

Introduction

Artificial intelligence (AI) has rapidly expanded within the field of medicine, with radiology being one of the most significantly impacted specialties. AI-powered models have demonstrated potential in improving diagnostic accuracy, increasing workflow efficiency, and supporting clinical decision-making. Recent developments in large language models (LLMs) have further expanded the potential applications of AI, particularly in areas like medical education.

Despite these advancements, limitations remain in the ability of current AI systems to interpret complex medical images. While many models demonstrate strong performance in text-based reasoning and knowledge recall, accurate image interpretation requires advanced pattern recognition and contextual understanding.

Evaluating AI performance on standardized medical examinations provides insight into both the strengths and limitations of these technologies. This study evaluates the performance of ChatGPT-4o on a standardized radiology training examination to further assess its potential role as a supportive tool in radiology education and clinical practice.

Aims and Objectives

- To determine the overall accuracy of ChatGPT-4o on a radiology training examination
- To compare the model's performance between image-dependent and image-independent questions.
- To analyze the model's performance across multiple radiology subspecialties.
- To identify current strengths and limitations of large language models in radiology-related tasks.

Methods

A total of 106 multiple-choice questions were obtained from the 2022 American College of Radiology (ACR) in-house training examination, which was available on the ACR website at the time of data collection. The exam included questions from multiple radiology subspecialties, including breast radiology, cardiac radiology, chest radiology, genitourinary radiology, musculoskeletal radiology, gastrointestinal radiology, neuroradiology, nuclear radiology, pediatric radiology, radiation physics, and ultrasound.

Each question was manually entered into ChatGPT-4o, and the model’s responses were recorded. Answers were then evaluated as either correct or incorrect based on the official answer key. Overall accuracy and subspecialty-specific accuracy were calculated.

To further analyze performance, the questions were categorized into two groups: image-dependent questions, which required interpretation of medical images, and image-independent questions, which could be answered using conceptual knowledge without image interpretation. Model performance between these categories was compared, and statistical analysis was performed using Fisher’s exact test.

Results

ChatGPT-4o achieved an overall accuracy of 69.8% (74/106). Accuracy was significantly higher for image-independent questions (95.2%, 40/42) compared to image-dependent questions (53.1%, 34/64; p < 0.001). Subspecialty performance varied, with the highest accuracy in ultrasound (90.0%) and radiation physics (88.9%), and the lowest in musculoskeletal (50.0%) and pediatric radiology (55.6%). Notably, the model achieved 100% accuracy on image-independent questions in nine subspecialties.

Subject	Correct	Incorrect	Total questions	Percent correct
Breast Radiology	6	4	10	60%
Cardiac Radiology	6	4	10	60%
Chest Radiology	7	3	10	70%
Genitourinary Radiology	8	2	10	80%
Musculoskeletal Radiology	5	5	10	50%
Gastrointestinal Radiology	5	3	8	62.50%
Neuroradiology	8	2	10	80%
Nuclear Radiology	7	3	10	70%
Pediatric Radiology	5	4	9	55.60%
Radiation Physics	8	1	9	88.90%
Ultrasound	9	1	10	90%
Total	74	32	106	69.80%

Table 1. Total performance of ChatGPT-4o on exam questions

Subject	Correct	Number of questions	Percent correct
Breast Radiology	1	5	20%
Cardiac Radiology	5	9	55.60%
Chest Radiology	3	6	50%
Genitourinary Radiology	7	9	77.80%
Musculoskeletal Radiology	3	7	42.90%
Gastrointestinal Radiology	3	6	50%
Neuroradiology	5	7	71.40%
Nuclear Radiology	2	4	50%
Pediatric Radiology	3	7	42.90%
Radiation Physics	1	2	50%
Ultrasound	1	2	50%
Total	34	64	53.10%

Table 2. ChatGPT-4o performance on image-dependent questions

Subject	Correct	Number of questions	Percent correct
Breast Radiology	5	5	100%
Cardiac Radiology	1	1	100%
Chest Radiology	4	4	100%
Genitourinary Radiology	1	1	100%
Musculoskeletal Radiology	2	3	66.70%
Gastrointestinal Radiology	2	2	100%
Neuroradiology	3	3	100%
Nuclear Radiology	5	6	83.30%
Pediatric Radiology	2	2	100%
Radiation Physics	7	7	100%
Ultrasound	8	8	100%
Total	40	42	95.20%

Table 3. ChatGPT-4o performance on image-independent questions

Conclusions

The results of this study demonstrate that ChatGPT-4o achieved an overall accuracy of 69.8% on a standardized radiology training examination. The model demonstrated significantly higher performance on image-independent questions, achieving an accuracy of 95.2%, compared to 53.1% accuracy on image-dependent questions.

These findings suggest that current large language models demonstrate strong capabilities in conceptual knowledge and text-based reasoning, but still face challenges when interpreting complex medical images.

Overall, ChatGPT-4o shows potential as a supportive tool for radiology education and knowledge-based tasks, while further advancements in AI development will be necessary to improve performance in medical image interpretation and diagnostic applications.

References

1.

Hameed BMZ, Prerepa G, Patil V, et al. Engineering and clinical use of artificial intelligence (AI) with machine learning and data science advancements: radiology leading the way for future. *Therapeutic Advances in Urology*. 2021;13. doi:10.1177/17562872211044880

2.

Alanazi MMF, Alghamdi A, Khan M, et al. Advancements in AI-driven diagnostic radiology: enhancing accuracy and efficiency. *Int J Health Sci (Qassim)*. 2021;5(S2):1402-1414. doi:10.53730/ijhs.v8nS1.14928

3.

Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics (Basel)*.2023;13(17):2760. doi:10.3390/diagnostics13172760

4.

Recht MP, Dewey M, Dreyer K, et al. Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. *Eur Radiol*. 2020;30:3576–3584. doi:10.1007/s00330-020-06672-5

5.

American College of Radiology. Diagnostic Radiology In-Training (DXIT) 2022 Exam Set. *American College of Radiology*. Published 2022. Accessed January 21, 2025. <https://www.acr.org/-/media/ACR/Files/DXIT-TXIT/DXIT-2022-Exam-Set.pdf>

6.

Payne DL, Purohit K, Borrero WM, et al. Performance of GPT-4 on the American College of Radiology In-Service Examination. *bioRxiv*. Published February 15, 2024. doi:10.1101/2024.02.15.580546

7.

Panayides AS, Amini A, Filipovic ND, et al. AI in medical imaging informatics: current challenges and future directions. *IEEE J Biomed Health Inform*. 2020;24(7):1837-1857. doi:10.1109/JBHI.2020.2991043

8.

Bhandari A. Revolutionizing radiology with artificial intelligence. *Cureus*. 2024;16(10):e72646. doi:10.7759/cureus.72646

9.

Tajmir SH, Alkasab TK. Toward augmented radiologists: changes in radiology education in the era of machine learning and artificial intelligence. *Acad Radiol*. 2018;25(6):747-750. doi:10.1016/j.acra.2018.03.007

10.

Sorantin E, Grasser MG, Hemmelmayr A, et al. The augmented radiologist: artificial intelligence in the practice of radiology. *Pediatr Radiol*. 2022;52:2074-2086. doi:10.1007/s00247-021-05177-7