Beyond Diagnosis: Evaluating Multimodal LLMs for Pathology Localization in Chest Radiographs

Advait Gosai[*†] ADVAIT_GOSAI@BERKELEY.EDU
University of California, Berkeley

Arun Kavishwar[*] ARUN_KAVISHWAR@ALUMNI.BROWN.EDU
Dana-Farber Cancer Institute

Stephanie L. McNamara SLMCNAMARA@MGH.HARVARD.EDU
Massachusetts General Hospital

Soujanya Samineni SOUJANYA_SAMINENI@DFCI.HARVARD.EDU
Dana-Farber Cancer Institute

Renato Umeton RENATO.UMETON@STJUDE.ORG
St. Jude Children's Research Hospital

Alexander Chowdhury ALEXANDER_CHOWDHURY@DFCI.HARVARD.EDU
Dana-Farber Cancer Institute

William Lotter LOTTERB@DS.DFCI.HARVARD.EDU
Dana-Farber Cancer Institute, Brigham and Women's Hospital, & Harvard Medical School

Abstract

Recent work has shown promising performance of frontier large language models (LLMs) and their multimodal counterparts in medical quizzes and diagnostic tasks, highlighting their potential for broad clinical utility given their accessible, general-purpose nature. However, beyond diagnosis, a fundamental aspect of medical image interpretation is the ability to localize pathological findings. Evaluating localization not only has clinical and educational relevance but also provides insight into a model's spatial understanding of anatomy and disease. Here, we systematically assess two general-purpose MLLMs (GPT-4 and GPT-5) and a domain-specific model (MedGemma) in their ability to localize pathologies on chest radiographs, using a prompting pipeline that overlays a spatial grid and elicits coordinate-based predictions. Averaged across nine pathologies in the CheXlocalize dataset, GPT-5 exhibited a localization accuracy of 49.7%, followed by GPT-4 (39.1%) and MedGemma (17.7%), all lower than a task-specific CNN baseline (59.9%) and a radiologist benchmark (80.1%). Despite modest performance, error analysis revealed that GPT-5's predictions were largely in anatomically plausible regions, just not always precisely localized. GPT-4 performed well on pathologies with fixed anatomical locations, but struggled with spatially variable findings and exhibited anatomically implausible predictions more frequently. MedGemma demonstrated the lowest performance on all pathologies, but showed improvements when provided examples through few shot prompting. Our findings highlight both the promise and limitations of current MLLMs in medical imaging and underscore the importance of integrating them with task-specific tools for reliable use.

Keywords: Multimodal LLMs, Chest Radiographs, Disease Localization

Data and Code Availability This study used the public CheXlocalize dataset (Saporta et al., 2022). Code is available at https://github.com/lotterlab/mllm_localization.

Institutional Review Board (IRB) This research does not require IRB approval.

1. Introduction

As AI advances from task-specific to generalist models, there is growing interest in evaluating frontier

[*] These authors contributed equally
[†] Work done while intern at Dana-Farber Cancer Institute

© 2025 A. Gosai, A. Kavishwar, S.L. McNamara, S. Samineni, R. Umeton, A. Chowdhury & W. Lotter.

models in medicine, including in medical imaging with the rise of multimodal large language models (MLLMs). Recent studies assessing GPT-4, a pioneering MLLM, on imaging-based quizzes and diagnostic tasks have yielded mixed results (Hayden et al., 2024; Brin et al., 2024; Strotzer et al., 2024; Zhou et al., 2024; Suh et al., 2024; Eriksen et al., 2023). For instance, Suh et al. (2024) reported that GPT-4 achieved clinical-level performance on *NEJM Image Challenges*, a multiple-choice, VQA dataset; however, Jin et al. (2024) found that the model's rationale for its choices were often flawed. Given the clinical potential of accessible, general-purpose MLLMs, these findings highlight the need to evaluate their performance across diverse tasks and to move beyond diagnostic accuracy toward a deeper understanding of their reasoning. Furthermore, it remains unclear if domain-specific MLLMs are required to address the challenges of general purpose models, or whether these challenges can instead be overcome with newer model versions.

A fundamental aspect of medical image interpretation is not only providing a diagnostic impression but also localizing findings associated with a diagnosis. In computer vision, diagnosis is typically framed as a classification task, whereas localization is approached through segmentation or bounding-box generation (i.e., detection) over relevant regions. In medical AI systems, these two tasks are treated as distinct regulatory categories: computer-aided diagnosis (CADx) for classification and computer-aided detection (CADe) for localization (McNamara et al., 2024). Critically, an AI model may perform well at classification, but struggle with precisely localizing the associated findings, or conversely struggle with distinguishing whether a finding is benign or pathological, but excel at identifying its location. While recent studies have evaluated the diagnostic performance of MLLMs, their ability to localize pathological findings (e.g., CADe) remains largely unexplored. Beyond potential clinical applications, evaluating MLLMs in localization can provide insights into their underlying clinical intuition and medical knowledge.

In this study, we systematically evaluate the ability of three MLLMs to localize pathologies on chest radiographs: two generalist models (GPT-4 (OpenAI, 2024) and GPT-5 (OpenAI, 2025)) and a domain-specific model (MedGemma (Sellergren et al., 2025)). To do so, we develop a prompting strategy that overlays a grid on the image and asks the model to predict coordinates corresponding to the pathology's location. We assess performance using the CheXlocalize dataset (Saporta et al., 2022), benchmarking against radiologists and a task-specific CNN model. To further explore the models' medical understanding, we analyze the distribution of predicted locations for each pathology and categorize mislocalizations based on their clinical plausibility. By benchmarking from a new angle–disease localization–and performing detailed error analysis, we sought to elucidate the spatial understanding of anatomy and disease of frontier MLLMs, probing both the potential for clinical utility and current limitations.

2. Related Work

Several recent studies have assessed general-purpose MLLMs such as GPT-4 in medical image interpretation (Jin et al., 2024; Hayden et al., 2024; Brin et al., 2024; Strotzer et al., 2024; Zhou et al., 2024; Suh et al., 2024; Eriksen et al., 2023). On 190 cases from the *Diagnosis Please* challenge, Suh et al. (2024) found that GPT-4 achieved diagnostic performance comparable to radiologists. GPT-4 also demonstrated strong accuracy on the *NEJM* Image Challenges, a multiple-choice diagnostic quiz for medical professionals (Jin et al., 2024; Eriksen et al., 2023). However, Jin et al. (2024) reported that while GPT-4 performed well on these tasks, its rationales were often flawed when explicitly prompted to explain its reasoning. On multiple-choice questions from the *American College of Radiology Diagnostic Radiology In-Training Examination*, GPT-4 achieved high accuracy on text-only questions (81.5%) but struggled with image-based ones (47.8%). Its diagnostic accuracy on real-world radiologic imaging cohorts has also shown considerable variability (Strotzer et al., 2024; Brin et al., 2024; Zhou et al., 2024).

3. Methods

While prior research has focused primarily on the diagnostic capabilities of MLLMs, our study examines spatial localization accuracy for pathologies in medical images that the models are explicitly prompted to recognize. We utilize chest radiographs from CheXlocalize (Saporta et al., 2022), a subset of the CheXpert (Irvin et al., 2019) dataset annotated with pixel-level masks for each of its classes. We focus on this task because it serves as both a common benchmark for medical AI and a clinically important application. Figure 1 contains an overview of the approach.

