the models.

Results: ChatGPT 4.0 demonstrated superior accuracy, correctly answering 83.5% (66/79) of the questions, compared to Google Gemini's 68.4% (54/79), with a statistically significant difference (p=0.0255, RR=1.221). No statistically significant differences were found between the models within individual subsections, with p-values ranging from 0.136 to 1.

Conclusion: ChatGPT 4.0 outperformed Google Gemini in overall accuracy for text-based pediatric radiology questions, highlighting its potential utility in medical education. However, the lack of significant differences within subsections and the exclusion of image-based questions underscore the need for further research with larger sample sizes and multimodal inputs to fully assess AI models' capabilities in radiology.

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

Introduction

The rapid innovation of artificial intelligence in the 21st century has brought about language learning models, also known as large language models (LLMs), as a tool with multiple capabilities [1-3]. These LLMs build on deep learning algorithms that allow them to analyze inputs and create text by predicting sequences of words [1, 4, 5]. The models are trained on substantial sets of data, enabling them to have an extensive understanding of language and perform various tasks based on their inputs, such as answering questions [2, 3]. Particularly, two leading language learning models are OpenAI's ChatGPT and Google's Gemini. ChatGPT, based on its recent GPT-4 architecture, excels in tasks that require logical text creation [6, 7]. Its development is part of OpenAI's refinements of its model, building on earlier versions such as ChatGPT 3.0 and ChatGPT 3.5 [2, 8]. Similarly, Google Gemini represents Google's AI model that is designed to complete multimodal tasks by integrating inputs from text, images, and video responses [1, 4, 9]. Although both models have their similarities, ChatGPT and Gemini have distinct differences. ChatGPT is often utilized to handle inputs based on reasoning and discussion (although it can still analyze images), whereas Google Gemini relies on its ability to both input and output a broader range of responses [1, 10, 11].

This study aims to compare ChatGPT and Google Gemini in their abilities to accurately answer questions from a pediatric radiology-specific resource. By evaluating their performance on a standardized set of questions, we aim to determine if there is a significant difference in accuracy between ChatGPT 4.0 and Google Gemini, providing insights into their respective strengths and weaknesses for their potential use in an educational setting. This analysis is intended to contribute a valuable perspective to the ongoing discussion about the optimal use of language learning models in various fields.

Materials And Methods

This study analyzed the accuracy of ChatGPT 4.0 and Google Gemini versions as of August 2024. Our primary objective was to determine if there was statistical significance in the overall accuracy of each AI model when answering standardized questions from a pediatric radiology question set [12]. To establish our

How to cite this article
Abdul Sami M, Abdul Samad M, Parekh K, et al. (October 05, 2024) Comparative Accuracy of ChatGPT 4.0 and Google Gemini in Answering Pediatric Radiology Text-Based Questions. Cureus 16(10): e70897. DOI 10.7759/cureus.70897

inclusion criteria, we included questions that were text-only and excluded image-based questions. From the textbook's total of 302 questions, we selected the 79 questions that met these criteria. Of these 79 questions, 38 were slightly modified by the authors to remove references to prior answers and adhere to the text-only format. For example, an unedited question from the textbook that read, "Regarding the most likely diagnosis of the patient in Question 8, which of the following is true?" was changed to "Regarding [specific medical condition], which of the following is true?" This alteration removed the reference to a previous question while maintaining the core concept being tested. These 79 questions were drawn from seven subsections, each representing a distinct field such as musculoskeletal system, and chest radiology (Figure *1*). Multi-disciplinary questions were excluded to create a better analysis of accuracy by a specific subsection. Accordingly, our secondary objective sought to determine the presence of statistical significance between each AI model's accuracy within individual subsections.

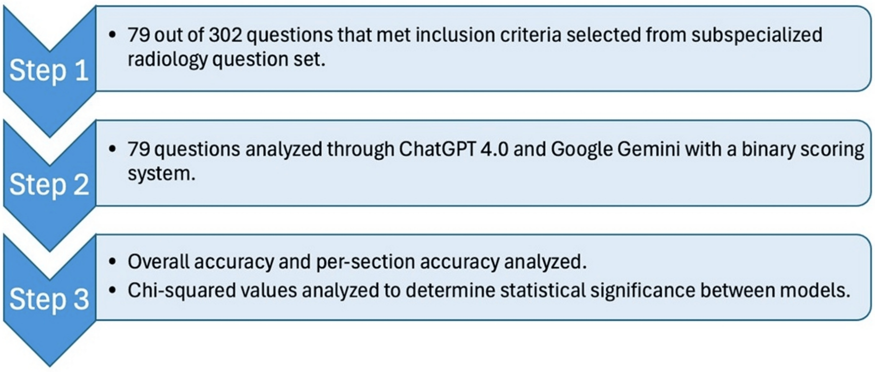
Step 1
• 79 out of 302 questions that met inclusion criteria selected from subspecialized radiology question set.
Step 2
• 79 questions analyzed through ChatGPT 4.0 and Google Gemini with a binary scoring system.
Step 3
• Overall accuracy and per-section accuracy analyzed.
• Chi-squared values analyzed to determine statistical significance between models.

