Radiology

ORIGINAL RESEARCH • COMPUTER APPLICATIONS

Toward Foundation Models in Radiology? Quantitative Assessment of GPT-4V's Multimodal and Multianatomic Region Capabilities

Quirin D. Strotzer, MD • Felix Nieberle, MD • Laura S. Kupke, MD • Gerardo Napodano, MD • Anna Katharina Muertz, MD • Stefanie Meiler, MD • Ingo Einspieler, MD • Janine Rennert, MD • Michael Strotzer, MD • Isabel Wiesinger, MD • Christina Wendl, MD • Christian Stroszczynski, MD • Okka W. Hamer, MD • Andreas Schicho, MD

From the Institute of Radiology (Q.D.S., L.S.K., G.N., A.K.M., S.M., I.E., J.R., C.W., C.S., O.W.H., A.S.) and Department of Cranio- and Maxillofacial Surgery (F.N.), University of Regensburg Medical Center, Franz-Josef-Strauss-Allee 11, 93053 Regensburg, Germany; Department of Radiology, Division of Neuroradiology, Massachusetts General Hospital, Harvard Medical School, Boston, Mass (Q.D.S.); Department of Radiology, Bayreuth Medical Center, Bayreuth, Germany (M.S.); Center of Neuroradiology, medbo District Hospital and University Medical Center Regensburg, Regensburg, Germany (I.W., C.W.); and Department of Radiology, Donaustauf Hospital, Donaustauf, Germany (O.W.H.). Received October 23, 2023; revision requested November 17; final revision received September 27, 2024; accepted October 2. **Address correspondence to** Q.D.S. (email: *quirin.strotzer@ukr.de*).

Q.D.S. supported by a Walter Benjamin Fellowship from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation [GZ: STR 1774/1-1]) and the ReForM A (Regensburg Research Funding in Medicine) program through the Faculty of Medicine, University of Regensburg, Germany.

Conflicts of interest are listed at the end of this article.

Radiology 2024; 313(2):e240955 • https://doi.org/10.1148/radiol.240955 • Content code: AI

Background: Large language models have already demonstrated potential in medical text processing. GPT-4V, a large vision-language model from OpenAI, has shown potential for medical imaging, yet a quantitative analysis is lacking.

Purpose: To quantitatively assess the performance of GPT-4V in interpreting radiologic images using unseen data.

Materials and Methods: This retrospective study included single representative abnormal and healthy control images from neuroradiology, cardiothoracic radiology, and musculoskeletal radiology (CT, MRI, radiography) to generate reports using GPT-4V via the application programming interface from February to March 2024. The factual correctness of free-text reports and the performance in detecting abnormalities in binary classification tasks were assessed using accuracy, sensitivity, and specificity. The binary classification performance was compared with that of a first-year nonradiologist in training and four board-certified radiologists.

Results: A total of 515 images in 470 patients (median age, 61 years [IQR, 44–71 years]; 267 male) were included, of which 345 images were abnormal. GPT-4V correctly identified the imaging modality and anatomic region in 100% (515 of 515) and 99.2% (511 of 515) of images, respectively. Diagnostic accuracy in free-text reports was between 0% (0 of 33 images) for pneumothorax (CT and radiography) and 90% (45 of 50 images) for brain tumor (MRI). In binary classification tasks, GPT-4V showed sensitivities between 56% (14 of 25 images) for ischemic stroke and 100% (25 of 25 images) for brain hemorrhage and specificities between 8% (two of 25 images) for brain hemorrhage and 52% (13 of 25 images) for pneumothorax, compared with a pooled sensitivity of 97.2% (1103 of 1135 images) and pooled specificity of 97.2% (1084 of 1115 images) for the human readers across all tasks. The model exhibited a clear tendency to overdiagnose abnormalities, with 86.5% (147 of 170 images) and 67.7% (151 of 223 images) false-positive rates for the free-text and binary classification tasks, respectively.

Conclusion: GPT-4V, in its earliest version, recognized medical image content and reliably determined the modality and anatomic region from single images. However, GPT-4V failed to detect, classify, or rule out abnormalities in image interpretation.

Supplemental material is available for this article.

Large language models (LLMs), such as OpenAI's ChatGPT, have significantly transformed text generation since gaining traction in 2022. On the basis of the transformer architecture and trained on massive—often undisclosed—amounts of text, LLMs can produce credible text output without needing many examples (few-shot learning and prompting, or zero-shot learning and prompting). LLMs have undergone extensive testing in medical settings, including for transforming free-text radiology reports into standardized reporting templates and for the data mining of free-text CT reports in lung cancer (1,2). The intrinsic "knowledge" of the models in answering text questions in a radiology board-style examination was demonstrated, and LLMs were recently shown to help simplify radiology reports (3,4).

More recently, GPT-4 with vision, called GPT-4V (OpenAI) (5), which is capable of processing image input, was introduced. Large vision-language models such as this could bring us closer to foundation models, which could be used for a broad spectrum of different tasks (6).

A qualitative assessment reported promising performance in creating radiology reports from single medical images but also highlighted current limitations of the model (7). "Red team" experts (ie, professionals tasked with testing the vulnerability of a system) also reported inconsistencies in interpreting radiologic images (8). Nevertheless, such widely available models will inevitably be misused for unintended applications, circumventing safeguards.

Given the potential and risks involved, a thorough analysis of these models is critical, but peer-reviewed literature on GPT-4V remains scarce. Promising performance was reported in a qualitative study on US image analysis and when GPT-4V was compared with LLMs without vision capabilities in medical case challenges (9,10). Other authors deemed the model unfit

Abbreviations

AI = artificial intelligence, LLM = large language model

Summary

OpenAI's GPT-4V reliably identified the imaging modality and anatomic region but could not safely detect, classify, or rule out abnormalities on single MRI, CT, and radiographic images.

Key Results

- The ability of OpenAI's GPT-4V, a large vision-language model, to analyze 515 CT, MRI, and radiographic images from neurologic, thoracic, and musculoskeletal regions was retrospectively assessed.
- The model detected the imaging modality and anatomic region with 100% and 99.2% accuracy, respectively.
- Limited diagnostic performance was observed for the model in free-text reports (accuracy, 36.5%) and binary classification tasks (sensitivity, 78%; specificity, 32.3%), with a tendency to overdiagnose abnormalities (86.5% and 67.7% false-positive rates for free-text and binary classification tasks, respectively).

for medical image interpretation and reported, for example, that the model was outperformed by commercial chest radiograph interpretation software (11–13). Most authors included only a limited number of publicly available images and no negative controls, thereby limiting model performance assessment. Public data may have been included in training GPT-4V, potentially biasing toward better results. Other works have yet to be peer-reviewed and have similar limitations (14–16). A quantitative analysis comparing the performance of GPT-4V to that of human readers on unseen data is still lacking.

Therefore, the aim of this study was to quantitatively assess whether GPT-4V, a model not explicitly optimized for medical applications, can interpret radiologic images as accurately as human readers using unseen data.

Materials and Methods

Data Acquisition

The institutional review board approved this retrospective Health Insurance Portability and Accountability Act–compliant bicenter study (vote no. 23–3545–104). The requirement to obtain informed consent was waived due to the retrospective nature and analysis of anonymized imaging data. The images originated from a university hospital and a tertiary hospital specializing in neurologic care.

