

Evaluating GPT-4V (GPT-4 with Vision) on Detection of Radiologic Findings on Chest Radiographs

Yiliang Zhou, MS • Hanley Ong, MD • Patrick Kennedy, MD • Carol C. Wu, MD • Jacob Kazam, MD • Keith Hentel, MD, MS • Adam Flanders, MD • George Shib, MD • Yifan Peng, PhD

From the Departments of Population Health Sciences (Y.Z., Y.P.) and Radiology (H.O., P.K., J.K., G.S.), Weill Cornell Medicine, 425 E 61st St, Ste 301, New York, NY 10065; Department of Thoracic Imaging, University of Texas MD Anderson Cancer Center, Houston, Tex (C.C.W.); and Department of Radiology, Thomas Jefferson University Hospital, Philadelphia, Pa (A.F.). Received December 4, 2023; revision requested January 11, 2024; revision received March 27; accepted March 29. Address correspondence to Y.P. (email: yip4002@med.cornell.edu).

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

Radiology 2024; 311(2):e233270 • <https://doi.org/10.1148/radiol.233270> • Content codes: CH AI

Background: Generating radiologic findings from chest radiographs is pivotal in medical image analysis. The emergence of OpenAI's generative pretrained transformer, GPT-4 with vision (GPT-4V), has opened new perspectives on the potential for automated image-text pair generation. However, the application of GPT-4V to real-world chest radiography is yet to be thoroughly examined.

Purpose: To investigate the capability of GPT-4V to generate radiologic findings from real-world chest radiographs.

Materials and Methods: In this retrospective study, 100 chest radiographs with free-text radiology reports were annotated by a cohort of radiologists, two attending physicians and three residents, to establish a reference standard. Of 100 chest radiographs, 50 were randomly selected from the National Institutes of Health (NIH) chest radiographic data set, and 50 were randomly selected from the Medical Imaging and Data Resource Center (MIDRC). The performance of GPT-4V at detecting imaging findings from each chest radiograph was assessed in the zero-shot setting (where it operates without prior examples) and few-shot setting (where it operates with two examples). Its outcomes were compared with the reference standard with regards to clinical conditions and their corresponding codes in the *International Statistical Classification of Diseases, Tenth Revision* (ICD-10), including the anatomic location (hereafter, laterality).

Results: In the zero-shot setting, in the task of detecting ICD-10 codes alone, GPT-4V attained an average positive predictive value (PPV) of 12.3%, average true-positive rate (TPR) of 5.8%, and average F1 score of 7.3% on the NIH data set, and an average PPV of 25.0%, average TPR of 16.8%, and average F1 score of 18.2% on the MIDRC data set. When both the ICD-10 codes and their corresponding laterality were considered, GPT-4V produced an average PPV of 7.8%, average TPR of 3.5%, and average F1 score of 4.5% on the NIH data set, and an average PPV of 10.9%, average TPR of 4.9%, and average F1 score of 6.4% on the MIDRC data set. With few-shot learning, GPT-4V showed improved performance on both data sets. When contrasting zero-shot and few-shot learning, there were improved average TPRs and F1 scores in the few-shot setting, but there was not a substantial increase in the average PPV.

Conclusion: Although GPT-4V has shown promise in understanding natural images, it had limited effectiveness in interpreting real-world chest radiographs.

© RSNA, 2024

Supplemental material is available for this article.

An earlier incorrect version appeared online. This article was corrected on May 8, 2024.

Generating radiologic findings from chest radiographs is pivotal in medical image analysis (1). Recent advancements in fine-tuned pretrained models have showcased their capability to translate image content into text (2). However, these models are often trained on extensive nonspecific data sets and may need more domain-specific tuning for chest radiographs. The emergence of OpenAI's generative pretrained transformer, GPT-4 with vision (GPT-4V) (3), a multimodal large language model (LLM) with visual recognition, has opened new perspectives on the potential for automated image-text pair generation in the medical care domain. Advanced multimodal LLMs, such as GPT-4V, can understand both text and images. While several studies have investigated the performance of GPT-4 in generating radiologic impressions (4) and summarizing clinical trials (5), the practical application of multimodal LLMs to the interpretation of real-world chest radiographs is yet to be thoroughly examined. Motivated

by this knowledge gap, the aim of this study was to investigate the capability of GPT-4V to generate radiologic findings from real-world chest radiographs.

Materials and Methods

Because of the publicly available nature of the data sets used in this study, the requirement to obtain written informed consent from all patients was waived by the institutional review board.

