Worse than Random? An Embarrassingly Simple Probing Evaluation of Large Multimodal Models in Medical VQA

Qianqi Yan[1], Xuehai He[1], Xiang Yue[2], Xin Eric Wang[1]

[1]University of California, Santa Cruz, [2]Carnegie Mellon University

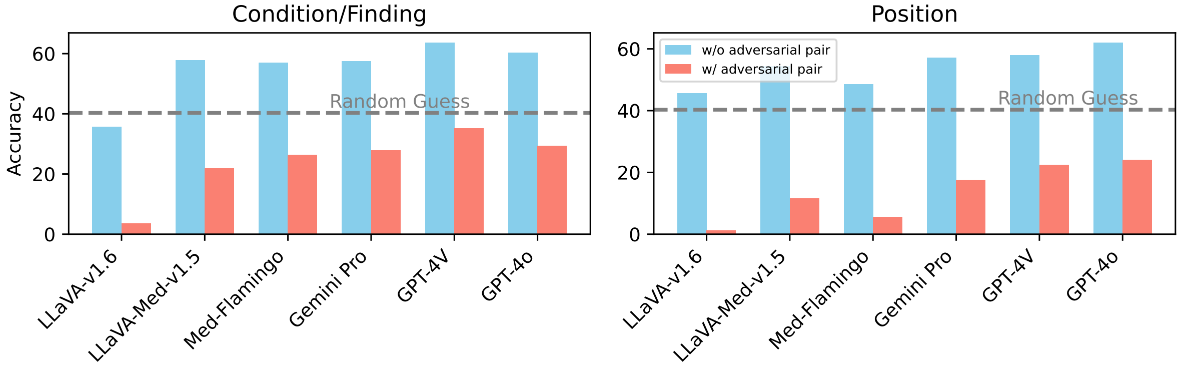

Figure 1: Accuracy of six LMMs on two types of specialized questions in medical diagnoses, with and without adversarial pairs. The significant drop in accuracy with adversarial pairs highlights the models' unreliability in handling medical diagnoses with our probing method.

Abstract

Large Multimodal Models (LMMs) have demonstrated impressive performance on existing medical Visual Question Answering (Med-VQA) benchmarks. However, high reported accuracy does not necessarily reflect their true diagnostic reliability in clinical settings. This study reveals that state-of-the-art models perform worse than random guessing on medical diagnosis questions when subjected to simple Probing Evaluation for Medical Diagnosis (ProbMed). ProbMed challenges models through *probing evaluation* and *procedural diagnosis*. Particularly, probing evaluation features pairing ground-truth questions with adversarial counterparts that feature negated and hallucinated attributes, while procedural diagnosis requires reasoning across various dimensions for each image, including modality recognition, organ identification, clinical findings, abnormalities, and positional grounding. Our evaluation reveals that even top-performing models like GPT-4o, GPT-4V, and Gemini Pro perform worse than random guessing on specialized diagnostic questions, indicating significant limitations in handling fine-grained medical inquiries. Furthermore, our ablation study on open-source models (e.g., LLaVA, LLaVA-Med, and Med-Flamingo) identifies poor visual understanding as a primary bottleneck—a limitation that can be partially mitigated by incorporating visual descriptions generated by GPT-4o, resulting in an average performance improvement of 9.44%. These findings underscore the urgent need for more robust evaluation methods and domain-specific expertise to ensure the reliability of LMMs in high-stakes medical applications.

1 Introduction

Foundation models, such as large language models (LLMs) (Achiam et al., 2023; Touvron et al., 2023; Jiang et al., 2023; Anil et al., 2023; Chung et al., 2024) and large multimodal models (LMMs) (OpenAI, 2024, 2023; Reid et al., 2024; Li et al., 2023; Liu et al., 2023a; Chen et al., 2023), have demonstrated impressive capabilities in understanding complex visual and text inputs, generating human-like language, and achieving high accuracy on various benchmarks. The integration of these foundation models into real-life medical practice holds immense potential given their advanced computational capabilities (Wu et al., 2023a; Yang et al., 2023) and promising progress on existing medical

Findings of the Association for Computational Linguistics: ACL 2025, pages 19188–19205
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

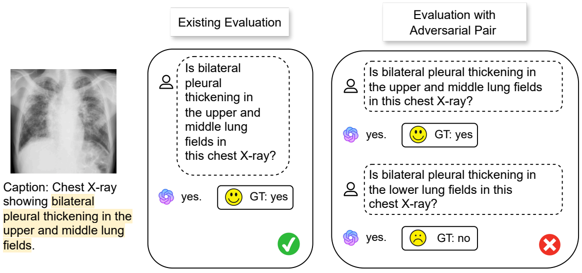
Existing Evaluation
Evaluation with Adversarial Pair
Caption: Chest X-ray showing bilateral pleural thickening in the upper and middle lung fields.
Is bilateral pleural thickening in the upper and middle lung fields in this chest X-ray?
yes.
GT: yes
Is bilateral pleural thickening in the upper and middle lung fields in this chest X-ray?
yes.
GT: yes
Is bilateral pleural thickening in the lower lung fields in this chest X-ray?
yes.
GT: no

Figure 2: An example illustrating the potential for misleading accuracy in existing evaluations. While the model correctly identifies the position of an existing finding in the standard evaluation, it fails to differentiate between actual and hallucinated positions when subjected to an adversarial evaluation.

Visual Question Answering (Med-VQA) benchmarks (Lau et al., 2018; Liu et al., 2021; He et al., 2020; Zhang et al., 2023). As we stand on the precipice of integrating these models into critical decision-making domains, one natural question appears: *how much can we trust these models in real-world scenarios, such as medicine and healthcare, where the stakes are high?*

Before discussing the reliability of LMMs in critical domains like Med-VQA, we must first address a fundamental question: *Are we evaluating LMMs correctly?* To address this question, we introduce a simple yet effective probing evaluation method that exposes the weaknesses of LMMs by creating binary questions with hallucination pairs over existing benchmarks. An example is shown in Figure 2. Despite the high accuracy reported on current Med-VQA tasks, our study reveals a significant vulnerability in LMMs when faced with adversarial questioning, as illustrated in Figure 1. The observed performance drops are alarming: even advanced models like GPT-4o, GPT-4V, and Gemini Pro perform worse than random guessing, with an average decrease of 27.78% across the tested models.

Based on this, we further analyze a critical question: *How reliable are LMMs in medical diagnosis, ranging from general questions to specialized diagnostic questions*? To address this question, we introduce ProbMed, which features procedural diagnosis designed to rigorously evaluate model performance across multiple diagnostic dimensions. We curated ProbMed from 6,303 images sourced from two widely-used biomedical datasets, MedICaT (Subramanian et al., 2020) and ChestX-ray14 (Wang et al., 2017). These images cover various modalities, including X-ray, MRI, and CT scans, and span multiple organs such as the abdomen, brain, chest, and spine. Using GPT-4 and a positional reasoning module, we generated metadata for each image, extracting information about abnormalities, condition names, and their corresponding locations. This metadata facilitated the automatic generation of 57,132 high-quality question-answer pairs, covering dimensions like modality recognition, organ identification, abnormalities, clinical findings, and positional reasoning.

Our systematic evaluation of twelve state-of-the-art LMMs on ProbMed revealed several critical insights. *First*, even the best-performing models, such as GPT-4o and Gemini Pro, performed close to random guessing on specialized diagnostic categories like Condition/Finding and Position, highlighting their limitations in handling fine-grained medical inquiries. *Second*, introducing adversarial pairs significantly reduced the accuracy of all models, with LLaVA-Med-v1.5's performance dropping by up to 28.22% and GPT-4o's accuracy decreasing by 20.71% in ProbMed. These findings emphasize the importance of adversarial testing in Med-VQA to uncover model weaknesses. *Third*, by incorporating chain-of-thought reasoning and adding visual descriptions generated by GPT-4o, we observe substantial improvements in model performance, suggesting that poor visual understanding is a critical bottleneck. The results indicate that augmenting these models with more accurate visual information could significantly improve their ability to handle complex medical tasks. Moreover, the CheXagent model, which was exclusively trained on chest X-rays, demonstrated that specialized domain knowledge is crucial. It showed that expertise gained on one particular organ could be transferable to another modality of the same organ in a zero-shot manner, highlighting the value of domain-specific training for improving model performance.

In summary, our work highlights significant gaps in the reliability of LMMs for medical diagnosis despite their impressive performance on existing general domain benchmarks. The insights from ProbMed underscore the urgent need for robust evaluation methodologies to ensure the accuracy and reliability of LMMs in real-world medical applications. Our findings also suggest that poor visual understanding is a key limitation for open-source models, which can be mitigated by incorporating chain-of-thought reasoning and accurate visual descriptions. This research inspires the development of more trustworthy AI systems in healthcare and beyond, ultimately contributing to better diagnostic outcomes and patient care.

2 Related Work

Large Multimodal Models in the Medical Domain The advancements in Large Multimodal Models (LMMs) have significantly enhanced the understanding and generation of medical content that integrates both visual and linguistic elements. Notable models include GPT-4o (OpenAI, 2024), GPT-4V (OpenAI, 2023), Gemini Pro (Reid et al., 2024), LLaVA (Liu et al., 2023a, 2024), and MiniGPT-v2 (Chen et al., 2023). The scalability and exceptional performance of these large foundation models have driven their application in the biomedical field.

Further progress has been made in fine-tuning general-domain LMMs for the biomedical field, resulting in specialized models like BiomedGPT (Zhang et al., 2024), LLaVA-Med (Li et al., 2024), Med-Flamingo (Moor et al., 2023), MedBLIP (Chen and Hong, 2024), RadFM (Wu et al., 2023b) and MedVInT (Zhang et al., 2023). Despite the promising results from these domain-specific LMMs, ongoing exploration exists into training smaller multimodal models to address specific clinical needs. For instance, models like LLaVA-RAD (Chaves et al., 2024) and CheXagent (Chen et al., 2024) have been developed for chest X-ray interpretation, aiming to bridge competency gaps in radiology tasks.