A diverse selection of common pathologic findings and imaging modalities frequently addressed by artificial intelligence (AI) models were intentionally chosen, as the objective of the work was to test the potential of GPT-4V as a foundation model. The selected categories included neuroradiology (ischemic stroke, brain hemorrhage, brain tumor, multiple sclerosis), cardiothoracic radiology (pneumothorax, pulmonary embolism, pneumonia, lung cancer), and musculoskeletal radiology (fracture). Pathologic subtypes were not differentiated. For example, brain tumor included meningioma, glioblastoma, and brain metastases. Brain hemorrhage included epidural, subdural, subarachnoid, and intraparenchymal hematomas.

A convenience sample of patients was created by querying radiology reports in the radiologic information systems of the hospitals. The reference standard diagnoses were manually confirmed on the basis of all available information (entire scan and report, follow-up imaging, and medical records).

Axial CT images, axial MRI scans, and radiographs were used, as applicable; for example, pulmonary CT angiograms for pulmonary embolism, lung or soft-tissue-window CT images and radiographs for lung cancer, fluid-attenuated inversion recovery MRI scans for multiple sclerosis, and contrast-enhanced

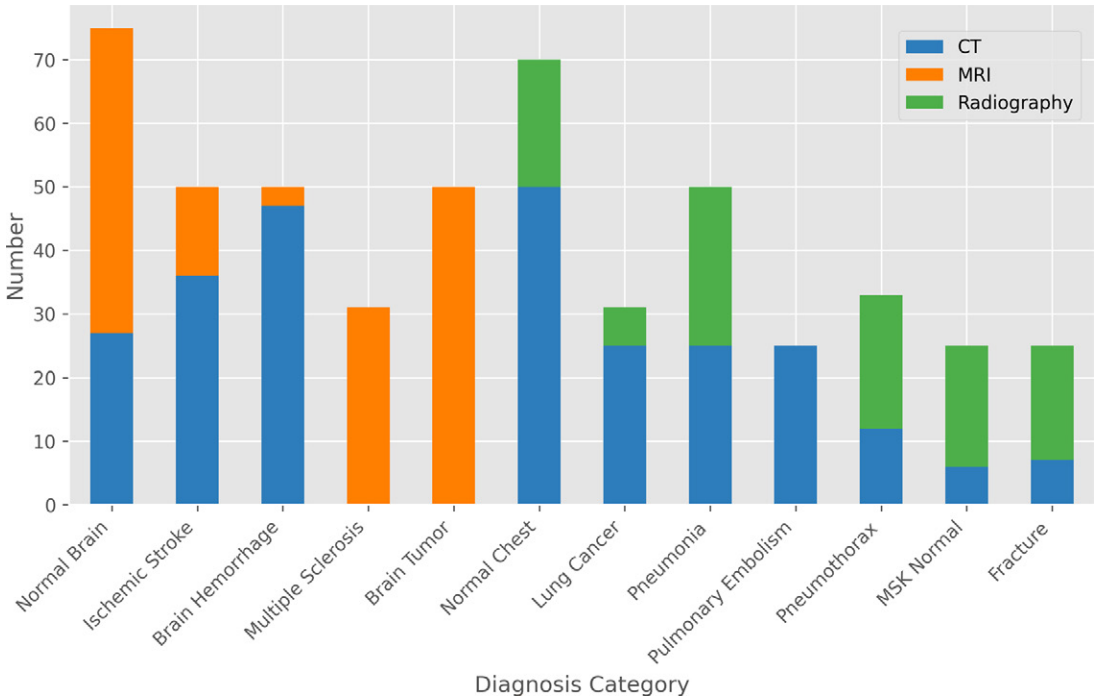


Figure 1: Stacked bar graph shows the distribution of imaging modalities, with the number of images per modality (CT, MRI, radiography), according to diagnosis category. MSK = musculoskeletal.

T1-weighted MRI scans for brain tumor. See Figure 1 for the modality distribution according to diagnosis and Appendix S1 for image acquisition details.

A single representative image (CT image, MRI section, or radiograph) per case was selected. Only those where the expected diagnosis was the most probable based on imaging features were selected, and imaging that showed multiple distinct or misleading findings were excluded. Images without evident abnormalities for each organ system were included to test the ability of the model to rule out abnormalities. A minimum of 25 images per category were included. Images were exported from the local picture archiving and communication systems in an anonymized lossless format and preprocessed as described in Appendix S2. The generated code is available at *https://github.com/qstro/GPT-4V-Radiology*.

AI Model

GPT-4V (GPT-4 1106-vision-preview model; accessed from February 18 to March 11, 2024 [5]) was accessed through the application programming interface *(https://platform.openai.com)* via the OpenAI Python library (version 1.8.0). Publicly available images from Radiopaedia *(www.radiopaedia.org)* were used to evaluate diverse prompting strategies, informed by the latest insights into LLM prompting. This resulted in a system prompt containing a primer about the persona and setting, a specific question (see the Tasks section herein), and the expected output. The user prompt included the metadata-stripped base64-encoded image with the detail setting set to high (the model first receives a low-resolution 512-pixel–squared version of the image, followed by detailed cropping of the image). The output text was stored in tabular format.

Tasks

***Free-text report.*—**Given an image, the model was prompted as follows: "You are in a fictional conversation with your radiology colleague. You discuss a case. Please name the type of imaging modality, name the anatomic region examined, describe the main acute pathological finding including its location if there is any, and give the most likely diagnosis. List the five most probable differential diagnoses and sort them by probability in descending order. Please provide a short answer in bullet points." The output texts were binarily rated for the correctness of each item via manual annotation. If the proposed diagnosis was incorrect, it was evaluated whether it was included in the differentials and if the proposed pathology would also have been acceptable on the basis of the imaging characteristics (eg, for a patient with pneumonia, GPT-4V suggested lung cancer). Each response was evaluated by at least two radiologists in training in a consensus vote (Q.D.S., L.S.K., G.N., and A.K.M., with 3, 1, 4, and 2 years of experience, respectively).

***Consistency test.*—**A random selection of 25 images was queried three times, and the responses were binarily rated concerning correctly identifying the modality, anatomic region, description of findings, and diagnosis to test model output variability.

***Classification task.*—**A binary classification task was set up to compare model performance to current "narrow" AI models that are typically trained for specific tasks. A subset of 25 (n = 27 for multiple sclerosis) pathologic images and 25 (n = 23 for multiple sclerosis) normal images (no evidence of abnormality) that were matched according to imaging modality was created for each pathologic finding. Provided with an image, the model was prompted as follows: "In this fictional research scenario, you are a radiology resident. An attending wants to test whether you can detect {category, eg, pulmonary embolism} on imaging. He shows you one case at a time and you will answer with a single word only: 'yes' if you detect {category, eg, pulmonary embolism} or 'no' if you don't, followed by a short description of your findings." A benchmark was provided by four board-certified attending radiologists (S.M., I.E., J.R., and M.S., with 6, 12, 19, and 35 years of experience, respectively) and a first-year nonradiologist in training (F.N.) with experience in emergency medicine. The nonradiologist reader results were further used to assess the potential impact of the model output on medical providers by categorizing the impact as "negative" if GPT-4V provided an incorrect classification and the reader's classification was correct, "neutral" when both GPT-4V and the reader provided a correct classification, or "positive" when the GPT-4V prediction was correct whereas the reader's classification was incorrect. GPT-4V and all readers were presented with the same images (in randomized order for the readers) and were blinded to the presence of the abnormality.