Study Design and Data Collection

In this retrospective study, a total of 100 chest radiographs and radiology reports were independently annotated by a cohort of radiologists that included two attending physicians and three residents to establish a reference standard (Fig 1). Of 100 chest radiographs, 50 were randomly selected from the National Institutes of Health (NIH) chest radiographic data set, and their corresponding reports

Abbreviations

ICD-10 = *International Statistical Classification of Diseases, Tenth Revision*; LLM = large language model; MIDRC = Medical Imaging and Data Resource Center; NIH = National Institutes of Health; PPV = positive predictive value; TPR = true-positive rate

Summary

This study examined the application of GPT-4 with vision (GPT-4V), a multimodal large language model with visual recognition, in detecting radiologic findings from a set of 100 chest radiographs and suggests that GPT-4V is currently not ready for real-world diagnostic usage in interpreting chest radiographs.

Key Results

- In this retrospective study, 100 chest radiographs with free-text radiology reports were annotated by two radiology attending physicians and three radiology residents to establish a reference standard; this reference standard was compared with the performance of GPT-4 with vision (GPT-4V) in generating imaging findings for 100 randomly selected radiographs from real-world data sets.
- The effectiveness of GPT-4V in interpreting chest radiographs is limited.
- When contrasting zero-shot and few-shot learning, there were improved average true-positive rates and F1 scores with few-shot learning, but there was not a substantial increase in the average positive predictive value.

were dictated by one radiology attending physician and three radiology residents (4). These 50 patients have been previously reported (5). The prior article dealt with generation of the impression section by using the findings section in the report, whereas the current article describes the generation of a table of radiologic findings from the image.

The remaining 50 chest radiographs and de-identified free-text radiology reports were randomly selected from the Medical Imaging and Data Resource Center (MIDRC) (4). Each report included a findings section and an impressions section.

Of 100 chest radiographs and reports, 10 cases were randomly selected (five from the NIH data set and five from the MIDRC data set), and two of these were randomly selected to serve as few-shot examples for the GPT-4V model. The remaining 90 cases (45 from each data set) were used to evaluate the performance of GPT-4V in a zero-shot learning setting, where it operates without prior examples, and in a few-shot learning setting, where it operates with two examples. Its outcomes were then compared with the reference standard as annotated by

radiologists with regard to clinical conditions and their corresponding codes in the *International Statistical Classification of Diseases, Tenth Revision* (ICD-10), including the anatomic location (hereafter, laterality).

GPT-4 with Vision

GPT-4V (accessed October 13, 2023; OpenAI) was used in this study (3). GPT-4V is a version of GPT-4 that allows users to instruct the LLM to analyze image inputs.

Experimental Setup

To obtain the reference standard tables, GPT-4 was used to convert each free-text radiology report into a table of radiologic findings by using a textual prompt (Appendix S1). This table included the radiologic findings, the corresponding ICD-10 diagnostic codes and their laterality, as well as descriptions of the ICD-10 codes (Table S1). Subsequently, each report was independently evaluated by three readers from a cohort of five board-certified radiologists and residents (H.O., P.K., C.C.W., J.K., G.S.). Two of the readers were 3rd-year radiology residents and the remaining three were radiology attending physicians, each with over 15 years of experience. Their radiology subspecialties cover chest, emergency department, bone, neurology, and body imaging. All readers had access to the image views and reports but not to additional clinical or patient data. Both data sets, NIH and MIDRC, were comprehensively reviewed, with the 50 reports of each data set being examined by three readers to maintain consistency and objectivity in the evaluation process. Findings were only included in the final tables if they were observed by at least two of the three readers; conversely, findings were excluded if two or more readers did not identify them. The majority vote principle was employed to provide a clear consensus for the presence or absence of radiologic findings and lead to the final reference standard table for each radiograph.

Evaluating Performance in the Zero-Shot Setting

To evaluate the performance of GPT-4V in the zero-shot setting, the chest radiographs (Fig 2A) were input into the GPT-4V model (accessed October 13, 2023) (2,3) along with the prompt (Fig 2B). The aim of this step was to generate a radiologic findings table (Fig 2C) comparable to the reference standard table. The analysis concentrated on aligning positive radiologic finding identification and laterality between GPT-4V-generated tables and the consensus reference standard tables. The positive predictive value (PPV), true-positive rate (TPR), and F1 score were

