Article published online: 2023-12-29

Assessing the Capability of ChatGPT, Google Bard, and Microsoft Bing in Solving Radiology Case Vignettes

Pradosh Kumar Sarangi[1] Ravi Kant Narayan[2] Sudipta Mohakud[3] Aditi Vats[3] Debabrata Sahani[3] Himel Mondal[4]

[1] Department of Radiodiagnosis, All India Institute of Medical Sciences, Deoghar, Jharkhand, India
[2] Department of Anatomy, ESIC Medical College & Hospital, Bihta, Patna, Bihar, India
[3] Department of Radiodiagnosis, All India Institute of Medical Sciences, Bhubaneswar, Odisha, India
[4] Department of Physiology, All India Institute of Medical Sciences, Deoghar, Jharkhand, India

Address for correspondence Himel Mondal, Department of Physiology, All India Institute of Medical Sciences, Deoghar, Jharkhand 814152, India (e-mail: himelmkcg@gmail.com).

Indian J Radiol Imaging 2024;34:276–282.

Abstract

Background The field of radiology relies on accurate interpretation of medical images for effective diagnosis and patient care. Recent advancements in artificial intelligence (AI) and natural language processing have sparked interest in exploring the potential of AI models in assisting radiologists. However, limited research has been conducted to assess the performance of AI models in radiology case interpretation, particularly in comparison to human experts.

Objective This study aimed to evaluate the performance of ChatGPT, Google Bard, and Bing in solving radiology case vignettes (Fellowship of the Royal College of Radiologists 2A [FRCR2A] examination style questions) by comparing their responses to those provided by two radiology residents.

Methods A total of 120 multiple-choice questions based on radiology case vignettes were formulated according to the pattern of FRCR2A examination. The questions were presented to ChatGPT, Google Bard, and Bing. Two residents wrote the examination with the same questions in 3 hours. The responses generated by the AI models were collected and compared to the answer keys and explanation of the answers was rated by the two radiologists. A cutoff of 60% was set as the passing score.

Results The two residents (63.33 and 57.5%) outperformed the three AI models: Bard (44.17%), Bing (53.33%), and ChatGPT (45%), but only one resident passed the examination. The response patterns among the five respondents were significantly different ($p = 0.0117$). In addition, the agreement among the generative AI models was significant (intraclass correlation coefficient [ICC] = 0.628), but there was no agreement between the residents (Kappa = –0.376). The explanation of generative AI models in support of answer was 44.72% accurate.

Keywords
- artificial intelligence
- Bard
- Bing
- ChatGPT
- natural language processing
- radiology
- FRCR2A
- fellowship

article published online
December 29, 2023

DOI https://doi.org/10.1055/s-0043-1777746.
ISSN 0971-3026.

Thieme Medical and Scientific Publishers Pvt. Ltd., A-12, 2nd Floor, Sector 2, Noida-201301 UP, India

Conclusion Humans exhibited superior accuracy compared to the AI models, showcasing a stronger comprehension of the subject matter. All three AI models included in the study could not achieve the minimum percentage needed to pass an FRCR2A examination. However, generative AI models showed significant agreement in their answers where the residents exhibited low agreement, highlighting a lack of consistency in their responses.

Introduction

Radiology plays a crucial role in diagnosing and monitoring various medical conditions through the use of imaging techniques such as X-rays, computed tomography (CT) scans, magnetic resonance imaging (MRI) scans, and ultrasounds. Accurate interpretation of radiological images requires in-depth knowledge, experience, and pattern recognition skills, making it a highly specialized field.[1] With rapid advancements in artificial intelligence (AI), doctors can now get assistance in diagnosis. Radiological images, acquired using diverse modalities, are subjected to data preprocessing to ensure optimal quality.[2,3] AI algorithms, including machine learning and deep learning models, facilitate precise localization of anatomical structures or pathological lesions. Subsequent analysis of the segmented regions allows for automated detection, characterization, and quantitative measurements, aiding radiologists in making informed diagnostic decisions.[4,5]

