Article published online: 2024-11-04

Evaluating ChatGPT-4's Performance in Identifying Radiological Anatomy in FRCR Part 1 Examination Questions

Pradosh Kumar Sarangi[1] Suvrankar Datta[2] Braja Behari Panda[3] Swaha Panda[4] Himel Mondal[5]

[1] Department of Radiodiagnosis, All India Institute of Medical Sciences, Deoghar, Jharkhand, India
[2] Department of Radiodiagnosis, All India Institute of Medical Sciences, New Delhi, India
[3] Department of Radiodiagnosis, Veer Surendra Sai Institute of Medical Sciences and Research, Burla, Odisha, India
[4] Department of Otorhinolaryngology and Head and Neck Surgery, All India Institute of Medical Sciences, Deoghar, Jharkhand, India
[5] Department of Physiology, All India Institute of Medical Sciences, Deoghar, Jharkhand, India

Address for correspondence Pradosh Kumar Sarangi, MD, PDF, EDiR, Department of Radiodiagnosis, All India Institute of Medical Sciences, Deoghar 814152, Jharkhand, India (e-mail: drpksarangi@gmail.com).

Indian J Radiol Imaging 2025;35:287–294.

Abstract

Background Radiology is critical for diagnosis and patient care, relying heavily on accurate image interpretation. Recent advancements in artificial intelligence (AI) and natural language processing (NLP) have raised interest in the potential of AI models to support radiologists, although robust research on AI performance in this field is still emerging.

Objective This study aimed to assess the efficacy of ChatGPT-4 in answering radiological anatomy questions similar to those in the Fellowship of the Royal College of Radiologists (FRCR) Part 1 Anatomy examination.

Materials and Methods We used 100 mock radiological anatomy questions from a free Web site patterned after the FRCR Part 1 Anatomy examination. ChatGPT-4 was tested under two conditions: with and without context regarding the examination instructions and question format. The main query posed was: "Identify the structure indicated by the arrow(s)." Responses were evaluated against correct answers, and two expert radiologists (>5 and 30 years of experience in radiology diagnostics and academics) rated the explanation of the answers. We calculated four scores: correctness, sidedness, modality identification, and approximation. The latter considers partial correctness if the identified structure is present but not the focus of the question.

Results Both testing conditions saw ChatGPT-4 underperform, with correctness scores of 4 and 7.5% for no context and with context, respectively. However, it identified the imaging modality with 100% accuracy. The model scored over 50% on the approximation metric, where it identified present structures not indicated by the arrow. However, it struggled with identifying the correct side of the structure, scoring approximately 42 and 40% in the no context and with context settings, respectively. Only 32% of the responses were similar across the two settings.

Keywords
- artificial intelligence
- ChatGPT-4
- large language model
- radiology
- FRCR
- anatomy
- fellowship

article published online
November 4, 2024

DOI https://doi.org/10.1055/s-0044-1792040.
ISSN 0971-3026.

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

Conclusion Despite its ability to correctly recognize the imaging modality, ChatGPT-4 has significant limitations in interpreting normal radiological anatomy. This indicates the necessity for enhanced training in normal anatomy to better interpret abnormal radiological images. Identifying the correct side of structures in radiological images also remains a challenge for ChatGPT-4.

Introduction

Anatomical knowledge is the cornerstone of radiology, serving as the essential framework for accurate image interpretation and diagnosis. Radiologists rely on a deep understanding of anatomical structures to distinguish between normal and pathological findings on various imaging modalities such as X-rays, computed tomography (CT) scans, magnetic resonance imaging (MRI), and ultrasound. Precise anatomical knowledge enables radiologists to identify the location, extent, and nature of abnormalities, which is critical for guiding clinical management and therapeutic interventions. Furthermore, a thorough grasp of anatomy helps in communicating findings effectively with other health care professionals, ensuring that patients receive the most appropriate and timely care.[1,2]

