Assessing GPT-4 Multimodal Performance in Radiological Image Analysis

Dana Brin, MD[1,2]; Vera Sorin, MD[1-3]; Yiftach Barash, MD[1-3]; Eli Konen, MD[1,2]; Girish Nadkarni, MD[4-5]; Benjamin S Glicksberg, MD[6]; Eyal Klang, MD[1-5]

1. Department of Diagnostic Imaging, Chaim Sheba Medical Center, Tel Hashomer, Israel
2. Faculty of Medicine, Tel-Aviv University, Israel
3. DeepVision Lab, Chaim Sheba Medical Center, Tel Hashomer, Israel
4. Division of Data-Driven and Digital Medicine (D3M), Icahn School of Medicine at Mount Sinai, New York, New York, USA
5. The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA.
6. Hasso Plattner Institute for Digital Health, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

Abstract

The integration of Artificial Intelligence (AI) in radiology presents opportunities to enhance diagnostic processes. As of recently, OpenAI's multimodal GPT-4 can analyze both images and textual data (GPT-4V). This study evaluates GPT-4V's performance in interpreting radiological images across a variety of modalities, anatomical regions, and pathologies. Fifty-two anonymized diagnostic images were analyzed using GPT-4V, and the results were compared with board-certified radiologists interpretations. GPT-4V correctly recognized the imaging modality in all cases. The model's performance in identifying pathologies and anatomical regions was inconsistent and varied between modalities and anatomical regions. Overall accuracy for anatomical region identification was 69.2% (36/52), ranging from 0% (0/16) in US images to 100% (15/15, 21/21) in X-ray and CT images. The model correctly identified pathologies in 30.5% of cases (11/36), ranging from 0% (0/9) in US images to 66.7% (8/12) for X-rays. The findings of this study indicate that despite its potential, multimodal GPT-4 is not yet a reliable tool for radiological images interpretation. Our study provides a baseline for future improvements in multimodal LLMs and highlights the importance of continued development to achieve reliability in radiology.

Introduction

Artificial Intelligence (AI) is transforming medicine, offering significant advancements especially in data-centric fields like radiology. Its ability to refine diagnostic processes and improve patient outcomes marks a revolutionary shift in medical workflows.

Radiology, heavily reliant on visual data, is a prime field for AI integration[1]. AI's ability to analyze complex images offers significant diagnostic support, potentially easing radiologist workloads by automating routine tasks and efficiently identifying key pathologies[2]. The increasing use of publicly available AI tools in clinical radiology has integrated these technologies into the operational core of radiology departments[3–5].

Among AI's diverse applications, Large Language Models (LLMs) have gained prominence, particularly GPT-4 from OpenAI, noted for its advanced language understanding and generation[6–15]. A notable recent advancement of GPT-4 is its multimodal ability to analyze images alongside textual data (GPT-4V)[16]. The potential applications of this feature can be substantial, specifically in radiology where the integration of imaging findings and clinical textual data is key to accurate diagnosis. Thus, the purpose of this study was to evaluate the performance of GPT-4V for the analysis of radiological images across various imaging modalities and pathologies.

Methods

A Sheba Medical Center Institutional Review Board (IRB) approval was granted for this study. The IRB committee waived informed consent.

Dataset Selection

We systematically reviewed all imaging examinations from one consecutive week as recorded in Sheba Medical Center's radiology information system (RIS). Our selection criteria aimed to include cases that would be considered resident-level in terms of diagnostic clarity and complexity. The inclusion of clear-cut cases was intended to ensure a focused evaluation of the AI's interpretive capabilities without the confounding variables of ambiguous or borderline findings.

A senior body imaging radiologist in conjunction with a radiology resident performed the case collection. We selected a total of 52 images, which represented a balanced cross-section of modalities including computed tomography (CT), ultrasound (US), and X-ray (**Table 1**). These images spanned various anatomical regions and pathologies, chosen to reflect a spectrum of common and critical findings appropriate for resident-level interpretation.

To uphold the ethical considerations and privacy concerns, each image was anonymized to maintain patient confidentiality prior to analysis. This process involved the removal of all identifying information, ensuring that the subsequent analysis focused solely on the clinical content of the images.

AI Interpretation with GPT-4 Multimodal

Using openAI's web interface, GPT-4V was prompted to analyze each image. The specific prompt used was *"We are conducting a study to evaluate GPT-4 image recognition abilities in healthcare. Identify the condition and describe key findings in the image."* This prompt was designed to elicit detailed interpretations of the imaging findings. The senior radiologist

and the resident reviewed the AI interpretations in consensus and compared them to the imaging findings.

