GPT-4V Cannot Generate Radiology Reports Yet

Yuyang Jiang*
University of Chicago
yuyang2001@uchicago.edu

Chacha Chen*
University of Chicago
chacha@uchicago.edu

Dang Nguyen
University of Chicago
dangnguyen@uchicago.edu

Benjamin M. Mervak
University of Michigan
bmervak@med.umich.edu

Chenhao Tan
University of Chicago
chenhao@uchicago.edu

Abstract

GPT-4V's purported strong multimodal abilities raise interests in using it to automate radiology report writing, but there lacks thorough evaluations. In this work, we perform a systematic evaluation of GPT-4V in generating radiology reports on two chest X-ray report datasets: MIMIC-CXR and IU X-RAY. We attempt to directly generate reports using GPT-4V through different prompting strategies and find that it fails terribly in both lexical metrics and clinical efficacy metrics. To understand the low performance, we decompose the task into two steps: 1) the **medical image reasoning** step of predicting medical condition labels from images; and 2) the **report synthesis** step of generating reports from (groundtruth) conditions. We show that GPT-4V's performance in image reasoning is consistently low across different prompts. In fact, the distributions of model-predicted labels remain constant regardless of which groundtruth conditions are present on the image, suggesting that the model is not interpreting chest X-rays meaningfully. Even when given groundtruth conditions in report synthesis, its generated reports are less correct and less natural-sounding than a finetuned LLaMA-2. Altogether, our findings cast doubt on the viability of using GPT-4V in a radiology workflow.

1 Introduction

Large language models (LLMs) are becoming multimodal, and GPT-4V represents the state-of-the-art [1]. Similar to the claimed general-purpose capabilities in LLMs [5, 22], large multimodal models (LMMs) are supposed to possess advanced skills across a wide range of domains, including high-stakes scenarios such as medicine [32]. However, in the field of radiology report generation, where relatively rich datasets are available, there has been inconclusive evidence regarding the performance of LMMs. Some studies [20, 32] claimed that GPT-4V performs well to some extent based on case studies and qualitative analysis. In contrast, [4] found that the model is not yet a reliable tool for radiological image interpretation on a small private dataset. [30] observed that GPT-4V can generate structured reports with incorrect content, as evidenced by case studies and qualitative analysis. To make sense of these results, we aim to perform a systematic and in-depth evaluation of GPT-4V beyond simply providing performance numbers.[2]

To do that, we perform three experiments as shown in Fig. 1 on two popular radiology report generation benchmarks, MIMIC-CXR and IU X-RAY. Our evaluation starts with Experiment 1:

*Equal contribution.

[2]We access GPT-4V (vision-preview 11/15/2023) through Azure OpenAI service.

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

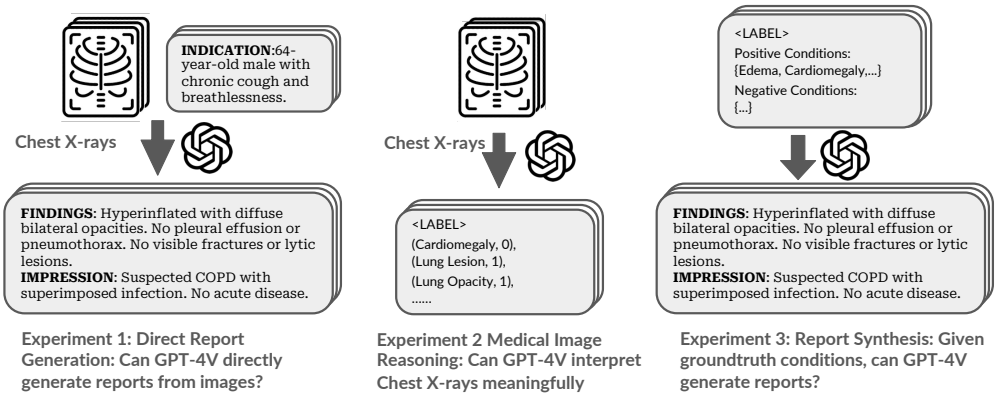

Figure 1: An overview of our evaluation. In Experiment 1, we evaluate the out-of-box capability of GPT-4V on radiology report generation. We further decompose the task into medical image reasoning (Experiment 2) and report synthesis (Experiment 3).

direct report generation. Different from previous works [20, 32], we conduct a thorough evaluation of GPT-4V's capability to directly generate reports from chest X-rays, utilizing different prompting strategies and assessing both lexical metrics, which measure how textually similar a generated report is to a reference report, and clinical efficacy metrics, which measure how clinically accurate it is. We experiment with various prompting strategies, including zero-shot, contextual enhancement, chain-of-thought (CoT) [29], and few-shot in-context learning. Despite our various attempts, the performance of GPT-4V is consistently low in both metrics.

To further investigate the reason for GPT-4V's poor performance, we break down report generation into two steps, **medical image reasoning** and **report synthesis given medical conditions**. For Experiment 2 (medical image reasoning), we first test whether GPT-4V can identify medical conditions from X-rays. Our findings indicate that GPT-4's performance in identifying medical conditions from images is unsatisfactory across different prompts. Based on limited capability results, we further compare the difference between distributions of predicted medical condition labels conditioned on different groundtruth image labels. We find that GPT-4V cannot interpret medical images meaningfully as the distribution of predicted labels does not vary depend on the groundtruth label.

Finally, in Experiment 3 (report synthesis), we explore whether bypassing the image reasoning bottleneck by providing groundtruth conditions enables GPT-4V to generate clinically usable reports. As expected, reports generated by GPT-4V achieve higher clinical efficacy; however, the limited improvement in lexical metrics suggests that GPT-4V-generated reports remain dissimilar to human-written reports in style. Most importantly, GPT-4V underperforms a finetuned LLaMA-2 in both lexical metrics and clinical efficacy metrics, calling into question its utility. We further validate our findings by conducting an additional human reader study with a board-certified radiologist to assess the clinical viability of GPT-4V-generated reports.