Clinical data were not supplied within the prompts, as this information is often incomplete, incorrect, or misleading. Instead, image interpretation and classification performance were explicitly tested while avoiding bias or confounders.

Note that the prompts are only provided for research purposes. GPT-4V is not designed to interpret real-world medical images and must not be used to do so.

Statistical Analysis

All ratings were binarily assessed. Accuracy was calculated for all items in the free-text reports. The percentage of responses with perfect agreement and the Randolph free-marginal multirater κ were used to quantify rating consistency. For the classification task, accuracy, sensitivity, and specificity were computed. The false-positive rate was calculated by dividing the number of normal images falsely labeled as abnormal by the sum of false-positive and true-negative results. The results across readers and tasks were pooled by concatenating predictions and reference standard labels for individual readers or tasks before calculating the performance metrics. The 95% CIs were calculated by bootstrapping with 1000 iterations. Agreement between the human readers was assessed with the Cohen κ statistic. Randolph and Cohen κ values were categorized as poor (<0.50), moderate (0.50–0.69), substantial (0.70–0.89), and (almost) perfect (≥0.90). Data were analyzed and visualized by an author (Q.D.S.) using the pandas (version 1.3.1), numpy (version 1.23.5), statsmodels (version 0.13.2), pingouin (version 0.5.3), seaborn (version 0.11.2), and matplotlib (version 3.4.3) Python libraries.

Results

General Results

A total of 515 unique representative images in 470 patients (median age, 61 years [IQR, 44–71 years]; 267 male, 203 female)

were included in the analysis. Duplicate patients occurred, for example, in a patient with lung cancer who had pneumothorax at another time point and a patient with multiple fractures in different regions. Regarding the imaging modality, 50.5% (260 of 515) of included images were acquired with CT, 28.3% (146 of 515) were acquired with MRI, and 21.2% (109 of 515) were acquired with radiography. Of the included images, 67% (345 of 515) were abnormal and 33% (170 of 515) were normal. Although GPT-4V is fine-tuned to refuse radiologic image interpretation, it only did so in 16 instances of free-text reports and was reprompted until it provided an interpretation.

Free-Text Results

The imaging modality was correctly identified in all 515 images, and the anatomic region was correctly identified in 99.2% (511 of 515). A pulmonary CT angiogram was identified as an abdominal CT image, although no abdominal structures were present in that section. A radiograph of the left elbow joint was labeled as a radiograph of the left knee, and two CT images of the wrist were classified as CT images of the foot or head.

The main finding was correctly identified in 36.5% (188 of 515) of images. Performance depended heavily on the pathologic finding and imaging modality. For example, GPT-4V missed all pneumothoraxes (0 of 33 correctly identified) and almost all pulmonary embolisms (one of 25 identified), but had an accuracy of 90% (45 of 50) in diagnosing brain tumors on MRI scans. Concerning inconspicuous chest imaging, 0% (0 of 50) of CT images but 60% (12 of 20) of radiographs were correctly identified as normal. The localization of findings was often incorrect, especially concerning laterality.

Imaging alone is not always unequivocal. For incorrect diagnoses, differential diagnoses were also considered, and the suggested diagnoses were analyzed. For multiple sclerosis, for example, the correct diagnosis was included in the differentials for 95.2% (20 of 21) of images. The suggested diagnosis could also have been acceptable in 71.4% (15 of 21) of images.

Overall, the model clearly tended to overdiagnose abnormalities, with a false-positive rate of 86.5% (147 of 170 images). Chest CT readings mainly included hallucinations of upper mediastinal masses, lung cancer, or pneumonia. Normal brain CT images were most often described as showing ischemic stroke or hemorrhage, and normal brain MRI scans as showing glioma or multiple sclerosis.

Table 1 provides results for each diagnosis category, and Figure 2 shows example free-text reports for two brain hemorrhage images.

Consistency Test

Agreement across the three generated output reports was perfect, with a rate of concordance of 100% (25 of 25) and a Randolph κ of 1.0 for items where the model showed high accuracy (modality, anatomic region). Items where the model performed poorly showed only moderate agreement (Randolph κ values of 0.68 for the description of findings and 0.52 for the diagnosis), which may indicate a higher level of randomness in model answers (Table 2).

Classification Tasks

GPT-4V showed overall poor performance in binary classification tasks (Table 3). The overall accuracy across all tasks was only slightly above chance (55.3% [95% CI: 51.1, 60.2]). The best results were observed for pneumonia, with accuracy, sensitivity, and specificity of 64% [95% CI: 50, 76], 96% [95% CI: 86.4, 100], and 32% [95% CI: 14.3, 50], respectively. The relatively high overall (pooled across tasks) sensitivity of 78% [95% CI: 72.6, 83.3] came at the expense of a low specificity of 32.3% [95% CI: 26.4, 38.2] and a false-positive rate of 67.7% (151 of 223 images). For example, only two of 25 normal brain CT images of the brain hemorrhage task were correctly classified.

In comparison, the human readers performed considerably better, with almost perfect agreement (Cohen κ ≥0.9 for all reader pairs) (Table 4). The performance ranged from an accuracy of 90.4% [95% CI: 86.8, 94], sensitivity of 90.4% [95% CI: 84.6, 95.4], and specificity of 90.4% [85.1, 95.2] for pulmonary embolism to perfect results for ischemic stroke and brain tumor. See Figure 3 and Appendix S3 (Table S1) for details. Owing to the high interrater agreement, the results of all the readers were pooled.

The hypothetical impact of the nonradiologist relying on GPT-4V would have been positive in 2% (nine of 450), neutral in 53.3% (240 of 450), and negative in 43.3% (195 of 450) of images.

In addition, multiple image inputs were tested for intracranial hemorrhage detection, with no relevant change in performance observed. Details are provided in Appendix S4 (Fig S1, Table S2).

Discussion

Large vision-language models can potentially solve imaging tasks without special training or fine-tuning. In our quantitative assessment of OpenAI's GPT-4V (5), the model accurately assessed the imaging modality (CT, MRI, or radiography) and anatomic region (brain, chest, or musculoskeletal system). However, detection and interpretation of pathologic findings were not reliable, and false-positive findings were frequent.

Obvious abnormalities, such as brain hemorrhage and brain tumor, were identified relatively often, but when tasked with generating free-text reports, GPT-4V missed more subtle findings such as pulmonary embolism and pneumothorax. Concerning radiography, one reason may be the limited input resolution. The image is automatically rescaled so that the shorter side is 768 pixels, whereas the matrix size for digital radiography is 2000 × 2500 pixels (17). Furthermore, the visual representations derived by the transformer encoder might render detection of subtle intensity changes more challenging than high-contrast variations, such as those found in brain hemorrhage.