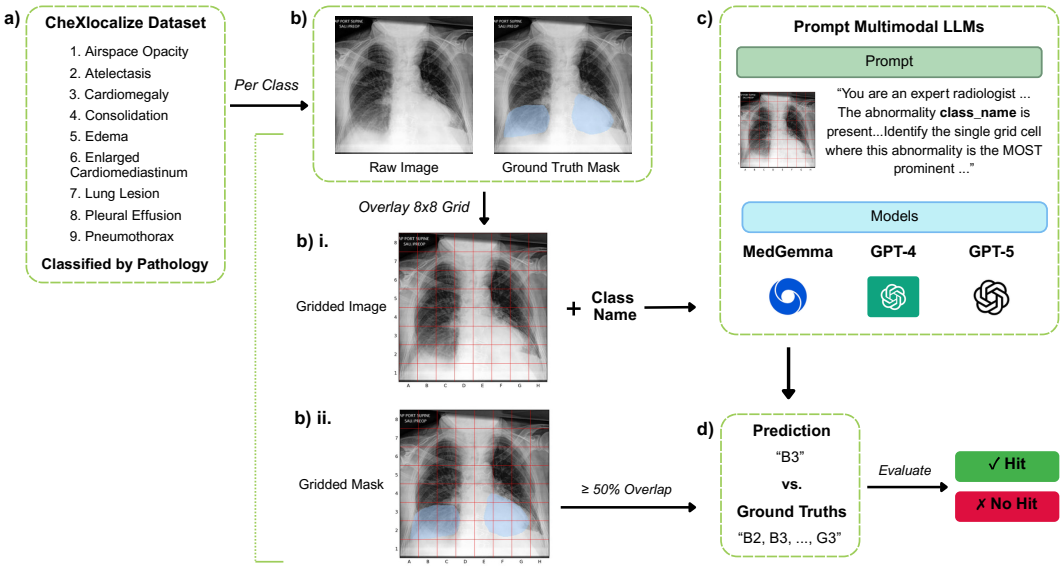

Figure 1: Study Overview. a) The nine pathologies in the CheXlocalize dataset were utilized. b) For each pathology class, a gridded image was produced for each radiograph where the pathology was present. c) Three multimodal LLMs were prompted to identify the grid cell where the stated pathology is most prominent. d) The predicted grid cells were compared to the ground truth cells, defined as cells with ≥50% overlap with the ground truth mask.

3.1. CheXlocalize Dataset

The CheXlocalize dataset includes chest radiograph images collected from Stanford Hospital in both inpatient and outpatient settings (Irvin et al., 2019). The dataset contains expert radiologist annotations for the localization of 10 distinct findings (Saporta et al., 2022). These ground-truth annotations come in the form of pixel-wise segmentation maps, enabling precise benchmarking. The validation and test splits of this dataset consist of 234 radiographs from 200 patients and 668 radiographs from 500 patients, respectively. All results reported in the current study are based on the test split. The image counts by pathology and view position are contained in the Appendix.

The 10 classes annotated in CheXlocalize are Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Lung Lesion, Lung Opacity, Pleural Effusion, Pneumothorax, and Support Devices. In this work, we include the first 9 classes, excluding Support Devices to focus exclusively on pathology localization. Our evaluation encompasses both frontal (anterior-posterior and posterior-anterior) and lateral view radiographs available in this dataset for a comprehensive assessment.

We compare our results on the test split to an expert human benchmark and a convolutional neural network (CNN) baseline reported alongside the dataset in Saporta et al. (2022). Both the ground truth annotations and the human benchmarking assessment were performed by radiologists, where the benchmarking effectively quantifies a performance ceiling of inter-reader variability. Specifically, ground truth annotations were performed by two board-certified radiologists. Three separate radiologists then also performed annotations, which were compared to the ground truth annotations to create the human performance benchmark. The CNN baseline consists of a DenseNet121 (Huang et al., 2017) classification model trained on CheXpert. Localization predictions were obtained through a GradCAM-based (Selvaraju et al., 2017) saliency pipeline. As detailed below, we compare the MLLMs to the 'hit rate' performance metric for the human and CNN benchmarks, which was quantified by Saporta et al. (2022) as whether the most representative point identified by the CNN saliency method or human benchmark annotations fell within the ground truth segmentation.

3.2. Prompting Technique

To evaluate MLLM localization capabilities, we designed a structured prompting pipeline by which each chest radiograph was digitally overlaid with a standardized grid. The grid divided each image into equally-sized cells, labeled alphanumerically to facilitate text-based spatial referencing. To do so, each radiograph image was first centrally cropped and resized to 256×256 pixels, a common pre-processing strategy for computer vision models. The MLLM was then tasked with identifying the specific grid cell that was most representative of the given pathology. For the core results, we use an 8x8 grid overlay (32×32 pixels per cell), which was chosen to balance the resolution size with the typical size of the pathology masks. Analysis using a higher resolution grid (16x16 cells) is also presented in the Appendix for sensitivity analysis.

We used the same prompt for all MLLMs, which explicitly stated that the given pathology was present in the image in order to make localization, not identification, the sole objective (see Appendix for full prompt). Because a single radiograph could exhibit multiple pathologies, we generated separate copies of each image for every present pathology and performed individual queries for each pathology-image pair. This approach enabled independent evaluation of localization performance for each pathology.

3.3. Evaluation Metrics

Performance evaluation was based on a 'hit rate' criterion, designed to mirror the pointing game methodology introduced by Saporta et al. (2022). A prediction (represented by a single grid cell returned by the MLLM) was considered a 'hit' if at least 50% of the grid cell overlapped with the ground truth segmentation provided in the CheXlocalize annotations for the queried pathology. Predictions falling below this overlap threshold were classified as misses. In the rare scenario where the ground truth mask contained no cells with $\geq$50% overlap (4.4% of all images), the prediction was deemed a hit if the cell contained any amount of ground truth mask. This could occur, for instance, if the segmentation mask was very small. Our defined hit rate metric enables meaningful comparison with the results reported by Saporta et al. (2022), wherein the pointing game methodology for their human benchmark awarded credit when the identified representative point fell within the ground-truth segmentation region. We report the average hit rate for each pathology and MLLM, alongside standard deviation error bars using bootstrapping (1,000 samples).

3.4. Error Analysis

Predictions not meeting the hit criteria were qualitatively reviewed by a radiologist and categorized into three distinct error types:

- **Partial hit**: Predicted cell partially overlapped ($\leq$50%) with the ground truth annotation.
- **Position error**: Predicted cell overlapped with anatomically plausible but incorrect regions for the pathology.
- **Anatomy error**: Predicted cell was anatomically implausible for the given pathology, suggesting a potential misunderstanding of chest anatomy and/or the pathology.

For consistency, the error analysis was only performed on frontal radiographs, which represent 87% of images in the test set. If a model produced more than 50 complete misses for a given pathology, only 50 were reviewed, and the proportion of position versus anatomy errors was estimated from this subset.

3.5. Models

GPT-4 and GPT-5 were executed using the OpenAI API. The model version for GPT-4 was `gpt-4o-2024-05-13`, configured at a temperature of 0 for reproducibility. The model version for GPT-5 was not obtainable via the API, but it was run within a month of its release on 2025-08-07, and the only relevant configurable parameter available was the reasoning effort, which was kept at the default level of medium. MedGemma was executed locally on 2 H100 GPUs using the Ollama platform. The 27B, instruction-tuned variant was used (accessed via puyangwang/medgemma-27b-it:q8). The temperature for MedGemma was left at a default value of 0.8. A temperature of 0 was also initially considered, but this resulted in almost no variability in the predicted grid cells (96.7% of predictions in either D4 or C4); thus the default temperature was deemed the most appropriate assessment for MedGemma.

