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

Qianqi Yan
University of California, Santa Cruz
qyan79@ucsc.edu

Xuehai He
University of California, Santa Cruz
xhe89@ucsc.edu

Xiang Yue
Carnegie Mellon University
xyue2@andrew.cmu.edu

Xin Eric Wang
University of California, Santa Cruz
xwang366@ucsc.edu

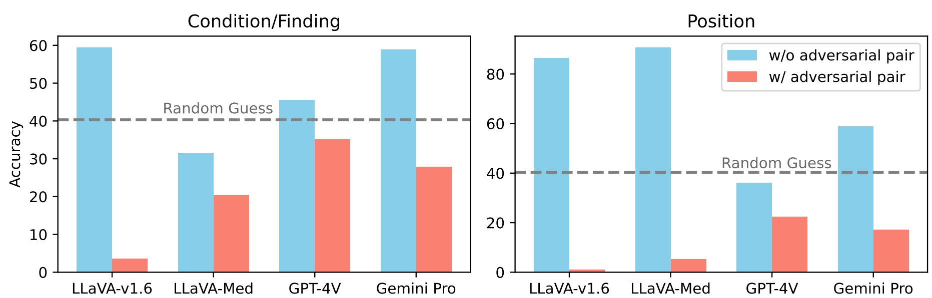

Figure 1: Accuracy of four LMMs on two types of specialized questions in medical diagnoses, with and without adversarial pairs. The significant drop in accuracy with adversarial pairs highlights the models' unreliability in handling medical diagnoses.

Abstract

Large Multimodal Models (LMMs) have shown remarkable progress in the field of medical Visual Question Answering (Med-VQA), achieving high accuracy on existing benchmarks. However, their reliability under robust evaluation is questionable. This study reveals that state-of-the-art models, when subjected to simple probing evaluation, perform worse than random guessing on medical diagnosis questions. To address this critical evaluation problem, we introduce the Probing Evaluation for Medical Diagnosis (ProbMed) dataset to rigorously assess LMM performance in medical imaging through *probing evaluation* and *procedural diagnosis*. Particularly, probing evaluation features pairing original questions with negation questions with hallucinated attributes, while procedural diagnosis requires reasoning across various diagnostic dimensions for each image, including modality recognition, organ identification, clinical findings, abnormalities, and positional grounding. Our evaluation reveals that top-performing models like GPT-4V and Gemini Pro perform worse than random guessing on specialized diagnostic questions, indicating significant limitations in handling fine-grained medical inquiries. Besides, models like LLaVA-Med struggle even with more general questions, and results from CheXagent demonstrate the transferability of expertise across different modalities of the same organ, showing that specialized domain knowledge is still crucial for improving performance. This study underscores the urgent need for more robust evaluation to ensure the reliability of LMMs in critical fields like medical diagnosis, and current LMMs are still far from applicable to those fields.
https://github.com/eric-ai-lab/ProbMed

Preprint. Under review.