Generative AI refers to a category of AI systems that have the ability to generate content, such as text, images, music, and more, in a way that mimics human creativity. A large language model (LLM) is a specific type of generative AI that is focused on natural language understanding and generation. Another is natural language processing (NLP), which is a field of AI that focuses on the interaction between computers and human language. NLP seeks to enable computers to understand, interpret, and generate human language in a valuable way. LLM is a subset of NLP. ChatGPT, Google Bard, and Bing are prominent AI-based models that have shown promise in assisting doctors and academicians in various domains.[6–8] However, their capabilities in solving radiology case vignettes have not been extensively investigated. Assessing the capability of AI models in radiology is essential to understand their potential clinical utility and identify areas where they may fall short. In a study assessing AI tools' responses to common lung cancer questions, ChatGPT exhibited higher accuracy compared to other tools. However, neither ChatGPT nor Bard or Bing consistently provided correct answers.[9] ChatGPT 3.5 demonstrated promising performance on a radiology board-style examination in lower-order thinking skills and clinical management knowledge but faced challenges with higher-order thinking tasks. On subsequent evaluation, GPT-4 showed overall improvement and better performance on higher-order thinking questions. However, GPT-4 had limited progress in addressing lower-order questions and occasionally provided incorrect responses.[10,11]

No previous study ascertained the capability of these three programs in solving questions that are asked in the examination of Fellowship of the Royal College of Radiologists 2A (FRCR2A). Hence, we conducted a comparative assessment of ChatGPT, Google Bard, and Bing by presenting them with 120 multiple-choice questions based on radiology case vignettes. The responses generated by these AI models were compared with the answers provided by two radiology residents. This research aimed to shed light on the performance of these AI models in radiology case interpretation and highlight the importance of human expertise in the field.

Materials and Methods

Type and Settings

This was a cross-sectional observational study. The study was conducted as a comparative evaluation of AI models (ChatGPT, Google Bard, and Bing) and radiology residents' responses in solving radiology case vignettes. This study was conducted in the World Wide Web and in the Department of Radiology at All India Institute of Medical Sciences, Deoghar and Bhubaneswar in June 2023. All the data collection was done on a personal computer connected to a personal broadband connection.

AI Tools Used

We used three AI programs—Google Bard, Microsoft Bing, and ChatGPT. The ChatGPT (GPT 3.5, free version) was developed by OpenAI and it utilizes a transformer-based architecture and NLP techniques to generate responses. Google Bard Experiment is an AI model developed by Google and it incorporates advanced language processing algorithms for generating responses to queries. Microsoft Bing (GPT 4, Creative) is a search engine developed by Microsoft and it also utilizes language understanding algorithms to generate relevant responses.

Case Vignette Preparation

A set of 120 radiology case vignettes with multiple-choice questions and answers were carefully crafted for this study by a radiologist with 7 years of experience with a postdoctorate degree. These vignettes covered a diverse range of diagnostic scenarios commonly encountered in radiology practice. Each vignette consisted of relevant clinical information as per the FRCR2A pattern (single best answer format) with findings of accompanying medical images (such as X-rays, CT scans, MRI scans, or ultrasounds), where necessary. Each question consists of the following three components: stem or clinical vignette, question, options (A–E) where all of which may be plausible in varying degrees but one is clearly the single best answer and other four are distractors.[12] The questions were subdivided into six broad categories: "cardiothoracic and vascular"; "musculoskeletal and trauma"; "gastrointestinal"; "genitourinary, adrenal, obstetrics and gynecology, and breast"; "pediatric"; and "central nervous system and head and neck." All the questions were reviewed by two subject experts for its validity and accuracy.

Data Collection from AI

The case vignettes were presented to ChatGPT, Google Bard, and Bing individually. The AI models were provided with the clinical information for each vignette and asked to generate responses from provided options to the accompanying questions. The responses generated by the AI models were collected and recorded for further analysis. A correct answer was scored 1 and a wrong answer was scored 0.

Score of Residents

Two experienced radiology residents who were preparing for FRCR2A participated in this study. They independently answered the questions of the case vignettes. They were provided a time of 3 hours for solving the questions. A correct answer was scored 1 and a wrong or no answer was scored 0.