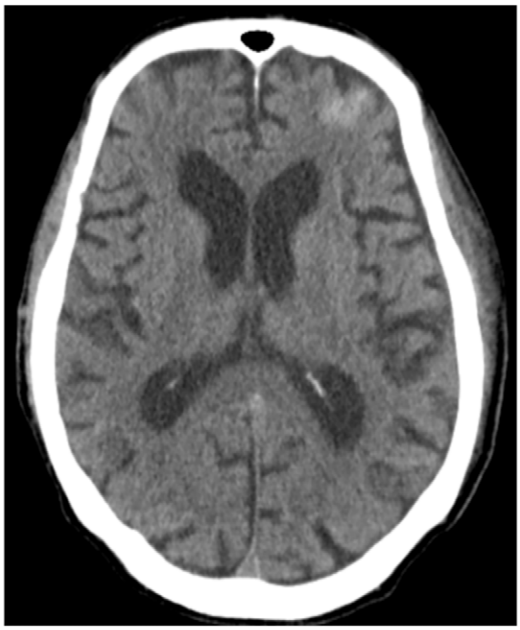
Even though GPT-4V theoretically knows the standard convention of how radiologic images are displayed and sometimes emphasizes it in its responses, many errors concerning the side of the abnormality could be observed. These errors may be due to limitations in contextual understanding (medical domain deviates from general convention, and GPT-4V might fail to make a connection between the image and the expected context), propagating errors through internal preprocessing steps (eg, cropping), or in model image perception (images are encoded into a numerical representation from which features are extracted). Additionally, confabulations about the exact anatomic location

Table 1: Accuracy of Free-Text Reports Generated by GPT-4V

Diagnosis Category and Modality	Modality Correct	Anatomic Region Correct	Exact Location Correct	Laterality Correct	Description Correct	Diagnosis Correct	Diagnosis Among Top 5 DDs	Suggested Abnormality Also Possible
All categories								
All modalities	100.0 (515/515)	99.2 (511/515)	30.8 (100/325)	50.5 (155/307)	47.4 (244/515)	36.5 (188/515)	20.8 (68/327)	26.3 (86/327)
Normal brain								
CT	100.0 (27/27)	100.0 (27/27)	NA	NA	3.7 (1/27)	0.0 (0/27)	0.0 (0/27)	3.7 (1/27)
MRI	100.0 (48/48)	100.0 (48/48)	NA	NA	8.3 (4/48)	2.1 (1/48)	2.1 (1/47)	4.3 (2/47)
Ischemic stroke								
CT	100.0 (36/36)	100.0 (36/36)	33.3 (11/33)	45.5 (15/33)	30.6 (11/36)	47.2 (17/36)	5.3 (1/19)	5.3 (1/19)
MRI	100.0 (14/14)	100.0 (14/14)	30.8 (4/13)	36.4 (4/11)	92.9 (13/14)	64.3 (9/14)	60.0 (3/5)	80.0 (4/5)
Brain hemorrhage								
CT	100.0 (47/47)	100.0 (47/47)	23.9 (11/46)	45.7 (21/46)	85.1 (40/47)	66.0 (31/47)	81.3 (13/16)	18.8 (3/16)
MRI	100.0 (3/3)	100.0 (3/3)	50.0 (1/2)	50.0 (1/2)	33.3 (1/3)	33.3 (1/3)	0.0 (0/2)	0.0 (0/2)
Multiple sclerosis								
MRI	100.0 (31/31)	100.0 (31/31)	50.0 (15/30)	39.1 (9/23)	93.6 (29/31)	32.3 (10/31)	95.2 (20/21)	71.4 (15/21)
Brain tumor								
MRI	100.0 (50/50)	100.0 (50/50)	34.8 (16/46)	48.8 (20/41)	90.0 (45/50)	90.0 (45/50)	80.0 (4/5)	80.0 (4/5)
Normal chest								
CT	100.0 (50/50)	100.0 (50/50)	NA	NA	0.0 (0/50)	0.0 (0/50)	0.0 (0/50)	2.0 (1/50)
Radiography	100.0 (20/20)	100.0 (20/20)	NA	NA	65.0 (13/20)	60.0 (12/20)	0.0 (0/8)	25.0 (2/8)
Lung cancer								
CT	100.0 (25/25)	100.0 (25/25)	36.0 (9/25)	38.1 (8/21)	60.0 (15/25)	36.0 (9/25)	25.0 (4/16)	56.3 (9/16)
Radiography	100.0 (6/6)	100.0 (6/6)	33.3 (2/6)	33.3 (2/6)	83.3 (5/6)	50.0 (3/6)	100.0 (3/3)	66.7 (2/3)
Pneumonia								
CT	100.0 (25/25)	100.0 (25/25)	32.0 (8/25)	48.0 (12/25)	80.0 (20/25)	52.0 (13/25)	66.7 (8/12)	91.7 (11/12)
Radiography	100.0 (25/25)	100.0 (25/25)	47.8 (11/23)	58.3 (14/24)	84.0 (21/25)	64.0 (16/25)	77.8 (7/9)	77.8 (7/9)
Pulmonary embolism								
CT	100.0 (25/25)	96.0 (24/25)	0.0 (0/25)	56.3 (9/16)	4.0 (1/25)	4.0 (1/25)	0.0 (0/24)	16.7 (4/24)
Pneumothorax								
CT	100.0 (12/12)	100.0 (12/12)	8.3 (1/12)	36.4 (4/11)	8.3 (1/12)	0.0 (0/12)	0.0 (0/12)	75.0 (9/12)
Radiography	100.0 (21/21)	100.0 (21/21)	31.6 (6/19)	47.4 (9/19)	23.8 (5/21)	0.0 (0/21)	4.8 (1/21)	33.3 (7/21)
Normal MSK								
CT	100.0 (6/6)	100.0 (6/6)	NA	NA	0.0 (0/6)	0.0 (0/6)	0.0 (0/6)	0.0 (0/6)
Radiography	100.0 (19/19)	100.0 (19/19)	NA	NA	57.9 (11/19)	57.9 (11/19)	0.0 (0/8)	25.0 (2/8)
Fracture								
CT	100.0 (7/7)	71.4 (5/7)	28.8 (2/7)	66.7 (2/3)	42.9 (3/7)	42.9 (3/7)	25.0 (1/4)	25.0 (1/4)
Radiography	100.0 (18/18)	94.4 (17/18)	23.1 (3/13)	92.7 (13/14)	27.8 (5/18)	33.3 (6/18)	16.7 (2/12)	8.3 (1/12)

Note.—Data are percentages, with numbers of images in parentheses. The percentage of correct answers concerning each scored item of the GPT-4V–generated (OpenAI [5]) free-text results is reported for each anatomic region, modality, and diagnosis category. "Diagnosis Among Top 5 DDs" and "Suggested Abnormality Also Possible" were only scored if the proposed diagnosis was incorrect. "Exact Location" and "Laterality" were only scored in abnormal cases where the location or laterality was mentioned. DDs = differential diagnoses, MSK = musculoskeletal, NA = not applicable.