To evaluate GPT-4V's performance, we checked for the accurate recognition of modality type, anatomical location, and pathology identification. Errors were classified as omissions, incorrect identifications, or hallucinations of pathology.

Data Analysis

The analysis was performed using Python version 3.10. Statistical significance was determined using a p-value threshold of less than 0.05. The primary metrics were the accuracies of modality, anatomical regions, and diagnoses identification, expressed as a percentage of correct identifications. A qualitative analysis of GPT-4V answers was also performed. A Fisher's exact test was employed to assess differences in the ability of GPT-4V to identify anatomical locations and pathologies across imaging modalities.

Results

Distribution of Imaging Modalities

The dataset consists of 52 diagnostic images categorized by modality (CT, X-ray, US), anatomical regions and pathologies. The results are summarized in **Table 1**. Overall, 36 images (69.2%) were pathological, 16 cases (30.8%) were normal.

GPT-4V Performance in Imaging Modality and Anatomical Region Identification

GPT-4V demonstrated a 100% (52/52) success rate for identification of the imaging modalities, across computed tomography (CT), ultrasound (US), and X-ray images, **Table 2**.

When analyzing GPT-4V's accuracy in anatomical regions identification, the model correctly identified all X-ray and CT images, and none of the US images (p<.001), **Table 2.**

Pathology Identification Accuracy

Pathology identification accuracy differed notably across imaging modalities (**Figure 1**).

CT scans demonstrated a pathology identification accuracy of 3/15 (20.0%), while no pathologies were identified using US 0/9 (0%).

X-rays showed a higher identification accuracy of 8/12 (66.7%). X-ray accuracy was significantly higher compared to both US (p = 0.005) and CT (p = 0.022).

Examples of cases from the GPT-4V image analysis are presented in **Figure 2**.

Error analysis across imaging modalities, detailed in **Table 3**, highlights specific trends. US images exhibited a notably high rate of false positive or hallucinated pathologies at 13/16 (81.3%), and a high overall mistake rate of 16/16 (100%). CT scans showed an overall mistake rate of 15/21 (71.4%). X-rays showed the lowest error rates across all mistake types (46.7%).

Error Analysis

A recurrent error in US imaging involved the misidentification of normal testicular structures as renal or liver pathologies. This error surfaced six times. For CT interpretations, GPT-4V three times hallucinated bladder-related pathologies when assessing scans for other conditions like ascites and metastases. X-ray analysis revealed a tendency towards over-diagnosis and mislocalization of opacities. This error was observed in three instances.

Discussion

This study offers a detailed evaluation of multimodal GPT-4 performance in radiological image analysis. GPT-4V correctly identified all imaging modalities. The model was inconsistent in identifying anatomical regions and pathologies, and wrongly identified anatomy and pathology in all US images. Consequently, GPT-4V, as it currently stands, cannot be relied upon for radiological interpretation.

However, the moments where GPT-4V accurately identified pathologies show promise, suggesting enormous potential with further refinement. The extraordinary ability to integrate textual and visual data is novel and has vast potential applications in healthcare, and radiology in particular. Radiologists interpreting imaging examinations rely on imaging findings alongside the clinical context of each patient. It has been established that clinical information and context can improve the accuracy and quality of radiology reports[17]. Similarly, the ability of LLMs to integrate clinical correlation with visual data marks a revolutionary step. This integration not only mirrors the decision making process of physicians, but also has the potential to ultimately surpass current image analysis algorithms which are mainly based on convolutional neural networks (CNNs)[18,19].

GPT-4V represents a new technological paradigm in radiology, characterized by its ability to understand context, learn from minimal data (zero-shot or few-shot learning), reason, and provide explanatory insights. These features mark a significant advancement from the traditional AI applications in the field. Furthermore, its ability to textually describe and explain images are awe-inspiring, and with the algorithm's improvement may eventually enhance medical education.

This study has several limitations. First, this was a retrospective analysis of patient cases, and the results should be interpreted accordingly. Second, there is potential for selection bias due to subjective case selection by the authors. Finally, we did not evaluate the performance

of GPT-4V in image analysis when textual clinical context was provided, this was outside the scope of this study.

To conclude, despite its vast potential, multimodal GPT-4 is not yet a reliable tool for clinical radiological images interpretation. Our study provides a baseline for future improvements in multimodal LLMs and highlights the importance of continued development to achieve clinical reliability in radiology.

