

**Research Letter**

# Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties

Felix Busch<sup>1</sup>, MD; Tianyu Han<sup>2</sup>, PhD; Marcus R Makowski<sup>3</sup>, MD, PhD; Daniel Truhn<sup>2</sup>, MSc, MD; Keno K Bressem<sup>4\*</sup>, MD; Lisa Adams<sup>3\*</sup>, MD

<sup>1</sup>Department of Neuroradiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Berlin, Germany

<sup>2</sup>Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany

<sup>3</sup>Department of Diagnostic and Interventional Radiology, Klinikum rechts der Isar, Technical University Munich, Munich, Germany

<sup>4</sup>Institute for Radiology and Nuclear Medicine, German Heart Center Munich, Technical University of Munich, Munich, Germany

\*these authors contributed equally

**Corresponding Author:**

Felix Busch, MD

Department of Neuroradiology

Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin  
Charitépl. 1

Berlin, 10117

Germany

Phone: 49 3045050

Email: [felix.busch@charite.de](mailto:felix.busch@charite.de)

**Related Articles:**

This is a corrected version. See correction statement in: <https://www.jmir.org/2024/1/e64411>

This is a corrected version. See correction statement in: <https://www.jmir.org/2026/1/e91415>

## Abstract

This study demonstrates that GPT-4V outperforms GPT-4 across radiology subspecialties in analyzing 207 cases with 1312 images from the Radiological Society of North America Case Collection.

(*J Med Internet Res* 2024;26:e54948) doi: [10.2196/54948](https://doi.org/10.2196/54948)

**KEYWORDS**

GPT-4; ChatGPT; Generative Pre-Trained Transformer; multimodal large language models; artificial intelligence; AI applications in medicine; diagnostic radiology; clinical decision support systems; generative AI; medical image analysis

## Introduction

The launch of GPT-4 has generated significant interest in the scientific and medical communities, demonstrating its potential in medicine with notable achievements such as an 83.76% zero-shot accuracy on the United States Medical Licensing Examination (USMLE) [1]. In radiology, GPT has spanned text-based tasks, including board exam question scoring, data mining, and report structuring [2,3]. The recent release of GPT-4's visual capabilities (GPT-4V) enables the combined analysis of text and visual data [4]. Our study focuses on evaluating the diagnostic capabilities of GPT-4V by comparing it to GPT-4 in advanced radiological tasks, benchmarking the

potential of this multimodal large language model in the medical imaging field.