Statistical Analysis

The collected data, including the responses from ChatGPT, Google Bard, Bing, and the radiology residents, were presented in number and percentage. A chi-squared test was used to compare the categorical values. As the data were not following normal distribution, a Kruskal–Wallis test with post hoc test was conducted to compare the median score among the respondents. Agreement of scores among the AI models and the two residents were tested by intraclass correlation coefficient (ICC) and Cohen's kappa, respectively. We used GraphPad Prism 9.5.0 (GraphPad Software Inc., United States) for statistical tests, and for any test, a p-value of less than 0.05 was considered statistically significant.

Ethical Issues

All the data used in the case vignettes were fictitious and prepared for teaching learning purposes. No patients' data were used. The study only involved audit of the public domain data. Hence, this study does not require ethics committee clearance as per the rule in the country guided by the Indian Council of Medical Research.

Results

►**Fig. 1** presents the accuracy rates of the three AI models (Bard, Bing, and ChatGPT) and the two residents (residents 1 and 2) in terms of the number of right and wrong answers. The AI models had a lower average accuracy rate (47.5%) compared to the residents (60.42%), indicating that the residents performed better in answering the questions.

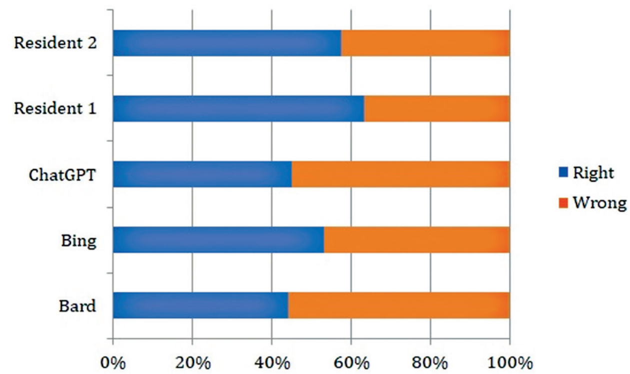


Fig. 1 Overall score of artificial intelligence–based language model and human participants.

Bing achieved the highest accuracy among the AI models, while Bard and ChatGPT had relatively lower accuracy rates.

In the Kruskal–Wallis test, we found that the pattern of answers among the five respondents were significantly different ($p = 0.0117$). However, in the post hoc test, only Bard versus resident 1 ($p = 0.03$) and ChatGPT versus resident 1 ($p = 0.04$) showed a statistically significant difference (►**Fig. 2**).

►**Table 1** presents the scores achieved by different AI models and two residents on various topics. The AI models showed varying performance across different topics, with some topics being better handled than others. The residents generally outperformed the AI models, demonstrating a higher level of understanding and accuracy in their responses.

►**Table 2** presents the ratings given by two raters for textual responses generated by "Bard," "Bing," and "ChatGPT." The textual responses for all three topics were generally rated as accurate (average, 44.72) or inaccurate, with a few instances of partial accuracy.

There was significant agreement of answers among three generative AI models; ICC average measures = 0.628 (95% CI: 0.495–0.73); $F(119) = 2.686$; $p < 0.0001$. However, there was

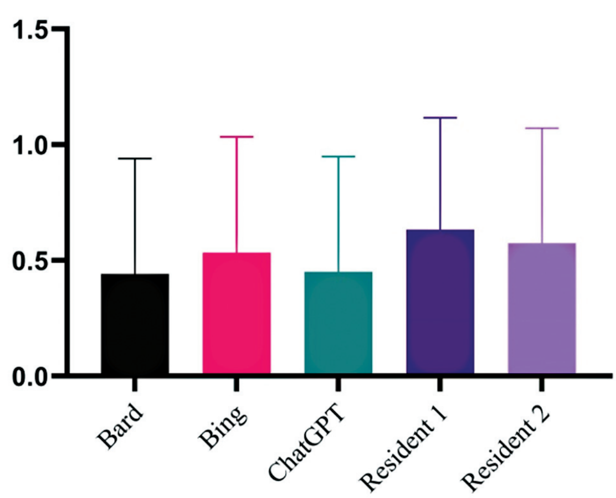


Fig. 2 Comparative scores among the artificial intelligence and residents.