Imaging Modality: Non-contrast CT (Computed Tomography) scan
Body Part Examined: Brain
Pathological Finding: There is an area of hyperdensity in the left cerebral hemisphere, consistent with a high attenuation lesion
Location: The lesion is located within the parenchyma, likely affecting the basal ganglia and surrounding white matter
Most Likely Diagnosis: Intracerebral hemorrhage (ICH)
Most Probable Differential Diagnoses:
1. Hypertensive hemorrhage (due to the location in the basal ganglia)
2. Cerebral amyloid angiopathy (typically in older patients)
3. Hemorrhagic transformation of an ischemic stroke
4. Arteriovenous malformation (AVM) rupture
5. Tumor presenting with hemorrhage (e.g., glioblastoma, metastasis



Imaging Modality: CT Scan (Computed Tomography)
Body Part Examined: Brain
Main Pathological Finding: There appears to be a hypodensity in the left cerebral hemisphere suggestive of a low attenuation area.
Location: Left cerebral hemisphere, possibly involving the left middle cerebral artery (MCA) territory based on the image provided.
Most Likely Diagnosis: Ischemic stroke
Differential Diagnoses (Descending Order of Probability):
1. Ischemic stroke due to thrombosis or embolism
2. Transient ischemic attack (TIA) with temporary symptoms and potential resolution on imaging
3. Glioma or other primary brain neoplasm with associated vasogenic edema
4. Brain abscess or infectious process with vasogenic edema
5. Demyelinating disease such as multiple sclerosis (less likely due to typical appearances and distribution of lesions)

Further clinical correlation and potentially more imaging with MRI and vascular imaging would be needed for definitive diagnosis and treatment planning.

Figure 2: Example images and GPT-4V (OpenAI)–generated free-text reports for brain hemorrhage. Left: Axial noncontrast head CT image in a 61-year-old female patient with a hematoma in the left basal ganglia, likely of hypertensive origin, and the corresponding model-generated free-text report show a correct main finding and diagnosis by GPT-4V. Right: Axial noncontrast head CT image in a 90-year-old male patient with a traumatic left frontal hematoma and the corresponding model-generated free-text report show an incorrect main finding and diagnosis by GPT-4V. Correct items in the response text are highlighted in green, while incorrect items are highlighted in red.

of a lesion within the imaged volume occurred frequently (eg, regarding the correct cerebral lobe). Overdiagnosis is most likely due to the massive overrepresentation of certain abnormalities in the training corpus.

Recent research reported an intrarater agreement (κ) of 0.84 when GPT-4V was used to assess images of malignancies via confocal laser endomicroscopy compared with 0.77 for human readers (18). We noted that the agreement strongly depended on the correctness of the answers, which declined as the responses became less accurate and more random. LLMs are designed to exert a certain degree of randomness to ensure output variability. However, this can potentially be a disadvantage if output accuracy is of interest.

We proposed a binary classification task for nine different abnormalities to compare the performance of GPT-4V in disease detection to that of published and commercially available AI tools. In all the tasks, GPT-4V yielded a subpar performance that specialized software can easily surpass. Examples are innumerable, including a sensitivity of 98.8% and specificity of 98.0% on test data for detecting hemorrhage on head CT images (19)

or areas under the receiver operating characteristic curve greater than 0.9 for pneumothorax detection on chest radiographs (20). In contrast to GPT-4V, these "narrow" applications can perform only single (or occasionally a few) tasks and rely on the correct input for inference.

Errors in AI systems can have various negative consequences for all stakeholders. Patients, in particular, can be harmed, either by the delay in treatment when a diagnosis is missed, by receiving inappropriate therapy if the wrong diagnosis is established, or by being subject to undue diagnostic measures and stress when a normal finding is declared pathologic. We assessed the hypothetical impact of the output of GPT-4V on the performance of a nonradiologist provider and found that this could have negatively affected performance in up to 43.3% of the cases and positively affected performance in only 2.0% of the cases. A recent study showed that errors in AI models can affect even trained radiologists' performance in chest radiograph interpretation (21).

One must keep in mind that GPT-4V is still under development, and detailed technical specifications remain undisclosed (8). Input is restricted to a limited number of two-dimensional images, with multisection image stacks (eg, those in MRI and CT) not being suitable. Although not explicitly optimized to interpret radiologic imagery, LLMs are unreliable, especially in complex and rare abnormalities (22). ChatGPT may refuse to interpret images when directly asked, but this safeguard can easily be bypassed.

Given these results, potential applications of GPT-4V in its current form include detecting out-of-distribution data. In a recent study, significant performance improvements were found, with an upstream in-distribution voting approach for chest radiograph classification (23). GPT-4V could render the acquisition of tailored data sets and training of specialized models obsolete.

The field of multimodal LLMs is rapidly evolving, and OpenAI is not the only competitor. Open-source models, such as LLaVA (Large Language and Vision Assistant) (24), are advantageous in the medical context. These open-source models can be deployed locally, which guarantees the protection of sensitive data. Fine-tuning these models toward radiology report generation promises better performance and has already been performed, including in LLaVa-Med (25), LlaVa-Rad (26), and CXR-LLAVA (27). This approach relies on large data sets of image-text pairs

Table 2: Consistency Test

Item	Reports with Perfect Agreement*	Randolph Free-Marginal κ
Modality	100 (25/25)	NA
Anatomic region	100 (25/25)	1.0
Description	76 (19/25)	0.68
Diagnosis	64 (16/25)	0.52

Note.—The percentage of free-text reports with perfect agreement and the Randolph free-marginal κ for testing consistency across three outputs are for correctly identifying the modality, anatomic region, description of findings, and diagnosis. The modality was correctly identified in all runs; however, free-marginal κ is not defined in cases with perfect agreement but only one label is present (zero division), leading to the reporting of NA. NA = not applicable.

* Data are percentages, with numbers of reports in parentheses.