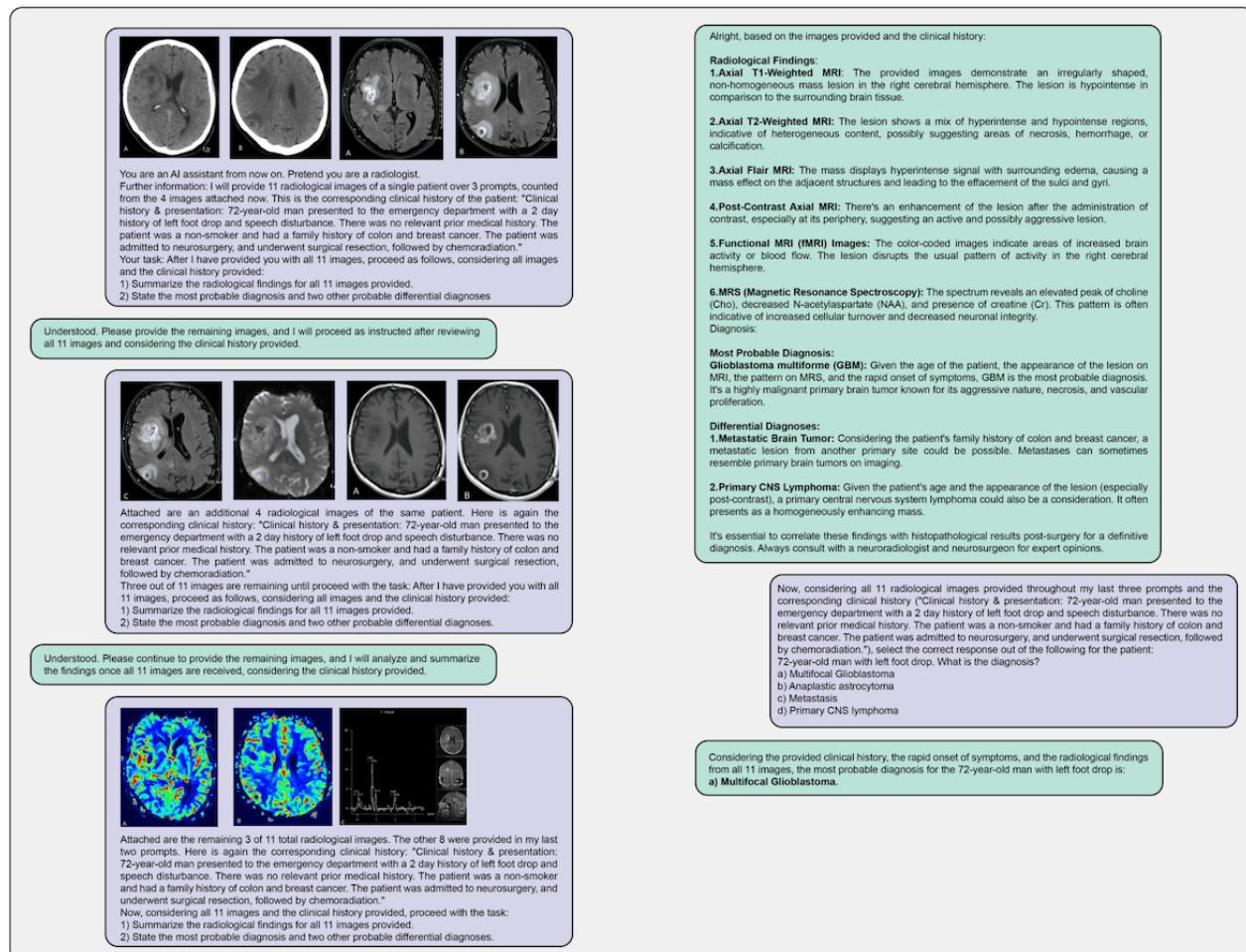
## Methods

We sourced 207 cases with 1312 images from the Radiological Society of North America (RSNA) Case Collection (accessible for RSNA members on the RSNA Case Collection website [5]), aiming to cover at least 10 cases for each of the 22 presented subspecialties. The cases within each subspecialty were chosen to present different pathologies. Each case had varying numbers of images and were usually labeled for more than 1 subspecialty, so that the total number of cases per subspecialty varied between 1 (for "Physics and Basic Science," no more than 1

case was available) and 43 (for “Gastrointestinal,” 10 cases in this category were chosen, with 33 additional cases from other subspecialties that were also labeled for “Gastrointestinal”).

GPT-4 and GPT-4V were accessed between November 6, 2023, and November 17, 2023. We utilized an application programming interface (API) account, which allowed us to use the models programmatically and ensure a consistent environment for each test. This access level was crucial, as it provided stable and repeatable interactions with the models, unlike what might be experienced with fluctuating conditions of regular account usage. The ground truth was established based on the final diagnoses stated in the RSNA case entries. We prompted each model 3 times via the API for the following two tasks: first, the models were asked to identify the diagnosis

**Figure 1.** An example conversation with GPT-4V showcasing the prompting style that was used if the question contained more than 4 images. Notably, GPT-4V often disregards the initial textual case description when additional image prompts are introduced, necessitating the description’s repetition for accurate responses. As the context window, image resolution limit, and embedding size limit of the web application version of GPT-4V are unknown, definitive conclusions regarding the cause of this forgetfulness could not be drawn. Nevertheless, the model’s consistent ability to recognize and recall images from the initial prompt (eg, axial FLAIR [fluid-attenuated inversion recovery] images) suggests that running out of context length is an unlikely explanation. Reproduced with permission from the Radiological Society of North America. Link to the displayed case: <https://cases.rsna.org/take-quiz/07c4b917-80fb-43c0-8b3b-59a0d8ceb203> (accessed 14th January 2026).



## Results

GPT-4 accurately identified the primary diagnosis in 18% (95% CI 12%-25%) of cases (first task). When including differential diagnoses, this accuracy increased to 28% (95% CI 22%-33%). In contrast, GPT-4V achieved a 27% (95% CI 21%-34%)