Table 1 Topic wise score of three generative AI models and two residents

Topic	Maximum achievable score	Bard	Bing	ChatGPT	Resident 1	Resident 2
	Number (%)					
Cardiothoracic and vascular	20 (100)	12 (60)	7 (35)	8 (40)	14 (70)	11 (55)
Musculoskeletal and trauma	20 (100)	4 (20)	7 (35)	8 (40)	12 (60)	12 (60)
Gastrointestinal	20 (100)	9 (45)	11 (55)	6 (30)	12 (60)	15 (75)
Genitourinary, adrenal, obstetrics and gynecology, and breast	20 (100)	9 (45)	15 (75)	15 (75)	13 (65)	14 (70)
Pediatric	20 (100)	8 (40)	12 (60)	8 (40)	16 (80)	9 (45)
Central nervous and head and neck	20 (100)	11 (55)	12 (60)	9 (45)	9 (45)	8 (40)
Median (Q1–Q3)	–	9 (7–11.25)	11.5 (7–12.75)	8 (7.5–10.5)	12.5 (11.25–14.5)	11.5 (8.75–14.25)

Note: Kruskal–Wallis test; $p = 0.122$.

Table 2 Analysis of the explanatory text of the questions by two individual raters

Large language model	Rater 1				Rater 2			
	Accurate	Partially accurate	Inaccurate	*p*-Value[a]	Accurate	Partially accurate	Inaccurate	*p*-Value[a]
Bard	50 (41.67)	8 (6.67)	62 (51.67)	< 0.001	52 (43.33)	9 (7.5)	59 (49.17)	< 0.001
Bing	61 (50.83)	7 (5.83)	52 (43.33)	< 0.001	60 (50)	6 (5)	54 (45)	< 0.001
ChatGPT	49 (40.83)	6 (5)	65 (54.17)	< 0.001	50 (41.67)	7 (5.83)	63 (52.5)	< 0.001

[a]The *p*-Value of chi-squared test where observed frequency was tested with expected equal distribution in three categories and significant *p*-Value indicates that occurrence was not by chance.

no agreement between the residents (agreement in 32.24%; Kappa = –0.376 (95% CI: –0.497 to –0.254).