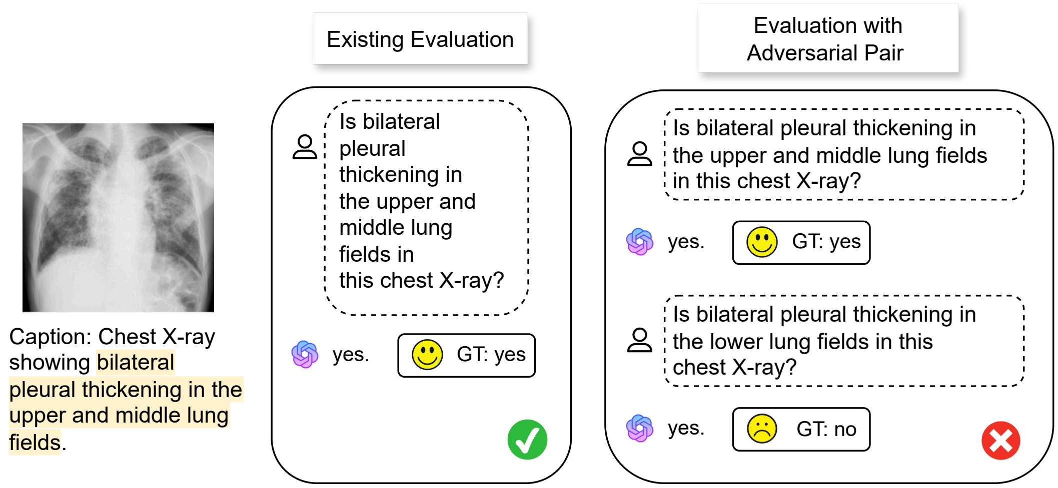

Figure 2: An example illustrating the potential for misleading accuracy in existing evaluations. While the model correctly identifies the position of an existing finding in the standard evaluation, it fails to differentiate between actual and hallucinated positions when subjected to an adversarial evaluation.

1 Introduction

Foundation models, such as large language models (LLMs) [OpenAI, 2023a, Touvron et al., 2023, Jiang et al., 2023, Anil et al., 2023, Chung et al., 2022] and large multimodal models (LMMs) [OpenAI, 2023b, Team et al., 2023, Li et al., 2023b, Liu et al., 2023a, Chen et al., 2023a], 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., 2023, Yang et al., 2023] and promising progress on existing medical 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 simple 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 performance drops observed are alarming: even advanced models like GPT-4V and Gemini Pro perform worse than random guessing, with an average decrease of 42% 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 seven state-of-the-art LMMs on ProbMed revealed several critical insights. *First*, even the best-performing models, such as GPT-4V 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's performance dropping by up to 57% and Gemini Pro's accuracy decreasing by 25% in ProbMed. These findings emphasize the importance of adversarial testing in Med-VQA to uncover model weaknesses. *Third*, 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 current 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. 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

2.1 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-4V [OpenAI, 2023b], Gemini Pro [Team et al., 2023], LLaVA [Liu et al., 2023a,b], and MiniGPT-v2 [Chen et al., 2023a]. 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 LLaVA-Med [Li et al., 2023a], MedBLIP [Chen et al., 2023b], and Med-Flamingo [Moor et al., 2023]. Despite the promising results from these domain-specific LMMs, there is ongoing exploration 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. 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.

2.2 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. ProbMed, as shown in Table 1, stands out by providing large-scale benchmarks and enabling categorical accuracy assessments across various diagnostic dimensions for each image, including modality recognition, organ recognition, clinical findings identification, and positional grounding. ProbMed uniquely incorporates adversarial negation pairs for each question-answer pair to ensure diagnostic specificity and reliability, setting it apart from existing benchmarks.

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

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

Table 1: Comparison ProbMed with a test set of existing medical VQA datasets, demonstrating our dataset's difference from existing benchmarks. For SLAKE, only the English subset is considered for head-to-head comparison with existing benchmarks.

Dataset	Images	Questions	Question Category	Procedural Diagnosis	Adversarial Pairs
VQA-RAD [Lau et al., 2018]	0.2k	0.4k	✓	✗	✗
SLAKE [Liu et al., 2021]	0.09k	1k	✓	✗	✗
PathVQA [He et al., 2020]	0.8k	6.7k	✗	✗	✗
PMC-VQA [Zhang et al., 2023]	50k	400k	✗	✗	✗
ProbMed (Ours)	6.3k	57k	✓	✓	✓

like ROUGE-L and BLEU-4. Recent works [Wang et al., 2024, Zheng et al., 2024] argue that multiple-choice questions may not be ideal due to inherent "selection bias." 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 ProbMed: Probing Evaluation for Medical Diagnosis

In this section, we design two evaluation principles and present a comprehensive analysis of state-of-the-art LMMs for Med-VQA using the created ProbMed dataset to address two 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?*

Our primary goal is to rigorously evaluate these models' readiness for real-life diagnostic tasks, particularly under adversarial conditions. Despite their high accuracy on existing benchmarks, the models struggle with simple probing evaluation. ProbMed is designed to expose these vulnerabilities and provide a more reliable assessment of model performance in real-world scenarios.