Comprehensive surveys of LLMs for healthcare highlight the progress, applications, and challenges in deploying LLMs in clinical settings (He et al., 2023; Zhou et al., 2024; Peng et al., 2023). Task-specific evaluations (Yan et al., 2023; Liu et al., 2023b) underline the potential and challenges of LMMs in the medical domain. As we move towards integrating these models into critical decision-making processes, it becomes imperative to assess their reliability in high-stakes environments like healthcare and medicine.

Medical Visual Question Answering Medical Visual Question Answering (Med-VQA) plays a crucial role in assessing the capabilities of models in interpreting and responding to queries about medical images. Some benchmarks, like VQA-RAD (Lau et al., 2018) and SLAKE (Liu et al., 2021), are manually constructed with categorical question types. While this method ensures high-quality question-answer pairs, it is labor-intensive and results in limited dataset scales.

Automated curation methods have been developed to address scalability. PathVQA (He et al., 2020) uses CoreNLP[1] tools, and PMC-VQA (Zhang et al., 2023) employs generative models to create larger datasets. However, these methods often sacrifice fine-grained question categories, and some require additionally trained models for question filtering.

Different evaluation methods are employed for assessing LMMs, including closed-ended VQA, multiple choice VQA, and open-ended generation tasks such as captioning and report generation. Open-ended VQA and report generation are typically considered more challenging and harder to evaluate, often requiring human or model evaluation alongside automated lexical similarity metrics like ROUGE-L and BLEU-4. Recent works (Wang et al., 2024; Zheng et al., 2024; Zong et al., 2024) argue that multiple-choice questions may not be ideal due to inherent selection bias and permutation sensitivity. In our work, we choose a relatively easy-to-evaluate method: closed-ended VQA augmented with adversarial evaluation methods featuring hallucinated attributes. By requiring the model to accurately distinguish relevant features, we enhance the reliability of the evaluation process. This method allows for clear and definitive assessments, improving the overall robustness of our findings in medical contexts.

3 Probing Evaluation for Medical Diagnosis

In this section, we introduce two complementary evaluation strategies that rigorously assess state-of-the-art LMMs for Med-VQA. Our approach is designed to answer the following research questions:

1. *Is the current evaluation of LMMs for Med-VQA reliable?*
2. *How reliable are LMMs on medical diagnosis, ranging from general questions to specialized diagnostic questions?*

Despite current high accuracy, we find that the models struggle with simple probing evaluation on existing benchmarks. Our evaluation framework, ProbMed (Probing Evaluation for Medical

[1] https://stanfordnlp.github.io/CoreNLP

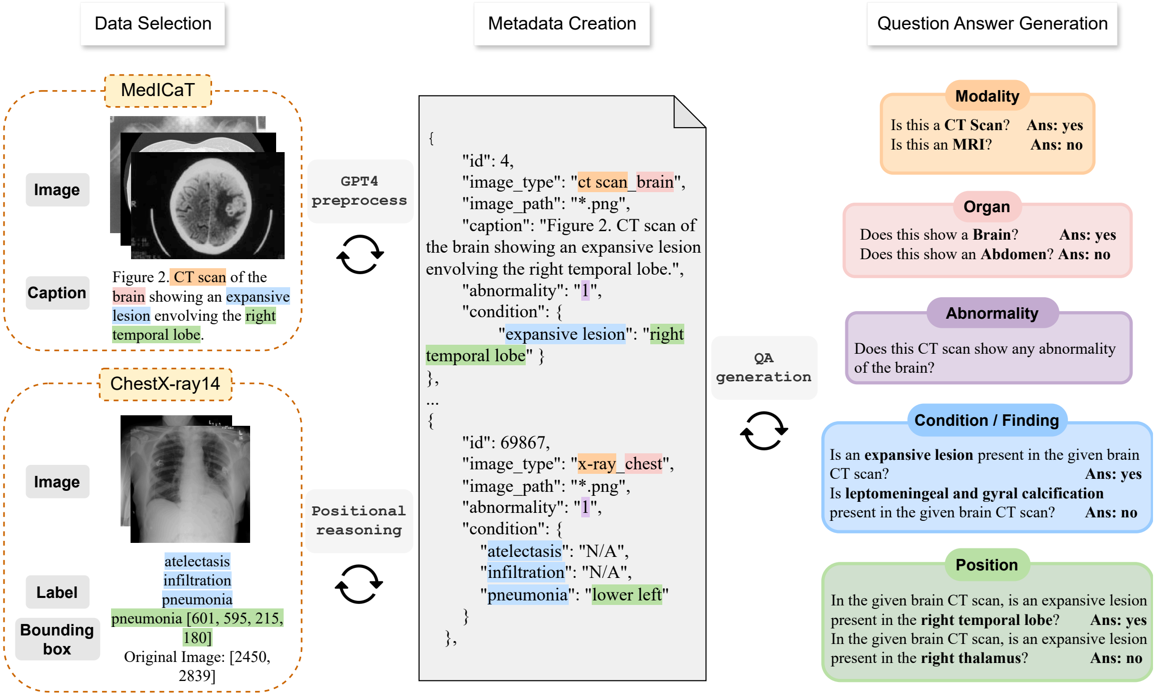

Figure 2. CT scan of the brain showing an expansive lesion envolving the right temporal lobe.

atelectasis
infiltration
pneumonia
pneumonia [601, 595, 215, 180]
Original Image: [2450, 2839]

```
{
    "id": 4,
    "image_type": "ct scan_brain",
    "image_path": "*.png",
    "caption": "Figure 2. CT scan of
the brain showing an expansive lesion
envolving the right temporal lobe.",
    "abnormality": "1",
    "condition": {
        "expansive lesion": "right
temporal lobe" }
},
...
{
    "id": 69867,
    "image_type": "x-ray_chest",
    "image_path": "*.png",
    "abnormality": "1",
    "condition": {
      "atelectasis": "N/A",
      "infiltration": "N/A",
      "pneumonia": "lower left"
    }
  },
```

Is this a **CT Scan**? **Ans: yes**
Is this an **MRI**? **Ans: no**

Does this show a **Brain**? **Ans: yes**
Does this show an **Abdomen**? **Ans: no**

Does this CT scan show any abnormality of the brain?

Is an **expansive lesion** present in the given brain CT scan? **Ans: yes**
Is **leptomeningeal and gyral calcification** present in the given brain CT scan? **Ans: no**

In the given brain CT scan, is an expansive lesion present in the **right temporal lobe**? **Ans: yes**
In the given brain CT scan, is an expansive lesion present in the **right thalamus**? **Ans: no**

Figure 3: Flow diagram of the ProbMed data curation process. Two comprehensive biomedical datasets were utilized to collect source data and construct a metadata file, enabling the automatic generation of high-quality question-answer pairs for the ProbMed dataset.

Diagnosis), is built on adversarial testing and multi-faceted diagnostic reasoning to further expose these vulnerabilities and provide a thorough analysis of model performance. We also explore enhancements through chain-of-thought reasoning and the incorporation of external visual descriptions (e.g., from GPT-4o) to address the noted limitations of open-sourced models.

3.1 Probing Evaluation with Adversarial Pairs

A core element of our framework is the use of adversarial pairs to test model robustness. For each image, we generate pairs of: **ground-truth questions** that query the presence of a specific entity (e.g., a particular clinical finding) with corresponding **adversarial questions** that introduce a negated or hallucinated attribute (e.g., a non-existent finding or an alternative organ). This pairing challenges the model to discern between actual diagnostic features and spurious details, thereby revealing its ability—or inability—to filter out irrelevant or misleading information. The performance drop observed under adversarial conditions highlights the fragility of current evaluation protocols and motivates the need for more robust assessment methods.

3.2 Procedural Diagnosis

Beyond binary question-answering, ProbMed incorporates *procedural diagnosis* to evaluate the models' diagnostic reasoning. Each image is associated with questions spanning multiple diagnostic dimensions, including **Modality Recognition:** Identifying the imaging technique (e.g., X-ray, MRI, CT). **Organ Identification:** Determining the anatomical region under investigation. **Clinical Findings and Abnormalities:** Detecting abnormal conditions. **Positional Reasoning:** Localizing findings spatially within the image. This multifaceted evaluation framework requires models to integrate diverse pieces of information for each test sample, thereby providing a more comprehensive measure of their diagnostic capabilities.

3.3 Data Filtering and Curation

ProbMed is curated from two widely recognized biomedical datasets: MedICaT and ChestX-ray14. The data curation process, summarized in Figure 3, involves the following steps:

Image Selection: From MedICaT (Subramanian et al., 2020), we extracted 4,543 image-caption pairs that focus on a single organ and modality with clear indications of normal or abnormal conditions. From ChestX-ray14 (Wang et al., 2017), we selected 1,760 frontal-view X-ray images balanced between healthy and abnormal cases, including those with bounding box annotations for abnormalities.

Metadata Generation: For each image, we generate a unified metadata record:

$$D_i = \{\texttt{mod}_i, \texttt{organ}_i, \{\texttt{condition}_j, \texttt{pos}_j\}_j^{n_i}\}$$

where $\texttt{mod}_i$ and $\texttt{organ}_i$ denote the imaging modality and anatomical region, respectively, and each $\{\texttt{condition}_j, \texttt{pos}_j\}$ pair represents a detected clinical finding and its positional description. For MedICaT, GPT-4 is leveraged via few-shot prompting to extract abnormality details and positional cues from captions. For ChestX-ray14, a dedicated positional reasoning module generates descriptive text based on bounding box coordinates.

3.4 Evaluation Protocol

For each diagnostic entity in the metadata, we automatically generate a **ground-truth** question Q_i, asking the model to confirm the presence of that specific entity. An **adversarial question** Q'_i, constructed by randomly selecting an alternative or hallucinated entity (e.g., an incorrect organ or false condition) and expecting a "no" response.

Crucially, our accuracy metric is defined in a strict manner: an entity is considered correctly predicted only if the model provides the correct answer for both Q_i and Q'_i. In other words, if a model answers "yes" to both questions for a given entity, it is deemed incorrect rather than receiving partial credit. For images containing more than one $\{\texttt{condition}_j, \texttt{pos}_j\}$ pair, the accuracy under the Condition/Finding and Position category is computed as the average accuracy over all n_i pairs—there is no bonus for partial correctness. This evaluation setup ensures that only unambiguous, fully correct responses are counted as hits, highlighting the models' true diagnostic reliability. (See Appendix B for detailed statistics on the number of questions in each category.)