In summary, our key contributions and conclusions are as follows:

- We perform the first systematic and in-depth evaluation to benchmark GPT-4V in radiology report generation. Our main conclusion is that GPT-4V cannot generate radiology reports yet.
- By decomposing the task into medical image reasoning and report synthesis, we demonstrate that GPT-4V cannot interpret chest X-ray images meaningfully in the image reasoning step, and further validate this finding through rigorous hypothesis testing.
- During report synthesis, we address the image reasoning bottleneck by providing groundtruth conditions. Nonetheless, both experimental results and human evaluations consistently show that GPT-4V performs worse than a finetuned LLaMA-2 baseline.

We include our code in the supplementary material.

2 Related Work

While there is an emerging line of work in investigating the direct application of GPT-4 in radiology report generation, there lacks a systematic evaluation. [20, 30, 32] tested capabilities for general medical applications through case studies, including selected examples of chest X-ray reports with qualitative analysis. [4] provided quantitative results on GPT-4V's accuracy in interpreting medical images, using a small private dataset that includes chest X-rays. But their evaluation only focused on identifying the imaging modality (e.g., CT, ultrasound, or MRI) and the anatomical region of the pathology, rather than assessing the overall quality of generated radiology reports. [15] evaluated GPT-4V on the public MIMIC-CXR dataset, but only used lexical and semantic metrics without assessing clinical efficacy. [6] included GPT-4V as one of the baselines. However, their focus is on proposing a new model. In contrast, we provide an in-depth evaluation across various metrics with different prompting strategies on two public datasets.

Prior work has also examined text-only applications of GPT-4 related to radiology report generation, such as summarizing findings [18, 26], handling various text processing tasks including sentence semantics, structural extraction, and summary of findings [19], radiology board-style examination [3], detecting errors in radiology reports [9], and refining human-written reports for better standardization and clarity [10]. Additionally, other related multimodal tasks include visual question answering based on radiology images [31] and biomedical image classification [20].

To the best of our knowledge, our work provides the first systematic and in-depth evaluation of GPT-4V's capabilities to generate radiology reports.

3 Experiment Setup

In this section, we provide an overview of our methods, datasets, and evaluation metrics.

Method. In Experiment 1 (Section 4.1), we evaluate GPT-4V's ability to directly generate radiology report given chest X-ray images. We consider five variations of prompts as outlined in Table 1. Prompt 1.1 (Basic generation) is a prompt to test the out-of-the-box capability of GPT-4V. We implement three additional prompting strategies leveraging insights in prompt engineering: (1) inspired by [21], we add relevant contextual information (i.e., the INDICATION) to derive Prompt 1.2 as "Indication enhancement", and add instructions on medical condition labels to Prompt 1.3 as "+instruction" enhancement; (2) we use a chain-of-thought (CoT) strategy in Prompt 1.4, eliciting the model with two steps: medical condition label prediction based on images followed by report synthesis based on the predicted labels; (3) We adopt few-shot in-context learning by adding a few example image-report pairs in Prompt 1.5. We compare these results with the state-of-the-art (SOTA) models.

In addition to evaluation of the end-to-end radiology report generation capability, we further evaluate on the decomposed tasks: Experiment 2 (Section 4.2): chest X-ray image reasoning; and Experiment 3 (Section 4.2): synthesizing a radiology report from given conditions. This decomposition allows us to look into the bottlenecks in the current generation performance. In Experiment 2, we prompt the model to directly output medical condition labels from images (Prompt 2.1). In Experiment 3, we bypass image reasoning to test GPT-4V's textual synthesis ability and provide groundtruth conditions to evaluate the model's report composition capability independently (Prompt 3.1). To contextualize the performance of GPT-4V, we also report the performance of a finetuned LLaMA-2 7B on groundtruth labels and groundtruth impressions following Alpaca [27].

Dataset and pre-processing. We use two chest X-ray datasets: MIMIC-CXR and IU X-RAY. The MIMIC-CXR dataset [14] contains chest X-ray images and their corresponding free-text radiology reports. The dataset includes 377,110 images from 227,835 studies. Each study has one radiology report and one or more chest X-rays. The IU X-RAYdataset [8] (also known as "Open-i") includes 3996 de-identified radiology reports and 8121 associated images from the Indiana University hospital network. For our evaluation, we randomly sample 300 studies from the MIMIC-CXR and IU

Table 1: An index to prompts used in all of our experiments.

Experiment 1: Direct Report Generation		
Prompt 1.1	Basic generation	Direct report generation based on chest X-ray images
Prompt 1.2	+Indication	Contextual enhancement by providing the indication section
Prompt 1.3	+Instruction	Contextual enhancement by providing instructions on medical conditions
Prompt 1.4	Chain-of-Thought (CoT)	Step 1 - medical condition labeling; Step 2 - report synthesis
Prompt 1.5	Few-shot	Few-shot: in-context learning given a few examples
Experiment 2: Medical Image Reasoning Capability		
Prompt 2.1	Image reasoning	Medical condition labeling directly from chest X-ray images
Experiment 3: Report Synthesis Given Medical Conditions		
Prompt 3.1	Report synthesis	Report generation using provided positive and negative conditions

X-RAY datasets after removing studies with empty impression or indication sections. More details about data processing can be checked in Appendix C.

Evaluation metrics. We evaluate the generated reports from two aspects:

- Lexical metrics. Lexical metrics focus on the surface form and the exact word matches between the generated and reference texts. We adopt common lexical metrics: BLEU [23] (1-gram and 4-gram), ROUGE-L [16], and METEOR [2].
- Clinical efficacy metrics. We first evaluate on **clinical correctness** based on labeler results on generated reports. Following existing works [11, 28, 21], we use the CheXbert automatic labeler [25] to extract labels for each of 14 Chexpert medical conditions [12]. We compute both positive F1 and negative F1 scores, where each condition has four labels: `present, absent, uncertain, unmentioned`. Positive F1 considers only positive labels against all others, while negative F1 considers negative labels as 1 and all other labels as 0. We report the macro-averaged F1 scores on all 14 conditions and on top 5 conditions (which only reports on the five most common conditions[3]). We also report RadGraph F1 [13], which captures the overlap in clinical entities and relations between a generated report and a reference report.

  Additionally, from a **pragmatic** viewpoint, commenting on negative observations is essential in radiology reports. Following [21], we compute Negative F1 and Negative F1-5, to evaluate whether the model can accurately identify negative conditions and include that in the generated reports. All reported F1 are macro-averaged. We also use the **hallucination** metric to quantify the proportion of uninferable information. Following [21], we define uninferable information to include previous studies, previous treatment details, recommendations, doctor communications, and image view descriptions.