FIGURE 1: Series of steps within methods followed in this study

During the collection of data, questions were answered using a binary scoring system where correct responses were marked as 1 and incorrect responses as 0. After running the questions as individual inputs, multiple chi-square analyses were utilized to determine the difference between the two models. We analyzed the results to identify statistical significance, qualifying p-values to be significant with a cut-off of 0.05 and calculating relative risk (RR) with 95% CI. The results were further analyzed across subsections to assess accuracy (Figure *2*). Since this study did not involve patient data, Institutional Review Board (IRB) approval was not required. This allowed for a streamlined research process, focused solely on evaluating the performance of the AI models under controlled conditions. All the study's data and analyses were securely recorded in a Microsoft Excel (Microsoft Corporation, Redmond, Washington, USA) spreadsheet and verified for accuracy. The data was verified for accuracy through a comprehensive process, where two researchers independently entered the data into separate Excel spreadsheets for calculations. We then used an automated comparison via Excel to identify any discrepancies between the two entries. Following this, the researchers reviewed all calculations and discrepancies manually. Finally, our senior researchers on the team reviewed the final results to ensure their accuracy and reliability.

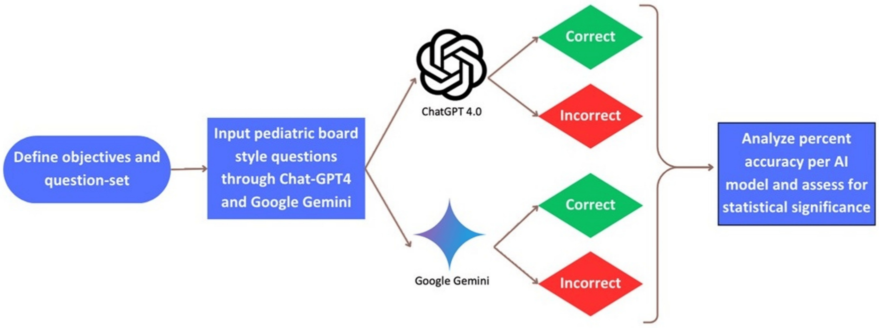
Define objectives and question-set
Input pediatric board style questions through Chat-GPT4 and Google Gemini
ChatGPT 4.0
Correct
Incorrect
Google Gemini
Correct
Incorrect
Analyze percent accuracy per AI model and assess for statistical significance

FIGURE 2: Illustrative diagram outlining sequences within the methodology

Results

The analysis of ChatGPT 4.0 and Google Gemini reflected a statistically significant difference in overall accuracy. ChatGPT 4.0 answered 83.5% of the 79 questions correctly, while Google Gemini answered 68.35% correctly (Figure 3). This difference in performance was evident in specific questions. For example, when asked "Regarding neuroblastoma stage IV-S, which of the following is TRUE?", ChatGPT 4.0 correctly answered "B. It affects the skin, liver, and bone marrow," while Google Gemini incorrectly responded, "D. There is no metastatic disease." Similar differences were further illustrated within other subsections as well.

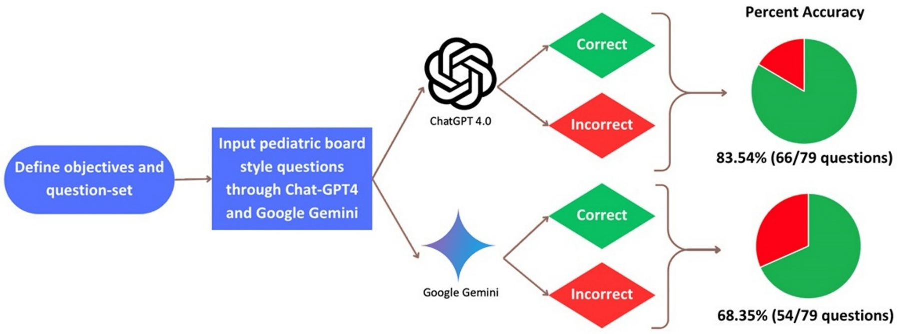

FIGURE 3: Overall diagnostic accuracy between models

The chi-square test yielded a chi-squared value of 4.99 with a p-value of 0.0255, and a RR of 1.221 (95% CI: 1.020 to 1.459), demonstrating that ChatGPT 4.0 was 22.1% more likely to provide correct answers than Google Gemini within the overall sample size (Table 1).