As artificial intelligence (AI) continues to evolve, its applications in medical education and examination preparation have expanded significantly. One such AI application, OpenAI's ChatGPT-4, has demonstrated impressive capabilities in natural language understanding and generation.[3] The capabilities of ChatGPT have been explored across various medical fields, demonstrating its potential as a versatile tool in health care.[4–7] In medical education, ChatGPT has been utilized to assist in teaching complex subjects, providing explanations, answering student queries, and generating practice questions.[8] In clinical practice, it has been employed to aid in diagnostic processes, offering differential diagnoses, and suggesting possible next steps based on patient symptoms and medical history. Additionally, ChatGPT has been tested in patient communication, where it helps in simplifying medical jargon, providing health information, and supporting patient engagement and adherence to treatment plans.[9] Research has also investigated its use in medical writing, including drafting research papers, summarizing articles, and even assisting in the systematic review process.[10–12]

Extensive research[13–19] has explored the applications of ChatGPT and other large language models (LLMs) in radiology, revealing promising innovations. These include supporting medical writing and research, organizing radiology reports, protocoling examinations, recommending screening procedures, addressing patient inquiries, simulating text-based radiology board examinations, offering differential diagnoses based on imaging patterns, providing impressions, and suggesting follow-up imaging in accordance with established guidelines.

The First Fellowship of the Royal College of Radiologists (FRCR) Part 1 examination[20] is a crucial milestone for radiology trainees, testing their knowledge of radiological anatomy among other foundational topics. No previous study has assessed the capability of ChatGPT in identifying radiological anatomy as per the FRCR pattern. Hence, this study aims to evaluate the performance of ChatGPT-4 in identifying radiological anatomy, specifically in the context of FRCR Part 1 examination questions. By assessing the GPT-4's accuracy and reliability in this domain, we can explore its potential as a supplementary tool for radiology trainees.

Materials and Methods

Type and Settings

This study employed a quantitative research design, conducted in a controlled setting using 100 mock radiological anatomy questions sourced from the Radiology Cafe Web site, specifically tailored to mirror the FRCR Part 1 Anatomy examination.[21] In the mock set, the questions are predominantly focused on cross-sectional imaging techniques such as CT and MRI, comprising 66% of the total. Radiographs account for 25% of the questions, while contrast-based imaging modalities like intravenous urography (IVU), barium studies, CT IVU, and CT enterography collectively make up 6%. Ultrasound questions constitute the remaining 3% of the set.

The evaluation was performed under two conditions: with context, where ChatGPT-4 was provided with detailed examination instructions and question formats, and without context, where the AI received only the primary query without additional guidance. The complete prompt with context and without context is provided in ►**Figs. 1** and **2**, respectively.

Questions

The questions used in the study were patterned after the FRCR Part 1 Anatomy examination and focused on identifying anatomical structures in radiological images. Each question posed the main query: "Identify the structure indicated by the arrow(s)." These questions were designed to assess the AI's ability to correctly recognize and name anatomical features.

Data Collection Method

ChatGPT-4 was prompted with each of the 100 mock questions twice, once with context and once without. The responses generated by the AI were recorded for subsequent evaluation. Two expert radiologists (PKS with >7 years and BBP with > 30 years of experience in radiology diagnostics and academics) independently reviewed and rated the

A

Act like a radiologist.
You will be presented with 100 questions, each accompanied by an image. Each image will have one or more arrows pointing to specific structures. Your task is to identify the structure(s) indicated by the arrows. Questions may ask you to identify the structure or describe its function, such as the nerve supply to a muscle or the drainage site of a gland's duct. Simply identifying the structure without addressing the question will result in a loss of one mark.
Radiology Exam Instructions:
- Answer all 100 questions.
- There is no negative marking.
- The images used demonstrate normal anatomy.
- Occasionally, minor age-related degenerative changes may be present on the image, but these will not be tested.
- Each question carries 2 marks, graded as follows:
- 0 marks: Incorrect answer
- 1 mark: Partially correct/less accurate answer
- 2 marks: Precise/accurate answer