3.1 Probing Evaluation with Adversarial Pairs

One of the main motivations behind ProbMed is to assess the models' ability to accurately distinguish between relevant and irrelevant features. ProbMed pairs original questions with negation questions containing hallucinated attributes. This method challenges the model robustness by requiring them to identify true conditions while disregarding false, hallucinated ones. For instance, a question about a specific finding is paired with a negated question featuring a different, non-existent finding to test if the model can exclusively identify the true finding.

3.2 Procedural Diagnosis

To ensure a comprehensive evaluation, ProbMed includes questions that require reasoning across multiple diagnostic dimensions for each image. These dimensions include modality recognition, organ identification, clinical findings, abnormalities, and positional reasoning. This multifaceted approach assesses a model's diagnostic capabilities beyond simple question-answer pairs, requiring it to integrate various pieces of information to form a coherent diagnostic picture.

3.3 Data Curation

As illustrated in Figure 3, ProbMed draws from two comprehensive biomedical datasets MedICaT and ChestX-ray14 to compile a diverse set of 6,303 images. MedICaT [Subramanian et al., 2020] contains 217k image-caption pairs from 131k open access biomedical papers. From this dataset, we selected 4,543 image-caption pairs focusing on a single organ and modality, with clear indications

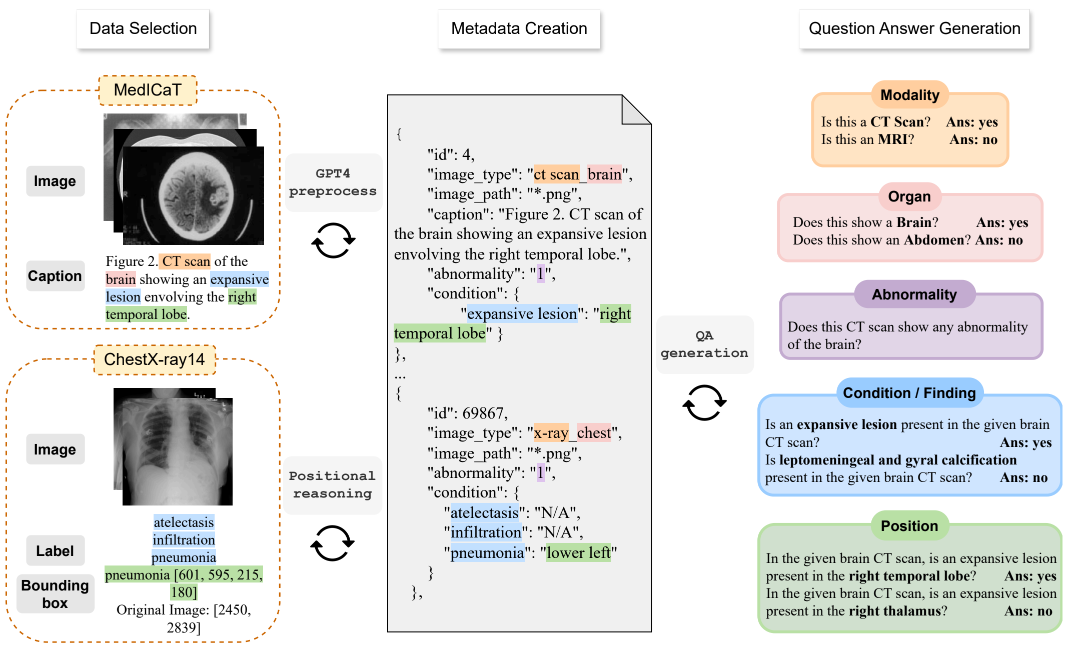

Figure 3: Flow diagram of the ProbMed data curation process. Two comprehensive biomedical datasets were utilized to collect source data and construct a metadata file, enabling the automatic generation of high-quality question-answer pairs for the ProbMed dataset.

of normal or abnormal conditions. These images span three modalities (X-ray, MRI, and CT scan) and four organs (abdomen, brain, chest, and spine). ChestX-ray14 [Wang et al., 2017] comprises 112k frontal-view X-ray images from 30k unique patients, including 880 images with abnormalities marked by bounding boxes. We selected 1,760 images, balanced between healthy and abnormal cases, with disease labels and bounding box annotations.