3.5 Expert Study

To validate the reliability of our metadata and the corresponding question-answer pairs, we conducted an expert verification study. Two medical experts independently reviewed 100 randomly sampled metadata entries out of 6,303 from ProbMed, as well as the 1,090 QA pairs corresponding to those metadata entries. The review process yielded an average accuracy of 94.00% for the metadata and 97.79% for the QA pairs. This rigorous validation underscores the quality and thorough curation of the ProbMed dataset. In addition, we conducted a study using the expert-verified data on GPT-4o to verify the robustness of our probing evaluation framework to dataset errors, detailed in Appendix C.

As of statistics reported in Table 1, our data curation process produced a total of 57,132 question-answer pairs (averaging 9 pairs per image), spanning a comprehensive set of diagnostic dimensions. These high-quality, balanced pairs provide a robust foundation for evaluating model performance.

Table 1: Dataset Statistics of ProbMed. There are 6.3k images and 57k VQA pairs in total. The dataset is balanced within each question type and image type.

Organ, Modality	Image	Question	Question with Answer "yes"	Unique Condition	Unique Positional Description
Abdomen MRI	84	757	375	107	75
Brain MRI	566	5,046	2,509	697	446
Chest MRI	40	382	189	52	38
Spine MRI	324	3,346	1,664	461	336
Abdomen CT scan	751	6,855	3,410	909	552
Brain CT scan	270	2,417	1,200	335	209
Chest CT scan	548	5,161	2,572	727	353
Spine CT scan	87	941	470	149	93
Abdomen X-ray	232	2,046	1,018	277	160
Brain X-ray	79	599	298	84	44
Chest X-ray	3,178	27,530	13,278	1,418	694
Spine X-ray	202	2,052	1,020	300	172
Total	6,303	57,132	28,003	/	/

4 Experiments and Analysis

We conducted a systematic evaluation and comprehensive analysis using ProbMed on twelve state-of-the-art LMMs to identify their strengths and weaknesses in imaging diagnostics. Apart from proprietary GPT-4o (OpenAI, 2024), GPT-4V (OpenAI, 2023) and Gemini Pro (Reid et al., 2024), we selected nine open-source models spanning across general models including LLaVA-v1 (Liu et al., 2024), LLaVA-v1.6 (Liu et al., 2023a), MiniGPT-v2 (Chen et al., 2023) and specialized models including LLaVA-Med-v1, LLaVA-Med-v1.5 (Li et al., 2024), Med-Flamingo (Moor et al., 2023), BiomedGPT (Zhang et al., 2024), RadFM (Wu et al., 2023b) and CheXagent (Chen et al., 2024). These models were chosen based on their computational cost, efficiency, and inference speed, making them practical for integration into medical practice.

4.1 RQ1: Reliability of Current Med-VQA Evaluation

Adversarial Evaluation in VQA-RAD. To assess whether current Med-VQA evaluations capture model vulnerabilities, we first introduce adversarial

Table 2: Model accuracy on the VQA-RAD test subset and ProbMed with adversarial pairs. Accuracy is reported in two ways: (1) averaged across individual questions in a pair and (2) requiring both the ground truth and adversarial questions for the same image to be answered correctly. The drop in accuracy across models demonstrates their vulnerability to adversarial questions, with percentage decreases shown in parentheses.

Models	VQA-RAD		ProbMed	
	Averaged Accuracy (%)	Accuracy (%) with Adversarial Pairs	Averaged Accuracy (%)	Accuracy (%) with Adversarial Pairs
LLaVA-v1	62.28	25.42 (-36.84)	55.82	19.30 (-36.51)
LLaVA-v1.6	44.06	8.47 (-35.59)	56.02	24.96 (-31.06)
MiniGPT-v2	66.10	46.61 (-19.49)	59.82	27.67 (-32.14)
LLaVA-Med-v1	43.22	3.38 (-39.83)	52.26	17.90 (-34.35)
LLaVA-Med-v1.5	48.30	15.25 (-33.05)	68.41	40.19 (-28.22)
CheXagent	55.50	21.18 (-34.32)	58.70	30.61 (-28.08)
BiomedGPT	56.35	17.79 (-38.55)	60.14	33.34 (-26.79)
Med-Flamingo	61.01	25.42 (-35.59)	64.13	35.66 (-28.47)
RadFM	67.79	38.98 (-28.81)	67.70	41.00 (-26.70)
Gemini Pro	63.13	44.91 (-18.22)	75.08	55.08 (-20.00)
GPT-4V	58.47	33.89 (-24.57)	75.70	55.28 (-20.42)
GPT-4o	69.91	55.08 (-14.83)	76.31	55.60 (-20.71)

pairs on the VQA-RAD test set (Lau et al., 2018). Because VQA-RAD provides only finalized QA pairs without detailed metadata, adversarial pairs were manually constructed by medical experts for 118 “yes” instances (yielding 236 QA pairs total). As shown in Table 2 (left), despite being based on limited information and scale, these adversarial questions lead to drastic accuracy drops. For example, models such as GPT-4o show a reduction from 69.91% to 55.08% (a 14.83% decrease), highlighting the need for robust evaluation protocols.

Adversarial Evaluation in ProbMed. The ProbMed dataset systematically incorporates adversarial pairs on scale for all 57k QA pairs. Here, each diagnostic entity is paired with a ground-truth question and a corresponding adversarial question. Table 2 (right) demonstrates a similar significant impact: even the best-performing models experience a minimum 20.00% drop in accuracy (with an average decrease of 27.78% across models) when evaluated under this rigorous scheme. This result emphasizes that high accuracy on standard benchmarks can be misleading without essential adversarial evaluation for uncovering model weaknesses.

4.2 How Reliable Are LMMs in Medical Diagnosis?

After "correcting" inflated model accuracy by introducing adversarial pairs, we continue to address the second research question. We conducted diagnostic probing ranging from general to specialized diagnostic questions using the ProbMed dataset.

Performance across Diagnostic Questions Table 3 shows the categorical accuracy of different models aggregated among all image types. While GPT-4o, GPT-4V, and Gemini Pro outperform other models and excel in general tasks such as recognizing image modality and organs, their low performance in specialized tasks like determining the existence of abnormalities and answering fine-grained questions about condition/finding and position highlights a significant gap in their ability to aid in real-life diagnosis.

On more specialized diagnostic questions, even top-performing models like GPT-4o, GPT-4V, and Gemini Pro performed close to random guessing. Their accuracy in identifying conditions and positions was alarmingly low, underscoring their limitations in handling fine-grained medical inquiries. RadFM, LLaVA-Med-v1.5, and Med-Flamingo outperform other specialized models in general questions, yet still struggle with specialized questions. LLaVA-Med-v1.5 achieves much higher accuracy among open-sourced models in identifying conditions/finding and their positions, but still performs around 10% worse than the proprietary models.

Among the open-sourced general-purpose models, MiniGPT-v2 performs the best, surpassing domain-specific models such as LLaVA-Med-v1 in determining positions of condition/finding without domain-specific training. A more detailed breakdown of the performance of different models on different image types across each question type

Table 3: Categorical and overall accuracy (%) of different models aggregated among all image types in ProbMed (averaging over three runs). The overall accuracy is weighted by the number of questions in each type. The best result in each question category is **in-bold**, and the second best is underlined.

Models	General Question		Specialized Question			Overall
	Modality	Organ	Abnormality	Condition/Finding	Position	
Random Choice	25.00	25.00	50.00	35.67	36.48	32.13
LLaVA-v1	25.30$_{\pm 1.18}$	41.92$_{\pm 1.21}$	50.00$_{\pm 2.01}$	0.35$_{\pm 0.03}$	0.14$_{\pm 0.06}$	19.30$_{\pm 0.18}$
LLaVA-v1.6	6.95$_{\pm 0.24}$	**80.33**$_{\pm 0.34}$	45.89$_{\pm 0.24}$	3.67$_{\pm 0.10}$	1.37$_{\pm 0.17}$	24.96$_{\pm 0.11}$
MiniGPT-v2	3.25$_{\pm 0.13}$	76.95$_{\pm 0.59}$	50.08$_{\pm 0.84}$	15.23$_{\pm 0.76}$	7.96$_{\pm 0.79}$	27.67$_{\pm 0.25}$
LLaVA-Med-v1	5.72$_{\pm 0.21}$	34.36$_{\pm 1.21}$	38.30$_{\pm 2.83}$	20.79$_{\pm 0.47}$	5.22$_{\pm 1.10}$	17.90$_{\pm 0.38}$
LLaVA-Med-v1.5	56.14$_{\pm 0.90}$	67.96$_{\pm 0.08}$	49.12$_{\pm 0.05}$	21.91$_{\pm 0.06}$	11.65$_{\pm 0.03}$	40.19$_{\pm 0.13}$
CheXagent	37.25$_{\pm 0.50}$	33.75$_{\pm 0.17}$	**73.31**$_{\pm 0.01}$	28.52$_{\pm 0.08}$	7.48$_{\pm 0.06}$	30.61$_{\pm 0.02}$
BiomedGPT	60.25$_{\pm 0.27}$	46.81$_{\pm 0.62}$	50.31$_{\pm 0.24}$	14.13$_{\pm 0.90}$	6.11$_{\pm 0.23}$	33.34$_{\pm 0.17}$
Med-Flamingo	44.38$_{\pm 0.20}$	62.02$_{\pm 0.54}$	50.00$_{\pm 0.01}$	26.17$_{\pm 0.13}$	5.72$_{\pm 0.06}$	35.66$_{\pm 0.14}$
RadFM	83.72$_{\pm 0.26}$	41.04$_{\pm 0.33}$	60.83$_{\pm 0.32}$	23.05$_{\pm 0.14}$	9.10$_{\pm 0.29}$	41.00$_{\pm 0.19}$
Gemini Pro	96.47$_{\pm 0.88}$	75.69$_{\pm 1.89}$	60.29$_{\pm 1.99}$	27.93$_{\pm 1.82}$	18.44$_{\pm 0.77}$	55.08$_{\pm 0.93}$
GPT-4V	92.51$_{\pm 1.10}$	71.73$_{\pm 2.45}$	53.30$_{\pm 1.90}$	**35.19**$_{\pm 1.16}$	22.40$_{\pm 1.89}$	55.28$_{\pm 0.98}$
GPT-4o	**97.03**$_{\pm 0.34}$	68.13$_{\pm 1.15}$	61.79$_{\pm 2.28}$	29.30$_{\pm 2.55}$	**24.06**$_{\pm 1.80}$	**55.60**$_{\pm 1.05}$

is available in Appendix A. Distribution plots of ground-truth answers and model responses within each question category are available in Appendix F.