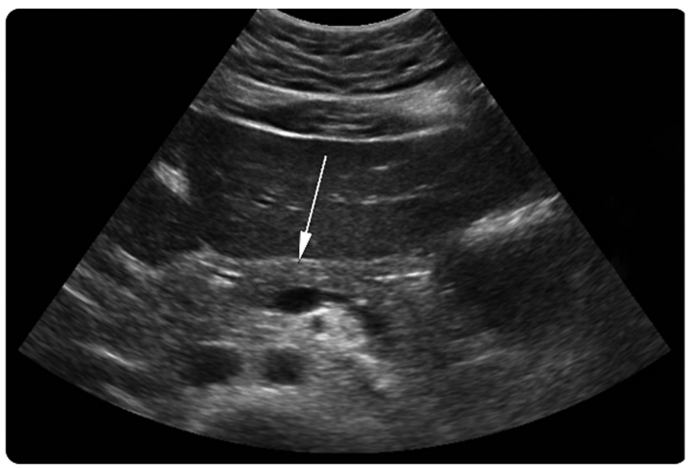
Be as specific as possible, and write a single answer. Always mention the side (e.g., left or right) wherever possible in your answer to avoid losing half a mark.
Example Question Format:
Question 1:
Identify the structure indicated by the arrow. [Image with arrow pointing to a structure]
Answer 1: The arrow is pointing to the left femoral head.

I'm ready to assist! Please provide the first question and its corresponding image, and I'll do my best to give you a precise and accurate answer.

Message ChatGPT

ChatGPT can make mistakes. Check important info.

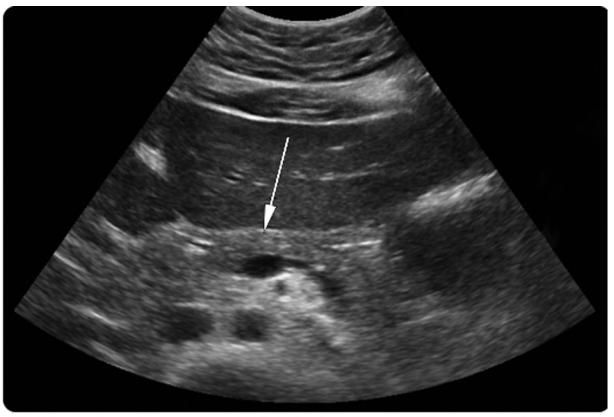
B



Identify the structure indicated by the arrow(s).

The arrow in the ultrasound image is pointing to the gallbladder.

Fig. 1 (**A**, **B**) Example of a prompt query including context for ChatGPT-4.



Identify the structure indicated by the arrow. Be as specific as possible, and write a single answer. Always mention the side (e.g., left or right) in your answers wherever possible.

The structure indicated by the arrow in the ultrasound image is the gallbladder.

Fig. 2 Example of a prompt query without context for ChatGPT-4.

explanations provided by ChatGPT-4, ensuring the reliability and accuracy of the assessment.

Scoring Method

The scoring method was designed to evaluate ChatGPT-4's performance across several dimensions. Each question carried a maximum of 2 marks for the correct answer, graded as follows: 0 marks for an incorrect answer, 1 mark for a partially correct or less accurate answer, and 2 marks for a precise and accurate answer, leading to a total possible score of 200. Additionally, a sidedness score was included, awarding 1 point for correctly identifying the side of a structure and 0 points for an incorrect identification, with an NA designation if the structure was not side specific. The modality identification score was also considered, granting 1 point for correctly identifying the imaging modality and 0 points for incorrect identification, with NA used if the modality was not mentioned in ChatGPT's response. Furthermore, for responses that scored 0 in structure identification, an approximation score was calculated to assess partial correctness; 1 point was given if the identified structure was present in the image but not the focus of the question, and 0 points if the structure was not present at all. These multifaceted scoring criteria provided a comprehensive evaluation of ChatGPT-4's ability to identify radiological anatomy accurately.

It is important to note that the ChatGPT-4 in this study was not pretrained with specific commands or questions. Each inquiry was made within a single chat session, without starting new chat tabs for individual questions. Sometimes, ChatGPT fails to generate a response on the first attempt, requiring the "Regenerate" option to be clicked to obtain an answer.

Data Analysis

Descriptive statistics were used to summarize the scores, and comparative analysis was performed to evaluate the differences between the context and no-context conditions. We used Microsoft Excel 2021 for the data analysis.