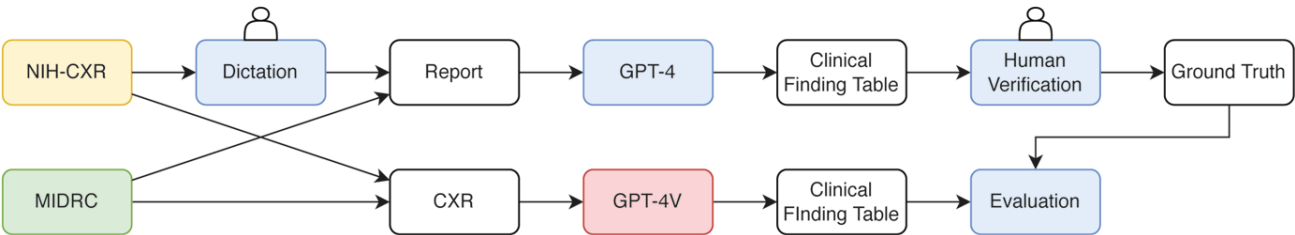
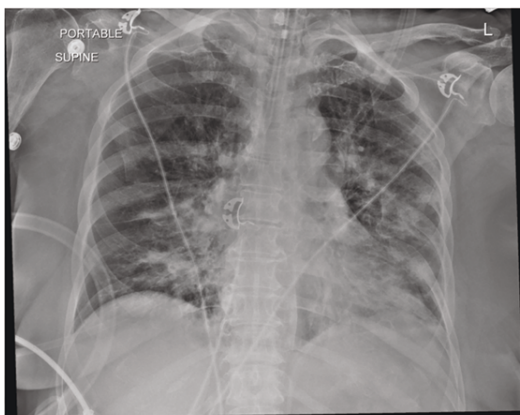


Figure 1: Diagram shows the study workflow, including construction of data and application of GPT-4 and GPT-4 with vision (GPT-4V). CXR = chest radiograph, MIDRC = Medical Imaging and Data Resource Center, NIH = National Institutes of Health.

A Chest radiograph**B Prompt to read a chest radiograph and generate a table of radiological findings**

Examine this image and look for any important clinical findings. Provide a summary in a table format where the positive clinical conditions are 1 and the negative clinical conditions are 0. Designate each condition as left side, right side, or bilateral. Provide an ICD-10 code in a separate column for positive findings only or N/A if not applicable.

Table columns include: [Exam No., Finding No., Clinical Finding, Left Side, Right Side, Bilateral, Midline, ICD-10 Code, ICD-10 Description]

Additional instructions:

1. Normal findings should be excluded from each table
2. Group similar findings together where possible for each table
3. Create a table

Figure 2: An example of GPT-4 with vision (GPT-4V) inputs and output, including the **(A)** chest radiograph, **(B)** prompt provided to GPT-4V to create a table of radiologic findings derived from the chest radiograph, and **(C)** resultant table of radiographic findings generated by GPT-4V. ICD-10 = *International Statistical Classification of Diseases, Tenth Revision*.

C GPT-4v-generated results

Radiological Finding	Location	ICD-10	ICD-10 Description
Pulmonary Infiltrate	Right Side	R91.8	Other nonspecific abnormal finding of lung field
Pleural Effusion	Right Side	J90	Pleural effusion, not elsewhere classified
Consolidation	Right Side	J18.9	Pneumonia, unspecified organism
Cardiomegaly	Midline	I51.7	Cardiomegaly
Medical Devices Present	Bilateral	Z96.0	Presence of urogenital implants