Error Analysis in Procedural Diagnosis An error analysis was conducted on two top-performing models (GPT-4V and Gemini Pro) across three specialized diagnostic question types: Abnormality, Condition/Finding, and Position. The procedural diagnosis is conducted by conditioning performance on later and more challenging questions on models' success on simple prior ones. For example, model accuracy under the Abnormality category is calculated within those test samples where the model correctly identifies both imaging modality and organ. And the accuracy under the Condition/Finding category is further conditioned on those success samples from prior categories, including Abnormality.

As shown in Table 4, both models show vulnerabilities to hallucination errors, particularly in the later stages of the diagnostic procedure. For example, errors under the Abnormality arise either from incorrect responses or a tendency to over-reject challenging questions. In the Condition/Finding and Position categories, errors are largely due to the acceptance of hallucinated entities. Notably, Gemini Pro is more prone to accepting false conditions and positions, which significantly lowers its strict accuracy in these areas.

Table 4: Error Analysis of GPT-4V and Gemini Pro on ProbMed. The table shows the accuracy and types of errors for three specialized question types: Abnormality, Condition/Finding, and Position. Errors are categorized into wrong answers, rejection to answer, denying ground truth, and accepting hallucinations, providing a detailed breakdown of model performance and failure modes.

Question Type	Accuracy and Error Type	Models	
		GPT-4V	Gemini Pro
Abnormality	Accuracy	66.06	**67.05**
	E_wrong_answer	67.47	100.00
	E_reject_to_answer	32.52	0.00
Condition/Finding	Accuracy	36.90	**39.97**
	E_deny_ground-truth	51.69	39.04
	E_accept_hallucination	42.12	59.69
	E_reject_to_answer	6.18	1.26
Position	Accuracy	**39.97**	26.40
	E_deny_ground-truth	39.04	23.31
	E_accept_hallucination	59.69	76.68
	E_reject_to_answer	1.26	0.00

4.3 Exploring Model Limitations and Potential Improvements

Impact of Chain-of-Thought Prompting and Visual Understanding To further investigate the underperformance of open-source models, we conducted an extensive ablation study on LLaVA-v1, LLaVA-v1.6, LLaVA-Med-v1, LLaVA-Med-v1.5, Med-Flamingo, and GPT-4o. In this study, we examined two additional experimental settings: (1) applying a Chain-of-Thought (CoT) approach where models first generate visual descriptions from the image on their own, then utilize the description to augment the question, (2) enhancing the models by providing external visual descriptions generated by GPT-4o in addition to the ques-

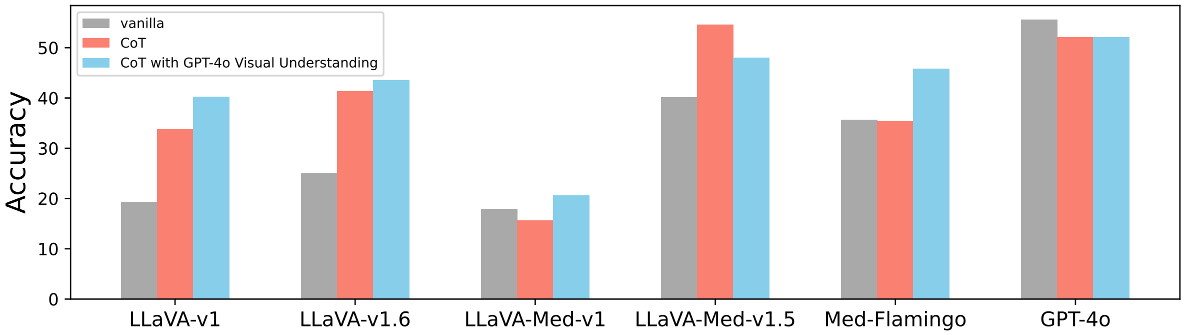

Figure 4: Accuracy comparison of LLaVA-v1, LLaVA-v1.6, LLaVA-Med-v1, LLaVA-Med-v1.5, Med-Flamingo, and GPT-4o under three different settings: vanilla (baseline performance), chain-of-thought (CoT) reasoning, and CoT with GPT-4o-generated visual descriptions. All models demonstrate significant performance improvement when visual descriptions from GPT-4o are included, indicating that poor visual understanding is a key factor limiting baseline performance. Chain-of-thought reasoning alone also leads to notable gains in accuracy, particularly in general-purpose models.

tion.

As shown in Figure 4, employing the chain-of-thought approach alone—without external visual descriptions—resulted in an average accuracy increase of 6.51%. In particular, LLaVA-Med-v1.5's accuracy improved from 40.19% to 54.55%, closing the gap to within 1.05% of the vanilla GPT-4o model. Interestingly, GPT-4o's performance decreased by 3.55% when the CoT mechanism was applied, potentially indicating that the model already internally employs its own chain-of-thought process.

Notably, all open-source models exhibited improved performance when augmented with visual descriptions generated by GPT-4o, suggesting that their baseline limitations stem primarily from poor visual comprehension. On average, these models showed an accuracy improvement of 9.44% across all question categories. This observation suggests that poor visual understanding is a major limitation of existing models, and augmenting them with external visual reasoning can lead to notable gains. Detailed performance changes of each model, organized by question category, can be found in Appendix D.

Transferability of Domain Expertise As shown in Table 14, CheXagent, a model trained exclusively on chest X-ray images, performs best in detecting abnormalities and identifying conditions/findings among all models when tested on chest X-ray images. We conducted a finer-grained analysis to explore whether the model's expertise in identifying features of a particular organ can be transferred to other imaging modalities.

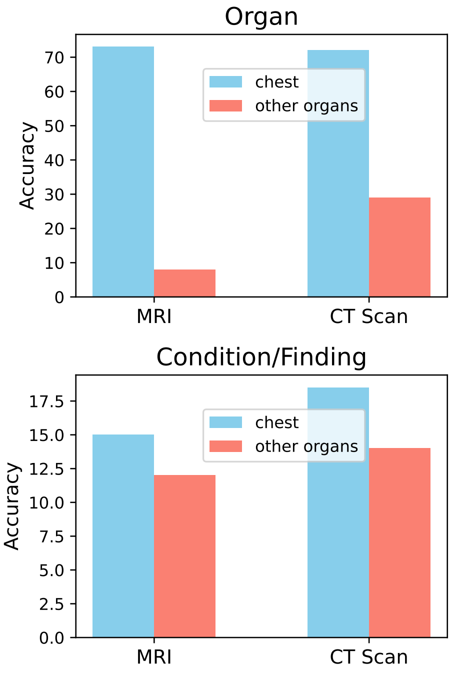

Figure 5: Accuracy comparison of CheXagent in identifying organs and conditions/findings across different modalities. The model demonstrates significantly higher accuracy in identifying organs on chest images compared to images of other organs for both MRI and CT scans. Additionally, CheXagent shows improved accuracy in identifying conditions/findings on chest images, indicating the transferability of its specialized knowledge from chest X-ray training to other imaging modalities.

As illustrated in Figure 5, CheXagent achieves significantly higher accuracy in identifying chest-related features compared to other organs, confirming our assumption that the model's pre-training on chest X-rays enhances its performance on recognizing chest images across different modalities. Interestingly, CheXagent also demonstrated higher accuracy in identifying conditions and findings in CT scans and MRIs of the chest, achieving a 3% increase in accuracy for MRIs and a 4% increase for CT scans compared with other organs within the same unseen modality. This indicates that specialized knowledge gained on chest X-rays can be transferred to other imaging modalities of the same organ in a zero-shot manner, highlighting the potential for cross-modality expertise transfer in real-life medical imaging diagnostics.

5 Conclusion

Evaluating the reliability of LMMs in the medical domain requires robust methods, and ProbMed, our newly introduced framework, addresses this by incorporating probing evaluation and procedural diagnosis. Our study reveals significant limitations in models like GPT-4o and Gemini Pro, which perform worse than random guessing on specialized diagnostic questions, while CheXagent's results highlight the critical importance of domain-specific knowledge. Furthermore, our additional experiments, which introduced Chain-of-Thought reasoning with external visual descriptions generated by GPT-4o, suggested that poor visual understanding is a major limitation of existing models and augmenting them with external visual reasoning can lead to notable gains.

6 Limitations

Despite the contributions, limitations such as the imbalanced image distribution favoring Chest X-rays (see Table 1) and the absence of open-ended evaluations, such as report generation, remain. The broader impact of our work includes the potential for improved diagnostic accuracy and better patient care, but it also highlights the risks of deploying unreliable models in healthcare. We recommend rigorous testing, continuous performance monitoring, and the incorporation of domain-specific expertise to mitigate these risks. Ultimately, our work aims to contribute to the development of trustworthy AI systems in healthcare, advancing diagnostic outcomes and patient safety.

Acknowledgments

This research project has benefited from the Microsoft Accelerate Foundation Models Research (AFMR) grant program.

References

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

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. 2024. Training small multimodal models to bridge biomedical competency gap: A case study in radiology imaging. *CoRR*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *ArXiv*, abs/2310.09478.