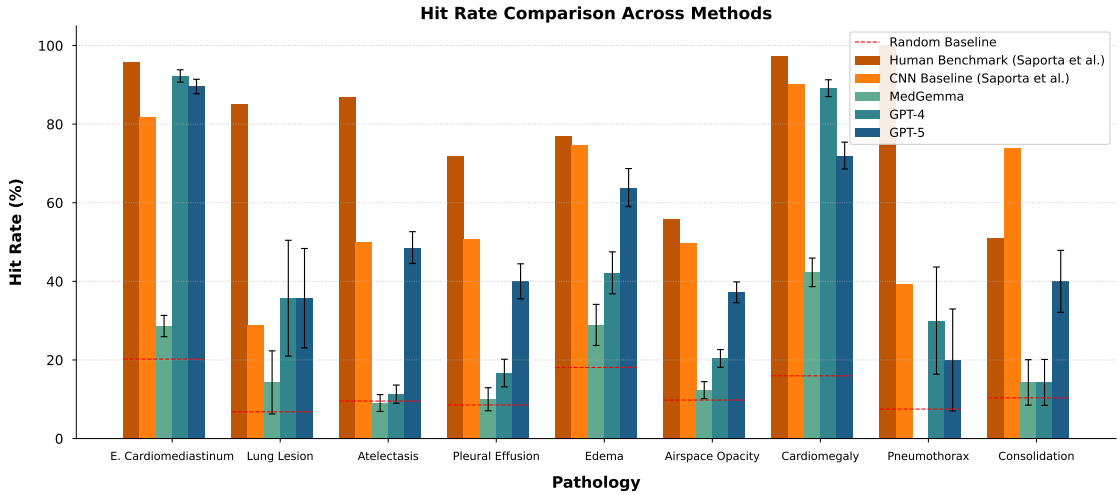

Figure 2: Hit rate by pathology. Pathologies are ordered by the relative performance difference between GPT-5 and the CNN baseline. Error bars represent standard deviation via bootstrapping.

4. Results

4.1. Hit Rate Performance

Figure 2 summarizes the hit rate performance of the MLLMs compared to the results reported by Saporta et al. (2022), consisting of an expert human (radiologist) benchmark and a convolutional neural network (CNN) baseline. A random baseline is also presented, which reflects the expected performance if cell selections were made uniformly at random over the grid. The random baseline is computed per pathology by first computing the chance level for each image (# ground truth cells / total cells) and then averaging across all images for that pathology.

Between the three MLLMs, GPT-5 had the highest average hit rate across pathologies (49.7%), followed by GPT-4 (39.1%), and MedGemma (17.7%). All models underperformed the CNN baseline and human benchmark on average, which showed 59.9% and 80.1% average hit rates, respectively. The average random baseline across pathologies was 11.9%.

Different trends were observed for different pathologies. GPT-4 and GPT-5 both outperformed the CNN baseline on enlarged cardiomediastinum (hit rate of 92.3% for GPT-4, 89.6% for GPT-5, 81.8% for CNN). GPT-5 showed the greatest improvements (≥20% higher hit rates) over GPT-4 on atelectasis, consolidation, edema, and pleural effusion, though it still underperformed the CNN baseline on these pathologies. GPT-4 exhibited its highest performance over GPT-5 for cardiomegaly (hit rate of 89.1% vs. 72.0%). MedGemma underperformed for each pathology, with its largest margin over chance occurring for cardiomegaly. These overall findings were robust to the resolution of the grid overlay, where halving the size of each cell showed similar patterns (Appendix Figure 6).

4.2. Error Categorization

To gain insights into the MLLMs' performance and general medical understanding, misses were reviewed and categorized as partial hits, position errors, or anatomy errors (see Methods for definitions). The breakdown of the MLLM predictions by error category is displayed in Fig. 3. For the two pathologies where GPT-4 and GPT-5 exhibited their highest performance (enlarged cardiomediastinum and cardiomegaly), the majority of misses were still partial hits for both models. The distribution of error categories varied more widely for the other pathologies. Despite a moderate average 'full hit' rate, only 6.3% of GPT-5's predictions represented an 'anatomy error' on average across pathologies, indicating that the majority of misses were either a partial hit or at least corresponded to a plausible anatomical region. GPT-4 exhibited a higher rate of anatomy errors (18.0% on average), with MedGemma exhibiting the most (29.9% on average). For both GPT-4 and GPT-5, the highest percentages of anatomy errors were observed for pleural effusion and pneumothorax, indicating that the models' predictions were relatively

often localized to an incorrect anatomical structure (i.e., non-lung regions) for these pathologies.

4.3. Visualization of Predictions

Heatmaps of the models' predictions were generated to further understand the error patterns (Fig. 4, 7, 8). A separate heatmap was generated per pathology and model, summarizing the model's predictions for the given pathology across the dataset. To provide an aggregate anatomical reference, the heatmaps are visualized over the average image across the test set. Only frontal views were considered in generating the heatmaps.

For cardiomegaly, GPT-4 consistently predicted central grid cells (Fig. 4). As cardiomegaly indicates an enlarged heart and the heart tends to appear towards the center of the image, the prediction heatmap aligns with the high performance of GPT-4 for this pathology. GPT-5 also predicted central grid cells for cardiomegaly, but demonstrated more variability that aligns with the ground truth distribution. Edema, indicating fluid in the lungs, varies more widely in its spatial position, though it often appears in the medial (towards center) regions of the lungs at the site of highest pulmonary vascular density. For this pathology, GPT-4 consistently predicted a central grid cell, often overlaying the heart/mediastinum (Fig. 4), whereas GPT-5's predictions were more distributed over the lungs, explaining its higher performance on this pathology. In general, the heatmaps of GPT-5 more faithfully represented the ground truth pathology, explaining its relatively low percentage of anatomy errors (Fig. 7, 8).

Examining anatomy errors on an individual image basis revealed a spectrum of model misunderstanding. GPT-5 exhibited no anatomy errors for consolidation, which is a pathology that localizes to the lungs, but GPT-4 and MedGemma sometimes predicted grid cells overlaying the heart/mediastinum (Figure 5). However, a dramatic error for GPT-5 is displayed in Figure 5, wherein both GPT-5 and GPT-4 predicted the shoulder as the location of a pneumothorax (collapsed lung).