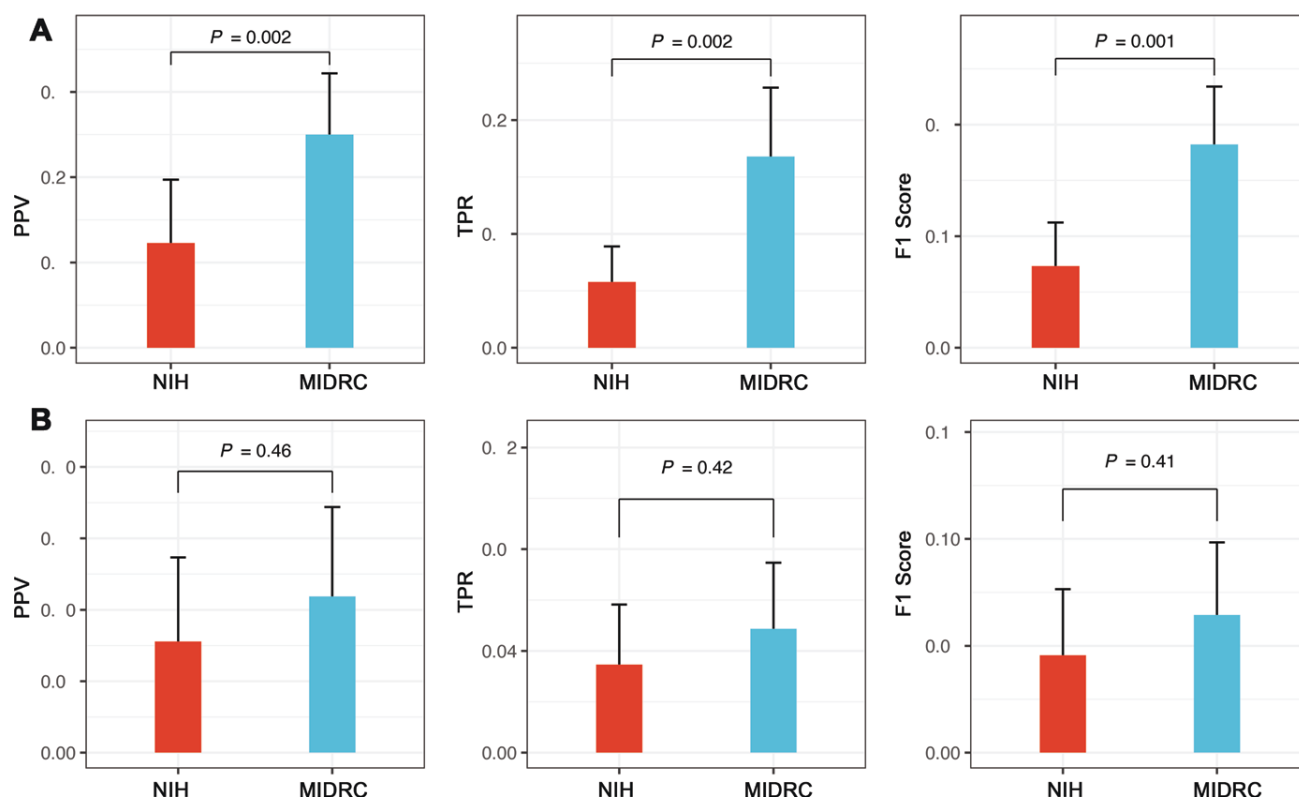


Figure 3: Bar graphs show the performance of GPT-4 with vision (GPT-4V) in the detection of radiologic findings from chest radiographs in the National Institutes of Health (NIH) and Medical Imaging and Data Resource Center (MIDRC) data sets in the zero-shot setting according to **(A)** the radiologic findings in *International Statistical Classification of Diseases, Tenth Revision* (ICD-10) codes only and **(B)** both the radiologic findings in ICD-10 codes and their corresponding lateralities. Statistical significance was assessed using the two-tailed *t* test. Error bars indicate the average mean value of the set of numbers with consideration of a confidence level within a normal distribution. PPV = positive predictive value, TPR = true-positive rate.