References

1. Langlotz CP. The Future of AI and Informatics in Radiology: 10 Predictions. *Radiology*. 2023;309(1):e231114. doi:10.1148/radiol.231114

2. Kühl J, Elhakim MT, Stougaard SW, et al. Population-wide evaluation of artificial intelligence and radiologist assessment of screening mammograms. *Eur Radiol*. Published online November 8, 2023. doi:10.1007/s00330-023-10423-7

3. Langius-Wiffen E, De Jong PA, Mohamed Hoesein FA, et al. Added value of an artificial intelligence algorithm in reducing the number of missed incidental acute pulmonary embolism in routine portal venous phase chest CT. *Eur Radiol*. Published online August 3, 2023. doi:10.1007/s00330-023-10029-z

4. Maiter A, Hocking K, Matthews S, et al. Evaluating the performance of artificial intelligence software for lung nodule detection on chest radiographs in a retrospective real-world UK population. *BMJ Open*. 2023;13(11):e077348. doi:10.1136/bmjopen-2023-077348

5. Tejani A, Dowling T, Sanampudi S, et al. Deep Learning for Detection of Pneumothorax and Pleural Effusion on Chest Radiographs: Validation Against Computed Tomography, Impact on Resident Reading Time, and Interreader Concordance. *J Thorac Imaging*. Published online September 29, 2023. doi:10.1097/RTI.0000000000000746

6. GPT-4 for Automated Determination of Radiologic Study and Protocol Based on Radiology Request Forms: A Feasibility Study | Radiology. Accessed November 11, 2023. https://pubs.rsna.org/doi/10.1148/radiol.230877?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed

7. Sorin V, Barash Y, Konen E, Klang E. Large language models for oncological applications. *J Cancer Res Clin Oncol*. Published online May 9, 2023. doi:10.1007/s00432-023-04824-w

8. Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. *J Am Coll Radiol JACR*. 2023;20(10):990-997. doi:10.1016/j.jacr.2023.05.003

9. Bajaj S, Gandhi D, Nayar D. Potential Applications and Impact of ChatGPT in Radiology. *Acad Radiol*. Published online October 5, 2023:S1076-6332(23)00460-9. doi:10.1016/j.acra.2023.08.039

10. Doo FX, Cook TS, Siegel EL, et al. Exploring the Clinical Translation of Generative Models Like ChatGPT: Promise and Pitfalls in Radiology, From Patients to Population Health. *J Am Coll Radiol JACR*. 2023;20(9):877-885. doi:10.1016/j.jacr.2023.07.007

11. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. *Nature*. 2023;619(7969):357-362. doi:10.1038/s41586-023-06160-y

12. Sorin V, Klang E, Sklair-Levy M, et al. Large language model (ChatGPT) as a support tool for breast tumor board. *Npj Breast Cancer*. 2023;9(1):44. doi:10.1038/s41523-023-00557-8

13. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. Published online April 12, 2023. Accessed June 29, 2023. http://arxiv.org/abs/2303.13375

14. Hasani AM, Singh S, Zahergivar A, et al. Evaluating the performance of Generative Pre-trained Transformer-4 (GPT-4) in standardizing radiology reports. *Eur Radiol*. Published online November 8, 2023. doi:10.1007/s00330-023-10384-x

15. Crimì F, Quaia E. GPT-4 versus Radiologists in Chest Radiography: Is It Time to Further Improve Radiological Reporting? *Radiology*. 2023;308(2):e231701. doi:10.1148/radiol.231701

16. Yang Z, Li L, Lin K, et al. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). Published online October 11, 2023. Accessed November 11, 2023. http://arxiv.org/abs/2309.17421

17. Leslie A, Jones AJ, Goddard PR. The influence of clinical information on the reporting of CT by radiologists. *Br J Radiol*. Published online May 29, 2014. doi:10.1259/bjr.73.874.11271897

18. Klang E. Deep learning and medical imaging. *J Thorac Dis*. 2018;10(3):1325-1328. doi:10.21037/jtd.2018.02.76

19. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional Neural Networks for Radiologic Images: A Radiologist's Guide. *Radiology*. 2019;290(3):590-606. doi:10.1148/radiol.2018180547

Table 1: Aggregated Data of Anatomical Regions and Pathologies by Imaging Modality

Modality	**Number of images**	**Anatomical Regions**	**Pathologies**
CT	21 (11 venous phase, 10 non-contrast)	Brain (10), Abdomen (7), Pelvis (4)	Normal (6) Parenchymal Hemorrhage (4), SBO (3), Subdural Hemorrhage (2), Gallstones (1), Ascites (1), Liver Metastases (1), Cholecystitis (1), Bowel Perforation (1), Subdural Hemorrhage with Additional Findings (1)
X-ray	15	Chest (15)	Normal (3) Mass (3), Cardiac Pacemaker (2), Pulmonary Edema (2), Various Forms of Consolidation (5), Pleural Effusion (1)*
US	16	Testicles (8), Kidney (6), Kidney & Liver (2)	Normal (7) Hydronephrosis (4), Testicular Mass (4), Fatty Liver (1)

* One case included both a consolidation and a pleural effusion.