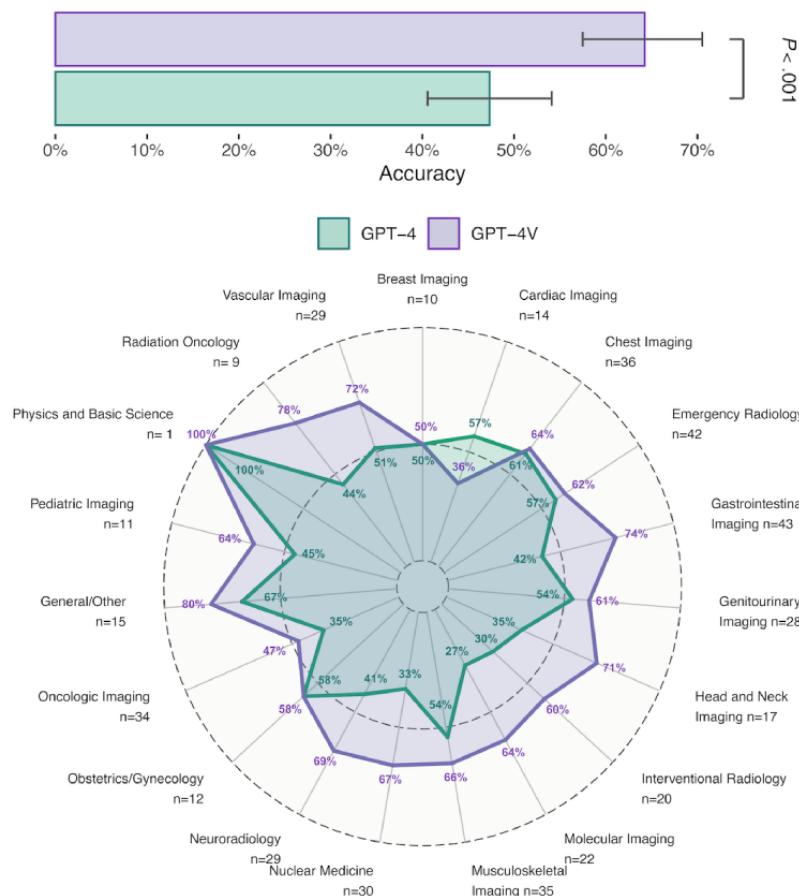
and 2 differentials (providing the patient history only for GPT-4 or patient history with images for GPT-4V); second, the models were asked to answer corresponding multiple-choice questions from the RSNA Case Collection. The GPT-4V assessment used a “chain-of-thought” prompt that guided the model through diagnostic reasoning (Figure 1), in contrast to the text-only assessment of GPT-4. For both tasks, a case was considered correctly diagnosed if the same correct result appeared for at least 2 of 3 prompts. Cases with no repeated correct diagnoses and cases with only false diagnoses across the 3 prompts were marked as incorrectly diagnosed. Mean accuracies and bootstrapped 95% CIs were calculated, and statistical significance was determined by using the McNemar test ( $P<.001$ ).

**Figure 1.** An example conversation with GPT-4V showcasing the prompting style that was used if the question contained more than 4 images. Notably, GPT-4V often disregards the initial textual case description when additional image prompts are introduced, necessitating the description’s repetition for accurate responses. As the context window, image resolution limit, and embedding size limit of the web application version of GPT-4V are unknown, definitive conclusions regarding the cause of this forgetfulness could not be drawn. Nevertheless, the model’s consistent ability to recognize and recall images from the initial prompt (eg, axial FLAIR [fluid-attenuated inversion recovery] images) suggests that running out of context length is an unlikely explanation. Reproduced with permission from the Radiological Society of North America. Link to the displayed case: <https://cases.rsna.org/take-quiz/07c4b917-80fb-43c0-8b3b-59a0d8ceb203> (accessed 14th January 2026).

64% (95% CI 59%-72%). The observed difference in performance was statistically significant ( $P<.001$ ). Across 15 subspecialties, GPT-4V outperformed GPT-4, with the sole

exception being in “Cardiac Imaging.” Figure 2 summarizes the accuracies across all subspecialties.

**Figure 2.** Comparison of GPT-4 and GPT-4V in various radiology subspecialties. Many cases spanned multiple subspecialties, and some subspecialties had very few cases. The number of images for individual cases ranged from 2 to 30, and the overall accuracy across all subspecialties, as shown in the bar plot, showed that GPT-4V performed significantly better than GPT-4. Error bars represent the 95% CIs. The radar plot shows the accuracy of GPT-4 (green line) and GPT-4V (purple line) across different radiology subspecialties. Each axis represents a specific radiology subspecialty, with the percentages indicating the accuracy of the model in that domain. Both models show varying levels of performance across subspecialties, with GPT-4V consistently performing better than GPT-4, except in “Cardiac Imaging” (cases: n=14; GPT-4V accuracy: 36%; GPT-4 accuracy: 57%). For “Physics and Basic Science” (cases: n=1), “Breast Imaging” (cases: n=10), and “Obstetrics/Gynecology” (cases: n=12), GPT-4V and GPT-4 showed on-par performance (accuracy: 100%, 50%, and 58%, respectively). Due to the small sample sizes in some categories, which ranged from 1 to 43 cases, these results should primarily be viewed as indicative trends rather than definitive conclusions about the models’ performance in these specific areas.