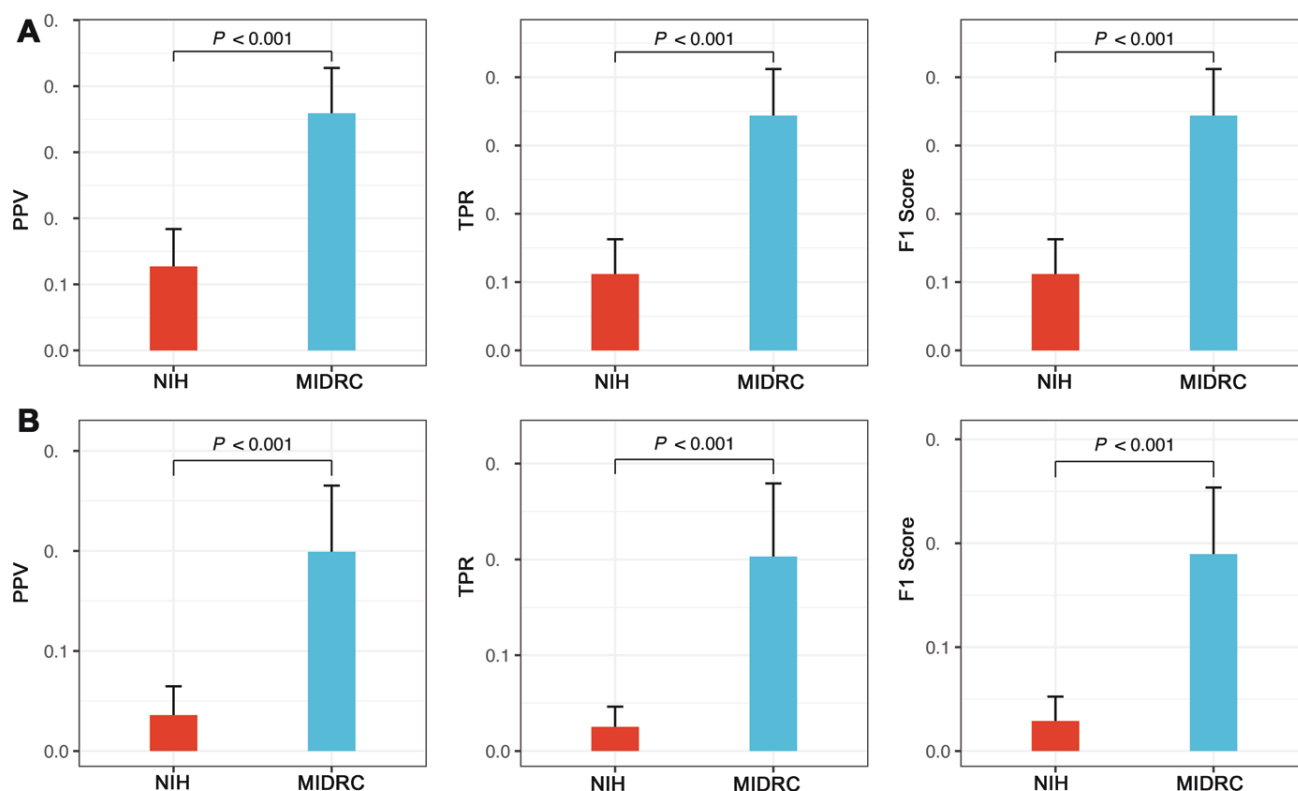


Figure 4: Bar graphs show the performance of GPT-4 with vision (GPT-4V) in the detection of radiologic findings from chest radiographs in the National Institutes of Health (NIH) and Medical Imaging and Data Resource Center (MIDRC) data sets in the few-shot setting according to (A) radiologic findings in *International Statistical Classification of Diseases, Tenth Revision* (ICD-10) codes only and (B) both the radiologic findings in ICD-10 codes and their corresponding lateralities. Statistical significance was assessed using the two-tailed *t* test. Error bars indicate the average mean value of the set of numbers with consideration of a confidence level within a normal distribution. PPV = positive predictive value, TPR = true-positive rate.

calculated at the report level (see Statistical Analysis). Notably, the per-report average F1 scores were used in this study as each case could have multiple diagnoses. These metrics were used to assess the accuracy of GPT-4V in detecting the ICD-10 codes and their respective laterality.

Evaluating Performance in the Few-Shot Setting

To evaluate the performance of GPT-4V in the few-shot setting, the input was extended to include two examples of chest radiographs with their corresponding radiologic findings tables before the prompt. From the pool of 10 chest radiographs, two were randomly selected to serve as few-shot examples for the GPT-4V model. Supplying the model with these examples helped to provide context, boosting the model's capacity to generate an accurate radiologic findings table. The same performance metrics (PPV, TPR, F1 score) were used at the report level to assess the effectiveness of few-shot learning.

Statistical Analysis

When obtaining the final reference standard tables, interrater agreement was assessed using the Cohen κ coefficient (6).

First, GPT-4V was employed to detect radiologic findings from each chest radiograph and these results were compared with the predicted findings obtained from the radiologists. To convey the performance evaluation, the PPV was used to denote the proportion of ICD-10 codes correctly predicted by

GPT-4V, while the TPR represented the ratio of true-positive predictions to the total number of ICD-10 codes identified by GPT-4V.

PPV, TPR, and F1 score were the metrics used to assess the performance of GPT-4V in detecting imaging findings from each chest radiograph. The F1 score is the harmonic mean of the PPV and TPR, per the following equation:

$$F1 = 2 \times \frac{PPV \times TPR}{PPV + TPR}$$

These predicted findings were then compared with those obtained from the radiologists.

In the evaluation of detecting ICD-10 codes, a radiologic finding was considered a true-positive finding if its ICD-10 code aligned with that in the reference standard table. In evaluating both the radiologic findings in ICD-10 codes and their corresponding lateralities, a radiologic finding was considered true positive if both its ICD-10 code and laterality matched those in the reference standard table.

As an example, when evaluating ICD-10 codes alone, there were two ICD-10 codes correctly predicted by GPT-4V (J90 and J18.9), with five findings predicted by GPT-4V (Fig 2) and four in the reference standard table (Table S1). Therefore, the PPV was 0.4 ($= 2/[2 + 3]$), TPR was 0.5 ($= 2/[2 + 2]$), and F1 score was 0.44.