To create high-quality, balanced question-answer pairs, we generated metadata for each image. For MedICaT images, GPT-4 [OpenAI, 2023a] was used to analyze captions, identify abnormalities, and extract positional descriptions using few-shot prompting. For ChestX-ray14 images, a positional reasoning module generated textual descriptors of abnormalities based on bounding boxes and image sizes. The standardized metadata across both subsets included unique condition names and positional descriptions for each organ-modality combination as candidates for later constructing adversarial question-answers.

Using the metadata, we generated a diverse set of high-quality questions for each image, covering various diagnostic dimensions. 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. This process included selecting alternative organs, modalities, and hallucinated conditions/findings. As shown in Table 2, this process resulted in a total of 57,132 question-answer pairs, averaging 9 pairs per image.

4 Experimental Analysis

We conducted a systematic evaluation and comprehensive analysis using the ProbMed dataset on seven state-of-the-art LMMs to identify their strengths and weaknesses in real-life imaging diagnostics. Apart from proprietary GPT-4V [OpenAI, 2023b] and Gemini Pro [Team et al., 2023], we selected five open-source models (7B parameters version), spanning across general models including LLaVA-v1 [Liu et al., 2023b], LLaVA-v1.6 [Liu et al., 2023a], MiniGPT-v2 [Chen et al., 2023a] and specialized models including LLaVA-Med [Li et al., 2023a] 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. For categorical accuracy, we only considered a

Table 2: 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	/	/

hit if the model correctly answered all questions within the category for an image (see Table 19). This means the model must correctly identify all existing entities and disregard all hallucinated entities.

4.1 Is Current Evaluation of LMMs for Med-VQA Reliable?

To address this first research question, we introduced adversarial pairs to the evaluation process to test the model's robustness and reliability. This strategy ensures that models must validate the absence of certain characteristics or findings rather than simply acknowledge existing conditions, thereby enhancing diagnostic specificity and reliability. To demonstrate the necessity of adversarial pairs for achieving valid and trustworthy accuracy scores in Med-VQA, we conduct an experimental analysis on the test set of an existing medical dataset, VQA-RAD [Lau et al., 2018], in addition to ProbMed.

4.1.1 Probing Evaluation with Adversarial Pairs in VQA-RAD

To construct challenging adversarial questions for an given image, ideally, we need full control over the ground truth information and a set of confusing candidates, as provided in ProbMed. However, since VQA-RAD [Lau et al., 2018] provides finalized question-answer pairs without metadata, we could only construct adversarial pairs for 118 test instances where the answer is "yes" out of 272 closed-ended question-answers pairs within its test set. Each adversarial pair was manually created such that, based on the limited information from the original question-answer pair, the answer to the adversarial question had to be negated. This process resulted in 236 question-answer pairs in total. The adversarial questions in this subset are less challenging than those in ProbMed, as they often involve a simple semantic negation of the original question due to limited information.

These results, as shown in Table 3, reveal the significant impact of adversarial pairs on model performance. Although the original accuracy appears very high for some underperforming models, the accuracy drops drastically after balancing the subset with adversarial pairs: 6.78% for GPT-4V and 16.95% for Gemini Pro, with an average decrease of 42.7% across the tested models.

4.1.2 Probing Evaluation with Adversarial Pairs in ProbMed

Table 4 demonstrates the similar significant impact of adversarial pairs in ProbMed. The accuracy of more capable models is generally less affected by the introduction of challenging adversarial pairs. However, even the most robust models experience a minimum drop of 10.52% in accuracy when tested with ProbMed's challenging questions, with an average decrease of 44.7% across the tested models, highlighting the critical role of probing evaluation in evaluating Med-VQA performance comprehensively.