Results

When evaluated without context, ChatGPT-4 achieved a correct answer score of 8 out of 200 possible marks, representing 4% accuracy. In terms of sidedness, it correctly identified the side of the anatomical structure in 26 out of 62 cases, yielding a score of 41.93%. We did not specifically ask ChatGPT to identify modality in the prompt. However, out of 46 instances, it responded with modality, and all responses were 100% correct regarding modality. For approximation, where partial correctness was considered, ChatGPT-4 scored 54 out of 96, equating to 56.25%.

With context, ChatGPT-4's performance improved slightly. The correct answer score increased to 15 out of 200, corresponding to 7.5% accuracy. The sidedness identification saw a marginal decrease, with 25 out of 62 correct identifications, resulting in a score of 40.03%. ►**Fig. 3** shows an example where ChatGPT correctly answered about the anatomy but failed to identify the side, resulting in a score of 1 according to the FRCR evaluation criteria. The modality identification remained perfect, with all 37 instances correctly identified, maintaining a 100% score. For approximation, the score was 48 out of 90, which translates to 53.33%. ►**Table 1** tabulates the scores of ChatGPT-4 under two different testing conditions.

►**Fig. 4** presents the visual representation of the pass score and ChatGPT's score in answering the questions. In the assessment, the pass score was set at 80, but the obtained score was only 4. Hence, ChatGPT's score was far less than the passing score. With context considered, although ChatGPT obtained a score of 7.5, it is far below the pass score.

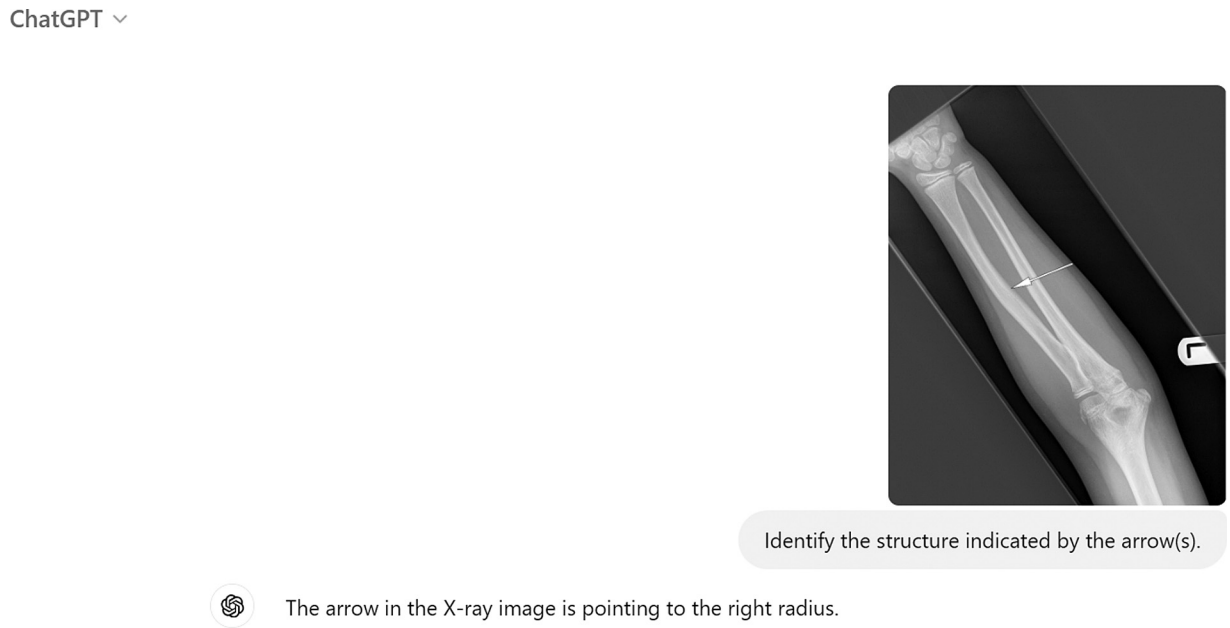


Fig. 3 An example where ChatGPT correctly answered about the anatomy but misidentified the side, scoring 1 out of 2.