4 Results

4.1 Experiment 1: Can GPT-4V directly generate reports from images?

We first evaluate the out-of-the-box capability of GPT-4V in generating radiology reports from chest X-ray images using basic generation (Prompt 1.1). Table 2 shows the results compared with existing state-of-the-art (SOTA) models. Overall, GPT-4V significantly underperforms the state-of-the-art models on both lexical and clinical efficacy metrics, with the exception of the METEOR score on the IU X-RAY dataset. The relatively better METEOR performance is due to its comprehensive evaluation criteria, which include synonymy and paraphrasing, not just exact word matches like BLEU and ROUGE. This allows METEOR to recognize semantic equivalents, even if the word choice differs. In other words, the generated report somewhat resembles a radiology report, although it fails at the exact word-level matching. For clinical efficacy metrics, the gaps to SOTA are consistently large. This suggests that GPT-4V struggles to accurately identify conditions in its generated reports from images alone.

[3]Top five conditions in the MIMIC-CXR are Pneumothorax, Pneumonia, Edema, Pleural Effusion, and Consolidation.

Table 2: Direct report generation performance comparison. GPT-4V shows a significant performance gap compared to SOTA, and the results are consistent across the five prompting strategies. Examples of generated reports across different prompts are shown in Appendix D.2.

Experiment	Lexical metrics				Clinic Efficacy Metrics					
	BLEU-1	BLEU-4	ROUGE	METEOR	Pos F1	Pos F1@5	Rad. F1	Neg F1*	Neg F1@5*	Hall.*↓
MIMIC-CXR										
Basic	0.299	0.035	0.214	0.279	0.117	0.124	0.135	0.004	0.001	0.687
+Indication	**0.323**	0.042	**0.227**	**0.294**	**0.181**	0.194	**0.159**	**0.037**	**0.096**	0.610
+Instruction	0.265	0.019	0.186	0.262	0.134	**0.236**	0.109	0.026	0.067	0.593
CoT	0.236	0.008	0.176	0.202	0.151	0.233	0.080	0.023	0.061	0.607
Few-shot	0.294	**0.053**	0.223	0.293	0.085	0.036	0.149	0.000	0.000	**0.578**
SOTA [ref.]	0.402 [17]	0.142 [11]	0.291 [17]	0.333 [11]	0.473 [17]	0.516 [28]	0.267 [28]	0.077 [21]	0.156 [21]	0.158 [21]
Δ(GPT-4V-SOTA)	-19.65%	-62.68%	-21.99%	-11.71%	-61.73%	-54.26%	-40.45%	-51.95%	-38.46%	42.00%
IU X-RAY										
Basic	0.278	0.038	0.218	0.326	0.030	0.024	0.178	0.000	0.000	0.494
+Indication	0.282	**0.042**	0.216	**0.328**	0.023	0.010	0.174	0.020	0.052	0.614
+Instruction	0.237	0.027	0.189	0.281	0.053	0.052	0.140	**0.041**	**0.106**	0.523
CoT	0.233	0.016	0.179	0.235	**0.072**	**0.119**	0.105	0.000	0.000	0.619
Few-Shot	**0.325**	0.037	**0.247**	0.318	0.061	0.080	**0.191**	0.026	0.067	**0.263**
SOTA [ref.]	0.499 [17]	0.184 [17]	0.390 [17]	0.208 [17]	-	-	-	-	-	-
Δ(GPT-4V-SOTA)	-53.54%	-77.17%	-36.67%	57.69%	-	-	-	-	-	-

To compare with SOTA numbers, all metrics, except for those marked with * (Neg F1, Neg F1@5, and Hall), are evaluated on the findings section. * columns are based on the impression section. A comprehensive table, including results for both the findings and impression sections, is provided in the Appendix D.1.
All numbers are only extracted from examples where GPT-4V successfully generated a report. Occasionally, GPT-4V responds that it "cannot provide a diagnostic report or interpretation for medical images". More details are available in Appendix D.1.

Table 3: Image reasoning performance of GPT-4V on chest X-ray images. The model performs poorly in identifying conditions from chest X-ray images across different prompting strategies. The results show positive F1 scores for correctly predicting the presence of medical conditions.

Metric	MIMIC-CXR		IU X-RAY	
	Chain-of-Thought (1st Step)	Image Reasoning	Chain-of-Thought (1st Step)	Image Reasoning
Positive F1	0.166	0.146	0.072	0.049
Positive F1@5	0.261	0.208	0.095	0.056

Our results are consistent across prompting strategies. Our prompting strategies include adding contextual information, chain-of-thought reasoning, and few-shot prompting. While indication enhancement (Prompt 1.2) provides indication section as input in addition to chest X-rays and improves many metrics for both datasets, it remains within the same range and does not significantly reduce the gap compared to SOTA. Instruction enhancement (Prompt 1.3) provides medical condition descriptions and improves the Positive F1-5 by 11.2% in MIMIC-CXR, the most effective so far, but there is still a significant gap to SOTA (54.26%). Chain-of-Thought (Prompt 1.4) performs similarly to instruction enhancement, as both follow the same labeling instructions. Few-Shot (Prompt 1.5) provides image-report pairs as context and generally improves only lexical metrics, RadGraph F1, and Hallucination, while clinical correctness remains consistently low across both datasets. This indicates that while few-shot prompting might help GPT-4V mimic the format of groundtruth reports, it still falls short in generating accurate reports.

4.2 Experiment 2: Can GPT-4V interpret chest X-rays meaningfully?

In this section, we probe GPT-4V's ability to reason about chest X-ray images alone. Specifically, we evaluate whether the model can meaningfully interpret chest X-ray images by measuring how accurately GPT-4V can label medical conditions present (positive F1). Table 3 provides an overview of GPT-4V's labeling performance under different prompting strategies.

We can see that GPT-4V cannot accurately specify positive conditions from given chest X-rays. This can be highlighted by consistently poor Positive F1 scores observed for both datasets under various

Table 4: χ^2-test for homogeneity of label distribution across different condition groups. When p-value is smaller than 0.0001, at 0.01% significance level, we can reject the null hypothesis that different groups follow the same label distribution.