Table 3: Performance in Binary Classification Tasks

Classification Task	GPT-4V			Readers		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Pathologic finding						
Ischemic stroke	46.0 (23/50) [34.0, 60.0]	56.0 (14/25) [36.0, 75.0]	36.0 (9/25) [17.4, 55.2]	100.0 (250/250) [100.0, 100.0]	100.0 (125/125) [100.0, 100.0]	100.0 (125/125) [100.0, 100.0]
Brain hemorrhage	54.0 (27/50) [40.0, 68.0]	100.0 (25/25) [100.0, 100.0]	8.0 (2/25) [0.0, 20.8]	99.6 (249/250) [98.8, 100.0]	100.0 (125/125) [100.0, 100.0]	99.2 (124/125) [97.4, 100.0]
Multiple sclerosis	60.0 (30/50) [46.0, 74.0]	74.1 (20/27) [55.6, 89.7]	43.5 (10/23) [23.8, 64.7]	96.8 (242/250) [94.4, 98.8]	97.8 (132/135) [94.9, 100.0]	95.7 (110/115) [91.3, 99.1]
Brain tumor	56.0 (28/50) [42.0, 70.0]	92.0 (23/25) [80.0, 100.0]	20.0 (5/25) [4.5, 36.0]	100.0 (250/250) [100.0, 100.0]	100.0 (125/125) [100.0, 100.0]	100.0 (125/125) [100.0, 100.0]
Lung cancer	56.0 (28/50) [42.0, 68.0]	64.0 (16/25) [42.9, 81.8]	48.0 (12/25) [29.2, 66.7]	99.6 (249/250) [98.8, 100.0]	100.0 (125/125) [100.0, 100.0]	99.2 (124/125) [97.4, 100.0]
Pneumonia	64.0 (32/50) [50.0, 76.0]	96.0 (24/25) [86.4, 100.0]	32.0 (8/25) [14.3, 50.0]	98.0 (245/250) [96.0, 99.6]	97.6 (122/125) [94.5, 100.0]	98.4 (123/125) [95.9, 100.0]
Pulmonary embolism	46.0 (23/50) [32.0, 60.0]	76.0 (19/25) [56.5, 91.7]	16.0 (4/25) [3.7, 33.3]	90.4 (226/250) [86.8, 94.0]	90.4 (113/125) [84.6, 95.4]	90.4 (113/125) [85.1, 95.2]
Pneumothorax	62.0 (31/50) [48.0, 74.0]	72.0 (18/25) [54.8, 88.5]	52.0 (13/25) [32.1, 70.0]	96.0 (240/250) [93.6, 98.4]	95.2 (119/125) [91.0, 98.4]	96.8 (121/125) [93.4, 99.2]
Fracture	54.0 (27/50) [40.0, 68.0]	72.0 (18/25) [54.8, 90.0]	36.0 (9/25) [17.9, 55.6]	94.4 (236/250) [91.6, 97.2]	93.6 (117/125) [88.8, 97.5]	95.2 (119/125) [90.8, 98.5]
Overall	55.3 (249/450) [51.1, 60.2]	78.0 (177/227) [72.6, 83.3]	32.3 (72/223) [26.4, 38.2]	97.2 (2187/2250) [96.5, 97.9]	97.2 (1103/1135) [96.1, 98.1]	97.2 (1084/1115) [96.2, 98.2]

Note.—Data are percentages, with numbers of images in parentheses and 95% CIs in brackets. Performance for all classification tasks (n = 50 images for each task) is reported separately and aggregated for all tasks as percentages. Results for readers and tasks were pooled by concatenating predictions and reference standards for individual readers or tasks before performance metric calculation. OpenAI's GPT-4V was used in this study (5).

Table 4: Interrater Agreement

Rater	Radiologist 1	Radiologist 2	Radiologist 3	Radiologist 4	Nonradiologist
Radiologist 1					
Radiologist 2	0.94				
Radiologist 3	0.94	0.94			
Radiologist 4	0.91	0.91	0.92		
Nonradiologist	0.9	0.92	0.92	0.9	

Note.—Cohen κ values were calculated between each pair of human readers of the binary classification task (pooled across all tasks) to quantify the level of interrater agreement.

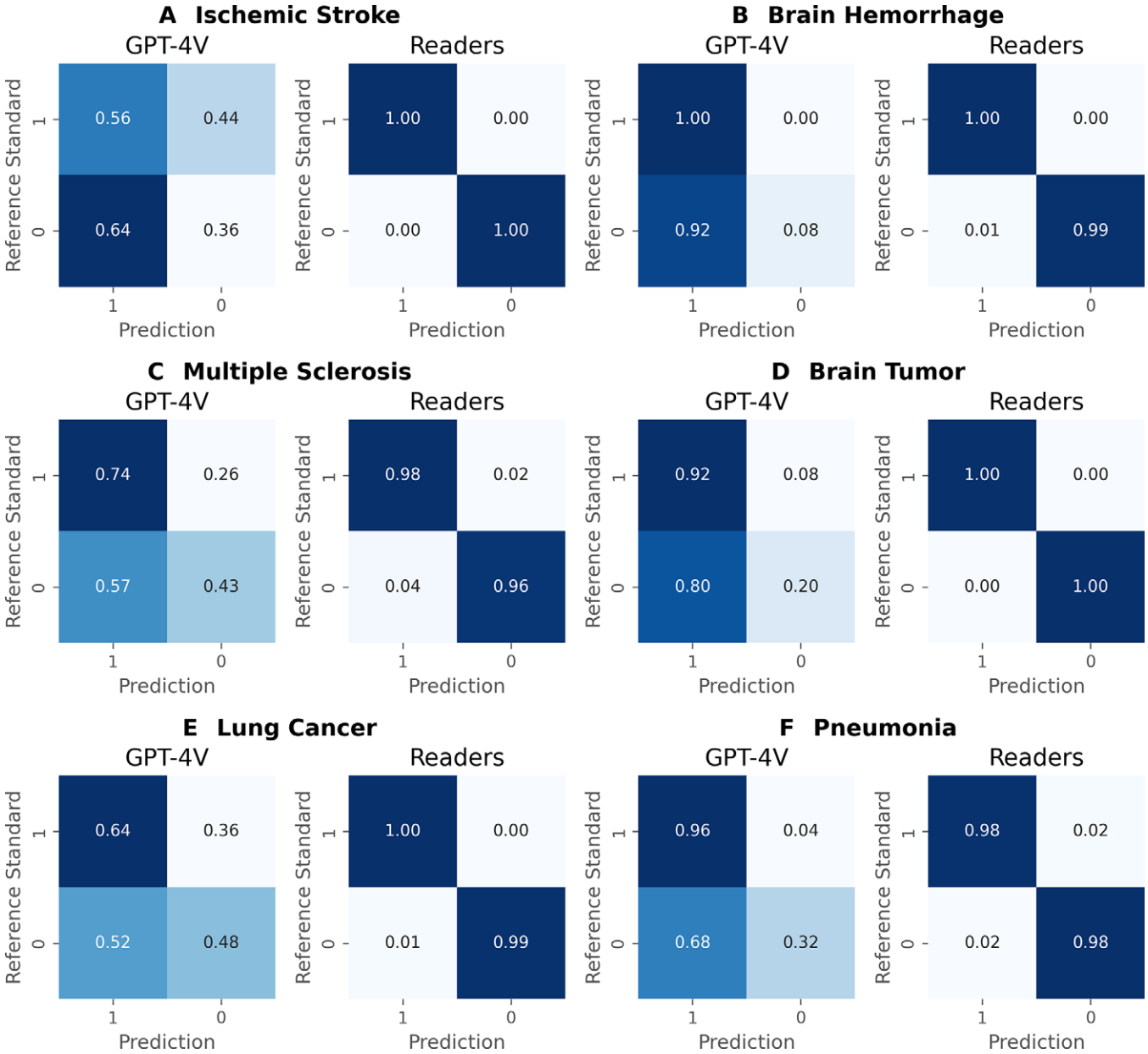


Figure 3: Performance in classification tasks for GPT-4V (OpenAI) and human readers. **(A–F)** Confusion matrices show the actual versus predicted classes for ischemic stroke **(A)**, brain hemorrhage **(B)**, multiple sclerosis **(C)**, brain tumor **(D)**, lung cancer **(E)**, and pneumonia **(F)**. Results of all readers were pooled. The y-axis represents the reference standard label, with 1 indicating the abnormality is present and 0 indicating the abnormality is not present. The x-axis shows whether the abnormality was predicted (reference standard 1) or not (reference standard 0) by GPT-4V or the readers. On this basis, the true-positive rate (sensitivity; top left), false-negative rate (top right), false-positive rate (bottom left), and true-negative rate (specificity; bottom right) were calculated. Squares are darker when the corresponding metric approaches 1 and lighter when it approaches 0 *(Fig 3 continues)*.