4.4. Prompt Variations

Given the moderate performance, we explored whether different prompting strategies could lead to improvements. We considered two additional approaches: chain-of-thought (CoT) prompting wherein the models were instructed to think step-by-step, and

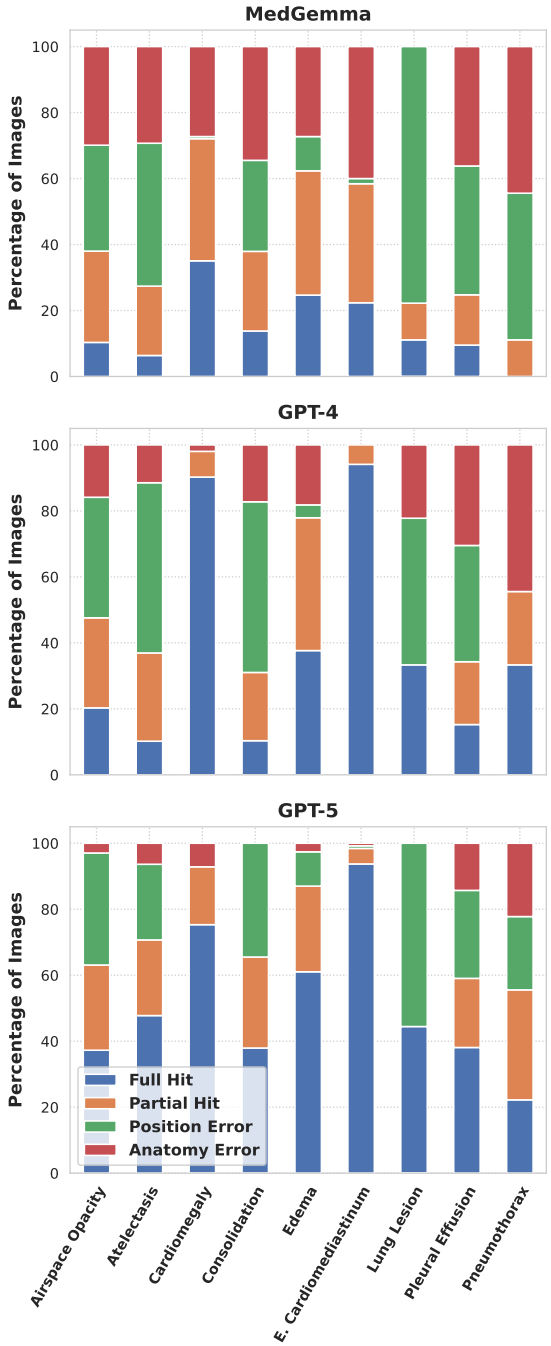

Figure 3: Error categorization. Each prediction on a frontal radiograph was categorized as a full hit (≥50% overlap), partial hit (0 < overlap < 50%), position error (no overlap but plausible anatomy), and anatomy error (implausible anatomy).

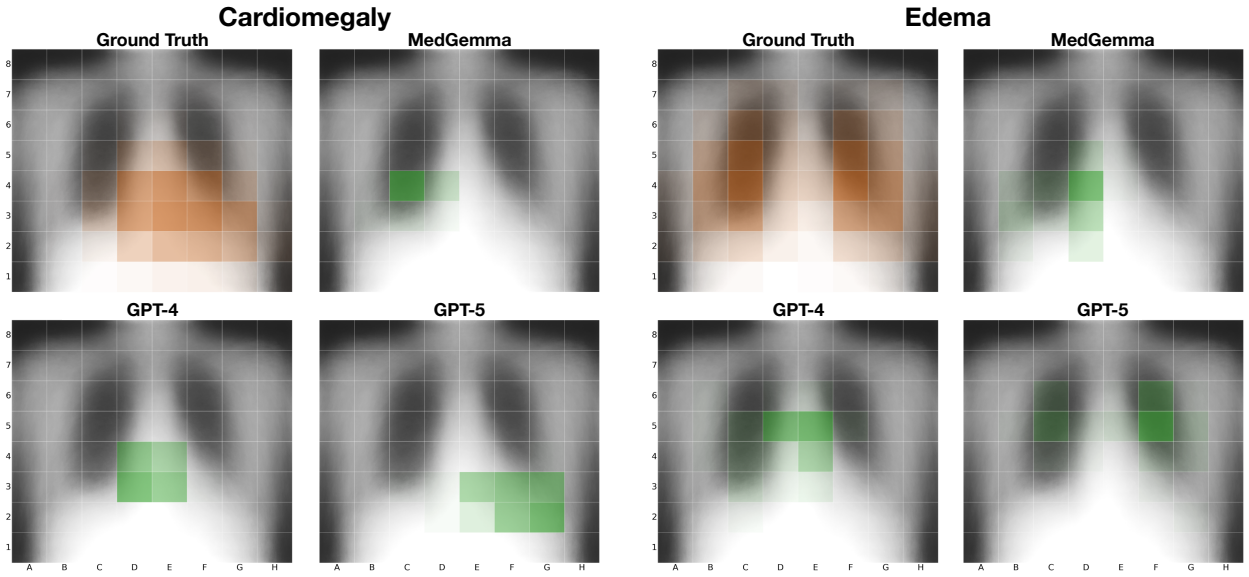

Figure 4: Example ground truth and prediction heatmaps. The heatmaps are computed over the frontal radiographs in the test set and are overlaid on the average image.

few shot prompting wherein the models were provided three randomly selected examples for the given pathology. Additional details and the full prompts are contained in the Appendix.

The mean performance across pathologies for each prompting strategy is contained in Table 1, and the performance for each pathology contained in Appendix Table 2. Overall, performance changes were modest. MedGemma exhibited the greatest relative improvement, from an average hit rate of 17.7% to 22.5% (27% increase) and 32.0% (81% increase) for CoT and few shot prompting, respectively. On a per-pathology basis, few-shot prompting improved MedGemma's performance for all pathologies except lung lesion, with particularly large gains for atelectasis, consolidation, and pleural effusion (>130% relative increases each). The improvement from CoT prompting was driven primarily by enlarged cardiomediastinum (191% increase). Few shot prompting also elicited non-trivial predictions for MedGemma even when the temperature was set to 0, resulting in an average hit rate of 32.1%, similar to that of the default temperature. Nonetheless, MedGemma's performances remained below the zero shot performances of GPT-4 and GPT-5. For GPT-4 and GPT-5, CoT prompting did not increase performance, but few shot prompting led to 4% and 5% relative increases, respectively. These increases were driven by edema and pleural effusion for GPT-4, and airspace opacity, cardiomegaly, and edema for GPT-5 (each with a ≥15% relative hit rate improvement).

Table 1: Mean hit rate across pathologies for each prompting strategy.

Model	**Prompt**	**Mean Hit Rate**
Random Baseline	-	11.9
Human Benchmark	-	80.1
CNN Baseline	-	59.9
MedGemma	Zero Shot	17.7
	CoT	22.5
	Few Shot	32.0
GPT-4	Zero Shot	39.1
	CoT	37.8
	Few Shot	40.8
GPT-5	Zero Shot	49.7
	CoT	47.4
	Few Shot	52.1

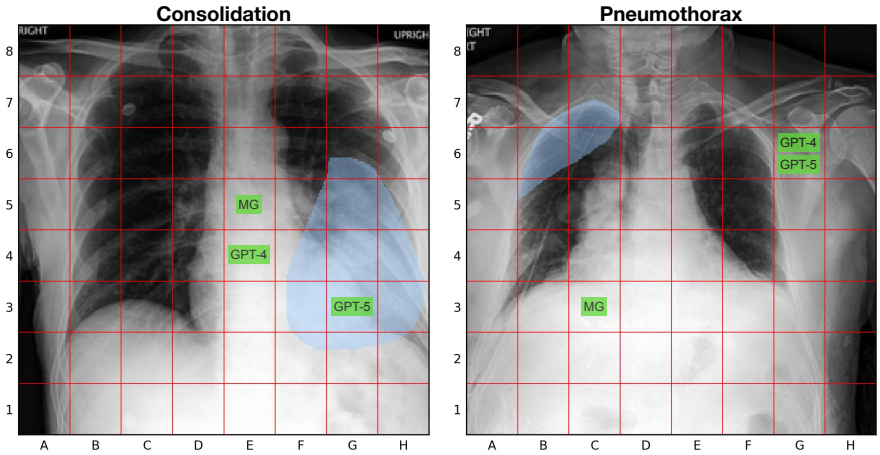

Figure 5: Example anatomy errors. a) A consolidation example where GPT-5 is correct but GPT-4's prediction overlays the heart/mediastinum. b) A pneumothorax example where both GPT-4 and GPT-5 predictions overlay the shoulder instead of the lungs.