Table 1 Scores of ChatGPT-4

Testing conditions	Correct answer score	Sidedness score	Modality identification score	Approximation score
ChatGPT without context	8/200 = 4%	26/62 = 41.93	46/46 = 100%	54/96 = 56.25%
ChatGPT with context	15/200 = 7.5%	25/62 = 40.03	37/37 =100%	48/90 =53.33%

Discussion

This study aimed to evaluate the efficacy of ChatGPT-4 in identifying radiological anatomy in the context of the FRCR Part 1 Anatomy examination. Our findings indicate that ChatGPT-4 demonstrates a modest improvement in performance when provided with contextual information. Specifically, the correctness scores were 7.5% with context and 4% without context. Notably, ChatGPT-4 showed perfect accuracy in identifying the imaging modality in both conditions. However, the AI struggled with identifying the correct side of the anatomical structures, with scores of approximately 40% in both settings, and had varying degrees of success in identifying related structures (approximation scores of 53.33% with context and 56.25% without context). The paradoxical response with context with response to approximation score might be caused by the additional information creating noise or confusion, leading to less accurate identification of related structures. The presence of context may introduce irrelevant details or ambiguity, hindering the AI's ability to focus on the key aspects necessary for accurate identification.[22]

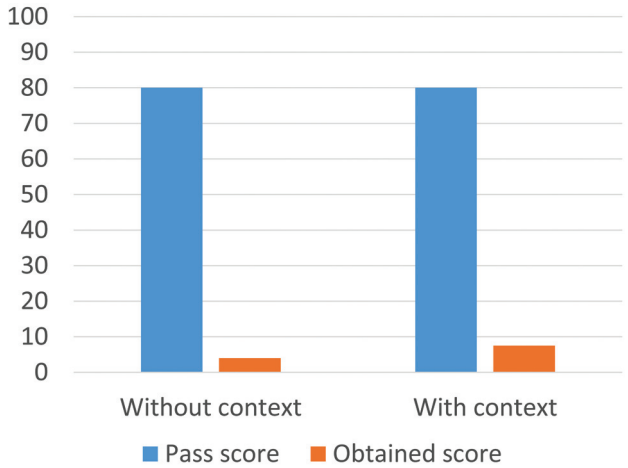


Fig. 4 Visual representation of the pass score versus ChatGPT's score on the questions.

The slight improvement in performance with context can be attributed to the additional guidance provided by the examination instructions and question formats. This suggests that ChatGPT-4 benefits from structured input, which helps it understand the specific requirements of the questions more clearly. The AI's consistent accuracy in modality identification underscores its capability to recognize imaging types effectively, likely due to comprehensive training on diverse medical datasets that include modality information. However, the challenges in identifying the correct side and partial correctness indicate areas where further refinement and targeted training are necessary. These findings suggest that while ChatGPT-4 can be a useful tool for supporting radiology education, its current limitations must be addressed to improve its diagnostic utility.

The consistent performance in modality identification suggests that ChatGPT-4 has a strong inherent capability to recognize imaging types, likely due to the extensive training on medical datasets that include modality information. The similarity in sidedness scores indicates that context had minimal impact on the AI's ability to determine the side of anatomical structures, which might suggest that this aspect relies more on inherent image interpretation skills than on additional instructions.

We recognize that the higher proportion of complex cross-sectional imaging anatomy (66%) may have impacted the overall results. However, we intentionally selected this distribution to mirror the actual composition of the FRCR Part 1 examination, where candidates are increasingly expected to interpret complex imaging. While this may have contributed to the lower performance observed, it also underscores the need to develop more advanced AI algorithms and vision models capable of handling the full spectrum of anatomical complexity encountered in real-world examinations across various modalities. Our study's findings align with and contrast various similar research efforts in the application of AI in medical fields. While specialized AI algorithms like CheXNet[23] have demonstrated high diagnostic accuracy comparable to radiologists, particularly due to their focused training on specific datasets, our results with ChatGPT-4, which is an LLM, reveal a lower accuracy in anatomical structure identification even with contextual information. This discrepancy underscores the difference between specialized diagnostic tools and general-purpose language models.