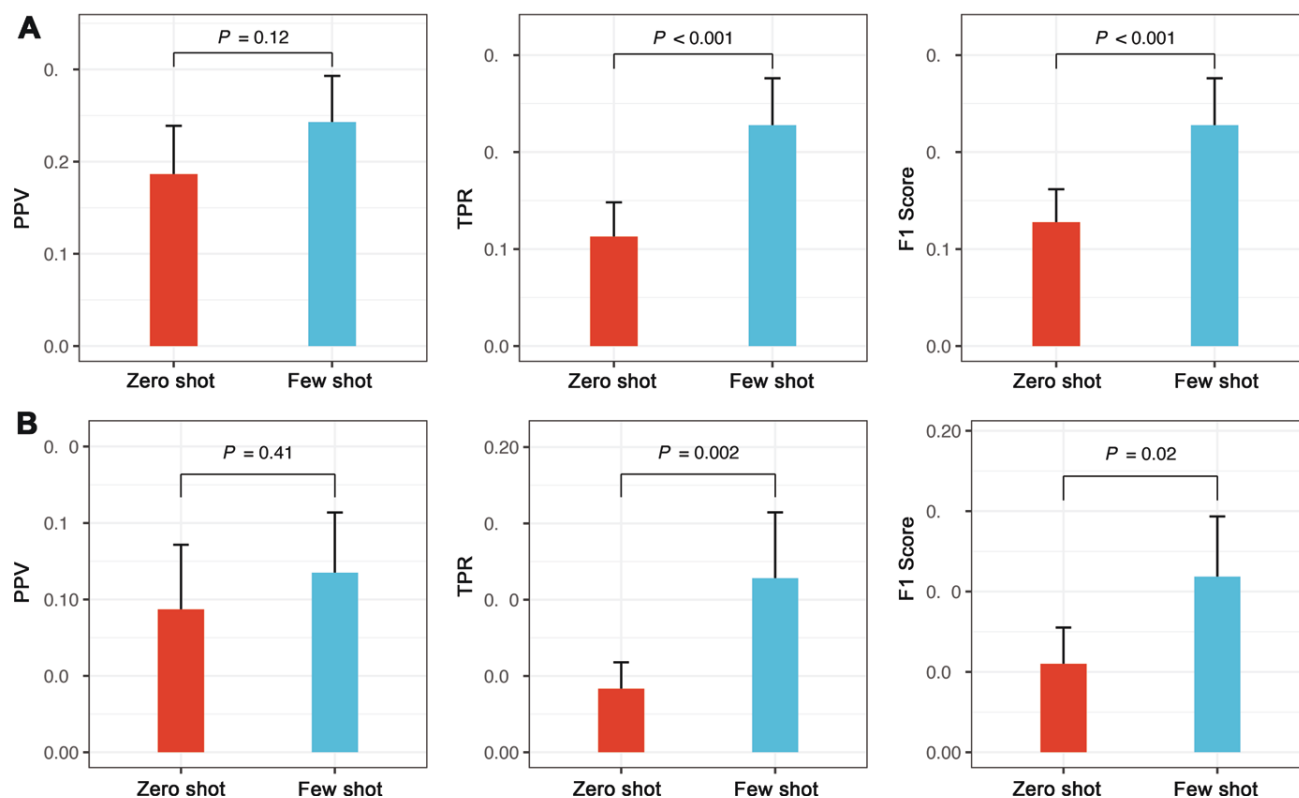


Figure 5: Bar graphs show the difference in performance of GPT-4V in the detection of radiologic findings from chest radiographs between zero-shot and few-shot settings according to **(A)** radiologic findings in *International Statistical Classification of Diseases, Tenth Revision (ICD-10)* codes only and **(B)** both the radiologic findings in ICD-10 codes and their corresponding lateralities. Statistical significance was assessed using the two-tailed *t* test. Error bars indicate the average mean value of the set of numbers with consideration of a confidence level within a normal distribution. PPV = positive predictive value, TPR = true-positive rate.

After obtaining the PPV, TPR, and F1 score for each chest radiograph, the macro averages were calculated for PPV, TPR, and F1 score across all 90 chest radiographs.

$P = .05$ was considered indicative of a statistically significant difference. Two-tailed *t* tests were used for calculating the *P* values for the performance metrics of GPT-4V in detecting radiologic findings from chest radiographs, specifically assessing the detection of findings represented by their associated lateralities and ICD-10 codes in both zero-shot and few-shot settings (Figs 3–5).

Results

Performance in the Zero-Shot Setting

The performance of the GPT-4V model in the zero-shot setting varied across the NIH and MIDRC data sets and for different scenarios (Fig 3). In the task of detecting ICD-10 codes alone, the model attained an average PPV of 5.53 of 45 radiographs (12.3%) (SD [0.25], SE [0.04], IQR [0.20]), average TPR of 2.60 of 45 radiographs (5.8%) (SD [0.10], SE [0.02], IQR [0.10]), and average F1 score of 3.30 of 45 radiographs (7.3%) (SD [0.13], SE [0.02], IQR [0.14]) on the NIH data set. Conversely, on the MIDRC data set, the model managed an average PPV of 11.25 of 45 radiographs (25.0%) (SD [0.24], SE [0.04], IQR [0.33]), average TPR of 7.56 of 45 radiographs (16.8%) (SD [0.20], SE [0.03], IQR [0.25]), and average F1 score of

Reader Agreement for the NIH and MIDRC Data Sets

Reader	NIH κ Coefficient	MIDRC κ Coefficient
A	0.78	NA
B	NA	0.96
C	-0.06	0.82
D	0.96	0.99

Note.—Interrater agreement between each reader and the reference standard table for both the NIH and MIDRC data sets was assessed with Cohen κ statistics. κ values were interpreted as follows: 1, indicates perfect agreement among readers and the reference standard; 0, indicates that agreement is no better than chance; and -1, indicates perfect disagreement. MIDRC = Medical Imaging and Data Resource Center, NA = not applicable, NIH = National Institutes of Health.