Table 3: Model accuracy on the VQA-RAD test subset before and after introducing adversarial pairs. The table demonstrates the significant drop in accuracy across various models when adversarial pairs are added, highlighting the models' vulnerabilities to adversarial questions. The percentage decrease in accuracy is noted in parentheses.

Models	Accuracy (%) without Adversarial Pairs	Accuracy (%) with Adversarial Pairs
LLaVA-v1-7B	96.61	25.42 (-71.19)
LLaVA-v1.6-7B	77.11	8.47 (-68.64)
LLaVA-Med	81.35	3.38 (-77.97)
MiniGPT-v2	72.03	46.61 (-25.42)
CheXagent	53.38	21.18 (-32.20)
GPT-4V	40.67	33.89 (-6.78)
Gemini Pro	61.86	44.91 (-16.95)

Table 4: The impact of adversarial pairs on model accuracy in ProbMed. This table illustrates the performance change of models with the introduction of adversarial pairs. The results indicate a substantial decline in accuracy, underlining the robustness of ProbMed in challenging LMMs. The percentage drop in accuracy is noted in parentheses.

Models	Accuracy (%) without Adversarial Pairs	Accuracy (%) with Adversarial Pairs
LLaVA-v1-7B	95.40	16.51 (-78.89)
LLaVA-v1.6-7B	83.79	21.97 (-61.82)
LLaVA-Med	71.91	14.77 (-57.14)
MiniGPT-v2	85.30	22.99 (-62.31)
CheXagent	39.55	22.18 (-17.37)
GPT-4V	66.27	55.75 (-10.52)
Gemini Pro	78.36	53.26 (-25.10)

4.2 How Reliable Are LMMs in Medical Diagnosis?

After correcting 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.

4.2.1 Performance across Diagnostic Questions

Table 5 shows the categorical accuracy of different models aggregated among all image types. While 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-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. CheXagent, trained exclusively on chest X-rays, achieved the highest accuracy in determining abnormalities and conditions. However, its performance in general tasks like identifying image modality and organs was lower, highlighting the need for domain-specific training. LLaVA-Med achieves much higher accuracy compared to both LLaVA-v1-7B and LLaVA-v1.6-7B in identifying conditions/finding and their positions but fell short on more general tasks like identifying image modality and organs as well as determining abnormality.

Among the open-sourced general-purpose models, MiniGPT-v2 performs the best, surpassing domain-specific LLaVA-Med in identifying organs and both LLaVA-Med and CheXagent in determining positions of condition/finding without domain-specific training. A more detailed breakdown of the

Table 5: Categorical and overall accuracy (%) of different models aggregated among all image types in ProbMed. The best result 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	25.00	25.00	50.00	35.67	36.48
LLaVA-v1-7B	25.28	40.53	50.00	0.34	0.10
LLaVA-v1.6-7B	6.77	**80.70**	46.18	3.57	1.07
LLaVA-Med	5.49	32.98	38.76	20.39	5.37
MiniGPT-v2	3.25	76.29	50.09	15.23	8.05
CheXagent	37.26	33.96	**73.32**	28.53	7.24
GPT-4V	92.50	71.71	53.30	**35.19**	**22.32**
Gemini Pro	**96.47**	75.69	62.59	27.92	17.17

performance of different models on different image types across each question type is available in Appendix A.

4.2.2 Error Analysis in Procedural Diagnosis

For models whose accuracy dropped drastically after introducing adversarial pairs, we observed a consistent accuracy pattern much lower than random guess performance for specialized questions. An error analysis focusing on GPT-4V and Gemini Pro across three specialized question types - Abnormality, Condition/Finding, and Position is further conducted. Each accuracy measurement is conditional on the model successfully answering the preceding diagnostic questions, reflecting a procedural diagnosis approach. This analysis reveals both models' vulnerabilities to hallucination errors, particularly as they progress through the diagnostic procedure, with Gemini Pro being more prone to accepting false conditions and positions.