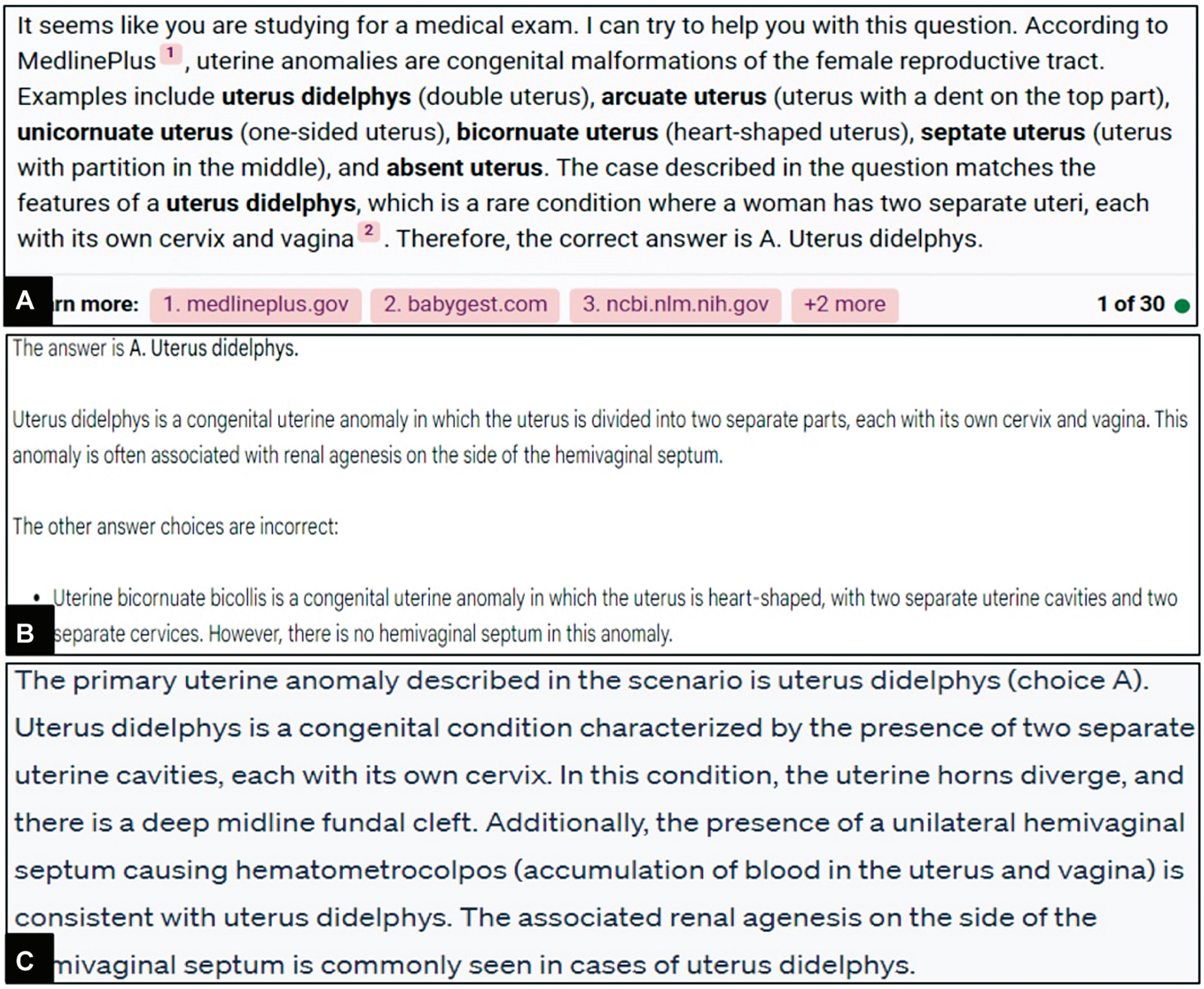
Discussion

The finding that the residents outperformed the AI models in terms of accuracy may be attributed to several factors. The residents possess a greater understanding of the questions due to their human cognitive abilities, contextual knowledge, and commonsense reasoning, which the AI models may struggle to replicate.[13] Additionally, the training data used for the AI models may have been limited or biased, hindering their ability to capture the nuances of the questions and topics accurately. The architecture and capabilities of the AI models, particularly Bard and ChatGPT, might have been insufficient to handle the complexity of the questions effectively. Language and linguistic challenges, such as idiomatic expressions or cultural references, could pose difficulties for the AI models in accurately comprehending and responding to the questions, whereas the residents' human intuition and linguistic skills give them an advantage.[14]

We found the explanation of the AI in supporting the answers was accurate in less than 50% of the cases. The AI models may have inherent limitations in understanding the specific nuances, details, or contextual elements of the topics being evaluated. Consequently, their responses could lack accuracy or consistency compared to human raters. Additionally, the topics themselves might have been inherently ambiguous or complex, making it challenging for the AI models to provide consistently accurate responses.[15] This complexity could be due to subjective elements, interpretational variations, or a lack of specific guidelines for determining accuracy.

In some of the cases, all the three AI models could point out the answers but residents could not. Such a question is available in ►**Table 3** and the answers of the generative AI models are partially presented in ►**Fig. 3**. In contrast, in some instances, the residents answered the question correctly, while all three AI models failed. ►**Table 3** (second row) shows an example question and the answers of the AI models are partially shown in ►**Fig. 4**.

In evaluating the performance of three language models in answering the FRCR2A single best answer questions, notable differences were observed.[16] Bing demonstrated a comprehensive approach by highlighting important findings in bold letters, providing in-text citations and references, and offering relevant images for better understanding. It exhibited a thorough analysis and occasionally provided multiple options as answers. However, Bing's responses took slightly longer compared to the other two LLMs. On the other hand,



It seems like you are studying for a medical exam. I can try to help you with this question. According to MedlinePlus [1], uterine anomalies are congenital malformations of the female reproductive tract. Examples include **uterus didelphys** (double uterus), **arcuate uterus** (uterus with a dent on the top part), **unicornuate uterus** (one-sided uterus), **bicornuate uterus** (heart-shaped uterus), **septate uterus** (uterus with partition in the middle), and **absent uterus**. The case described in the question matches the features of a **uterus didelphys**, which is a rare condition where a woman has two separate uteri, each with its own cervix and vagina [2]. Therefore, the correct answer is A. Uterus didelphys.