8.20 of 45 radiographs (18.2%) (SD [0.17], SE [0.03], IQR [0.29]). The notable differences in measurements between the two data sets were primarily due to the MIDRC data set having fewer missing GPT-4V-generated ICD-10 codes than the NIH data set. On the MIDRC data set, GPT-4V generated 144 radiologic findings, while the reference standard comprised 261 findings. However, on the NIH data set, GPT-4V produced 102 radiologic findings, while the reference standard comprised 220 findings. Nevertheless, when both the ICD-10 codes and their

corresponding laterality were taken into account, the GPT-4V model in the zero-shot setting produced an average PPV of 3.5 of 45 radiographs (7.8%) (SD [0.20], SE [0.03], IQR [0.0]), average TPR of 1.56 of 45 radiographs (3.5%) (SD [0.01], SE [0.01], IQR [0.0]), and average F1 score of 2.05 of 45 radiographs (4.5%) (SD [0.10], SE [0.02], IQR [0.0]) on the NIH data set; and an average PPV of 4.92 of 45 radiographs (10.9%) (SD [0.21], SE [0.03], IQR [0.25]), average TPR of 2.19 of 45 radiographs (4.9%) (SD [0.09], SE [0.01], IQR [0.10]), and average F1 score of 2.90 of 45 radiographs (6.4%) (SD [0.11], SE [0.02], IQR [0.15]) on the MIDRC data set.

Performance in the Few-Shot Setting

With few-shot learning, GPT-4V showed improved performance on both the NIH and MIDRC data sets (Fig 4). When the model was provided with two illustrative chest radiographs and their corresponding radiologic findings tables, there was a marked improvement on the NIH data set, with the average PPV increasing to 5.72 of 45 radiographs (12.7%) (SD [0.19], SE [0.03], IQR [0.25]). The average TPR also enhanced to 4.69 of 45 radiographs (10.4%) (SD [0.17], SE [0.03], IQR [0.22]), while the average F1 score reached 5.03 of 45 radiographs (11.1%) (SD [0.17], SE [0.03], IQR [0.22]). On the MIDRC data set, the average PPV improved to 16.15 of 45 radiographs (35.9%) (SD [0.23], SE [0.03], IQR [0.50]), the average TPR improved to 16.68 of 45 radiographs (37.1%) (SD [0.23], SE [0.03], IQR [0.50]), and the average F1 score improved to 15.47 of 45 radiographs (34.3%) (SD [0.23], SE [0.03], IQR [0.50]). When tasked with detecting both ICD-10 codes and their corresponding lateralities, the GPT-4V model demonstrated improved efficacy. On the NIH data set, it displayed an average PPV of 1.62 of 45 radiographs (3.5%) (SD [0.10], SE [0.01], IQR [0.0]), average TPR of 1.14 of 45 radiographs (2.5%) (SD [0.07], SE [0.01], IQR [0.0]), and average F1 score of 1.30 of 45 radiographs (2.8%) (SD [0.08], SE [0.01], IQR [0.0]). On the MIDRC data set, GPT-4V achieved an average PPV of 8.96 of 45 radiographs (19.9%) (SD [0.22], SE [0.03], IQR [0.33]), average TPR of 9.14 of 45 radiographs (20.3%) (SD [0.25], SE [0.04], IQR [0.33]), and average F1 score of 8.53 of 45 radiographs (19.0%) (SD [0.21], SE [0.03], IQR [0.31]).

When contrasting zero-shot and few-shot learning approaches on both data sets (Fig 5), there were improved average TPR and F1 scores with few-shot learning in both scenarios (ICD-10 codes only and ICD-10 codes with lateralities). Nonetheless, there was not a substantial increase in the average PPV, suggesting that while few-shot learning may enhance the model's capacity to detect findings, it does not noticeably enhance the precision of the predictions.

Interrater Agreement

The study compared each reader's outcomes to the reference standard to evaluate their interpretations (Table). On the NIH data set, the interrater agreement was 0.78 for reader A, -0.06 for reader C, and 0.96 for reader D. On the MIDRC data set, the interrater agreement was 0.96 for reader B, 0.82 for reader C, and 0.99 for reader D.