As shown in Table 6, for Abnormality questions, conditioned on correctly identifying both image modality and organ, GPT-4V's errors arise from both incorrect answers and its tendency to reject challenging questions, while Gemini Pro attained a slightly higher accuracy of 67.05%, with all errors resulting from incorrect answers.

More specialized questions in identifying conditions and their positions, conditioned on successful abnormality detection, reveal both models' vulnerabilities to hallucination errors, particularly as they progress through the diagnostic procedure, with Gemini Pro more prone to accepting false conditions and positions. For questions on condition/finding, GPT-4V's accuracy dropped to 36.9%, with roughly even error distribution between denying ground-truth conditions and accepting hallucinated conditions, while most of the errors of Gemini Pro were from accepting hallucinations. For questions on position, further conditioned on correctly identifying conditions/findings, Gemini Pro had a lower accuracy of 26.4%, with 76.68% of its errors due to accepting hallucinated positions.

4.2.3 Transferability of Domain Expertise

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. As shown in Table 16, CheXagent, a model trained exclusively on chest X-rays images, performs best in detecting abnormalities and identifying conditions/findings among all seven models when tested on chest X-ray images. We analyzed its performance to explore the transferability of expertise across the rest modalities.

As illustrated in Figure 4, 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 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.

Table 6: 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

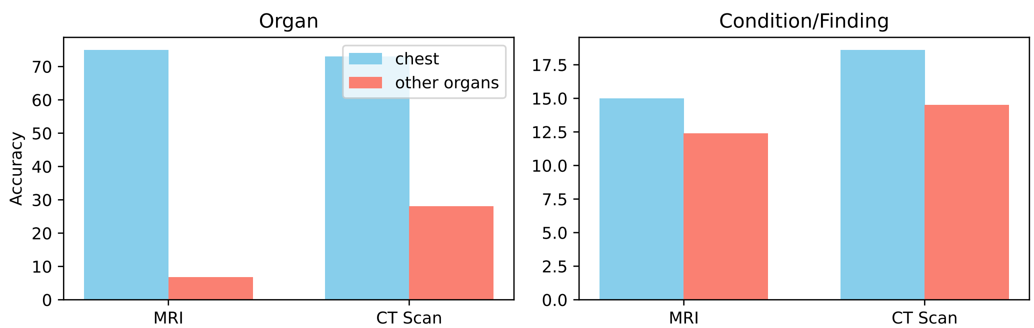

Figure 4: Accuracy comparison of CheXagent in identifying organs and conditions/findings across different modalities. The model demonstrates significantly higher accuracy in identifying organs on chest images compared to images of other organs for both MRI and CT scans. Additionally, CheXagent shows 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.

5 Conclusion

Evaluating the reliability of LMMs in the medical domain requires robust methods, and ProbMed, our newly introduced dataset, addresses this by incorporating probing evaluation and procedural diagnosis. Our study reveals significant limitations in models like GPT-4V and Gemini Pro, which perform worse than random guessing on specialized diagnostic questions, while CheXagent's results highlight the importance of domain-specific knowledge. These findings underscore the urgent need for robust evaluation methodologies to ensure the reliability and accuracy of LMMs in medical imaging. Despite the contributions, limitations such as the imbalanced image distribution favoring Chest X-rays (see Table 2) and the lack of open-ended evaluations like report generation remain. The broader impact of our work includes improved diagnostic accuracy and patient care, but also highlights the risks of deploying unreliable models in healthcare. To mitigate these risks, we recommend rigorous testing, continuous performance monitoring, and incorporation of domain-specific expertise. Our work aims to enhance the development of trustworthy AI systems in healthcare, ultimately improving patient outcomes.