Many authors have conducted studies on pretrained ChatGPT-4 and align with our study findings.[24–30] GPT-4V's limited effectiveness in interpreting real-world chest radiographs, as found by Zhou et al, underscores the need for ongoing development to achieve dependable performance in radiology diagnostics.[24]

Another relevant study by Brin et al[25] assessed GPT-4V's performance in radiological image analysis and found significant variability across different imaging modalities and anatomical regions. This variability is consistent with our findings that ChatGPT-4, a general-purpose AI, may not yet be reliable for specialized radiological tasks. Furthermore, Hirano et al[26] reported no significant enhancement in accuracy when comparing GPT-4 Turbo with Vision to text-only GPT-4 Turbo in the context of the Japan Diagnostic Radiology Board Examination. These findings reinforce the need for targeted AI training to enhance performance in specific medical applications.

Additionally, Bera et al[27] highlighted the challenges in the reproducibility of ChatGPT-4's responses in a radiology board–style examination, which poses reliability issues despite the model's promising capabilities. Sultan et al[28] discussed the potential of ChatGPT-4 in ultrasound image analysis, showing that while it can aid in diagnostics, its performance still requires improvement for clinical reliability. Senkaiahliyan et al[29] evaluated GPT-4V's suitability for clinical care and education, concluding that its diagnostic accuracy and decision-making abilities are currently insufficient for safe clinical use.

Finally, Kelly et al[30] highlighted the potential of GPT-4V to disrupt AI radiology research while cautioning against its current limitations, such as misclassified cases and overly cautious nonanswers. This aligns with our observation that ChatGPT-4 shows potential as a supplementary tool in medical education, but its general-purpose nature and broad training limit its accuracy in specialized tasks compared with highly specialized AI systems.

Recent study by Wu et al[31] needs a special mention. They demonstrated that integrating ChatGPT-4 with image-to-text models improves consistency and diagnostic accuracy, highlighting a potential avenue for enhancing AI performance in medical imaging. They developed a generative image-to-text model with multitask learning classification to interpret American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS) features, based on a large dataset of over 20,000 thyroid nodule images with pathologic diagnoses. The model employs an encoder and a Multigate Mixture of Experts (MMoE) framework. Their approach involved three strategies: (1) human–LLM interaction, where human readers initially interpreted images, followed by an LLM diagnosis based on recorded TI-RADS features; (2) image-to-text–LLM, which utilized an image-to-text model for analysis prior to LLM diagnosis; and (3) a convolutional neural network (CNN) model for end-to-end image analysis and diagnosis. The image-to-text–LLM strategy with ChatGPT-4.0 demonstrated performance comparable to the human–LLM interaction involving two senior readers and one junior reader, and it outperformed the human–LLM interaction involving only a junior reader. Although this approach was less efficient than the CNN method, the incorporation of LLMs enhanced interpretability and supported junior readers in diagnosing thyroid nodules.

The findings of this study have significant implications for medical education and practice. ChatGPT-4's ability to accurately identify imaging modalities suggests that it can be a valuable supplementary tool in radiology education, helping medical students understand different imaging types and their applications. However, the AI's limitations in anatomical structure identification and sidedness indicate that it should be used with caution and not as a sole diagnostic or reporting tool. Enhancing AI training with more focused and specialized datasets could improve its accuracy and reliability in medical settings, thereby increasing its utility as a supportive educational resource.

Several limitations of our study must be acknowledged. The relatively small sample size of 100 questions may limit the generalizability of our findings. Additionally, the use of mock examination questions, while beneficial for standardization, may not fully capture the complexity and variability of real-world clinical scenarios. We utilized a single publicly available LLM, GPT-4, through the ChatGPT-4 platform, which features multimodal capabilities. This general-purpose LLM is not specifically trained for medical contexts. GPT-4V, accessible via the ChatGPT interface, was the model used in our study. While medicine-specific multimodal LLMs, such as Google's Med-PaLM M and Microsoft's LLaVA-Med, have shown potential in early research, they

are not as widely accessible to the broader medical community as GPT-4.[32]

We recognize the concerns about the relevance of using professional examinations to assess LLMs. Professional examinations may not fully reflect the complexities of AI performance in real-world clinical settings.[33,34] Therefore, professional examinations should not be used as benchmarks for evaluating LLMs due to their brittle nature (where small changes in questions or the order of multiple-choice options can alter LLM responses) and their stochastic nature (repeated inquiries can produce different outcomes). The primary objective was to assess ChatGPT-4's performance in a controlled environment to establish a baseline understanding of its capabilities in radiological anatomy identification. We recognize that radiologists are required to pass professional examinations like the The American Board of Radiology (ABR) and FRCR to demonstrate their proficiency, and our study sought to replicate a similar testing mechanism for this purpose.