Statistics	Overall		Top 6 Conditions	
	Groundtruth	GPT-4V	Groundtruth	GPT-4V
χ^2 statistic	1770.38	74.25	317.86	6.11
p-value	p < 0.0001	1.0000	p < 0.0001	1.0000
df.	144	144	25	25

prompting strategies. Furthermore, this inability to accurately interpret images may directly contribute to GPT-4V's failure in generating high-quality reports, as confirmed by similar Positive F1 score of 0.151 (MIMIC-CXR) and 0.072 (IU X-RAY) from the report synthesis phase of Chain-of-Thought (see Table 2), compared to 0.166 (MIMIC-CXR) and 0.072 (IU X-RAY) from the initial label generation phase of Chain-of-Thought.

Overall, these results indicate GPT-4V's limited ability in identifying medical conditions from chest X-ray images, regardless of whether labels are derived from CoT 1st step or direct prompting.

Testing whether GPT-4V generates labels based on given chest X-rays. Considering the failure of GPT-4V to accurately label medical conditions, we would like to investigate to what extent can GPT-4V predict meaningful labels given a specific chest X-ray image. To test this, we group chest X-rays by their groundtruth conditions and then analyze the generated label distribution for each group. If the label distributions are similar across different condition groups, it would suggest that GPT-4V is not meaningfully identifying labels from the chest X-rays but rather assigning labels randomly without proper image interpretation. For example, if the model's generated label probabilities are roughly the same regardless of whether the groundtruth condition of the given image is Edema or Cardiomegaly, it indicates a limited capability in medical image reasoning.

Formally, let X_{ij} be a binary random variable that takes the value 1 if GPT-4V labels the j-th condition as positive for the chest X-ray image associated with the i-th study, and 0 otherwise, where $i = 1, 2, \ldots, 300$ and $j = 1, 2, \ldots, 13$. We exclude the "No Findings" condition from this study. We define $Y_j = \sum_{i=1}^{300} X_{ij}$ as the sum of positive mentions for the j-th condition across all 300 studies, and $\mathbf{Y} = [Y_1, \ldots, Y_{13}]$ as the count vector. Next, we categorize the study pool into 13 condition groups, where group k consists all studies that are ground truth positive for the k-th condition based on the associated radiology report. Note that there might be overlaps between these groups, as a single study can be positive for multiple conditions. For each group k, GPT-4V's labeling process given the chest X-ray image from i-th study can be modeled as:

$$\begin{cases} X_{ij}^{(k)} \sim \text{Bernoulli}(P_j^{(k)}) \text{ for } i \in \text{group } k \text{ and } j = 1, \ldots, 13 \\ \mathbf{Y}_k \sim \text{Multinomial}(n_k; \mathbf{P}_k) \text{ with } \mathbf{P}_k = [P_1^{(k)}, \ldots, P_{13}^{(k)}] \end{cases} \tag{1}$$

where n_k is the number of studies in group k, and $P_j^{(k)}$ is the probability that GPT-4V labels the j-th condition as positive for the chest X-ray images associated with the studies in group k.

We first use a **χ^2-test** to test if GPT-4V follows the same label distribution across different groups, i.e., testing the null hypothesis (H_0) that $\mathbf{P}_k = \mathbf{P}_{k'}$ for any groups k and k'. Additionally, we use **bootstrap confidence interval** [7] to test if GPT-4V labels one certain condition independently of the groundtruth condition group. Specifically, we test the null hypothesis (H_0) that $P_j^{(k)} = P_j$ for any condition j and group k. More test details can be found in Appendix D.3.

Table 4 presents χ^2-test results for the homogeneity of label distribution across different groups. For both the overall and top 6 conditions[4], at 0.01% significance level, we can both reject the null

[4]Due to the sparsity of the original study pool, we report results for two different tables: (1) A modified table with zero elements replaced by 0.001; (2) A reduced table with only the six most frequent conditions in the subsample.

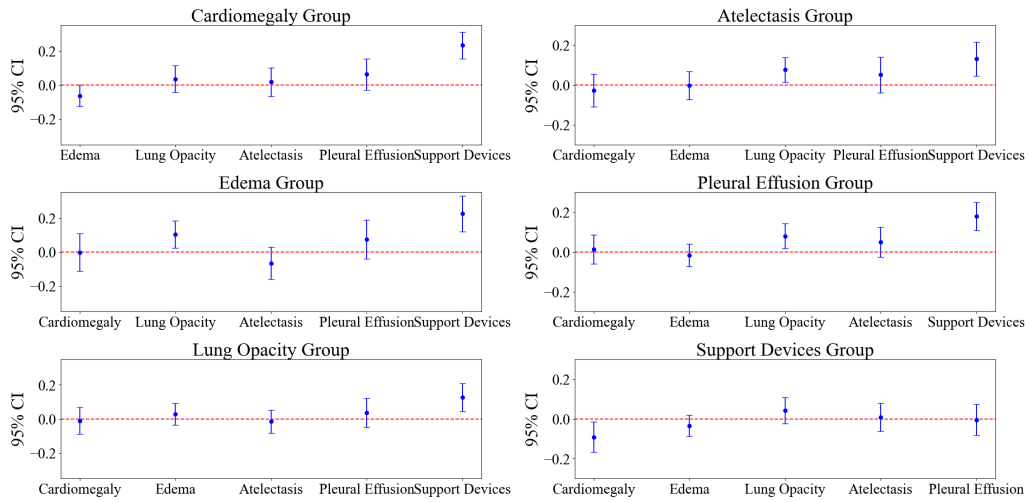

Figure 2: 95% Bootstrap confidence interval of top 6 conditions in our sample. When zero falls into the interval, at 95% confidence level, we cannot reject the null hypothesis that GPT-4V labels j-th condition independent of which condition group this study belongs to.

Table 5: Performance in report generation with groundtruth conditions. Although GPT-4V's performance improves significantly, it still underperforms finetuned LLaMA-2, especially in matching the writing style of groundtruth reports.