A

rn more: 1. medlineplus.gov 2. babygest.com 3. ncbi.nlm.nih.gov +2 more **1 of 30**

The answer is **A. Uterus didelphys.**

Uterus didelphys is a congenital uterine anomaly in which the uterus is divided into two separate parts, each with its own cervix and vagina. This anomaly is often associated with renal agenesis on the side of the hemivaginal septum.

The other answer choices are incorrect:

- Uterine bicornuate bicollis is a congenital uterine anomaly in which the uterus is heart-shaped, with two separate uterine cavities and two separate cervices. However, there is no hemivaginal septum in this anomaly.

B

The primary uterine anomaly described in the scenario is uterus didelphys (choice A). Uterus didelphys is a congenital condition characterized by the presence of two separate uterine cavities, each with its own cervix. In this condition, the uterine horns diverge, and there is a deep midline fundal cleft. Additionally, the presence of a unilateral hemivaginal septum causing hematometrocolpos (accumulation of blood in the uterus and vagina) is consistent with uterus didelphys. The associated renal agenesis on the side of the mivaginal septum is commonly seen in cases of uterus didelphys.

C

Fig. 3 Examples of a question that three artificial intelligence (AI) models—(**A**) Bard, (**B**) Bing, and (**C**) ChatGPT—answered correctly and both residents answered incorrectly.

Table 3 Example of two cases, related question, and response options

Case	Question	Answer option	Correct answer
A 14-year-old adolescent girl comes in with complaint of recurring pelvic pain that occurs in a cyclic pattern. During a speculum vaginal examination, a protruding vaginal mass is observed. An MRI of the pelvic region reveals a primary uterine anomaly characterized by divergent uterine horns with a prominent midline fundal cleft, two distinct uterine cavities, two separate cervices, and a hemivaginal septum on one side causing hematometrocolpos. Renal agenesis is also present on the same side as the hemivaginal septum	What is the primary uterine anomaly?	A. Uterus didelphys B. Uterine bicornuate bicollis C. Septate uterus D. Arcuate uterus E. Imperforate hymen	A
A 28-year-old male reports a previous incident of his patella dislocating and then spontaneously relocating during a football game. An MRI of the knee has been recommended.	Identify the MRI findings that align with the reported clinical history of patellar dislocation.	A. Bone oedema involving medial facet of patella and medial femoral condyle. B. Bone oedema involving posterior patella and anterior aspect of the tibial plateau. C. Bone oedema involving the lateral facet of patella and lateral femoral condyle. D. Bone oedema involving the lateral facet of patella and medial femoral condyle. E. Bone oedema involving the medial facet of patella and lateral femoral condyle.	E

Abbreviations: MRI, magnetic resonance imaging; STIR, short tau inversion recovery.