Additionally, it is not entirely accurate to compare the performance of an LLM on a practice set with the pass percentage of a real examination for several reasons. One key difference lies in the context and format between practice set versus real examination. Real examinations often contain more complex or nuanced questions that demand critical thinking, problem-solving, and interpretation, whereas practice sets may focus more on rote learning or specific types of questions. LLMs may excel in controlled settings, particularly if the questions resemble their training data. However, real examinations typically require the application of knowledge in unfamiliar contexts, which can be more challenging for the model. Therefore, further validation using real examination data is necessary to gain a more comprehensive assessment of the model's performance.

We recognize that data contamination from GPT-4's training set could potentially affect the results of our study.[33] We acknowledge that the Radiology Cafe Web site, from which we sourced the 100 mock radiological anatomy questions, may have been part of the training data, potentially influencing the model's performance. However, given the unexpectedly poor results from ChatGPT, it seems less likely that data contamination played a significant role in this case.

It is crucial to note that the stochastic nature of LLMs like GPT-4 can result in variability in responses across repeated inquiries.[33] In our study, we conducted the test twice—once with context and once without—to provide a preliminary evaluation of the model's performance. However, we acknowledge the importance of assessing reproducibility through multiple runs to ensure the reliability of our findings.

To overcome limitations and improve LLM performance in radiology-specific applications, strategies like grounding LLMs in external data, using prompt engineering, and implementing fine-tuning should be considered. These approaches are vital for creating effective radiology tools that fully leverage the potential of LLMs.[32] There needs to be further research exploring the potential of ChatGPT-4 and other AI models in interpreting larger and more diverse datasets, including real-world clinical scenarios. Investigating the impact of additional contextual information and refining training protocols could help overcome the limitations in this study. Moreover, developing specialized training datasets that focus on anatomical structure identification and sidedness could enhance the performance of AI models in these areas, especially if fine-tuned on large well-annotated datasets.

Conclusion

ChatGPT-4 consistently falls short of achieving the pass mark of 80% in both evaluation methods. Despite this, it demonstrates perfect accuracy in identifying the study modality, achieving a score of 100% in this aspect. Moreover, when assessing approximation, it manages to correctly identify structures present in the image more than 50% of the time. However, it sometimes misses the mark when it comes to correctly pinpointing these structures with arrow marks. Thus, while it excels in certain aspects of analysis, its overall performance in meeting the passing criteria remains a significant challenge.

Data Availability Statement
The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

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 Pathiraja F, Little D, Denison AR. Are radiologists the contemporary anatomists? Clin Radiol 2014;69(05):458–461
3 Open AI. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. Accessed July 14, 2024 at: https://openai.com/index/gpt-4/
4 Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388(13):1233–1239
5 Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Beyond ChatGPT: what does GPT-4 add to healthcare? The dawn of a new era. Cardiol J 2023;30(06):1018–1025
6 Mu Y, He D. The potential applications and challenges of ChatGPT in the medical field. Int J Gen Med 2024;17:817–826
7 Montazeri M, Galavi Z, Ahmadian L. What are the applications of ChatGPT in healthcare: gain or loss? Health Sci Rep 2024;7(02):e1878
8 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
9 Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023;25:e48568
10 Mondal H, Mondal S, Podder I. Using ChatGPT for writing articles for patients' education for dermatological diseases: a pilot study. Indian Dermatol Online J 2023;14(04):482–486
11 Doyal AS, Sender D, Nanda M, Serrano RA. ChatGPT and artificial intelligence in medical writing: concerns and ethical considerations. Cureus 2023;15(08):e43292