Qiuhui Chen and Yi Hong. 2024. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. In *Proceedings of the Asian Conference on Computer Vision*, pages 2404–2420.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. 2024. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Preprint*, arXiv:2310.05694.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *Preprint*, arXiv:2003.10286.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jason Lau, Sagnik Gayen, Asma Ben Abacha, et al. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *Preprint*, arXiv:2102.09542.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Zhengliang Liu, Hanqi Jiang, Tianyang Zhong, Zihao Wu, Chong Ma, Yiwei Li, Xiaowei Yu, Yutong Zhang, Yi Pan, Peng Shu, Yanjun Lyu, Lu Zhang, Junjie Yao, Peixin Dong, Chao Cao, Zhenxiang Xiao, Jiaqi Wang, Huan Zhao, Shaochen Xu, Yaonai Wei, Jingyuan Chen, Haixing Dai, Peilong Wang, Hao He, Zewei Wang, Xinyu Wang, Xu Zhang, Lin Zhao, Yiheng Liu, Kai Zhang, Liheng Yan, Lichao Sun, Jun Liu, Ning Qiang, Bao Ge, Xiaoyan Cai, Shijie Zhao, Xintao Hu, Yixuan Yuan, Gang Li, Shu Zhang, Xin Zhang, Xi Jiang, Tuo Zhang, Dinggang Shen, Quanzheng Li, Wei Liu, Xiang Li, Dajiang Zhu, and Tianming Liu. 2023b. Holistic evaluation of gpt-4v for biomedical imaging. *Preprint*, arXiv:2312.05256.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

OpenAI. 2023. Gpt-4v(ision) technical work and authors. *Technical report.*

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. 2023. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Ben jamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem W. Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Oscar Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomás Kociský, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, J Christopher Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Ying-Qi Miao, Lukás Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontan'on, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu, YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy, Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayana Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, A.E. Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Venkatesh Ramasesh, Anton Briukhov, Da-Woon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matt Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le

Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara N. Sainath, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela de Castro Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rrustemi, Elena Gribovskaya, Aurko Roy, Wiktor Gworek, S'ebastien M. R. Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Joshua Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost R. van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya B Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xiance Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, S'ebastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Michael B. Chang, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Rosca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravichandra Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Luvci'c, Rajkumar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjosund, Elena Buchatskaya, Elspeth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos L. Araya, Nikhil Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Zhufeng Pan, Zachary Nado, Stephanie Winkler, Dian Yu, Mohammad Saleh, Lorenzo Maggiore, Iain Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Chung-Cheng Chiu, Zoe Ashwood, Khuslen Baatarsukh, Sina Samangooei, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruibo Liu, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxi aoyu Feng, Matthew Mauger, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabriel Barth-Maron, Craig Swanson, Dominika Rogozi'nska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Ren shen Wang, Dave Lacey, Anastasija Ili'c, Yao Zhao, Woohyun Han, Lora Aroyo, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Raphael Lopez Kaufman, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David Reitter, Kingshuk Dasgupta, Shourya Sarcar, T. Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anais White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Dangyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Kiran Vodrahalli, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mahmoud Alnahlawi, Christopher Yew, Priya Ponnapalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, cCauglar Unlu, David Reid, Zora Tung, Daniel F. Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Rui Zhu, Ricardo Aguilar, Mai Gim'enez, Jiawei Xia, Olivier Dousse, Willi Gierke, Soheil Hassas Yeganeh, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexander Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Albert Cui, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Daniel Niels Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Daniel Toyama, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nicholas Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, Donghyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alexey Yakubovich, Nilesh Tripuraneni, James Manyika, Ha roon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Anna Bulanova, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Clément Farabet, Pedro Valenzuela, Quan Yuan, Christoper A. Welty, Ananth Agarwal, Mianna Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Elnaz Davoodi, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Rebeca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert, Clara Huiyi Hu, A. Ya. Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Ku-

mar, Alejandro Lince, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecnikowski, Jiří ima, Anna Koop, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Kalpesh Krishna, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Adrian Goedeckemeyer, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas Fitzgerald, Aviral Kumar, Sanjay Ghemawat, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Repina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Ilia Shumailov, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yuchung Cheng, Tolga Bolukbasi, Kather ine Lee, Saaber Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel El Kaed, Jing Li, Jakub Sygnowski, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Põder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Junwen Bai, Slav Petrov, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, Oriol Vinyals, and Alexandra Chronopoulou. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530.

Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. Medicat: A dataset of medical images, captions, and textual references. *ArXiv*, abs/2010.06000.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. In *Annual Meeting of the Association for Computational Linguistics*.

Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Can gpt-4v(ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *Preprint*, arXiv:2310.09909.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023b. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Preprint*, arXiv:2308.02463.

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

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

Kai Zhang, Jun Yu, Eashan Adhikarla, Rong Zhou, Zhiling Yan, Yixin Liu, Zhengliang Liu, Lifang He, Brian Davison, Xiang Li, Hui Ren, Sunyang Fu, James Zou, Wei Liu, Jing Huang, Chen Chen, Yuyin Zhou, Tianming Liu, Xun Chen, Yong Chen, Quanzheng Li, Hongfang Liu, and Lichao Sun. 2024. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *Preprint*, arXiv:2305.17100.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *Preprint*, arXiv:2305.10415.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. *Preprint*, arXiv:2309.03882.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. 2024. A survey of large language models in medicine: Progress, application, and challenge. *Preprint*, arXiv:2311.05112.

Yongshuo Zong, Tingyang Yu, Ruchika Chavhan, Bingchen Zhao, and Timothy Hospedales. 2024. Fool your (Vision and) language model with embarrassingly simple permutations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 62892–62913. PMLR.

A Breakdown Results on Different Image Modality and Organ.

This section shows a detailed model performance breakdown by each (imaging modality, organ) pair under five categories.

A.1 Brain CT Scan

Table 5: Results of different models on Brain CT scan in ProbMed. The best-performing model in each question category is **in-bold**, and the second best is underlined.

		General Question		Specialized Question		
		Modality	Organ	Abnormality	Condition/Finding	Position
Random Choice	Acc. with adv. pairs	25	25	50	35.28	35.01
LLaVA-v1	Acc. with adv. pairs	25.18	52.59	50	0	0
	Avg. acc.	62.59	72.22	/	46.57	49.60
LLaVA-v1.6	Acc. with adv. pairs	10.74	72.22	23.52	0	0.52
	Avg. acc.	55.37	84.44	/	30.79	41.91
MiniGPT-v2	Acc. with adv. pairs	1.11	**92.59**	50	17.77	8.42
	Avg. acc.	50.55	96.29	/	51.20	54.25
LLaVA-Med-v1	Acc. with adv. pairs	4.81	10.74	8.82	11.85	3.15
	Avg. acc.	50.18	33.88	/	40.71	49.78
LLaVA-Med-v1.5	Acc. with adv. pairs	50.37	80.37	44.11	11.85	15.26
	Avg. acc.	74.81	89.62	/	52.98	54.83
BiomedGPT	Acc. with adv. pairs	24.44	5.18	58.82	14.44	2.63
	Avg. acc.	62.03	52.59	/	53.88	35.84
Med-Flamingo	Acc. with adv. pairs	3.70	9.62	50	18.14	5.26
	Avg. acc.	51.85	47.03	/	50.16	47.85
CheXagent	Acc. with adv. pairs	11.85	0	47.05	12.96	5.26
	Avg. acc.	40.55	23.88	/	53.00	51.46
GPT-4o	Acc. with adv. pairs	**94.81**	**93.70**	61.76	35.92	26.31
	Avg. acc.	97.22	96.66	/	68.76	64.83
GPT-4V	Acc. with adv. pairs	94.07	84.07	61.76	**37.03**	**31.05**
	Avg. acc.	96.85	91.48	/	67.01	65.00
Gemini Pro	Acc. with adv. pairs	84.44	85.18	**70.58**	34.81	21.05
	Avg. acc.	92.03	92.40	/	68.01	60.16
	num	270	270	34	270	270

A.2 Chest CT Scan

Table 6: Results of different models on Chest CT Scan in ProbMed. The best-performing model in each question category is **in-bold**, and the second best is underlined.

		General Question		Specialized Question		
		Modality	Organ	Abnormality	Condition/Finding	Position
Random Choice	Acc. with adv. pairs	25	25	50	32.69	33.76
LLaVA-v1	Acc. with adv. pairs	27.55	46.35	50	0.36	0.23
	Avg. acc.	63.77	73.08	/	48.54	50.11
LLaVA-v1.6	Acc. with adv. pairs	2.73	**76.82**	50	0.54	0.46
	Avg. acc.	51.18	86.58	/	41.42	45.75
MiniGPT-v2	Acc. with adv. pairs	0.54	53.28	50	10.21	3.22
	Avg. acc.	50.27	75.82	/	51.11	51.49
LLaVA-Med-v1	Acc. with adv. pairs	5.47	39.78	29.41	14.41	4.37
	Avg. acc.	51.18	68.06	/	45.50	51.72
LLaVA-Med-v1.5	Acc. with adv. pairs	51.09	61.86	41.17	14.78	9.21
	Avg. acc.	75.54	80.10	/	52.60	54.64
BiomedGPT	Acc. with adv. pairs	15.51	2.91	52.94	7.11	2.30
	Avg. acc.	56.93	50.63	/	50.93	34.65
Med-Flamingo	Acc. with adv. pairs	22.26	70.98	50	19.16	7.14
	Avg. acc.	60.31	85.49	/	51.11	48.89
CheXagent	Acc. with adv. pairs	6.75	72.99	50	18.61	7.83
	Avg. acc.	32.93	86.49	/	56.80	51.55
GPT-4o	Acc. with adv. pairs	**97.62**	65.99	67.64	27.60	19.58
	Avg. acc.	98.72	81.90	/	63.54	61.67
GPT-4V	Acc. with adv. pairs	97.07	72.94	67.64	32.9	**20.78**
	Avg. acc.	98.44	85.74	/	65.01	59.54
Gemini Pro	Acc. with adv. pairs	95.62	58.21	**82.35**	**34.48**	14.28
	Avg. acc.	97.71	78.37	/	65.62	56.84
	num	548	548	34	548	548

A.3 Spine CT Scan

Table 7: Results of different models on Spine CT Scan in ProbMed. The best-performing model in each question category is **in-bold**, and the second best is underlined.