Abbreviations: ultrasound (US), computed tomography (CT).

Table 2: GPT-4 Modality and Anatomy Identification Accuracy. Identified/Total (%).

Modality (total)	Modality	Anatomical region
CT (21)	21/21 (100%)	21/21 (100%)
X-ray (15)	15/15 (100%)	15/15 (100%)
US (16)	16/16 (100%)	0/16 (0%)
Total (52)	52/52 (100%)	36/52 (69.2%)

Abbreviations: ultrasound (US), computed tomography (CT).

Table 3: GPT-4 mistake types across different modalities**.** Identified/Total (%).

Modality	Ignores a Pathology	Identifies Differently	Hallucinates	Total Mistakes*
CT	8/21 (38.1%)	4/21 (19.0%)	5/21 (23.8%)	15/21 (71.4%)
X-ray	4/15 (26.7%)	2/15 (13.3%)	1/15 (6.7%)	7/15 (46.7%)
US	2/16 (12.5%)	3/16 (18.8%)	13/16 (81.3%)	16/16 (100%)
Total	14/52 (26.9%)	9/52 (17.3%)	19/52 (36.5%)	38/52 (73.1%)

*Some cases included more than one mistake, for example GPT-4V ignored a small bowel obstruction and hallucinated a bladder. In the calculation of “Total mistakes” we counted every incorrect image interpretation once.

Abbreviations: computed tomography (CT), ultrasound (US).