12 Ahaley SS, Pandey A, Juneja SK, Gupta TS, Vijayakumar S. ChatGPT in medical writing: a game-changer or a gimmick? Perspect Clin Res 2024;15(04):165–171
13 Bera K, O'Connor G, Jiang S, Tirumani SH, Ramaiya N. Analysis of ChatGPT publications in radiology: literature so far. Curr Probl Diagn Radiol 2024;53(02):215–225
14 Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2024;34(05):2817–2825
15 Sarangi PK, Lumbani A, Swarup MS, et al. Assessing ChatGPT's proficiency in simplifying radiological reports for healthcare professionals and patients. Cureus 2023;15(12):e50881
16 Sarangi PK, Narayan RK, Mohakud S, Vats A, Sahani D, Mondal H. Assessing the capability of ChatGPT, Google Bard, and Microsoft Bing in solving radiology case vignettes. Indian J Radiol Imaging 2023;34(02):276–282
17 Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J. Use of ChatGPT, GPT-4, and Bard to improve readability of ChatGPT's answers to common questions about lung cancer and lung cancer screening. AJR Am J Roentgenol 2023;221(05):701–704
18 Sarangi PK, Irodi A, Panda S, Nayak DSK, Mondal H. Radiological differential diagnoses based on cardiovascular and thoracic imaging patterns: perspectives of four large language models. Indian J Radiol Imaging 2023;34(02):269–275
19 Sarangi PK, Datta S, Swarup MS, et al. Radiologic decision-making for imaging in pulmonary embolism: accuracy and reliability of large language models: Bing, Claude, ChatGPT, and Perplexity. Indian J Radiol Imaging 2024;34(04):653–660
20 The Royal College of Radiologists. FRCR 1 (Radiology) - CR1. Accessed July 14, 2024 at: https://www.rcr.ac.uk/exams-training/rcr-exams/clinical-radiology-exams/frcr-part-1-radiology-cr1/
21 Radiology Cafe. Mock anatomy. Accessed July 14, 2024 at: https://www.radiologycafe.com/exams/mock-anatomy-exams/
22 Liu NF, Lin K, Hewitt J, et al. Lost in the middle: how language models use long contexts. CoRR abs/2307.03172. arXiv preprint arXiv:2307.03172. 2023:10
23 Rajpurkar P, Irvin J, Zhu K, et al. Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225. 2017
24 Zhou Y, Ong H, Kennedy P, et al. Evaluating GPT-V4 (GPT-4 with vision) on detection of radiologic findings on chest radiographs. Radiology 2024;311(02):e233270
25 Brin D, Sorin V, Barash Y, et al. Assessing GPT-4 multimodal performance in radiological image analysis. medRxiv. 2023:2023–11
26 Hirano Y, Hanaoka S, Nakao T, et al. GPT-4 Turbo with Vision fails to outperform text-only GPT-4 Turbo in the Japan Diagnostic Radiology Board Examination. Jpn J Radiol 2024;42(08):918–926
27 Bera K, Gupta A, Jiang S, et al. Assessing performance of multimodal ChatGPT-4 on an image based Radiology Board-style examination: an exploratory study. medRxiv. 2024:2024–01
28 Sultan LR, Mohamed MK, Andronikou S. ChatGPT-4: a breakthrough in ultrasound image analysis. Radiol Adv 2024;1(01):6
29 Senkaiahliyan S, Toma A, Ma J, et al. GPT-4V (ision) unsuitable for clinical care and education: a clinician-evaluated assessment. medRxiv. 2023:2023–11
30 Kelly BS, Duignan S, Mathur P, et al. Spot the difference: can ChatGPT4-vision transform radiology artificial intelligence? medRxiv. 2023:2023–11
31 Wu SH, Tong WJ, Li MD, et al. Collaborative enhancement of consistency and accuracy in US diagnosis of thyroid nodules using large language models. Radiology 2024;310(03):e232255
32 Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology 2024;310(01):e232756
33 Kim W. Seeing the unseen: advancing generative AI research in radiology. Radiology 2024;311(02):e240935
34 Narayanan A, Kapoor S. GPT-4 and professional benchmarks: the wrong answer to the wrong question. AI Snake Oil 2023. Accessed October 10, 2024 at: https://aisnakeoil.substack.com/p/gpt-4-and-professional-benchmarks