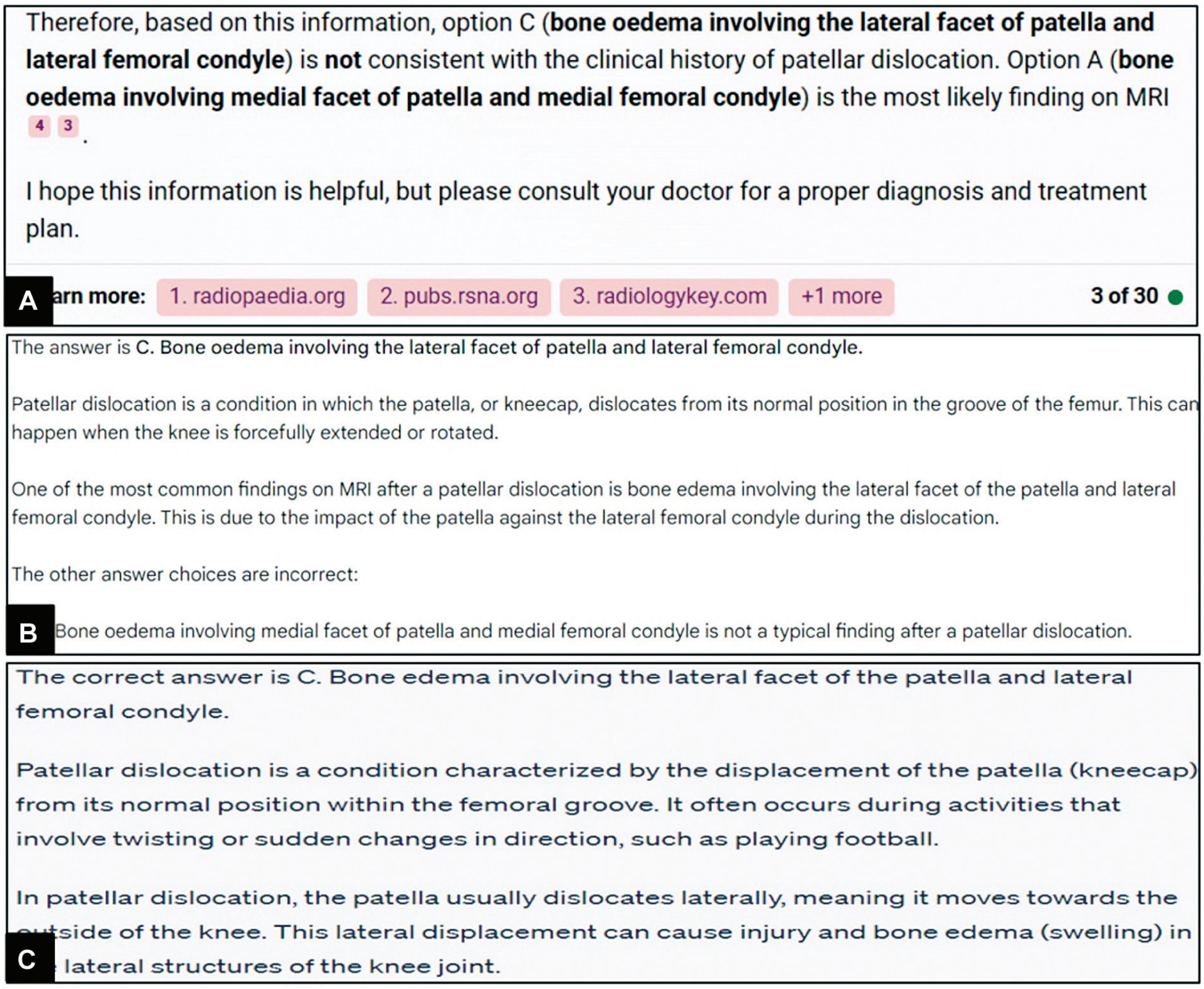


Fig. 4 Example of a question that both residents answered correctly and three artificial intelligence (AI) models—(**A**) Bard, (**B**) Bing, and (**C**) ChatGPT—answered incorrectly.

Bard sometimes acknowledged its limitations as a language model and admitted its inability to process certain questions. It occasionally provided irrelevant responses; however, it occasionally presented descriptions in a user-friendly tabular format.[17] As for ChatGPT, it stood out for its prompt responses and was observed to analyze options sequentially. Overall, among the three LLMs, Bing exhibited the highest success rate in answering the FRCR2A single best answer questions.

This study contributes to the field by examining the capability of three generative AI models in solving the FRCR2A type questions. It highlights the performance differences between AI models and sheds light on the residents' superior accuracy rates.[18,19] The analysis of response patterns and agreement metrics provides valuable insights into the strengths and limitations of both AI and human raters in this particular domain.

There are a few limitations to consider in this study. The sample size of residents and generative AI models may be limited, potentially affecting the generalizability of the findings. Additionally, the study focused on a specific topic or set of questions, which may not represent the entire range of tasks or topics AI models and residents could encounter. Furthermore, we did not explore the underlying causes of the agreement or disagreement among the AI models and residents as it was beyond our capabilities.