		General Question		Specialized Question		
		Modality	Organ	Abnormality	Condition/Finding	Position
Random Choice	Acc. with adv. pairs	25	25	50	30.85	31.06
LLaVA-v1	Acc. with adv. pairs	22.98	44.82	50	0	0
	Avg. acc.	61.49	70.68	/	49.47	50.00
LLaVA-v1.6	Acc. with adv. pairs	4.59	72.41	0	0	1.28
	Avg. acc.	52.29	83.90	/	37.66	41.07
MiniGPT-v2	Acc. with adv. pairs	1.14	41.37	0	12.64	5.12
	Avg. acc.	50.57	58.62	/	54.41	51.21
LLaVA-Med-v1	Acc. with adv. pairs	2.29	11.49	50	11.49	6.41
	Avg. acc.	48.27	30.45	/	46.37	48.77
LLaVA-Med-v1.5	Acc. with adv. pairs	32.18	67.81	50.0	9.19	14.10
	Avg. acc.	65.51	83.33	/	55.23	51.27
BiomedGPT	Acc. with adv. pairs	28.73	8.04	0	6.89	2.56
	Avg. acc.	63.79	53.44	/	50.00	33.27
Med-Flamingo	Acc. with adv. pairs	6.89	39.08	50	14.94	8.97
	Avg. acc.	53.44	68.39	/	53.92	52.22
CheXagent	Acc. with adv. pairs	4.59	27.58	50	10.34	2.56
	Avg. acc.	34.48	58.04	/	49.45	50.20
GPT-4o	Acc. with adv. pairs	**87.35**	76.74	0	30.23	20.77
	Avg. acc.	93.10	88.37	/	66.01	60.08
GPT-4V	Acc. with adv. pairs	81.39	69.76	0	**33.73**	**25.97**
	Avg. acc.	89.53	84.30	/	65.77	63.13
Gemini Pro	Acc. with adv. pairs	87.2	**77.9**	50	22.09	**25.97**
	Avg. acc.	92.44	88.95	/	61.64	64.94
	num	86	86	2	86	86

A.4 Abdominal CT Scan

Table 8: Results of different models on Abdominal CT Scan in ProbMed. The best-performing model in each question category is **in-bold**, and the second best is underlined.

		General Question		Specialized Question		
		Modality	Organ	Abnormality	Condition/Finding	Position
Random Choice	Acc. with adv. pairs	25	25	50	35.53	37.03
LLaVA-v1	Acc. with adv. pairs	26.49	54.19	50	0.53	0
	Avg. acc.	63.24	77.09	/	47.70	50.00
LLaVA-v1.6	Acc. with adv. pairs	1.86	**82.82**	41.42	1.06	0.66
	Avg. acc.	50.93	91.07	/	38.36	45.82
MiniGPT-v2	Acc. with adv. pairs	0	37.15	48.57	6.12	2.14
	Avg. acc.	50.00	66.97	/	48.49	50.22
LLaVA-Med-v1	Acc. with adv. pairs	5.05	45	30	15.44	5.28
	Avg. acc.	51.53	70.90	/	45.13	49.24
LLaVA-Med-v1.5	Acc. with adv. pairs	51.93	67.64	48.57	11.31	16.03
	Avg. acc.	75.96	83.42	/	52.86	65.61
BiomedGPT	Acc. with adv. pairs	67.77	12.38	57.14	15.31	4.62
	Avg. acc.	83.75	55.52	/	54.49	45.06
Med-Flamingo	Acc. with adv. pairs	1.73	35.55	50	20.37	8.26
	Avg. acc.	50.86	67.57	/	51.03	49.46
CheXagent	Acc. with adv. pairs	25.03	38.21	52.85	15.57	6.61
	Avg. acc.	51.46	65.71	/	51.08	50.19
GPT-4o	Acc. with adv. pairs	97.99	65.28	51.42	23.12	**28.23**
	Avg. acc.	98.93	81.50	/	58.24	64.59
GPT-4V	Acc. with adv. pairs	95.72	72.72	45.71	27	23.25
	Avg. acc.	97.72	85.56	/	58.92	60.02
Gemini Pro	Acc. with adv. pairs	**98.31**	69.19	**65.71**	**28.79**	20.39
	Avg. acc.	99.00	84.20	/	61.03	59.27
	num	750	750	70	750	750

A.5 Brain MRI

Table 9: Results of different models on Brain MRI in ProbMed. The best-performing model in each question category is **in-bold**, and the second best is underlined.

		General Question		Specialized Question		
		Modality	Organ	Abnormality	Condition/Finding	Position
Random Choice	Acc. with adv. pairs	25	25	50	36.7	36.64
LLaVA-v1	Acc. with adv. pairs	1.23	32.86	50	0.53	0
	Avg. acc.	49.29	65.37	/	47.85	49.87
LLaVA-v1.6	Acc. with adv. pairs	17.49	88.51	28.57	0.53	0.48
	Avg. acc.	58.74	93.10	/	31.73	37.46
MiniGPT-v2	Acc. with adv. pairs	1.94	96.64	50	15.72	4.37
	Avg. acc.	50.88	98.32	/	52.16	50.51
LLaVA-Med-v1	Acc. with adv. pairs	3	8.12	23.21	14.66	2.91
	Avg. acc.	47.08	32.50	/	47.72	48.35
LLaVA-Med-v1.5	Acc. with adv. pairs	75.61	84.98	42.85	13.78	13.62
	Avg. acc.	87.80	92.40	/	53.37	53.52
BiomedGPT	Acc. with adv. pairs	15.37	12.36	44.64	11.48	2.67
	Avg. acc.	54.41	56.00	/	51.26	42.06
Med-Flamingo	Acc. with adv. pairs	0.35	13.60	50	10.77	3.16
	Avg. acc.	47.61	51.32	/	48.27	50.01
CheXagent	Acc. with adv. pairs	0	0	50	10.77	6.81
	Avg. acc.	20.40	21.99	/	50.37	51.87
GPT-4o	Acc. with adv. pairs	**97.69**	**97.34**	66.07	25.84	**30.24**
	Avg. acc.	98.58	98.67	/	61.05	66.13
GPT-4V	Acc. with adv. pairs	96.99	94.33	58.92	**36.1**	27.8
	Avg. acc.	98.40	97.07	/	65.89	62.38
Gemini Pro	Acc. with adv. pairs	95.22	94.87	**78.57**	35.51	19.7
	Avg. acc.	97.26	97.34	/	65.59	59.67
	num	566	566	56	566	566

A.6 Chest MRI

Table 10: Results of different models on Chest MRI in ProbMed. The best-performing model in each question category is **in-bold**, and the second best is underlined.

		General Question		Specialized Question		
		Modality	Organ	Abnormality	Condition/Finding	Position
Random Choice	Acc. with adv. pairs	25	25	50	34.18	34.11
LLaVA-v1	Acc. with adv. pairs	0	35	50	0	0
	Avg. acc.	41.25	66.25	/	45.00	50.00
LLaVA-v1.6	Acc. with adv. pairs	5	32.5	37.5	0	0
	Avg. acc.	51.24	56.25	/	31.35	43.01
MiniGPT-v2	Acc. with adv. pairs	0	35	50	10	8.82
	Avg. acc.	47.50	62.50	/	47.91	49.50
LLaVA-Med-v1	Acc. with adv. pairs	5	45	12.5	12.5	5.88
	Avg. acc.	43.75	68.75	/	49.06	46.32
LLaVA-Med-v1.5	Acc. with adv. pairs	50.00	35.00	50.00	12.5	11.76
	Avg. acc.	72.5	62.5	/	53.75	53.92
BiomedGPT	Acc. with adv. pairs	0.00	5.00	50.00	10.00	2.94
	Avg. acc.	40.00	51.24	/	51.04	49.01
Med-Flamingo	Acc. with adv. pairs	2.50	45.00	50	10.00	8.82
	Avg. acc.	43.75	72.50	/	48.75	47.79
CheXagent	Acc. with adv. pairs	0	**75**	50	15	0
	Avg. acc.	17.50	87.50	/	44.58	47.05
GPT-4o	Acc. with adv. pairs	**90.00**	35.89	**62.50**	17.94	**24.24**
	Avg. acc.	93.75	65.38	/	54.80	61.36
GPT-4V	Acc. with adv. pairs	76.92	51.28	37.5	**25.64**	18.18
	Avg. acc.	86.25	71.79	/	58.11	61.74
Gemini Pro	Acc. with adv. pairs	87.5	**62.5**	37.5	17.5	11.76
	Avg. acc.	91.25	77.50	/	54.89	56.86
	num	40	40	8	40	40

A.7 Spine MRI

Table 11: Results of different models on Spine MRI in ProbMed. The best-performing model in each question category is **in-bold**, and the second best is underlined.

		General Question		Specialized Question		
		Modality	Organ	Abnormality	Condition/Finding	Position
Random Choice	Acc. with adv. pairs	25	25	50	31.51	31.52
LLaVA-v1	Acc. with adv. pairs	0	32.09	50	50	0.3
	Avg. acc.	49.22	65.58	/	47.15	49.88
LLaVA-v1.6	Acc. with adv. pairs	3.08	86.72	27.77	0.30	0
	Avg. acc.	51.54	92.74	/	30.59	34.37
MiniGPT-v2	Acc. with adv. pairs	0.3	49.69	52.77	6.79	2.02
	Avg. acc.	50.15	70.52	/	48.97	50.01
LLaVA-Med-v1	Acc. with adv. pairs	1.54	5.24	36.11	12.96	5.4
	Avg. acc.	45.06	24.07	/	48.52	48.04
LLaVA-Med-v1.5	Acc. with adv. pairs	70.67	84.56	50.00	11.11	11.14
	Avg. acc.	84.72	91.97	/	52.40	51.89
BiomedGPT	Acc. with adv. pairs	0.30	5.86	50.00	7.71	3.04
	Avg. acc.	45.06	52.77	/	51.31	44.04
Med-Flamingo	Acc. with adv. pairs	0.30	29.93	50 “	17.90	5.40
	Avg. acc.	50.00	64.50	/	50.54	50.14
CheXagent	Acc. with adv. pairs	0	13.58	47.22	15.43	2.7
	Avg. acc.	22.53	44.44	/	51.28	48.54
GPT-4o	Acc. with adv. pairs	**98.44**	84.52	**63.88**	19.50	**24.40**
	Avg. acc.	98.91	91.95	/	55.46	63.70
GPT-4V	Acc. with adv. pairs	96.28	**90.71**	55.55	22.6	15.59
	Avg. acc.	97.51	94.73	/	58.89	57.52
Gemini Pro	Acc. with adv. pairs	98.13	88.81	57.14	**24.53**	14.91
	Avg. acc.	98.75	94.09	/	59.19	58.20
	num	332	332	35	332	332

A.8 Abdominal MRI

Table 12: Results of different models on Abdominal MRI in ProbMed. The best-performing model in each question category is **in-bold**, and the second best is underlined.