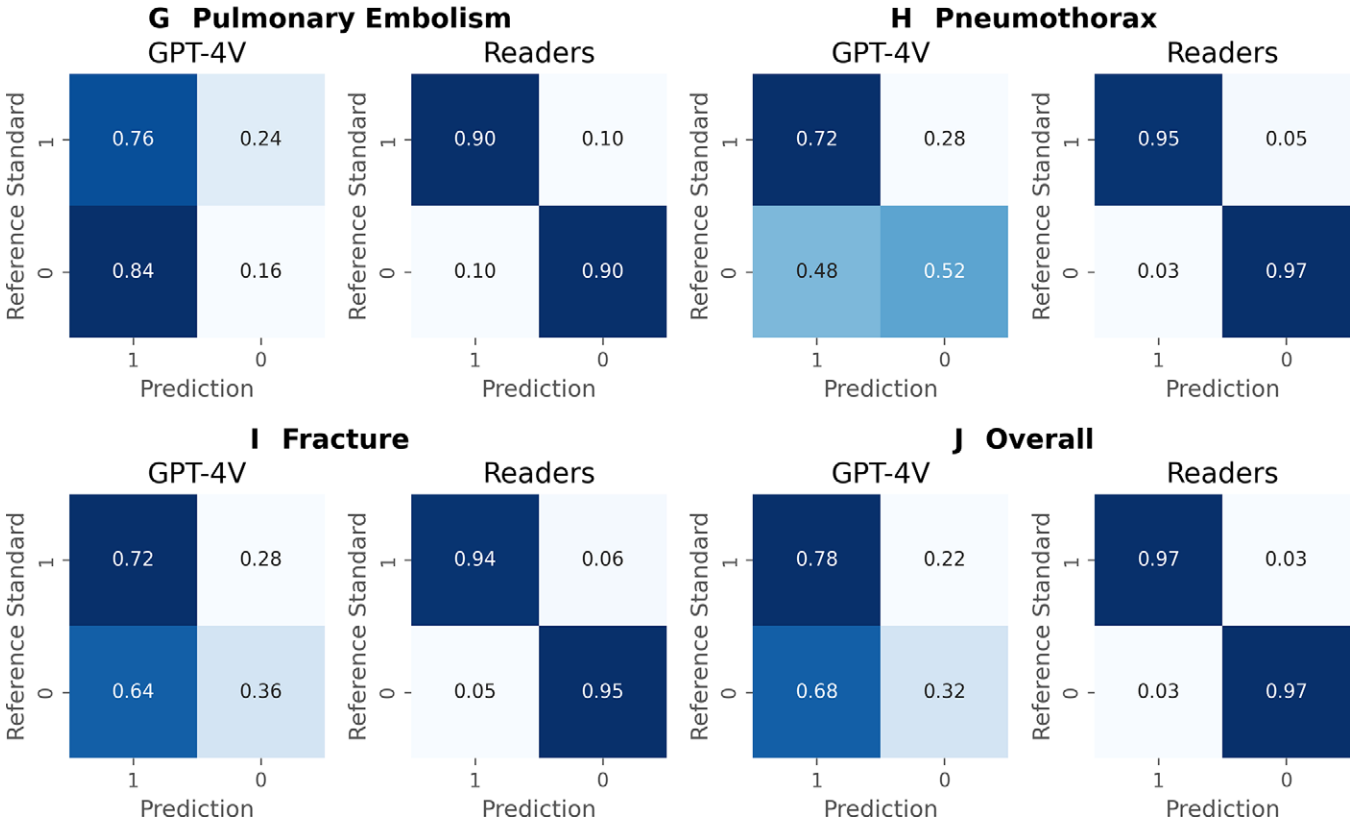


Figure 3 *(continued)*: Performance in classification tasks for GPT-4V (OpenAI) and human readers. **(G–J)** Confusion matrices show the actual versus predicted classes for pulmonary embolism **(G)**, pneumothorax **(H)**, fracture **(I)**, and overall (aggregated for all tasks) **(J)**. Results of all readers were pooled. The y-axis represents the reference standard label, with 1 indicating the abnormality is present and 0 indicating the abnormality is not present. The x-axis shows whether the abnormality was predicted (reference standard 1) or not (reference standard 0) by GPT-4V or the readers. On this basis, the true-positive rate (sensitivity; top left), false-negative rate (top right), false-positive rate (bottom left), and true-negative rate (specificity; bottom right) were calculated. Squares are darker when the corresponding metric approaches 1 and lighter when it approaches 0.

from the radiology domain. These are most commonly available for chest radiographs, limiting the broad generalizability in radiologic imaging. LLaVa-Med uses a subset of the PMC-15M (15 million image-text pairs extracted from PubMed Central) data set for fine-tuning with radiographic, CT, and MRI images (and gross pathologic and histopathologic images), substantially expanding the scope of the model (28). Fundamental limitations still exist. The model is trained only on two-dimensional data, not accounting for the complexity of findings in three-dimensional medical data. Pathologic images are generally overrepresented in medical publications, which may exacerbate the problem of false-positive findings. The image resolution of the model is 224 × 224 pixels (even lower than that of GPT-4V), whereby small details in the images may be missed. Additionally, these models often rely on fine-tuning data synthesized by other LLMs. Careful attention must be paid to quality assurance. Otherwise, a vicious cycle of declining quality could occur.

Our study had several limitations. First, we used a cross-section of patients and refrained from collecting detailed clinical data. Second, although GPT-4V supports multi-image input, we provided only one image per case to ensure uniformity. However, we also tested multiple image inputs for intracranial hemorrhage detection and did not observe a relevant change in performance. Third, we queried the model only once for each image (except for the subset for consistency testing and if the model refused interpretation). Our approach involved submitting a singular prompt without follow-up questions. Although carefully crafted, the prompts used may affect response quality, and further research on prompt strategies for medical imaging is needed. Fourth, no clinical data were provided with the prompts. Fifth, we calculated only the hypothetical impact on a nonradiologist provider, but we did not provide the model output during the readings to assess the actual influence of the model. Further studies that specifically address this issue are essential. Finally, the human readers performed only the binary classification task. This methodologic decision ensured that each reader assessed each image only once, preventing recall bias. The classification task was chosen over generating free-text reports, as it allowed for a standardized assessment and the calculation of more detailed performance metrics.

In conclusion, GPT-4V, in its earliest form, reliably determined the imaging modality and anatomic region from single radiologic images. However, despite providing convincing-sounding output, it cannot interpret medical images, as it failed to detect, classify, or rule out abnormalities safely. Nevertheless, large vision-language models show potential as foundation models in radiology. Adapting to field-specific requirements, such as stack inputs, and carefully fine-tuning the model for medical image analysis (eg, by leveraging representative data of a large health care system) should improve performance. Thoroughly testing diagnostic capabilities and challenging U.S. Food and Drug Administration–approved algorithms is mandated before considering multimodal large language models in medical practice.