Discussion

The emergence of multimodal large language models (LLMs) that can understand both text and images, such as OpenAI's GPT-4V (3), shows potential for automated image-text pair generation. However, applying these models to real-world data is yet to be thoroughly examined. This study assessed the feasibility of using GPT-4V to detect radiologic findings from chest radiographs in both zero-shot and few-shot learning contexts. The results (average PPV, 5.53 of 45 radiographs [12.3%]; average TPR, 2.60 of 45 radiographs [5.8%]) demonstrated that radiologic findings tables generated by GPT-4V still need further preparation for use in clinical practice. We acknowledge a limitation in employing GPT-4 for converting radiologic reports into a structured table, where inaccuracies in the *International Statistical Classification of Diseases, Tenth Revision* (ICD-10) code assignments and distinguishing between radiologic findings and conclusions may impact the reliability and interpretability of the data. A notable limitation of the GPT-4V output was its failure to detect several clinical conditions based on corresponding ICD-10 codes. Overall, the top three findings that GPT-4V could not detect were "endotracheal tube," "central venous catheter," and "degenerative changes of osseous structures." Conversely, GPT-4V most accurately detected findings such as "chest drain," "air-space disease," and "lung opacity."

Although GPT-4V has shown promise in understanding real-world images (7), its effectiveness in interpreting real-world chest radiographs was limited.

Our study had limitations. First, few-shot learning may be more prone to generating ICD-10 codes already in the provided examples, potentially reducing the diversity of ICD-10 codes generated. Second, we were unable to access other multimodal LLMs that support image inputs; consequently, we lacked comparative data with respect to the results of GPT-4V. Finally, due to the limited size of the data set and relatively low level of interrater agreement among the radiologists, analysis by GPT-4V may have been challenging.

In conclusion, GPT-4V has shown promise in understanding natural images but had limited effectiveness in interpreting real-world chest radiographs. Our results highlight the need for additional comprehensive development and assessment prior to incorporating the GPT-4V model into clinical practice routines. Task-specific, fine-tuned, multimodal large language models or foundation models are urgently needed for this purpose, although it is not necessarily the best solution. To yield robust and generalizable results, we plan to explore larger and more diverse data sets using real-world data in future studies. This will involve including multiple modalities, such as CT and MRI of the brain, to conduct a more thorough evaluation of the performance of GPT-4V.

Deputy Editor: Linda Moy

Scientific Editor: Sarah Atzen

Acknowledgement: We acknowledge parts of this article were generated by OpenAI's GPT-4 and GPT-4V (<http://openai.com>).

Author contributions: Guarantors of integrity of entire study, Y.Z., G.S., Y.P.; 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, **Y.Z.**, **P.K.**, **K.H.**, **G.S.**, **Y.P.**; clinical studies, **Y.Z.**, **C.C.W.**, **J.K.**; experimental studies, **Y.Z.**, **H.O.**, **P.K.**, **J.K.**, **K.H.**, **G.S.**, **Y.P.**; statistical analysis, **Y.Z.**, **Y.P.**; and manuscript editing, **Y.Z.**, **H.O.**, **K.H.**, **A.F.**, **G.S.**, **Y.P.**

Disclosures of conflicts of interest: **Y.Z.** No relevant relationships. **H.O.** No relevant relationships. **P.K.** No relevant relationships. **C.C.W.** No relevant relationships. **J.K.** No relevant relationships. **K.H.** Patents planned, issued, or pending with Weill Cornell Medicine; board member of New York State Radiological Society. **A.F.** No relevant relationships. **G.S.** No relevant relationships. **Y.P.** Recipient of a National Science Foundation CAREER Award (2145640).

References

1. Speets AM, van der Graaf Y, Hoes AW, et al. Chest radiography in general practice: indications, diagnostic yield and consequences for patient management. *Br J Gen Pract* 2006;56(529):574–578.
2. 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>. Published September 29, 2023. Accessed November 15, 2023.
3. GPT-4V. OpenAI. <https://openai.com/research/gpt-4v-system-card>. Accessed October 13, 2023.
4. Wang X, Peng Y, Lu L, Bagheri M, Lu Z, Summers R. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017, 3462–3471.
5. Sun Z, Ong H, Kennedy P, et al. Evaluating GPT-4 on Impressions Generation in Radiology Reports. *Radiology* 2023;307(5):e231259.
6. Waisberg E, Ong J, Masalkhi M, et al. GPT-4: a new era of artificial intelligence in medicine. *Ir J Med Sci* 2023;192(6):3197–3200.
7. Zhang X, Lu Y, Wang W, et al. GPT-4V(ision) as a Generalist Evaluator for Vision-Language Tasks. *arXiv 2311.01361* [preprint] <https://arxiv.org/abs/2311.01361>. Published November 2, 2023. Accessed November 15, 2023.