		General Question		Specialized Question		
		Modality	Organ	Abnormality	Condition/Finding	Position
Random Choice	Acc. with adv. pairs	25	25	50	37.13	38.26
LLaVA-v1	Acc. with adv. pairs	0	39.28	50.00	2.38	0
	Avg. acc.	48.22	69.64	/	46.42	50.00
LLaVA-v1.6	Acc. with adv. pairs	2.38	73.8	35.71	1.19	0
	Avg. acc.	51.19	85.11	/	35.46	44.06
MiniGPT-v2	Acc. with adv. pairs	0	36.9	50	8.33	4.54
	Avg. acc.	50.00	67.26	/	47.51	51.70
LLaVA-Med-v1	Acc. with adv. pairs	2.38	47.61	50.00	14.28	9.09
	Avg. acc.	41.66	72.61	/	47.42	46.46
LLaVA-Med-v1.5	Acc. with adv. pairs	51.19	65.47	50.00	13.09	16.66
	Avg. acc.	75.59	81.54	/	54.31	56.37
BiomedGPT	Acc. with adv. pairs	1.19	3.57	50.00	14.28	1.51
	Avg. acc.	38.69	50.00	/	51.33	46.46
Med-Flamingo	Acc. with adv. pairs	2.38	27.38	50.00	20.23	3.03
	Avg. acc.	50.59	62.50	/	49.55	50.50
CheXagent	Acc. with adv. pairs	0	26.19	50.00	11.9	10.6
	Avg. acc.	19.04	56.54	/	49.20	49.62
GPT-4o	Acc. with adv. pairs	**91.66**	67.85	64.28	21.42	**39.39**
	Avg. acc.	95.83	81.54	/	55.30	70.51
GPT-4V	Acc. with adv. pairs	86.9	**75**	50	27.38	25.75
	Avg. acc.	92.26	85.71	/	58.58	58.77
Gemini Pro	Acc. with adv. pairs	89.28	72.61	**85.71**	**28.57**	25.75
	Avg. acc.	94.04	86.30	/	63.39	60.98
	num	84	84	14	84	84

A.9 Brain X-ray

Table 13: Results of different models on Brain X-ray in ProbMed. The best-performing model in each question category is **in-bold**, and the second best is underlined.

		General Question		Specialized Question		
		Modality	Organ	Abnormality	Condition/Finding	Position
Random Choice	Acc. with adv. pairs	25	25	50	44.77	47.08
LLaVA-v1	Acc. with adv. pairs	45.56	26.58	50	0	0
	Avg. acc.	72.78	51.89	/	48.10	50.00
LLaVA-v1.6	Acc. with adv. pairs	11.39	13.92	16.66	8.86	4.44
	Avg. acc.	55.06	48.10	/	45.04	48.88
MiniGPT-v2	Acc. with adv. pairs	18.98	**83.54**	50	18.98	17.77
	Avg. acc.	59.49	89.87	/	51.37	52.22
LLaVA-Med-v1	Acc. with adv. pairs	8.86	8.86	0	20.25	4.44
	Avg. acc.	54.43	31.01	/	51.16	48.33
LLaVA-Med-v1.5	Acc. with adv. pairs	49.36	31.64	50.00	8.86	13.33
	Avg. acc.	73.41	56.96	/	53.16	55.55
BiomedGPT	Acc. with adv. pairs	12.65	6.32	50.00	11.39	2.22
	Avg. acc.	53.16	49.36	/	52.95	43.33
Med-Flamingo	Acc. with adv. pairs	8.86	0	50	22.78	8.88
	Avg. acc.	54.43	15.18	/	50.73	48.33
CheXagent	Acc. with adv. pairs	84.81	0	50	12.65	8.88
	Avg. acc.	92.40	29.74	/	51.16	55.00
GPT-4o	Acc. with adv. pairs	**94.93**	52.56	**66.66**	37.17	**40.90**
	Avg. acc.	96.20	73.71	/	62.07	69.31
GPT-4V	Acc. with adv. pairs	82.05	8.97	33.33	**43.58**	22.72
	Avg. acc.	90.38	47.43	/	68.48	59.09
Gemini Pro	Acc. with adv. pairs	89.87	51.89	50	31.64	31.11
	Avg. acc.	93.03	74.05	/	61.81	63.88
	num	79	79	6	79	79

A.10 Chest X-ray

Table 14: Results of different models on Chest X-ray in ProbMed. The best-performing model in each question category is **in-bold**, and the second best is underlined.

		General Question		Specialized Question		
		Modality	Organ	Abnormality	Condition/Finding	Position
Random Choice	Acc. with adv. pairs	25	25	50	37.59	37.08
LLaVA-v1	Acc. with adv. pairs	28.75	36.57	50	0.12	0.11
	Avg. acc.	64.37	68.25	/	34.41	50.05
LLaVA-v1.6	Acc. with adv. pairs	7.11	83.97	47.94	5.89	1.52
	Avg. acc.	53.49	91.61	/	34.52	48.85
MiniGPT-v2	Acc. with adv. pairs	4.93	**94.07**	50.05	18.78	11.94
	Avg. acc.	52.46	96.98	/	46.09	53.15
LLaVA-Med-v1	Acc. with adv. pairs	6.25	39.77	40.24	26.28	6.14
	Avg. acc.	52.62	67.19	/	50.78	51.34
LLaVA-Med-v1.5	Acc. with adv. pairs	55.44	65.48	49.53	31.82	9.78
	Avg. acc.	77.67	82.69	/	62.70	54.22
BiomedGPT	Acc. with adv. pairs	91.34	86.05	50.00	16.92	9.08
	Avg. acc.	95.46	92.93	/	43.00	41.46
Med-Flamingo	Acc. with adv. pairs	80.92	90.00	50	35.83	5.24
	Avg. acc.	90.46	95.00	/	63.47	48.00
CheXagent	Acc. with adv. pairs	53.68	39.64	**76.59**	**42.75**	9.38
	Avg. acc.	76.84	69.82	/	70.80	54.00
GPT-4o	Acc. with adv. pairs	97.97	62.98	62.01	32.13	**21.81**
	Avg. acc.	98.81	81.39	/	59.35	59.95
GPT-4V	Acc. with adv. pairs	91.53	67.51	53.18	39.35	21.35
	Avg. acc.	95.62	83.37	/	64.69	55.64
Gemini Pro	Acc. with adv. pairs	**98.07**	76.74	61.29	25.83	15.31
	Avg. acc.	98.94	88.32	/	52.22	54.97
	num	3120	3120	1948	3120	3120

A.11 Spine X-ray

Table 15: Results of different models on Spine X-ray in ProbMed. The best-performing model in each question category is **in-bold**, and the second best is underlined.

		General Question		Specialized Question		
		Modality	Organ	Abnormality	Condition/Finding	Position
Random Choice	Acc. with adv. pairs	25	25	50	30.95	30.99
LLaVA-v1	Acc. with adv. pairs	44.55	45.04	50	0.49	0
	Avg. acc.	72.27	71.78	/	47.32	49.42
LLaVA-v1.6	Acc. with adv. pairs	4.45	**82.67**	33.33	1.48	0.57
	Avg. acc.	52.22	90.84	/	35.87	42.02
MiniGPT-v2	Acc. with adv. pairs	2.97	52.47	58.33	16.33	4.02
	Avg. acc.	51.48	71.78	/	53.84	51.07
LLaVA-Med-v1	Acc. with adv. pairs	8.41	7.92	33.33	17.82	5.74
	Avg. acc.	52.72	28.96	/	52.82	47.58
LLaVA-Med-v1.5	Acc. with adv. pairs	46.53	71.78	50.00	14.85	13.32
	Avg. acc.	73.01	85.89	/	55.78	54.79
BiomedGPT	Acc. with adv. pairs	40.09	16.83	58.33	12.37	2.87
	Avg. acc.	68.06	55.19	/	50.27	40.77
Med-Flamingo	Acc. with adv. pairs	14.35	25.24	50	14.85	5.17
	Avg. acc.	57.17	62.12	/	51.09	48.38
CheXagent	Acc. with adv. pairs	82.17	20.29	62.5	16.83	0.57
	Avg. acc.	91.08	50.74	/	52.70	48.70
GPT-4o	Acc. with adv. pairs	**95.54**	**79.70**	47.82	**34.15**	**25.86**
	Avg. acc.	97.02	89.60	/	68.99	66.03
GPT-4V	Acc. with adv. pairs	85.57	72.13	47.82	29.85	18.49
	Avg. acc.	92.03	85.32	/	65.20	57.18
Gemini Pro	Acc. with adv. pairs	95.02	70.14	**70.83**	17.91	19.07
	Avg. acc.	96.76	84.82	/	58.04	61.72
	num	201	201	24	201	201

A.12 Abdominal X-ray

Table 16: Results of different models on Abdominal X-ray in ProbMed. The best-performing model in each question category is **in-bold**, and the second best is underlined.

		General Question		Specialized Question		
		Modality	Organ	Abnormality	Condition/Finding	Position
Random Choice	Acc. with adv. pairs	25	25	50	36.55	37.46
LLaVA-v1	Acc. with adv. pairs	53.87	53.01	50	2.15	0.56
	Avg. acc.	76.93	76.50	/	49.14	50.00
LLaVA-v1.6	Acc. with adv. pairs	5.17	56.46	46	6.46	1.12
	Avg. acc.	52.15	75.64	/	47.63	48.16
MiniGPT-v2	Acc. with adv. pairs	4.74	38.79	50	18.53	5.64
	Avg. acc.	52.37	67.24	/	53.65	50.23
LLaVA-Med-v1	Acc. with adv. pairs	7.75	42.24	60	14.65	4.51
	Avg. acc.	53.23	68.96	/	47.47	50.87
LLaVA-Med-v1.5	Acc. with adv. pairs	52.58	50.86	50.00	6.46	14.68
	Avg. acc.	76.07	73.49	/	52.02	54.75
BiomedGPT	Acc. with adv. pairs	35.77	1.29	50.00	10.34	4.51
	Avg. acc.	65.30	37.50	/	52.94	46.79
Med-Flamingo	Acc. with adv. pairs	28.01	34.48	50	14.65	4.51
	Avg. acc.	64.00	66.37	/	52.52	46.25
CheXagent	Acc. with adv. pairs	77.15	23.70	70	12.93	2.25
	Avg. acc.	88.57	52.80	/	51.30	49.64
GPT-4o	Acc. with adv. pairs	**98.26**	61.47	70	27.27	21.46
	Avg. acc.	99.13	79.22	/	61.83	59.81
GPT-4V	Acc. with adv. pairs	84.84	50.21	60	**31.16**	**23.16**
	Avg. acc.	92.42	71.42	/	59.63	57.03
Gemini Pro	Acc. with adv. pairs	97.14	**63.36**	**85**	27.15	19.2
	Avg. acc.	98.70	80.81	/	59.97	58.80
	num	232	232	20	232	232

B Dataset Statistics

In this section, we detail the dataset statistics.