Subsection	Number of questions	Chi-squared value	P-value	RR	95% CI
Overall	79	4.99	0.0255	1.221	1.020-1.459
GI tract	13	0.866	0.352	1.223	0.795-1.879
GU tract	14	2.052	0.152	1.393	0.862-2.252
MSK	15	0	1	1	0.827-1.209
Chest	2	1.334	0.248	2	0.501-7.997
Neuroradiology	5	0	1	1	0.827-1.209
Vascular	15	2.222	0.136	1.57	0.842-2.916
Cardiac	15	0.6	0.439	1.222	0.731-2.040

TABLE 1: Table delineating sample sizes per subsection, chi-squared values, p-values, RRs, and 95% CI

RR, relative risk

However, there was no statistically significant difference in accuracy between the two models when the results were analyzed individually by subsection. The closest to statistical significance were the genitourinary tract and vascular radiology, with p-values of 0.152 and 0.136, respectively, and RR values of 1.393 (95% CI: 0.862, 2.252) and 1.570 (95% CI: 0.842, 2.916). In subsequent order of increasing p-values were chest (p-value: 0.248, RR: 2.000 (95% CI: 0.501, 7.997)), gastrointestinal tract (p-value: 0.352, RR: 1.223 (95% CI: 0.795, 1.879)), cardiac radiology (p-value: 0.439, RR: 1.222 (95% CI: 0.731, 2.040)), musculoskeletal system (p-value: 1.000, RR: 1.000 (95% CI: 0.827, 1.209)), and neuroradiology (p-value: 1.000, RR: 1.000, with no meaningful CI due to identical performance) (Figures 4, 5 and Table 1).

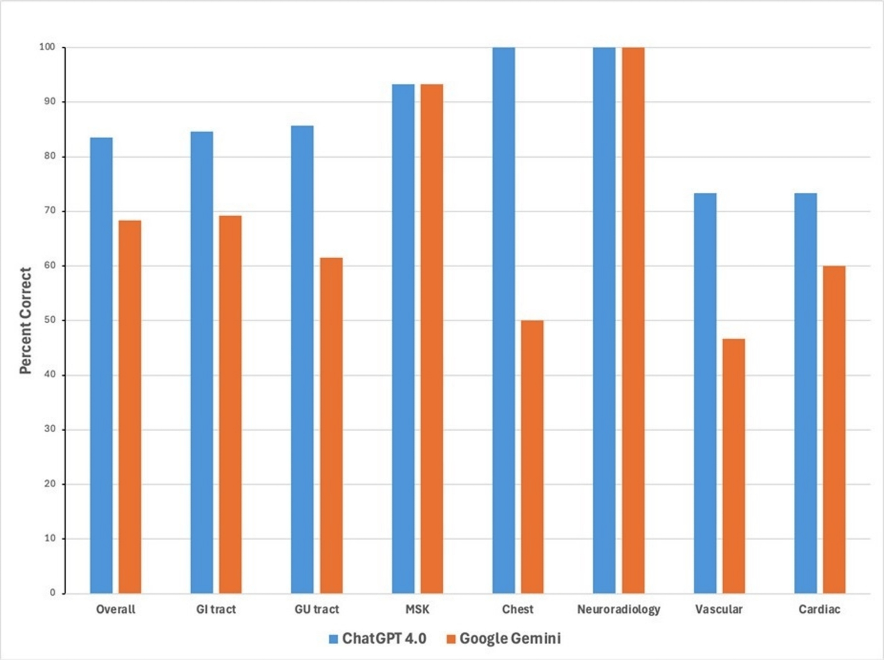

FIGURE 4: Diagram comparing diagnostic accuracy between ChatGPT 4.0 and Google Gemini overall and within subsections