Future research could explore the factors that contribute to the residents' higher accuracy rates, such as their cognitive processes, decision-making strategies, or domain-specific knowledge. Investigating the specific challenges that AI models face in this context could guide improvements in their training methodologies, data selection, or algorithmic enhancements.[20] Moreover, expanding the study to include a larger and more diverse sample of residents and AI models would help in assessing the generalizability of the findings across different populations and AI systems.

Conclusion

The residents consistently outperformed the AI models in terms of accuracy. While Bing exhibited the highest accuracy among the AI models, all three models had lower accuracy rates compared to the residents. These findings underscore the ongoing need for human expertise in providing accurate and reliable responses. Although AI models have potential, they currently lack the comprehension and accuracy levels of human professionals. Advancements in NLP and AI algorithms are crucial to enhance the capabilities of AI models in question answering tasks. This study sheds light on the existing limitations and highlights the importance of continued research and development in this field.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author(s) used ChatGPT (Free Research Preview May 24 Version) in order to edit the grammar and language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Funding
None.

Conflict of Interest
None declared.

References

1 Rathan R, Hamdy H, Kassab SE, Salama MNF, Sreejith A, Gopakumar A. Implications of introducing case based radiological images in anatomy on teaching, learning and assessment of medical students: a mixed-methods study. BMC Med Educ 2022;22(01):723
2 Meng F, Kottlors J, Shahzad R, et al. AI support for accurate and fast radiological diagnosis of COVID-19: an international multicenter, multivendor CT study. Eur Radiol 2023;33(06):4280–4291
3 Nijiati M, Zhang Z, Abulizi A, et al. Deep learning assistance for tuberculosis diagnosis with chest radiography in low-resource settings. J XRay Sci Technol 2021;29(05):785–796
4 Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. J Digit Imaging 2017;30(04):449–459
5 Esmaeili M, Vettukattil R, Banitalebi H, Krogh NR, Geitung JT. Explainable artificial intelligence for human-machine interaction in brain tumor localization. J Pers Med 2021;11(11):1213
6 Sinha RK, Deb Roy A, Kumar N, Mondal H. Applicability of ChatGPT in assisting to solve higher order problems in pathology. Cureus 2023;15(02):e35237
7 Vaishya R, Misra A, Vaish A. ChatGPT: is this version good for healthcare and research? Diabetes Metab Syndr 2023;17(04):102744
8 Korb KB, Nyberg EP, Oshni Alvandi A, et al. Individuals vs. BARD: experimental evaluation of an online system for structured, collaborative bayesian reasoning. Front Psychol 2020;11:1054
9 Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology 2023;307(05):e230922
10 Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023;307(05):e230582
11 Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: improvements in advanced reasoning. Radiology 2023;307(05):e230987
12 McCoubrie P, McKnight L. Single best answer MCQs: a new format for the FRCR part 2a exam. Clin Radiol 2008;63(05):506–510
13 Surry LT, Torre D, Trowbridge RL, Durning SJ. A mixed-methods exploration of cognitive dispositions to respond and clinical reasoning errors with multiple choice questions. BMC Med Educ 2018;18(01):277
14 Tilmatine M, Hubers F, Hintz F. Exploring individual differences in recognizing idiomatic expressions in context. J Cogn 2021;4(01):37
15 Khan B, Fatima H, Qureshi A, et al. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. Biomed Mater Devices 2023:1–8
16 Agarwal M, Sharma P, Goswami A. Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. Cureus 2023;15(06):e40977
17 Williams MC, Shambrook J. How will artificial intelligence transform cardiovascular computed tomography? A conversation with an AI model. J Cardiovasc Comput Tomogr 2023;17(04):281–283
18 Korteling JEH, van de Boer-Visschedijk GC, Blankendaal RAM, Boonekamp RC, Eikelboom AR. Human- versus artificial intelligence. Front Artif Intell 2021;4:622364
19 Booth TC, Martins RDM, McKnight L, Courtney K, Malliwal R. The Fellowship of the Royal College of Radiologists (FRCR) examination: a review of the evidence. Clin Radiol 2018;73(12):992–998
20 Ferrell B, Raskin SE, Zimmerman EB. Calibrating a transformer-based model's confidence on community-engaged research studies: decision support evaluation study. JMIR Form Res 2023;7:e41516