Table 17 shows the number of questions across each question type. For each image sample, ground-truth questions were created based on available metadata with "yes" answers. For each ground-truth question, we also created a corresponding adversarial question by selecting random adversarial entities and assigning "no" answers. For an image showing a normal organ without abnormality, since there is no ground-truth information on the existence of the condition and position, we only construct hallucinated questions for the condition/finding question type. For an image showing abnormality, the number of question pairs per category equals the number of existing conditions or positions.

Table 17: Number of questions across each question type for each image.

Question type	Image with Normal Organ	Image with Abnormality
Modality	2	2
Organ	2	2
Abnormality	1	1
Condition/Finding	1	2 x number of existing conditions
Position	0	2 x number of existing positions

C Robustness Study

We conducted an additional robustness analysis using GPT-4o—the best-performing model from our evaluations—on the subset of 100 images and 1,090 VQA pairs used in our expert validation. The following table presents results on randomly sampled test samples from the expert-validated subset (20%, 40%, 60%, 80%, and 100%) and compares accuracy across all samples versus manually validated ("valid") samples:

Table 18: Valid percentage and accuracy under all samples vs. valid samples on different test sample splits. N_i denotes the number of images in the split, and N denotes the number of corresponding question-answer pairs.

N_i, N	Valid QA (%)	Acc. all samples (%)	Acc. valid samples (%)
20, 218	100.00	52.63	52.63
40, 421	96.19	54.36	55.19
60, 596	96.81	55.32	55.25
80, 821	95.73	54.21	54.55
100, 1090	97.79	55.63	55.59
6303, 57132	/	55.60	/

From these results, we observe minimal differences between the accuracy of all samples versus the validated subset. This indicates our evaluation framework is highly robust to the approximately 6% error rate identified during pre-processing. Furthermore, the accuracy of GPT-4o on the full ProbMed dataset (55.60%) closely aligns with the performance on the subset (55.63%), confirming the generalizability of our robustness findings.

D Impact of Chain-of-Thought Prompts and Visual Descriptions on Model Performance

In this section, we illustrate model performance under different categories under the three prompting settings: Vanilla, Chain-of-Thought, and Chain-of-Thought with GPT-4o Visual Understanding.

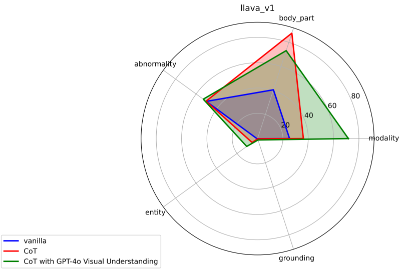

Figure 6: Accuracy of the LLaVA-v1 model across five diagnostic categories under three settings: vanilla (blue), chain-of-thought (CoT, red), and CoT with GPT-4o Visual Understanding (green).

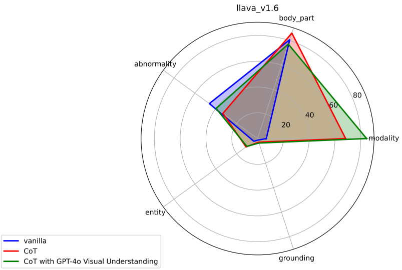

Figure 7: Accuracy of the LLaVA-v1.6 model across five diagnostic categories under three settings: vanilla (blue), chain-of-thought (CoT, red), and CoT with GPT-4o Visual Understanding (green).

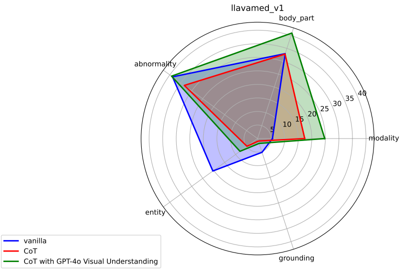

Figure 8: Accuracy of the LLaVA-Med-v1 model across five diagnostic categories under three settings: vanilla (blue), chain-of-thought (CoT, red), and CoT with GPT-4o Visual Understanding (green).

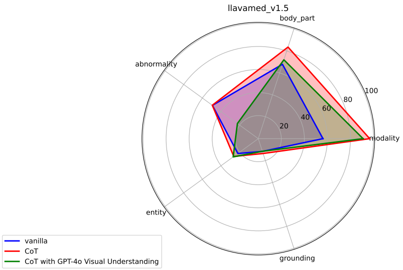

Figure 9: Accuracy of the LLaVA-Med-v1.5 model across five diagnostic categories under three settings: vanilla (blue), chain-of-thought (CoT, red), and CoT with GPT-4o Visual Understanding (green).

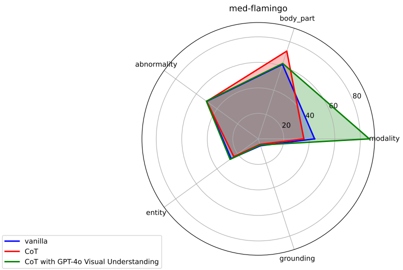

Figure 10: Accuracy of the Med-Flamingo model across five diagnostic categories under three settings: vanilla (blue), chain-of-thought (CoT, red), and CoT with GPT-4o Visual Understanding (green).

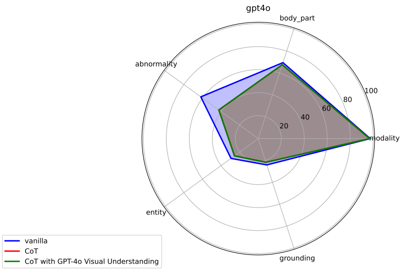

Figure 11: Accuracy of the GPT-4o model across five diagnostic categories under three settings: vanilla (blue), chain-of-thought (CoT, red), and CoT with GPT-4o Visual Understanding (green).

E Prompt Details

Below is the few-shot model prompt used for GPT-4 to extract medical conditions and their locations from image captions:

```
You are a helpful assistant and you are given a caption
      describing a medical image. Extract medical conditions
      and diseases, along with their locations, if specified.
       Do not include any information that cannot be directly
       inferred from the image, for example, patient status
      or patient history. Outputs should be in the format: "<
      condition/disease1> : <location1>, <condition/disease2>
       : <location2>...". The term "<location>" should
      include at least one positional descriptor and should
      be explicit in the original caption along with the
      condition/disease. Otherwise, it should be replaced
      with "None".

For example, consider the caption: "Fig. 1. MRI abdomen and
      pelvis showing the cervical mass." The output should be
       "<cervical mass> : None". For the caption: "Chest
      radiograph shows enlargement of the hilar mass with
      spread into the left lower lobe." The output should be
      "<enlargement of the hilar mass> : <left lower lobe>".
      Similarly, for the caption: "Abdominal CT scan reveals
      an enhancing rounded pseudo-aneurysm in the cystic
      artery, alongside high-density material within the
      gallbladder's lumen and near the gastrohepatic ligament
      ." The correct output is "<enhancing rounded pseudo-
      aneurysm> : <cystic artery>, <high-density material> :
      <lumen of the gallbladder and region of the
      gastrohepatic ligament>".

Make sure that the response contains only the information in
       the original caption without adding extra details.
```

F Response Distribution Visualization within each Category

Below, we illustrate the distribution of model responses within each category, valid "yes" responses are in red, and valid "no" responses are in blue.

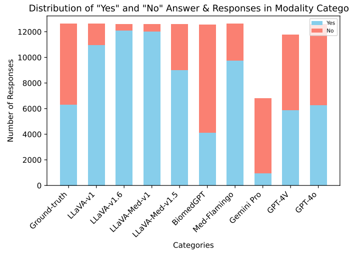

Figure 12: Distribution plot of "yes and "no" ground-truth answers and model responses within the Modality category.

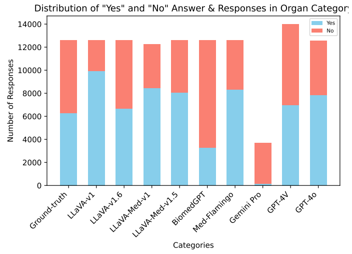

Figure 13: Distribution plot of "yes and "no" ground-truth answers and model responses within the Organ category.

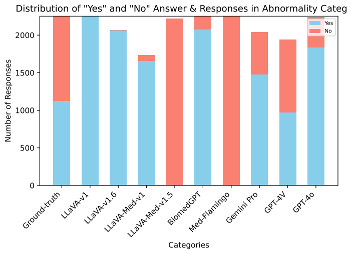

Figure 14: Distribution plot of "yes and "no" ground-truth answers and model responses within the Abnormality category.

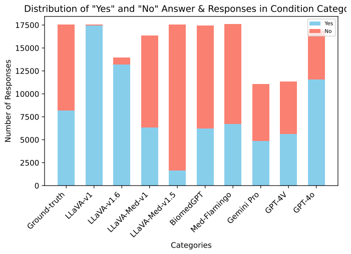

Figure 15: Distribution plot of "yes and "no" ground-truth answers and model responses within the Condition/Finding category.

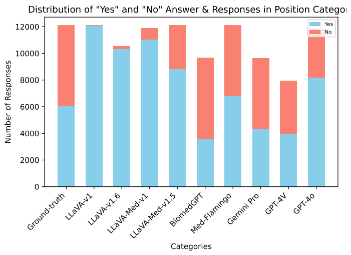

Figure 16: Distribution plot of "yes and "no" ground-truth answers and model responses within the Position category.