Experiment	Lexical metrics				Clinic Efficacy Metrics					
	BLEU-1	BLEU-4	ROUGE	METEOR	Pos F1	Pos F1@5	Rad. F1	Neg F1	Neg F1@5	Hall.↓
MIMIC-CXR										
GPT-4V	0.135	0.018	0.119	0.161	0.118	0.160	0.071	0.004	0.001	0.687
GPT-4V (gt)	0.176	0.007	0.185	0.179	0.885	**0.977**	0.103	0.584	**0.958**	**0.431**
LLaMA-2 (gt)	**0.301**	**0.094**	**0.330**	**0.348**	**0.923**	0.957	**0.286**	**0.703**	0.941	0.710
IU X-RAY										
GPT-4V	0.219	0.019	0.232	0.295	0.036	0.041	0.155	0.000	0.000	0.275
GPT-4V (gt)	0.216	0.003	0.229	0.207	0.852	0.919	0.089	**0.630**	0.868	0.235
LLaMA-2 (gt)	**0.454**	**0.124**	**0.460**	**0.441**	**0.871**	**0.928**	**0.297**	0.627	**0.963**	**0.110**

All metrics are evaluated on the impression section.

hypothesis for groundtruth reports that different groups follow the same label distribution, but not for GPT-4V's generated reports.

Figure 2 illustrates the 95% bootstrap confidence intervals for top 6 conditions[5]. If zero falls within the interval, we cannot reject the null hypothesis that GPT-4V labels the j-th condition independently of the condition group at 95% confidence level. The figure shows that, in 21 out of 30 cases (70%), we cannot reject the null hypothesis. The condition that consistently depends on the group is "support devices", which is not actually the groundtruth medical condition.

In summary, the results show that GPT-4V labels conditions independently of the groundtruth condition, and there is no significant difference in label distributions across groups in GPT-4V's generated reports, unlike the groundtruth reports.

4.3 Experiment 3: Given groundtruth conditions, can GPT-4V generate reports?

Given that GPT-4V cannot perform image reasoning, we next investigate whether GPT-4V can produce high-quality radiology reports when provided with accurate medical conditions. We conduct an experiment on report synthesis (Prompt 3.1) on GPT-4V and use a finetuned LLaMA-2 model as a baseline for comparison.

[5]Due to the sparsity of the original study pool, we limit our analysis to the six most frequent conditions in our subsample.

Table 6: Human evaluation of radiology report quality. From the perspective of radiologists, GPT-4V underperforms relative to the finetuned LLaMA-2, particularly in fine-grained likert scale metrics of diagnostic accuracy, completeness, and clartiy/readability.

	Binary	**Likert Scale (1-5)**		
	Clinically Usable	**Diagnostic Accuracy**	**Completeness**	**Clarity/Readability**
Groundtruth	50/50 (100%)	**4.72**	**4.84**	**4.84**
LLaMA-2	42/50 (84%)	4.12	4.62	4.88
GPT-4V	43/50 (86%)	4.06	4.04	3.68

All metrics are evaluated on the impression section.

Table 7: Comparison of example reports from human annotation.

Groundtruth labels	**Model**	**Report**	**Usable**	**Acc.**	**Comp.**	**Clarity**
Positive: Enlarged Cardiomediastinum, Lung Opacity, Atelectasis Negative: Pneumonia	Groundtruth	In comparison with study of ___, there is little change in the cardiomediastinal silhouette and pacer leads. Continued elevation of the right hemidiaphragmatic contour. Opacification medially above the elevated hemidiaphragm most likely represents atelectatic changes. No definite acute focal pneumonia.	✓	5	5	4
	GPT-4V	The chest radiograph demonstrates an enlarged cardiomediastinum, lung opacity which may be due to various etiologies, and atelectasis. There is no clear evidence of pneumonia on the radiograph.	✓	3	4	3
	LLaMA-2	1. No evidence of pneumonia. 2. Stable postoperative appearance of the chest. 3. Increased retrocardiac opacity likely reflects atelectasis.	✓	5	5	5

Table 5 shows that while using groundtruth conditions significantly enhances GPT-4V's clinical accuracy, it still does not perform as well as the finetuned LLaMA-2, particularly in matching the content of groundtruth reports. Progress in clinical accuracy is evidenced by large improvements in F1 scores for both datasets compared to basic generation (Prompt 1.1). However, there are only minor changes in lexical metrics and RadGraph F1, which focus on entity relation matching in groundtruth reports, along with consistently large gaps with finetuned LLaMA-2, suggest that groundtruth conditions are insufficient to align GPT-4V's writing closely with that of groundtruth reports. The higher scores of the finetuned LLaMA-2 in lexical metrics also indicate that finetuning open models is an effective way to leverage existing datasets.

Human Evaluation To further evaluate the quality of GPT-4V-generated reports beyond automatic metrics, we collaborate with a board-certified radiologist to conduct a human evaluation. From our testing set of 300 studies, we randomly select 50 cases for blind human evaluation. The radiologist is provided with anonymized chest X-ray images and randomly ordered IMPRESSION sections from groundtruth reports, as well as reports generated by LLaMA-2 and GPT-4V. Both LLaMA-2 and GPT-4V are prompted with groundtruth medical conditions. The evaluation involves a detailed review of three reports per study case, assessing each report's clinical usability with a binary label as the first step. Then, the radiologist rates each report on two dimensions: clinical efficacy (diagnostic accuracy and completeness) and lexical performance (clarity/readability). Reports are rated on a Likert scale, where a score of 5 denotes superior performance and a score of 1 denotes poor performance. We compute and report the average scores for each metric across different report types.

Table 6 shows that, from the perspective of radiologists, GPT-4V still underperforms the finetuned LLaMA-2. Groundtruth reports are indeed of high quality, rated as clinically usable in 50 out of 50 cases, compared to 42 out of 50 for LLaMA-2 and 43 out of 50 for GPT-4V. While the difference

in clinical usability between LLaMA-2 and GPT-4V is not large, LLaMA-2 outperforms GPT-4V across all other Likert scale metrics, especially in completeness and clarity/readability.

Table 7 presents an example study with three different reports. While groundtruth reports offer detailed clinical insights and varied descriptors, GPT-4V tends to provide vague statements, only stating "lung opacity which may be due to various etiologies" without specifying its location, severity, or offering a differential diagnosis. LLaMA-2 performs slightly better by offering some specific diagnoses, yet still lacks detailed descriptions.

In short, human annotation corroborates with the findings from our Experiment 3. Given groundtruth conditions, GPT-4V-generated reports still do not meet the standards of human-written reports. They lack comprehensive coverage of all relevant clinical findings and do not effectively summarize and organize the patient's condition in a readable manner.