## Discussion

Our study shows that GPT-4V has improved performance over GPT-4 in solving complex radiological problems, indicating its potential to detect pathological features in medical images and thus its radiological domain knowledge. The RSNA Case Collection, which is aimed at expert-level professional radiologists, highlights the promise of GPT-4V in specialized medical contexts.

However, the use of GPT-4V warrants a cautious approach. At this time, it should be considered, at best, as a supplemental tool to augment—not replace—the comprehensive analyses performed by trained medical professionals.

Extending the initial research by Yang et al [6], our study explores the medical image analysis capabilities of GPT-4V in more complex scenarios and with a wider range of cases. The ongoing development of multimodal models, such as

Med-Flamingo, for medical applications signals a growing interest in this area [7].

One challenge is the scarcity of specialized medical data sets. As our study used RSNA member-exclusive cases, it was unlikely that these cases were in GPT-4V’s training data; thus, the risk of data contamination was minimized. However, the corresponding images for each case were indented to highlight specific pathologies, and this does not fully replicate clinical practice, where one would have to analyze each separate image to identify potential pathologies—a task that specialized deep learning models would be better suited to perform.

Future efforts should focus on detailed performance comparisons between generalist models (like GPT-4V) and emerging, radiological domain-specialized, artificial intelligence diagnostic models to clarify the clinical relevance and applicability of generalist models in clinical practice.

Our results encourage conducting further performance evaluations of multimodal models in different radiologic subdisciplines, as well as using larger data sets, to gain a more holistic understanding of their role in radiology.

## Data Availability

The cases analyzed in this study are available from the Radiological Society of North America (RSNA) Case Collection. This repository can be accessed by RSNA members on the RSNA Case Collection website [5], where each case is presented with detailed clinical information, imaging data, questions, multiple-choice answers, and diagnostic conclusions. The cases from the RSNA Case Collection were reproduced with permission from the Radiological Society of North America. These cases were not used for model training, nor were they retained by any tools or systems employed in this study. No additional unpublished data from these cases were utilized in this study. Researchers and readers are encouraged to directly access the RSNA Case Collection for further information.

## Conflicts of Interest

None declared.

## References

1. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online on Apr 12, 2023. [FREE Full text] [doi: [10.48550/arXiv.2303.13375](https://doi.org/10.48550/arXiv.2303.13375)]
2. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology*. Jun 2023;307(5):e230582. [doi: [10.1148/radiol.230582](https://doi.org/10.1148/radiol.230582)] [Medline: [37191485](https://pubmed.ncbi.nlm.nih.gov/37191485/)]
3. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology*. May 2023;307(4):e230725. [doi: [10.1148/radiol.230725](https://doi.org/10.1148/radiol.230725)] [Medline: [37014240](https://pubmed.ncbi.nlm.nih.gov/37014240/)]
4. GPT-4V(ision) system card. OpenAI. Sep 25, 2023. URL: <https://openai.com/research/gpt-4v-system-card> [accessed 2023-10-14]
5. RSNA Case Collection. Radiological Society of North America. URL: <https://cases.rsna.org/> [accessed 2024-04-24]
6. Yang Z, Li L, Lin K, Wang J, Lin CC, Liu Z, et al. The dawn of LMMs: preliminary explorations with GPT-4V(ision). arXiv. Preprint posted online on Oct 11, 2023. [FREE Full text] [doi: [10.48550/arXiv.2309.17421](https://doi.org/10.48550/arXiv.2309.17421)]
7. Moor M, Huang Q, Wu S, Yasunaga M, Zakka C, Dalmia Y, et al. Med-Flamingo: a multimodal medical few-shot learner. arXiv. Preprint posted online on Jul 27, 2023. [FREE Full text] [doi: [10.48550/arXiv.2307.15189](https://doi.org/10.48550/arXiv.2307.15189)]

## Abbreviations

**API:** application programming interface

**RSNA:** Radiological Society of North America

**USMLE:** United States Medical Licensing Examination

*Edited by G Eysenbach; submitted 28.Nov.2023; peer-reviewed by L Zhu, S Kommireddy, H Younes; comments to author 06.Feb.2024; revised version received 10.Feb.2024; accepted 20.Mar.2024; published 01.May.2024*

*Please cite as:*

*Busch F, Han T, Makowski MR, Truhn D, Bressem KK, Adams L*

*Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties*  
*J Med Internet Res* 2024;26:e54948

*URL: <https://www.jmir.org/2024/1/e54948>*

*doi: [10.2196/54948](https://doi.org/10.2196/54948)*

*PMID: [38691404](https://pubmed.ncbi.nlm.nih.gov/38691404/)*

©Felix Busch, Tianyu Han, Marcus R Makowski, Daniel Truhn, Keno K Bressem, Lisa Adams. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 01.May.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.