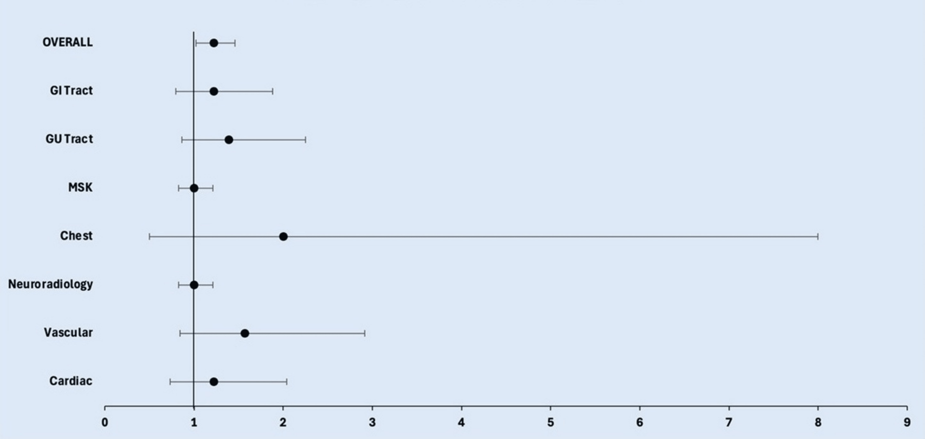

FIGURE 5: Forest plot (RR with 95% CI) reflecting the lack of statistical significance within subsections between models but reflecting the presence of statistical significance in overall performance

RR, relative risk

Discussion

Overall accuracy and performance variability

The significant difference in overall accuracy between ChatGPT 4.0 and Google Gemini indicates that ChatGPT 4.0 may be more reliable in answering text-based questions. This reliability could be due to ChatGPT's specific datasets that help it better answer specific types of questions, compared to Google Gemini's multimodal training [1, 13, 14]. The RR analysis also supports the data that ChatGPT 4.0 is more likely to answer correctly overall, with an RR of 1.221. However, the lack of statistical significance within subsections, such as in MSK and neuroradiology, may suggest that true significant differences were seen on an aggregate scale that could not be seen perhaps due to the smaller sample sizes within sections.

Additionally, the exclusion of image-based questions, which are vital in fields such as radiology, may have influenced the results. Including these questions could have provided a more comprehensive comparison, although significant improvements are yet necessary in AI models for an accurate assessment [15, 16].

Subgroup performance and analysis

Despite the overall significant difference in accuracy between ChatGPT 4.0 and Google Gemini, our subgroup analysis did not reveal statistically significant differences when examining the results by specific subspecialties. For instance, in the genitourinary tract and vascular radiology, ChatGPT 4.0 demonstrated a trend toward better performance, with RR of 1.393 and 1.570, respectively. However, these differences were not statistically significant, as indicated by p-values of 0.152 and 0.136. Similarly, in chest radiology, ChatGPT 4.0 appeared to outperform Google Gemini, with an RR of 2.000, but this difference also lacked statistical significance (p-value: 0.248).

For other subspecialties, such as the gastrointestinal tract, cardiac radiology, musculoskeletal system, and neuroradiology, the performance of the two models was nearly identical. This was reflected in p-values ranging from 0.352 to 1.000 and RR values that indicated no meaningful differences in accuracy.

These findings suggest that while ChatGPT 4.0 may hold an advantage in overall accuracy, the differences within specific subspecialties were not as pronounced. The lack of statistical significance in these areas could be attributed to smaller sample sizes, which might not have been sufficient to detect subtle differences between the models. This highlights the need for future research to include larger sample sizes within each subspecialty and potentially incorporate image-based questions, which may provide a more comprehensive comparison of these AI models' strengths and weaknesses in specific radiology fields.

Limitations and potential biases

The limitations of this study include the exclusion of image-based questions, utilizing a single subspecialized question set within pediatric radiology, and the low sample size for certain sections, all of which rendered it difficult to determine statistically significant differences in accuracy where they might exist. Moreover, this study did not provide a longitudinal assessment of the AI models' performance, which is relevant given the rapid pace of improvement in these models. The absence of evaluation of the models' performances might evolve over time limiting our understanding of their potential and reliability in clinical settings. Additionally, the lack of comparison to a human radiologist's performance further limits the ability to contextualize these AI models. The implications of these limitations include an incomplete assessment of the models' true performance, particularly given that this study was conducted with a controlled question set rather than real-life patient scenarios involving actual patient data [1, 17, 18].

Future studies should aim to increase the sample size and include an increased number of AI models to provide a more accurate assessment of each model's strengths and weaknesses.

Implications within education

The educational implications of this study are significant in fields such as general or specialized medical education. AI models like ChatGPT and Google Gemini could be utilized in training students with both text and image-based questions [19, 20]. For example, ChatGPT 4.0’s current performance suggests that it could be used to better understand case studies or accurately summarize educational content [18]. The potential of AI in medical education extends beyond radiology, with applications in other specialties where analyzing text is critical [20, 21]. Integrating AI models could further support medical residents or physicians in breaking down complex topics or creating personalized learning experiences [22].