5 Limitations

In this paper, we use GPT-4V, one of the most capable LMMs across various domains, to conduct a systematic evaluation of its capabilities in generating radiology reports. Comparisons with other general-domain LMMs, including Google's Gemini and OpenAI's newer GPT-4o, are reserved for future research. Note that at the time of our submission, GPT-4o API was not available via Microsoft Azure platform.

Additionally, we employ four common prompting strategies in our study and encourage future research to explore additional techniques, such as Self-Critique [24], to verify the robustness of our findings. Due to resource constraints, we randomly select a 300-sample subset for overall evaluation and choose 50 samples for a human study. Besides, the human study is limited to a single radiologist's subjective assessment, potentially influenced by their personal style and preferences. While our human evaluation could be improved by recruiting more radiologists, we believe that GPT-4V's poor performance may not justify a significantly larger human evaluation. That said, our results suggest that finetuned open models may hold the potential of fitting into the current radiologist workflow if we can leverage medical image reasoning abilities of other models.

Despite these limitations, we believe the findings from this paper are well-supported by our comprehensive and detailed evaluation framework. Results from our work raise serious concerns about how to safely integrate general-domain LMMs into real-world radiology workflows. It is worth noting that OpenAI itself restricts the medical use of GPT-4V. In our experiments, especially with the few-shot prompt, GPT-4V tends to return "I'm sorry, but I cannot provide a diagnostic report or interpretation for medical images. If you have any medical concerns, please consult a qualified healthcare professional who can provide a proper examination and diagnosis."

6 Conclusions

We perform a systematic evaluation of GPT-4V in radiology report generation on two chest X-ray benchamarks. We find that GPT-4V cannot generate radiology reports, even across different prompting strategies. To understand the low performance, we decompose the main task into image reasoning and report synthesis. The results demonstrate that GPT-4V struggles significantly with interpreting chest X-rays meaningfully, which directly impacts its ability to generate reports. Furthermore, even when we bypass this problem by providing groundtruth conditions, GPT-4V still underperforms a finetuned LLaMA-2 baseline and consistently fails to replicate the writing style of groundtruth reports or meet the preferences of radiologists. Overall, our study highlights substantial concerns regarding the feasibility of integrating GPT-4V into real radiology workflows.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[3] Rajesh Bhayana, Robert R Bleakney, and Satheesh Krishna. Gpt-4 in radiology: improvements in advanced reasoning. *Radiology*, 307(5):e230987, 2023.

[4] Dana Brin, Vera Sorin, Yiftach Barash, Eli Konen, Benjamin S. Glicksberg, Girish Nadkarni, and Eyal Klang. Assessing gpt-4 multimodal performance in radiological image analysis. *medRxiv*, nov 2023.

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[6] Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation, 2024.

[7] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Number 1. Cambridge university press, 1997.

[8] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[9] Roman Johannes Gertz, Thomas Dratsch, Alexander Christian Bunck, Simon Lennartz, Andra-Iza Iuga, Martin Gunnar Hellmich, Thorsten Persigehl, Lenhard Pennig, Carsten Herbert Gietzen, Philipp Fervers, et al. Potential of gpt-4 for detecting errors in radiology reports: Implications for reporting accuracy. *Radiology*, 311(1):e232714, 2024.

[10] Amir M Hasani, Shiva Singh, Aryan Zahergivar, Beth Ryan, Daniel Nethala, Gabriela Bravomontenegro, Neil Mendhiratta, Mark Ball, Faraz Farhadi, and Ashkan Malayeri. Evaluating the performance of generative pre-trained transformer-4 (gpt-4) in standardizing radiology reports. *European Radiology*, pages 1–9, 2023.

[11] Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023.

[12] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[13] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

[14] A Johnson, T Pollard, R Mark, S Berkowitz, and Steven Horng. Mimic-cxr database (version 2.0. 0). *physionet*, 2:5, 2019.

[15] Yingshu Li, Yunyi Liu, Zhanyu Wang, Xinyu Liang, Lingqiao Liu, Lei Wang, Leyang Cui, Zhaopeng Tu, Longyue Wang, and Luping Zhou. A comprehensive study of gpt-4v's multimodal capabilities in medical imaging. *medRxiv*, pages 2023–11, 2023.

[16] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[17] Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18635–18643, 2024.

[18] Fenglin Liu, Hongjian Zhou, Yining Hua, Omid Rohanian, Lei Clifton, and David Clifton. Large language models in healthcare: A comprehensive benchmark. *medRxiv*, pages 2024–04, 2024.

[19] Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Maria Teodora Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, et al. Exploring the boundaries of gpt-4 in radiology. *arXiv preprint arXiv:2310.14573*, 2023.

[20] Zhengliang Liu, Hanqi Jiang, Tianyang Zhong, Zihao Wu, Chong Ma, Yiwei Li, Xiaowei Yu, Yutong Zhang, Yi Pan, Peng Shu, et al. Holistic evaluation of gpt-4v for biomedical imaging. *arXiv preprint arXiv:2312.05256*, 2023.

[21] Dang Nguyen, Chacha Chen, He He, and Chenhao Tan. Pragmatic radiology report generation. In *Machine Learning for Health (ML4H)*, pages 385–402. PMLR, 2023.

[22] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[24] Noah Shinn, Beck Labash, and Ashwin Gopinath. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*, 2023.

[25] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

[26] Zhaoyi Sun, Hanley Ong, Patrick Kennedy, Liyan Tang, Shirley Chen, Jonathan Elias, Eugene Lucas, George Shih, and Yifan Peng. Evaluating gpt-4 on impressions generation in radiology reports. *Radiology*, 307(5):e231259, 2023.

[27] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

[28] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024.

[29] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[30] Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, et al. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909*, 2023.

[31] Zhiling Yan, Kai Zhang, Rong Zhou, Lifang He, Xiang Li, and Lichao Sun. Multimodal chatgpt for medical applications: an experimental study of gpt-4v. *arXiv preprint arXiv:2310.19061*, 2023.