References

R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, E. Chu, J. H. Clark, L. E. Shafey, Y. Huang, K. Meier-Hellstern, G. Mishra, E. Moreira, M. Omernick, K. Robinson, S. Ruder, Y. Tay, K. Xiao, Y. Xu, Y. Zhang, G. H. Abrego, J. Ahn, J. Austin, P. Barham, J. Botha, J. Bradbury, S. Brahma, K. Brooks, M. Catasta, Y. Cheng, C. Cherry, C. A. Choquette-Choo, A. Chowdhery, C. Crepy, S. Dave, M. Dehghani, S. Dev, J. Devlin, M. Díaz, N. Du, E. Dyer, V. Feinberg, F. Feng, V. Fienber, M. Freitag, X. Garcia, S. Gehrmann, L. Gonzalez, G. Gur-Ari, S. Hand, H. Hashemi, L. Hou, J. Howland, A. Hu, J. Hui, J. Hurwitz, M. Isard, A. Ittycheriah, M. Jagielski, W. Jia, K. Kenealy, M. Krikun, S. Kudugunta, C. Lan, K. Lee, B. Lee, E. Li, M. Li, W. Li, Y. Li, J. Li, H. Lim, H. Lin, Z. Liu, F. Liu, M. Maggioni, A. Mahendru, J. Maynez, V. Misra, M. Moussalem, Z. Nado, J. Nham, E. Ni, A. Nystrom, A. Parrish, M. Pellat, M. Polacek, A. Polozov, R. Pope, S. Qiao, E. Reif, B. Richter, P. Riley, A. C. Ros, A. Roy, B. Saeta, R. Samuel, R. Shelby, A. Slone, D. Smilkov, D. R. So, D. Sohn, S. Tokumine, D. Valter, V. Vasudevan, K. Vodrahalli, X. Wang, P. Wang, Z. Wang, T. Wang, J. Wieting, Y. Wu, K. Xu, Y. Xu, L. Xue, P. Yin, J. Yu, Q. Zhang, S. Zheng, C. Zheng, W. Zhou, D. Zhou, S. Petrov, and Y. Wu. Palm 2 technical report, 2023.

J. M. Z. Chaves, S.-C. Huang, Y. Xu, H. Xu, N. Usuyama, S. Zhang, F. Wang, Y. Xie, M. Khademi, Z. Yang, H. Awadalla, J. Gong, H. Hu, J. Yang, C. Li, J. Gao, Y. Gu, C. Wong, M. Wei, T. Naumann, M. Chen, M. P. Lungren, S. Yeung-Levy, C. P. Langlotz, S. Wang, and H. Poon. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation, 2024.

J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning, 2023a.

Q. Chen, X. Hu, Z. Wang, and Y. Hong. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts, 2023b.

Z. Chen, M. Varma, J.-B. Delbrouck, M. Paschali, L. Blankemeier, D. V. Veen, J. M. J. Valanarasu, A. Youssef, J. P. Cohen, E. P. Reis, E. B. Tsai, A. Johnston, C. Olsen, T. M. Abraham, S. Gatidis, A. S. Chaudhari, and C. Langlotz. Chexagent: Towards a foundation model for chest x-ray interpretation, 2024.

H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models, 2022.

X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie. Pathvqa: 30000+ questions for medical visual question answering, 2020.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023.

J. Lau, S. Gayen, A. Ben Abacha, et al. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251, 2018. doi: 10.1038/sdata.2018.251. URL `https://doi.org/10.1038/sdata.2018.251`.

C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, 2023a.

J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023b.

B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering, 2021.

H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning, 2023a.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning, 2023b.

M. Moor, Q. Huang, S. Wu, M. Yasunaga, C. Zakka, Y. Dalmia, E. P. Reis, P. Rajpurkar, and J. Leskovec. Med-flamingo: a multimodal medical few-shot learner, 2023.

OpenAI. Gpt-4 technical report. *Technical report.*, 2023a. URL `https://arxiv.org/abs/2303.08774`.

OpenAI. Gpt-4v(ision) technical work and authors. *Technical report.*, 2023b. URL `https://cdn.openai.com/contributions/gpt-4v.pdf`.

S. Subramanian, L. L. Wang, S. Mehta, B. Bogin, M. van Zuylen, S. Parasa, S. Singh, M. Gardner, and H. Hajishirzi. Medicat: A dataset of medical images, captions, and textual references, 2020.

G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.

X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.