4.5. Comparing MedGemma to Gemma

Although MedGemma is specifically designed for the medical domain, it is a relatively smaller model (27B parameters) compared to GPT-4 and GPT-5 (estimated to be at least 10x larger). To disentangle the effects of model size from domain-specific training, we evaluated the performance of Gemma 3 (27B), the base model on which MedGemma is built. Gemma performed similarly to MedGemma for zero shot prompting (average hit rate of 17.6% vs. 17.7%, respectively) and CoT prompting (25.0% and 22.5%); however, MedGemma's few shot performance was moderately higher (32.0% vs. 25.5%) with improvements observed for 7 out of 9 pathologies (Appendix Table 2). Thus, MedGemma's low performance on the studied task may partly be attributable to its smaller size, where the domain-specific training appears to help with few shot prompting but not baseline performance.

5. Discussion

Beyond diagnosis, we evaluated the ability of frontier MLLMs to localize findings in chest radiographs. We found that performance was mixed, with GPT-4 and GPT-5 performing relatively strong on pathologies that appear in consistent anatomical locations, but underperforming the CNN and human benchmark overall. Nonetheless, GPT-5 showed strong improvements compared to GPT-4 on the more spatially variable pathologies, and its predictions were largely anatomically plausible even when it did not precisely localize the pathology. MedGemma, a medicine-specific model that includes chest radiographs in its training, exhibited the lowest performance overall and had a high proportion of anatomically-implausible predictions. Few shot prompting brought MedGemma's performance closer to the other models, with modest improvements for GPT-4 and GPT-5 as well.

Our findings have several actionable implications for clinical practice and AI development. First, while (M)LLMs have recently shown strong performance on question-answer challenges and diagnostic tasks (Eriksen et al., 2023; Suh et al., 2024; Nori et al., 2023), our results show that leading models struggle with fine-grained spatial reasoning. Smaller models such as MedGemma (and Gemma 3 27B on which it was built) may especially be limited in the ability to generalize to novel tasks, such as the one introduced here, compared to traditional benchmarks used in the field. Thus, the observed results may be partly attributable to the task itself rather than inherent medical knowledge, highlighting the importance of rigorous evaluation and adaptation strategies (e.g., few shot prompting) for specific use cases in clinical practice. While GPT-5's performance remained well below the human benchmark, the improvements in accuracy and medical understanding observed with GPT-5 compared to GPT-4 are encouraging. It re-

mains to be seen whether these gains will continue to occur with future model versions, but these results at least suggest that large, general-purpose MLLMs are a promising approach compared to domain-specific vision-language models. In the meantime, agentic strategies that integrate the flexibility of LLMs with task-specific tools may be the best strategy, where recent work in the context of chest radiograph interpretation has demonstrated this possibility (Fallahpour et al., 2025).

Limitations Our study has several limitations. Our analysis focused on one dataset given the need for localized annotations and meaningful benchmark comparisons. Future work could extend our analysis pipeline to other datasets and tasks. Furthermore, some pathologies in the CheXlocalize dataset have a small image count, leading to large error bars for the results for those pathologies. We included these pathologies because they still reveal meaningful model behavior, as illustrated in Fig. 5 where GPT-4 and GPT-5 localize a lung pathology to the shoulder. Another challenge lies in the variability of ground truth annotation sizes across images and pathologies, which makes it difficult to define a hit rate metric that robustly captures all scenarios. Additionally, while interpretable, hit rate metrics sacrifice the flexibility and granularity offered by other localization metrics, such as IoU. Finally, although we explored different prompting strategies, more extensive prompt engineering could be explored in the future.

Conclusion Systematically evaluating foundation models will be increasingly important as AI is integrated into clinical settings. Beyond our current findings, the proposed analysis pipeline can be leveraged to probe the spatial reasoning of foundation models more broadly. A deeper understanding of the strengths and limitations of emerging models will be essential for guiding algorithm improvements and ensuring safe, effective clinical use.

Acknowledgments

W.L. acknowledges funding support from NIBIB award R21EB035247 and NLM award R01LM014775.

References

Dana Brin, Vera Sorin, Yiftach Barash, Eli Konen, Benjamin S Glicksberg, Girish N Nadkarni, and Eyal Klang. Assessing GPT-4 multimodal performance in radiological image analysis. *Eur. Radiol.*, August 2024.

Alexander V Eriksen, Sören Möller, and Jesper Ryg. Use of GPT-4 to diagnose complex clinical cases. *NEJM AI*, November 2023.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. Medrax: Medical reasoning agent for chest x-ray, 2025. URL https://arxiv.org/abs/2502.02673.

Nolan Hayden, Spencer Gilbert, Laila M Poisson, Brent Griffith, and Chad Klochko. Performance of GPT-4 with vision on text- and image-based ACR diagnostic radiology in-training examination questions. *Radiology*, September 2024.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269. IEEE Computer Society, 2017. ISBN 978-1-5386-0457-1. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2017.html#HuangLMW17.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, 2019.

Qiao Jin, Fangyuan Chen, Yiliang Zhou, Ziyang Xu, Justin M Cheung, Robert Chen, Ronald M Summers, Justin F Rousseau, Peiyun Ni, Marc J Landsman, Sally L Baxter, Subhi J Al'Aref, Yijia Li, Alexander Chen, Josef A Brejt, Michael F Chiang, Yifan Peng, and Zhiyong Lu. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *NPJ Digit Med*, 7(1):190, July 2024.

Stephanie L McNamara, Paul H Yi, and William Lotter. The clinician-AI interface: intended use and explainability in FDA-cleared AI devices for medical image interpretation. *NPJ Digit Med*, 7(1):80, March 2024.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of

gpt-4 on medical challenge problems, 2023. URL https://arxiv.org/abs/2303.13375.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

OpenAI. Gpt-5. https://openai.com/gpt-5, 2025.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven Q H Truong, Chanh D T Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, Matthew P Lungren, and Pranav Rajpurkar. Benchmarking saliency methods for chest X-ray interpretation. *Nat. Mach. Intell.*, 4(10):867–878, October 2022.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Baptiste Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. MedGemma Technical Report. arXiv preprint arXiv:2507.05201, 2025. URL https://arxiv.org/abs/2507.05201.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626. IEEE Computer Society, 2017. ISBN 978-1-5386-1032-9. URL http://dblp.uni-trier.de/db/conf/iccv/iccv2017.html#SelvarajuCDVPB17.

Quirin D Strotzer, Felix Nieberle, Laura S Kupke, Gerardo Napodano, Anna Katharina Muertz, Stefanie Meiler, Ingo Einspieler, Janine Rennert, Michael Strotzer, Isabel Wiesinger, Christina Wendl, Christian Stroszczynski, Okka W Hamer, and Andreas Schicho. Toward foundation models in radiology? quantitative assessment of GPT-4V's multimodal and multianatomic region capabilities. *Radiology*, 313(2):e240955, November 2024.