[32] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Experiment setup can be checked in Section 3 and codes are included in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Error bars are reported for bootstrap confidence interval. We follow commone practices and report performance numbers in the tables.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix B.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 3.
   (b) Did you mention the license of the assets? [Yes] See Appendix C.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appendix C.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] Both MIMIC and IU X-RAY are already de-identifiable datasets and do not contain offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Appendix E.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No] Our human evaluation includes only a human reader study and the involved radiologist is an author.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No]

A Prompts

Prompt 1.1 Basic generation: direct report generation based on chest X-ray images.

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Write a report that contains only the FINDINGS and IMPRESSION sections based on the attached images. Provide only your generated report, without any additional explanation and special format. Your answer is for reference only and is not used for actual diagnosis.

Prompt 1.2 Indication enhancement: providing the indication section.

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Below is INDICATION related to chest X-ray images. INDICATION: {} Write a report that contains only the FINDINGS and IMPRESSION sections based on the attached images and INDICATION. Provide only your generated report, without any additional explanation and special format. Your answer is for reference only and is not used for actual diagnosis.

Prompt 1.3 Instruction enhancement: providing information on medical condition labels.

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Below is an observation plan consisting of 14 conditions: "No Finding", "Enlarged Cardiomediastinum", "Cardiomegaly", "Lung Lesion", "Lung Opacity", "Edema", "Consolidation", "Pneumonia", "Atelectasis", "Pneumothorax", "Pleural Effusion", "Pleural Other", "Fracture", "Support Devices". Based on attached images, assign labels for each condition except "No Finding": "1", "0", "-1", "2". It is noted that "No Finding" is either "2" or "1". These labels have the following interpretation: 1 - The observation was clearly present on the chest X-ray image. 0 - The observation was absent on the chest X-ray image and was mentioned as negative. -1 - The observation was unclear if it exists. 2 - The observation was absent but not explicitly mentioned. Based on labels you choose for each condition, write a report that contains only the FINDINGS and IMPRESSION sections. Don't return any of your assigned labels. Provide only your generated report, without any additional explanation and special format. Your answer is for reference only and is not used for actual diagnosis.

Prompt 1.4 Chain-of-Thought: step 1 - medical condition labeling; step 2 - report synthesis.

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Below is an observation plan consisting of 14 conditions: “No Finding”, “Enlarged Cardiomediastinum”, “Cardiomegaly”, “Lung Lesion”, “Lung Opacity”, “Edema”, “Consolidation”, “Pneumonia”, “Atelectasis”, “Pneumothorax”, “Pleural Effusion”, “Pleural Other”, “Fracture”, “Support Devices”. Based on attached images, assign labels for each condition except “No Finding”: “1”, “0”, “-1”, “2”. It is noted that “No Finding” is either “2” or “1”. These labels have the following interpretation: 1 - The observation was clearly present on the chest X-ray image. 0 - The observation was absent on the chest X-ray image and was mentioned as negative. -1 - The observation was unclear if it exists. 2 - The observation was absent but not explicitly mentioned. The first step is to return one list of your assigned labels. For multiple images, assign the labels based on all images and return only one list of labels for the given 14 conditions. The second step is to write a report that contains only the FINDINGS and IMPRESSION sections based on labels you choose for each condition. Your answer is for reference only and is not used for actual diagnosis. Strictly follow the format below to provide your output. <LABEL> [(“No Finding”, “1”|“2”), (“Enlarged Cardiomediastinum”, “0”|“1”|“2”|“-1”), (“Cardiomegaly”, “0”|“1”|“2”|“-1”), (“Lung Lesion”, “0”|“1”|“2”|“-1”), (“Lung Opacity”, “0”|“1”|“2”|“-1”), (“Edema”, “0”|“1”|“2”|“-1”), (“Consolidation”, “0”|“1”|“2”|“-1”), (“Pneumonia”, “0”|“1”|“2”|“-1”), (“Atelectasis”, “0”|“1”|“2”|“-1”), (“Pneumothorax”, “0”|“1”|“2”|“-1”), (“Pleural Effusion”, “0”|“1”|“2”|“-1”), (“Pleural Other”, “0”|“1”|“2”|“-1”), (“Fracture”, “0”|“1”|“2”|“-1”), (“Support Devices”, “0”|“1”|“2”|“-1”)] </LABEL> <REPORT> FINDINGS: <findings> IMPRESSION: <impression> </REPORT>

Prompt 1.5 Few-shot: few-shot in-context learning given a few examples (MIMIC).

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Write a report that contains only the FINDINGS and IMPRESSION sections based on the attached images. Provide only your generated report, without any additional explanation and special format. Your answer is for reference only and is not used for actual diagnosis. [.JPEG] FINDINGS: Single portable view of the chest is compared to previous exam from ___. Enteric tube is seen with tip off the inferior field of view. Left PICC is seen; however, tip is not clearly delineated. Persistent bibasilar effusions and a right pigtail catheter projecting over the lower chest. There is possible right apical pneumothorax. Superiorly, the lungs are clear of consolidation. Cardiac silhouette is within normal limits. Osseous and soft tissue structures are unremarkable. IMPRESSION: No significant interval change with bilateral **pleural effusions** with right pigtail catheter in the lower chest. Possible small right apical pneumothorax. [.JPEG] FINDINGS: Frontal and lateral radiographs of the chest show hyperinflated lungs with flattened diaphragm, consistent with emphysema. Asymmetric opacity in the right middle lobe is concerning for pneumonia. No pleural effusion or pneumothorax is seen. The cardiomediastinal contours are within normal limits aside from a tortuous aorta. IMPRESSION: Right middle lobe opacity concerning for **pneumonia**. [.JPEG] FINDINGS: PA and lateral views of the chest provided. Midline sternotomy wires and mediastinal clips again noted. Suture is again noted in the right lower lung with adjacent rib resection. There is mild scarring in the right lower lung as on prior. There is no focal consolidation, large effusion or pneumothorax. No signs of congestion or edema. The heart remains moderately enlarged. The mediastinal contour is stable. IMPRESSION: Postsurgical changes in the right hemithorax. Mild **cardiomegaly** unchanged. No edema or pneumonia. [.JPEG] FINDINGS: PA and lateral views of the chest provided. Biapical pleural parenchymal scarring noted. No focal consolidation concerning for pneumonia. No effusion or pneumothorax. No signs of congestion or edema. Cardiomediastinal silhouette is stable with an unfolded thoracic aorta and top-normal heart size. Bony structures are intact. IMPRESSION: **No acute findings.** Top-normal heart size. [.JPEG]

Prompt 1.5 Few-shot: few-shot in-context learning given a few examples (IU X-RAY).