Deputy Editor: Linda Moy
Scientific Editor: Shannyn Wolfe (AJE)

Acknowledgments: The VolkswagenStiftung provided Open Access funding for this publication (Grant Number: 98 525). We thank Florian Zeman, MSc (Center for Clinical Studies, University of Regensburg Medical Center, Regensburg, Germany), for helping with the statistical analysis. Q.D.S thanks Rajiv Gupta, MD, PhD (Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA), for his invaluable mentorship. We acknowledge that OpenAI's GPT-4V *(platform.openai.com)* was used to generate outputs in this study.

Author contributions: Guarantors of integrity of entire study, **Q.D.S., G.N., A.K.M., C.S., A.S.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **Q.D.S., F.N., G.N., I.E., C.W., O.W.H.**; clinical studies, **Q.D.S., L.S.K., G.N., I.E., J.R., M.S., C.S., O.W.H.**; experimental studies, **Q.D.S., F.N., G.N., A.K.M., A.S.**; statistical analysis, **Q.D.S., G.N., A.S.**; and manuscript editing, **Q.D.S., F.N., G.N., S.M., I.E., I.W., C.W., C.S., O.W.H., A.S.**

Data sharing: Data generated or analyzed during the study are available from the corresponding author by request.

Disclosures of conflicts of interest: **Q.D.S.** Travel support from RSNA Introduction to Research for International Young Academics (IRIYA) 2022. **F.N.** No relevant relationships. **L.S.K.** No relevant relationships. **G.N.** No relevant relationships. **A.K.M.** No relevant relationships. **S.M.** No relevant relationships. **I.E.** No relevant relationships. **J.R.** No relevant relationships. **M.S.** No relevant relationships. **I.W.** No relevant relationships. **C.W.** Advisory board for BALT (new study on flow diversion). **C.S.** No relevant relationships. **O.W.H.** Consulting fees and/or advisory board, Boehringe Ingelheim; lecture payments, Boehringer Ingelheim and Roche. **A.S.** No relevant relationships.

References

1. Adams LC, Truhn D, Busch F, et al. Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. Radiology 2023;307(4):e230725.
2. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. Radiology 2023;308(3):e231362.
3. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. Radiology 2023;307(5):e230582.
4. Doshi R, Amin KS, Khosla P, Bajaj SS, Chheang S, Forman HP. Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis. Radiology 2024;310(3):e231593.
5. GPT-4V (GPT-4 1106-vision-preview model). OpenAI. https://platform.openai.com/docs/overview. Accessed February 18 to March 11, 2024.
6. Willemink MJ, Roth HR, Sandfort V. Toward Foundational Deep Learning Models for Medical Imaging in the New Era of Transformer Networks. Radiol Artif Intell 2022;4(6):e210284.
7. Yang Z, Li L, Lin K, et al. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). arXiv 2309.17421 [preprint] https://arxiv.org/abs/2309.17421. Posted September 29, 2023. Updated October 11, 2023. Accessed October 12, 2023.
8. GPT-4V(ision) System Card. OpenAI. https://cdn.openai.com/papers/GPTV_System_Card.pdf. Published September 25, 2023. Accessed October 14, 2023.
9. Han T, Adams LC, Bressem KK, Busch F, Nebelung S, Truhn D. Comparative Analysis of Multimodal Large Language Model Performance on Clinical Vignette Questions. JAMA 2024;331(15):1320–1321.
10. Sultan LR, Mohamed MK, Andronikou S. ChatGPT-4: a breakthrough in ultrasound image analysis. Radiol Advances 2024;1(1):umae006.
11. Deng J, Heybati K, Shammas-Toma M. When vision meets reality: Exploring the clinical applicability of GPT-4 with vision. Clin Imaging 2024;108:110101.
12. 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(2):e233270.
13. Lee KH, Lee RW, Kwon YE. Validation of a Deep Learning Chest X-ray Interpretation Model: Integrating Large-Scale AI and Large Language Models for Comparative Analysis with ChatGPT. Diagnostics (Basel) 2023;14(1):90.
14. Yan Z, Zhang K, Zhou R, et al. Multimodal ChatGPT for Medical Applications: an Experimental Study of GPT-4V. arXiv 2310.19061 [preprint] https://arxiv.org/abs/2310.19061. Posted October 29, 2023. Accessed March 20, 2024.
15. Busch F, Han T, Makowski M, et al. From Text to Image: Exploring GPT-4Vision's Potential in Advanced Radiological Analysis across Subspecialties. arXiv 2311.14777 [preprint] http://arxiv.org/abs/2311.14777. Posted November 24, 2023. Accessed February 17, 2024.
16. Brin D, Sorin V, Barash Y, et al. Assessing GPT-4 Multimodal Performance in Radiological Image Analysis. Eur Radiol 2024. 10.1007/s00330-024-11035-5. Published online August 30, 2024.
17. Huda W, Abrahams RB. X-ray-based medical imaging and resolution. AJR Am J Roentgenol 2015;204(4):W393-W397.
18. Sievert M, Aubreville M, Mueller SK, et al. Diagnosis of malignancy in oropharyngeal confocal laser endomicroscopy using GPT 4.0 with vision. Eur Arch Otorhinolaryngol 2024;281(4):2115–2122.
19. Salehinejad H, Kitamura J, Ditkofsky N, et al. A real-world demonstration of machine learning generalizability in the detection of intracranial hemorrhage on head computerized tomography. Sci Rep 2021;11(1):17051.
20. Thian YL, Ng D, Hallinan JTPD, et al. Deep Learning Systems for Pneumothorax Detection on Chest Radiographs: A Multicenter External Validation Study. Radiol Artif Intell 2021;3(4):e200190.
21. Yu F, Moehring A, Banerjee O, Salz T, Agarwal N, Rajpurkar P. Heterogeneity and predictors of the effects of AI assistance on radiologists. Nat Med 2024;30(3):837–849.
22. 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.
23. Wollek A, Willem T, Ingrisch M, Sabel B, Lasser T. Out-of-distribution detection with in-distribution voting using the medical example of chest x-ray classification. Med Phys 2024;51(4):2721–2732.
24. Liu H, Li C, Wu Q, Lee YJ. Visual Instruction Tuning. arXiv 2304.08485 [preprint] https://arxiv.org/abs/2304.08485. Posted April 17, 2023. Updated December 11, 2023. Accessed March 21, 2024.
25. Li C, Wong C, Zhang S, et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. arXiv 2306.00890 [preprint] https://arxiv.org/abs/2306.00890. Posted June 1, 2023. Accessed March 21, 2024.
26. Chaves JMZ, Huang S-C, Xu Y, et al. Training Small Multimodal Models to Bridge Biomedical Competency Gap: A Case Study in Radiology Imaging. arXiv 2403.08002 [preprint] https://arxiv.org/abs/2403.08002. Posted March 12, 2024. Updated June 27, 2024. Accessed March 20, 2024.
27. Lee S, Youn J, Kim H, et al. CXR-LLAVA: a multimodal large language model for interpreting chest X-ray images. arXiv 2310.18341 [preprint] https://arxiv.org/abs/2310.18341. Posted October 22, 2023. Updated January 14, 2024. Accessed March 20, 2024.
28. Zhang S, Xu Y, Usuyama N, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv 2303.00915 [preprint] https://arxiv.org/abs/2303.00915. Posted March 2, 2023. Updated January 16, 2024. Accessed March 21, 2024.