Pae Sun Suh, Woo Hyun Shim, Chong Hyun Suh, Hwon Heo, Chae Ri Park, Hye Joung Eom, Kye Jin Park, Jooae Choe, Pyeong Hwa Kim, Hyo Jung Park, Yura Ahn, Ho Young Park, Yoonseok Choi, Chang-Yun Woo, and Hyungjun Park. Comparing diagnostic accuracy of radiologists versus GPT-4V and gemini pro vision using image inputs from diagnosis please cases. *Radiology*, 312(1):e240273, July 2024.

Yiliang Zhou, Hanley Ong, Patrick Kennedy, Carol C Wu, Jacob Kazam, Keith Hentel, Adam Flanders, George Shih, and Yifan Peng. Evaluating GPT-V4 (GPT-4 with vision) on detection of radiologic findings on chest radiographs. *Radiology*, 311(2): e233270, May 2024.

Appendix

Prompts

Baseline zero shot prompt

System: You are an expert chest radiologist specializing in analyzing {view} chest X-rays. Your task is to precisely localize abnormalities using a grid overlay.
User: [Image]
This is a gridded {view} view of a chest X-ray. The abnormality '{condition}' is confirmed to be present in this image. Your task:

1. Identify the single grid cell where this abnormality - '{condition}' is the MOST prominent.
2. Provide only the grid coordinate for this most representative cell. A grid coordinate is defined as a letter followed by a number. If the abnormality spans multiple cells, choose the cell that is most representative.
3. Do not include any explanations or additional text in your response.

Chain-of-Thought Prompt

System: You are an expert chest radiologist specializing in analyzing view chest X-rays. Your task is to precisely localize abnormalities using a grid overlay. Approach this task methodically, using your expertise to analyze the image step-by-step.
User: [Image]
Your task:

1. Carefully examine the entire chest X-ray image, paying close attention to areas where '{condition}' typically manifests.
2. Identify all grid cells where you can observe signs of '{condition}'. List these cells in your thought process.
3. For each identified cell, briefly note the specific features or abnormalities you observe that are consistent with '{condition}'.
4. Compare the identified cells and their features. Consider which cell contains the most prominent or representative manifestation of '{condition}'.
5. If the condition spans multiple cells, determine which single cell encompasses the most significant part of the abnormality.
6. Based on your analysis, select the single grid coordinate (e.g., D5) that best represents the location of '{condition}' in this X-ray.
7. Provide your final answer as a single grid coordinate without any additional explanation or text.

Example thought process: "I observe signs of '{condition}' in cells C4, D4, and D5. In C4, there's [specific feature]. D4 shows [another feature]. D5 contains [feature]. Comparing these, D5 appears to be the most representative location for '{condition}'. Final answer: D5"

Fewshot Prompt

System: You are an expert chest radiologist specializing in analyzing chest X-rays. You are given a gridded {view} image of a chest X-ray with the abnormality '{condition}' confirmed to be present.
User: Zeroshot Prompt with [Example Image 1]
Assistant: Ground Truth grid coordinate for Example 1
User: Zeroshot Prompt with [Example Image 2]
Assistant: Ground Truth grid coordinate for Example 2
User: Zeroshot Prompt with [Example Image 3]
Assistant: Ground Truth grid coordinate for Example 3

User: Zeroshot Prompt with [Test Image]

Few shot selection: Three random examples from the validation split with the same view position and pathology were selected. If there were not enough examples in the validation split, the remainder were selected randomly from the test split (excluding the image being evaluated).

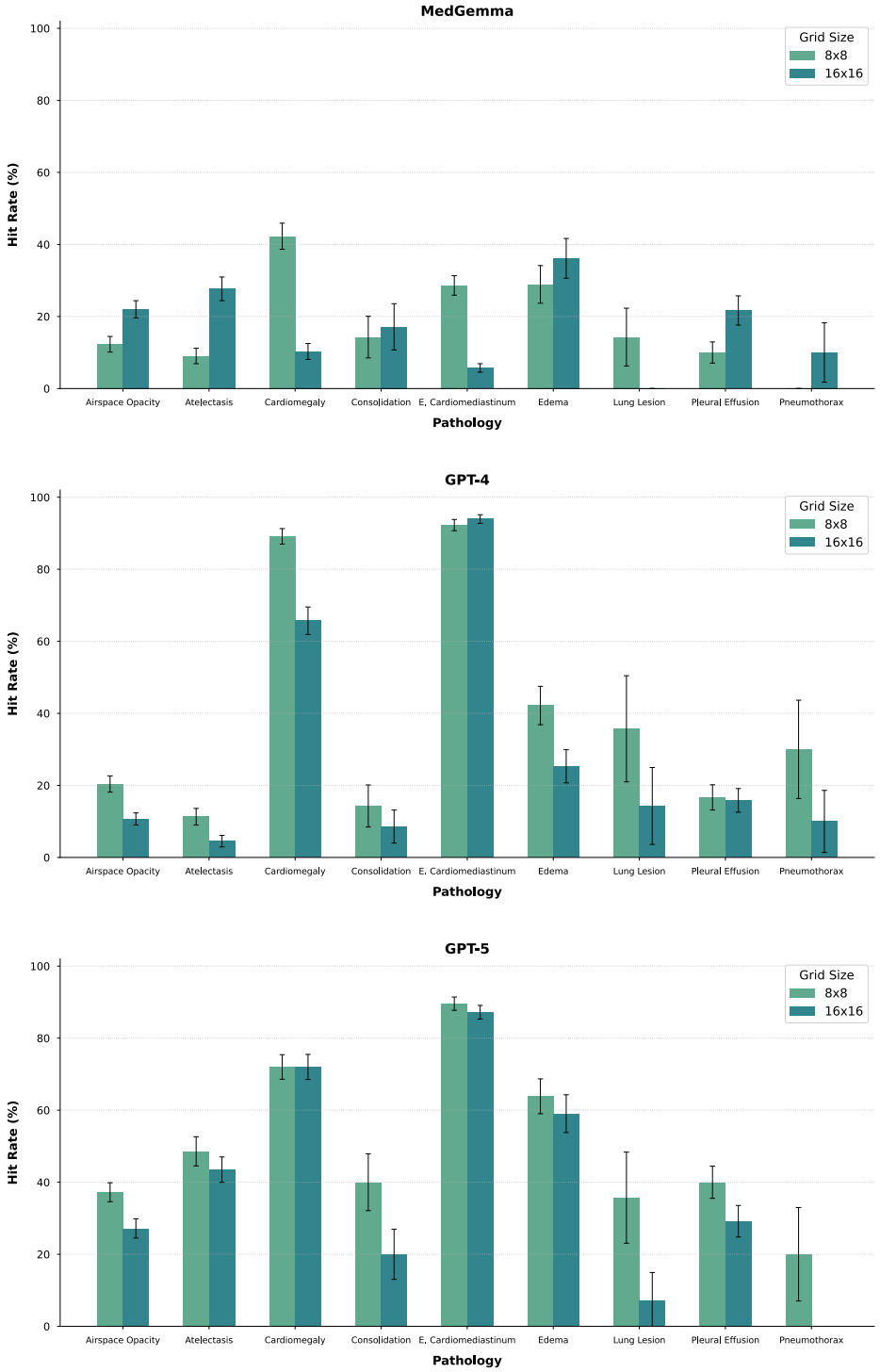

Figure 6: Performance by grid size. An 8x8 grid size was used for the main analysis. A higher resolution grid size (16x16) showed similar patterns between models and did not increase performance.