Ethical considerations

The ethical considerations of AI models in educational settings are important, especially in fields such as radiology. Public concern over the use of AI in healthcare has grown, with recent studies indicating that while people realize the potential benefits of AI, they are apprehensive about its accuracy, data privacy, and the potential for bias [16, 19]. In radiology specifically, the concern remains around using AI for decision-making in a field where errors could have serious consequences [19, 22]. This study's findings underscore the importance of methodically evaluating various AI models before their adoption in fields such as healthcare or general education [23].

Furthermore, the integration of AI in radiology education (or within any field) raises questions about developing the critical thinking skills of trainees. There is a need to find a balance between utilizing AI as an educational tool and ensuring that future medical professionals develop and maintain the skills needed for independent clinical judgment.

Conclusions

The study demonstrates a statistically significant difference in accuracy between ChatGPT 4.0 and Google Gemini when answering standardized radiology-related questions, with ChatGPT 4.0 achieving an accuracy rate of 83.5% compared to Google Gemini's 68.4%, suggesting that ChatGPT 4.0 may be more reliable for certain text-based tasks in medical education. However, the observed variability across different pediatric radiology subspecialties and the exclusion of image-based questions indicate that both AI models have distinct strengths and weaknesses that should be carefully considered. The findings emphasize the potential role of AI in enhancing medical education and diagnostic capabilities, particularly in radiology, while also underscoring the need for responsible integration of these technologies to complement human expertise. Future research should focus on refining these models and developing guidelines for their ethical and effective use in healthcare and educational contexts.

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

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

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

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

During the preparation of this work, the authors used ChatGPT 4.0 and Google Gemini. 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. What’s the Difference? TechRepublic. (2023). Accessed: August 10, 2024: https://www.techrepublic.com/article/chatgpt-vs-google-gemini/.
2. 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
3. Juluru K, Shih HH, Keshava Murthy KN, et al.: Integrating al algorithms into the clinical workflow . Radiol Artif Intell. 2021, 3:e210013. 10.1148/ryai.2021210013
4. 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
5. 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
6. Rossettini G, Rodeghiero L, Corradi F, 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 Med Educ. 2024, 24:694. 10.1186/s12909-024-05630-9
7. 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
8. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W: ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the specialty certificate Examination in Dermatology. Clin Exp Dermatol. 2024, 49:686-91. 10.1093/ced/llad255
9. Brin D, Sorin V, Vaid A, et al.: Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments . Sci Rep. 2023, 13:16492. 10.1038/s41598-023-43436-9

10. 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. 10.2196/46482
11. Patil NS, Huang RS, van der Pol CB, Larocque N: Comparative performance of ChatGPT and Bard in a text-based radiology knowledge assessment. Can Assoc Radiol J. 2024, 75:344-50. 10.1177/08465371231193716
12. Blumer SL, Halabi SS, Biko DM: Pediatric Imaging: A Core Review . Lippincott Williams & Wilkins (LWW), 2023.
13. Masalkhi M, Ong J, Waisberg E, Lee AG: Google DeepMind's Gemini AI versus ChatGPT: a comparative analysis in ophthalmology. Eye (Lond). 2024, 38:1412-7. 10.1038/s41433-024-02958-w
14. Baytak A: The content analysis of the lesson plans created by ChatGPT and Google Gemini . RESSAT. 2024, 9:329-50.
15. Moglia A, Georgiou K, Cerveri P, et al.: Large language models in healthcare: from a systematic review on medical examinations to a comparative analysis on fundamentals of robotic surgery online test. Artif Intell Rev. 2024, 57:231.
16. Imran M, Almusharraf N: Google Gemini as a next generation AI educational tool: a review of emerging educational technology. Smart Learn Environ. 2024, 11:22.
17. 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
18. 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
19. Sun Z, Ong H, Kennedy P, et al.: Evaluating GPT4 on impressions generation in radiology reports . Radiology. 2023, 307:e231259. 10.1148/radiol.231259
20. 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
21. Cozzi A, Pinker K, Hidber A, et al.: BI-RADS category assignments by GPT-3.5, GPT-4, and Google Bard: a multilanguage study. Radiology. 2024, 311:e232133. 10.1148/radiol.232133
22. 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
23. Lee TJ, Campbell DJ, Patel S, Hossain A, Radfar N, Siddiqui E, Gardin JM: Unlocking health literacy: The ultimate guide to hypertension education from ChatGPT versus Google Gemini. Cureus. 2024, 16:e59898. 10.7759/cureus.59898