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Write a report that contains only the FINDINGS and IMPRESSION sections based on the attached images. Provide only your generated report, without any additional explanation and special format. Your answer is for reference only and is not used for actual diagnosis. [.PNG] FINDINGS: 2 images. Heart size upper limits of normal. Mediastinal contours are maintained. The patient is mildly rotated. There is a small to moderate sized right apical pneumothorax which measures approximately 2.0 cm. No focal airspace consolidation is seen. Left chest is clear. No definite displaced bony injury is seen. Results called XXXX. XXXX XXXX p.m. XXXX, XXXX. IMPRESSION: Small to moderate right apical **pneumothorax**. [.PNG] FINDINGS: The heart is normal in size and contour. There is focal airspace disease in the right middle lobe. There is no pneumothorax or effusion. IMPRESSION: Focal airspace disease in the right middle lobe. This is most concerning for **pneumonia**. Recommend follow up to ensure resolution. [.PNG] FINDINGS: Stable cardiomegaly with vascular prominence without overt edema. No focal airspace disease. No large pleural effusion or pneumothorax. The XXXX are intact. IMPRESSION: Stable **cardiomegaly** without overt pulmonary edema. [.PNG] FINDINGS: Heart is enlarged. There is prominence of the central pulmonary vasculature. Mild diffuse interstitial opacities bilaterally, predominantly in the bases, with no focal consolidation, pleural effusion, or pneumothoraces. XXXX and soft tissues are unremarkable. IMPRESSION: Cardiomegaly with pulmonary interstitial edema and XXXX bilateral **pleural effusions**. [.PNG] FINDINGS: The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no XXXX of a pleural effusion. There is no evidence of pneumothorax. IMPRESSION: **Normal chest x-XXXX.** [.PNG] FINDINGS: IMPRESSION: **Presumed closure device** at the level of the ligamentum arteriosum. Normal cardiac silhouette and clear lungs, with no evidence of left-to-right shunt. [.PNG]

Prompt 2.1 Image reasoning: medical condition labeling from chest X-ray images (2-class).

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Below is an observation plan consisting of 14 conditions: "No Finding", "Enlarged Cardiomediastinum", "Cardiomegaly", "Lung Lesion", "Lung Opacity", "Edema", "Consolidation", "Pneumonia", "Atelectasis", "Pneumothorax", "Pleural Effusion", "Pleural Other", "Fracture", "Support Devices". Based on attached images, assign labels for each condition: "1", "0". If the observation was clearly present on the chest X-ray image, assign "1" to the condition. Otherwise, assign "0" to the condition. For multiple images, assign the labels based on all images and return only one list of labels for the given 14 conditions. Your answer is for reference only and is not used for actual diagnosis. Strictly follow the format below to provide your output. <LABEL> [("No Finding", "0"|"1"), ("Enlarged Cardiomediastinum", "0"|"1"), ("Cardiomegaly", "0"|"1"), ("Lung Lesion", "0"|"1"), ("Lung Opacity", "0"|"1"), ("Edema", "0"|"1"), ("Consolidation", "0"|"1"), ("Pneumonia", "0"|"1"), ("Atelectasis", "0"|"1"), ("Pneumothorax", "0"|"1"), ("Pleural Effusion", "0"|"1"), ("Pleural Other", "0"|"1"), ("Fracture", "0"|"1"), ("Support Devices", "0"|"1")] </LABEL>

Prompt 2.2 Image reasoning: medical condition labeling from chest X-ray images (4-class).

User

Below is an observation plan consisting of 14 conditions: “No Finding”, “Enlarged Cardiomediastinum”, “Cardiomegaly”, “Lung Lesion”, “Lung Opacity”, “Edema”, “Consolidation”, “Pneumonia”, “Atelectasis”, “Pneumothorax”, “Pleural Effusion”, “Pleural Other”, “Fracture”, “Support Devices”.

Based on attached images, assign labels for each condition except “No Finding”: “1”, “0”, “-1”, “2”. It is noted that “No Finding” is either “2” or “1”. These labels have the following interpretation:
1 - The observation was clearly present on the chest X-ray image.
0 - The observation was absent on the chest X-ray image and was mentioned as negative.
-1 - The observation was unclear if it exists.
2 - The observation was absent but not explicitly mentioned.

For multiple images, assign the labels based on all images and return only one list of labels for the given 14 conditions. Your answer is for reference only and is not used for actual diagnosis. Strictly follow the format below to provide your output.

<LABEL>
[
(“No Finding”, “1”|“2”),
(“Enlarged Cardiomediastinum”, “0”|“1”|“2”|“-1”),
(“Cardiomegaly”, “0”|“1”|“2”|“-1”),
(“Lung Lesion”, “0”|“1”|“2”|“-1”),
(“Lung Opacity”, “0”|“1”|“2”|“-1”),
(“Edema”, “0”|“1”|“2”|“-1”),
(“Consolidation”, “0”|“1”|“2”|“-1”),
(“Pneumonia”, “0”|“1”|“2”|“-1”),
(“Atelectasis”, “0”|“1”|“2”|“-1”),
(“Pneumothorax”, “0”|“1”|“2”|“-1”),
(“Pleural Effusion”, “0”|“1”|“2”|“-1”),
(“Pleural Other”, “0”|“1”|“2”|“-1”),
(“Fracture”, “0”|“1”|“2”|“-1”),
(“Support Devices”, “0”|“1”|“2”|“-1”)
]
</LABEL>

Prompt 3.1 Report synthesis: report generation using provided positive and negative conditions.

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Below is a given observation plan: <LABEL> Positive Conditions: {} Negative Conditions: {} </LABEL> Write a report that contains only the FINDINGS and IMPRESSION sections based on given labels rather than images. For positive conditions, you should clearly mention it in the report. For negative conditions, you should clearly mention in the report that there is no clear evidence of this condition. You should not mention any other conditions not listed above. Your answer is for reference only and is not used for actual diagnosis. Strictly follow the format below to provide your output. <REPORT> FINDINGS: <findings> IMPRESSION: <impression> </REPORT>