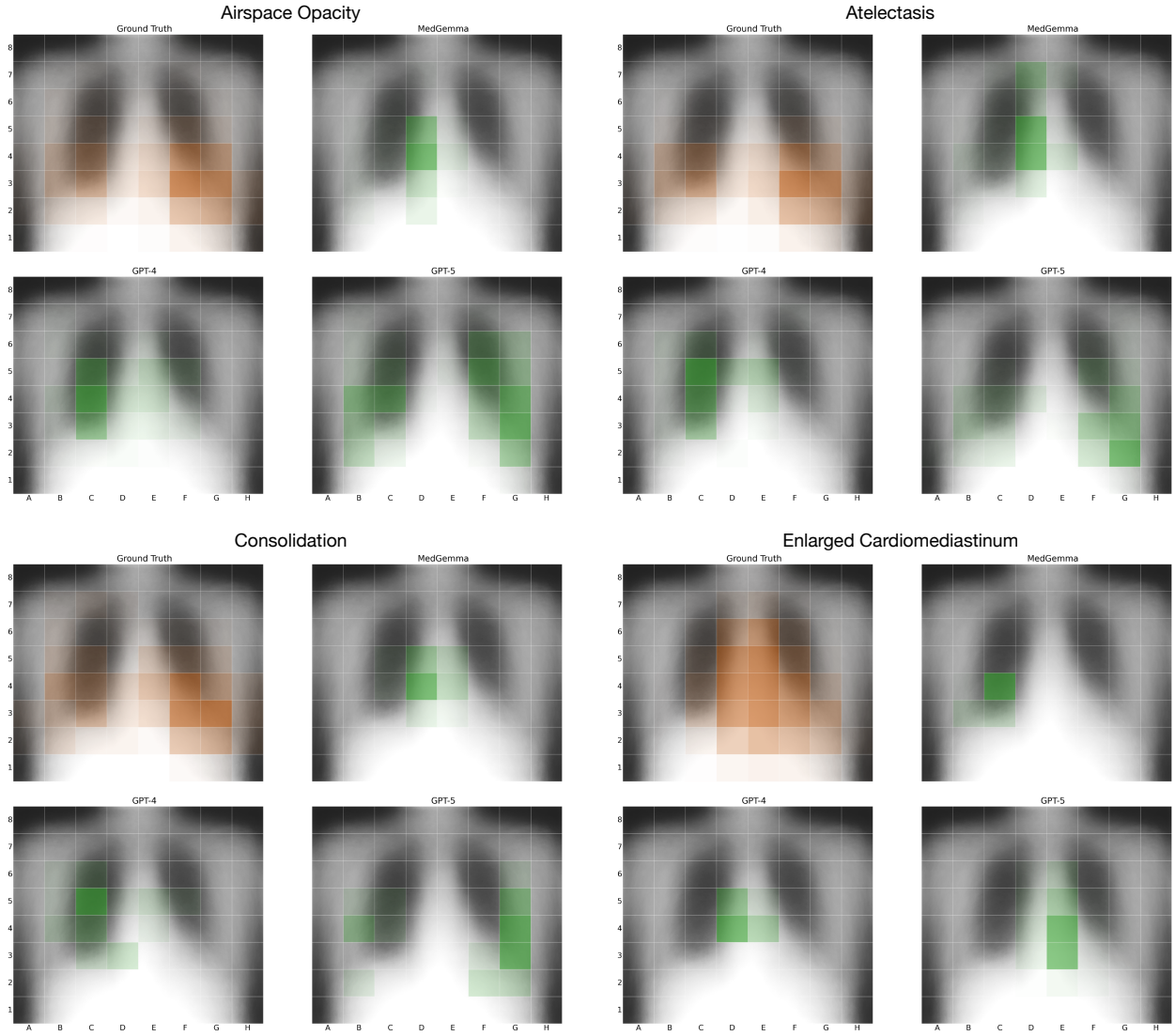

Figure 7: Ground truth and prediction heatmaps. The heatmaps are computed over the frontal radiographs in the test set and are overlaid on the average image.

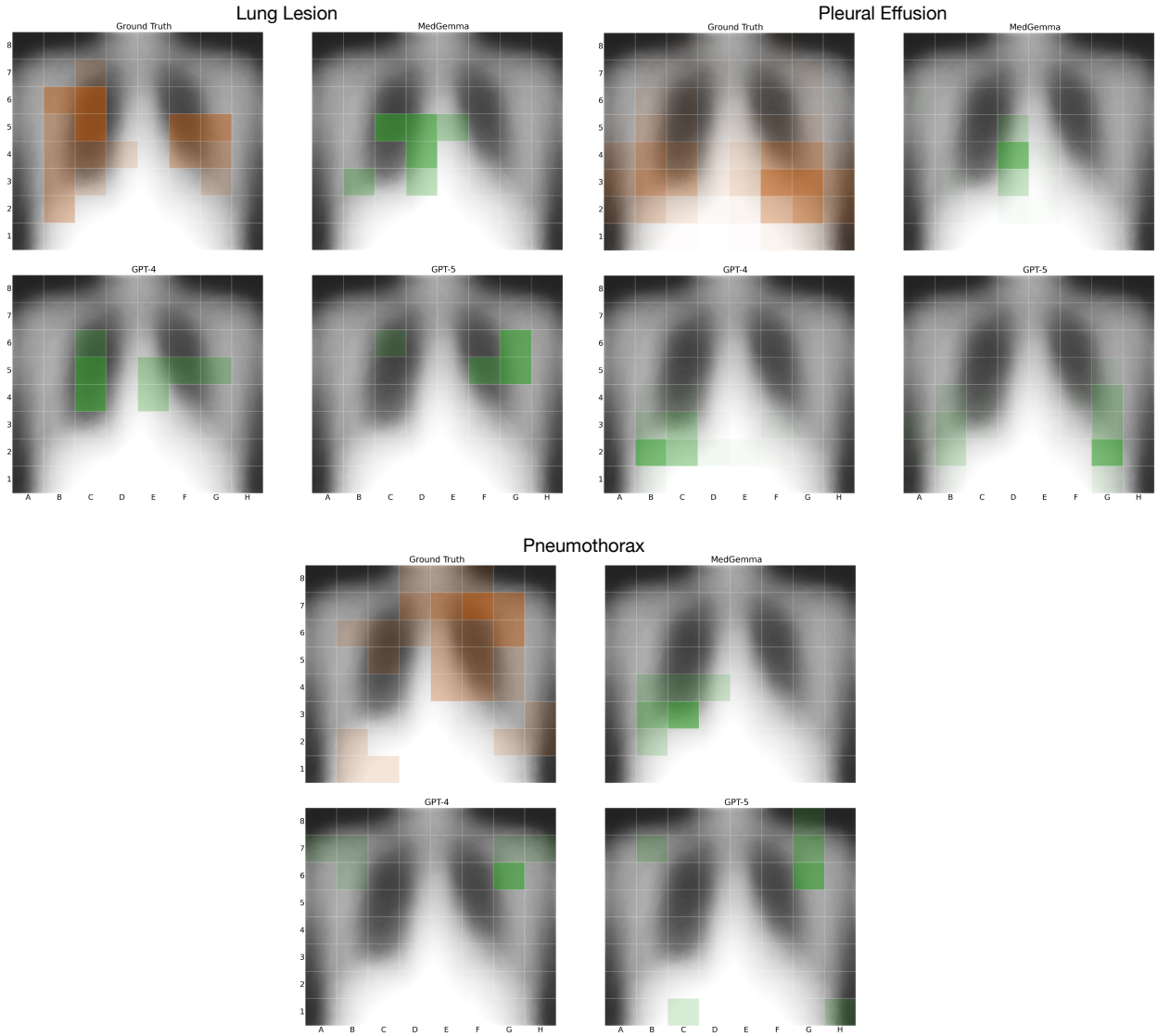

Figure 8: Ground truth and prediction heatmaps. The heatmaps are computed over the frontal radiographs in the test set and are overlaid on the average image.