Figure 1: Pathology Accuracy by Imaging Modality

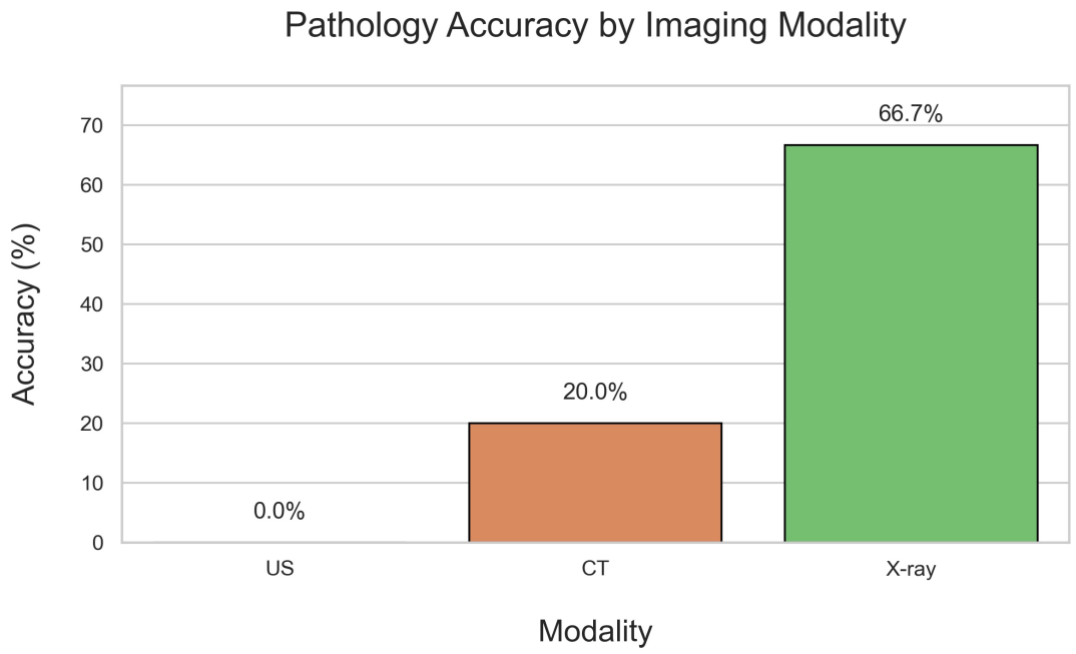
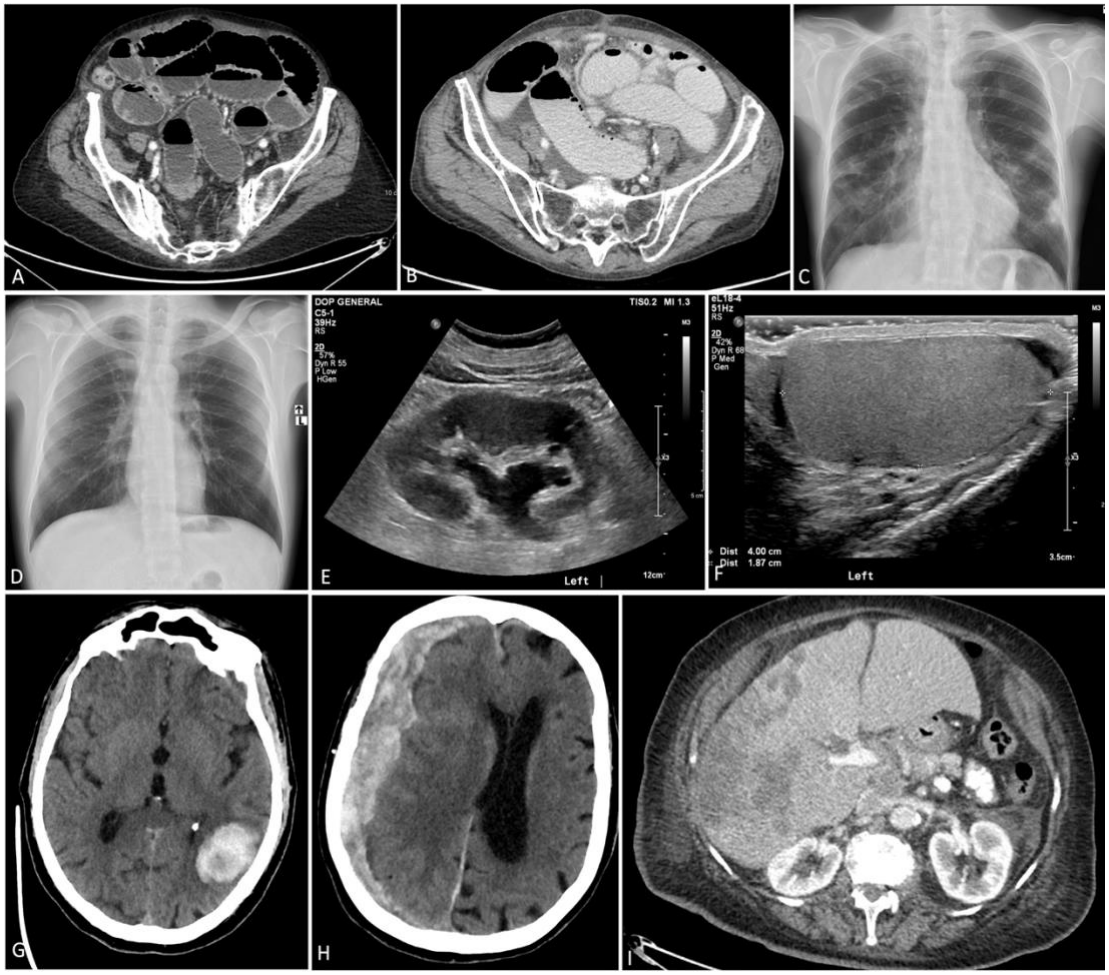


Figure 2: Illustrative Cases from GPT-4V Image Analysis in Radiology



Six radiological images (A-F) processed by GPT-4 in response to the following prompt: "We are conducting a study to evaluate GPT-4 image recognition abilities in healthcare. Identify the condition and describe key findings in the image." **A** and **B**. Axial contrast-enhanced CT scans of the pelvis displaying dilated bowel loops suggestive of bowel obstruction. GPT-4 accurately recognized the obstruction in Image A but failed to detect any pathology in Image B. **C**. Posteroanterior (PA) chest X-ray demonstrating patchy bilateral opacities, which were correctly identified by GPT-4 as indicative of pulmonary pathology. **D**. Normal PA chest X-ray, where GPT-4 incorrectly reported a "prominent mass or opacity in the right middle to lower lung zone." **E**. Ultrasound (US) of the left kidney displaying features of hydronephrosis, which GPT-4V incorrectly labeled as an ovarian cyst. **F**. Longitudinal US of a normal testicle and epididymal head, misinterpreted by GPT-4V as a dilated bile duct within the liver. **G.** Axial non-contrast CT of the head where GPT-4V correctly detected the presence of intracranial hemorrhage. **H.** Axial CT of the head displaying a midline shift which GPT-4V recognized; however, it failed to note the right subdural hemorrhage. **I.** Axial contrast-enhanced CT of the abdomen where GPT-4V correctly noted the liver but overlooked multiple hypodense lesions indicative of metastases and incorrectly reported visibility of the bladder.