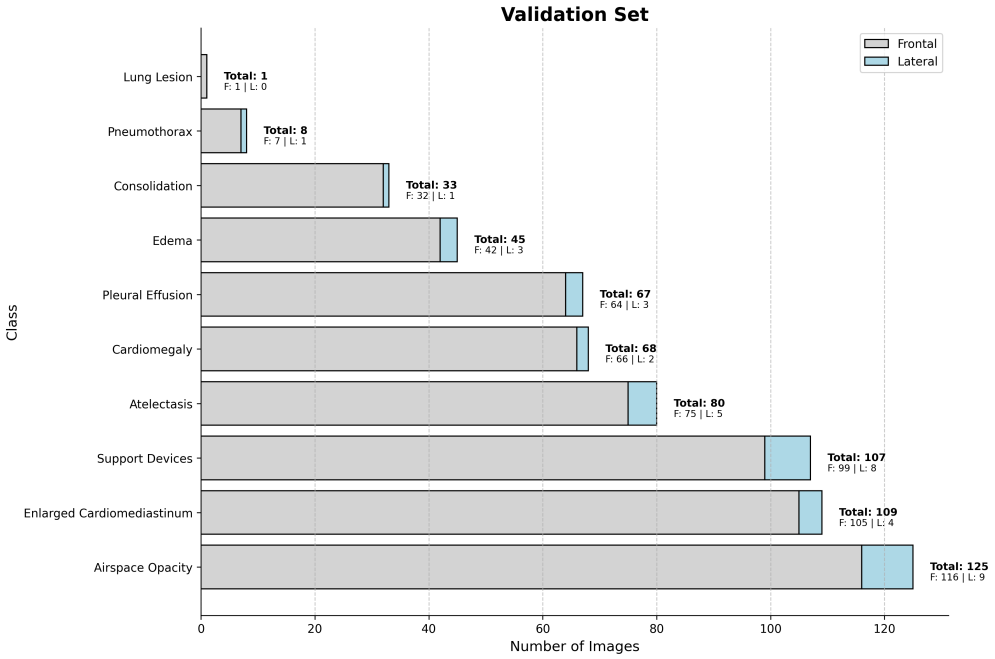

Figure 9: Image counts in CheXlocalize Validation Set.

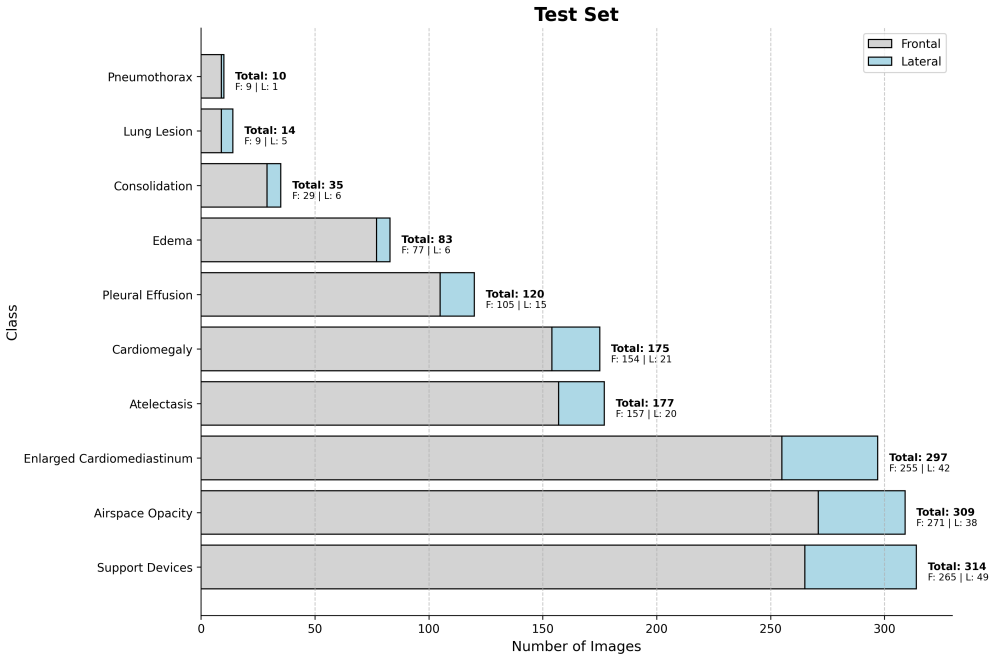

Figure 10: Image counts in CheXlocalize Test Set.

Table 2: Hit rate per pathology for each prompting strategy. Bold indicates the highest performance for the pathology; underline indicates the second highest.

Model	Prompt	Airspace Opacity	Atelect.	Cardiomeg.	Consol.	Edema	E. Cardio.	Lung Lesion	Pleural Effusion	Pneumo.
Random Baseline	-	9.8	9.5	16.0	10.4	18.1	20.2	6.8	8.6	7.5
Human Benchmark	-	**55.9**	**87.0**	**97.2**	51.0	**76.9**	**95.7**	**85.0**	**71.8**	**100.0**
CNN Baseline	-	49.8	50.1	90.3	**73.8**	74.6	81.8	29.0	50.7	39.2
MedGemma	Zero Shot	12.3	9.0	42.3	14.3	28.9	28.6	14.3	10.0	0.0
	CoT	6.1	7.9	37.7	17.1	31.3	83.2	0.0	9.2	10.0
	Few Shot	23.3	23.7	57.7	37.1	43.4	62.6	7.1	23.3	10.0
GPT-4	Zero Shot	20.4	11.3	89.1	14.3	42.2	92.3	35.7	16.7	30.0
	CoT	20.1	9.0	81.1	22.9	45.8	90.2	35.7	25.0	10.0
	Few Shot	17.8	11.3	85.7	20.0	54.2	92.9	35.7	19.2	30.0
GPT-5	Zero Shot	37.2	48.6	72.0	40.0	63.9	89.6	35.7	40.0	20.0
	CoT	36.2	44.6	72.6	31.4	61.4	91.9	21.4	36.7	30.0
	Few Shot	44.7	50.3	84.6	45.7	74.7	91.6	21.4	45.8	10.0
Gemma3 27B	Zero Shot	12.0	11.3	19.4	17.1	20.5	53.9	7.1	7.5	10.0
	CoT	6.1	9.6	33.1	17.1	50.6	83.5	21.4	3.3	0.0
	Few Shot	16.8	21.5	42.9	5.7	34.9	73.1	14.3	20.0	0.0
Image Count	-	309	177	175	35	83	297	14	120	10