X. Wang, B. Ma, C. Hu, L. Weber-Genzel, P. Röttger, F. Kreuter, D. Hovy, and B. Plank. "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models, 2024.

C. Wu, J. Lei, Q. Zheng, W. Zhao, W. Lin, X. Zhang, X. Zhou, Z. Zhao, Y. Zhang, Y. Wang, and W. Xie. Can gpt-4v(ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis, 2023.

Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023.

X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering, 2023.

C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang. Large language models are not robust multiple choice selectors, 2024.

A Breakdown Results on Different Image Modality and Organ.

A.1 Brain CT Scan

Table 7: 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	25	25	50	35.28	35.01
	acc w.o. adv. pair	50	50	/	40.6	40.4
LLaVA-v1-7B	acc	25.18	52.59	50	0	0
	acc w.o. adv. pair	99.62	87.03	/	92.96	98.94
LLaVA-v1.6-7B	acc	10.74	72.22	23.52	0	0.52
	acc w.o. adv. pair	99.62	75.18	/	55.18	77.36
LLaVA-Med	acc	4.81	10.74	8.82	11.85	3.15
	acc w.o. adv. pair	95.55	37.4	/	26.66	92.63
MiniGPT-v2	acc	1.11	**92.59**	50	17.77	8.42
	acc w.o. adv. pair	100	98.88	/	64.81	89.47
CheXagent	acc	11.85	0	47.05	12.96	5.26
	acc w.o. adv. pair	33.7	0	/	16.66	14.21
GPT-4V	acc	**94.07**	84.07	<u>61.76</u>	**37.03**	**31.05**
	acc w.o. adv. pair	94.81	84.44	/	55.92	49.47
Gemini Pro	acc	<u>84.44</u>	<u>85.18</u>	**70.58**	<u>34.81</u>	<u>21.05</u>
	acc w.o. adv. pair	95.92	91.48	/	45.18	37.36
	num	270	270	34	270	270

A.2 Chest CT Scan

Table 8: 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	25	25	50	32.69	33.76
	acc w.o. adv. pair	50	50	/	38.89	39.6
LLaVA-v1-7B	acc	27.55	46.35	50	0.36	0.23
	acc w.o. adv. pair	99.81	99.81	/	96.35	100
LLaVA-v1.6-7B	acc	2.73	**76.82**	50	0.54	0.46
	acc w.o. adv. pair	99.08	86.67	/	76.45	87.32
LLaVA-Med	acc	5.47	39.78	29.41	14.41	4.37
	acc w.o. adv. pair	96.71	90.51	/	35.58	95.62
MiniGPT-v2	acc	0.54	53.28	50	10.21	3.22
	acc w.o. adv. pair	99.45	96.53	/	77.55	90.09
CheXagent	acc	6.75	<u>72.99</u>	50	18.61	7.83
	acc w.o. adv. pair	19.7	99.08	/	35.76	30.41
GPT-4V	acc	**97.07**	72.94	<u>67.64</u>	<u>32.9</u>	**20.78**
	acc w.o. adv. pair	97.62	93.96	/	47.34	39.72
Gemini Pro	acc	<u>95.62</u>	58.21	**82.35**	**34.48**	<u>14.28</u>
	acc w.o. adv. pair	99.27	97.26	/	60.6	61.05
	num	548	548	34	548	548

A.3 Spine CT Scan

Table 9: 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 <u>underlined</u>.

		General Question		Specialized Question		
		Modality	Organ	Abnormality	Condition/Finding	Position
Random Choice	acc	25	25	50	30.85	31.06
	acc w.o. adv. pair	50	50	/	36.42	36.62
LLaVA-v1-7B	acc	22.98	44.82	50	0	0
	acc w.o. adv. pair	98.85	91.95	/	97.7	100
LLaVA-v1.6-7B	acc	4.59	<u>72.41</u>	0	0	1.28
	acc w.o. adv. pair	98.85	85.05	/	65.51	70.51
LLaVA-Med	acc	2.29	11.49	50	11.49	6.41
	acc w.o. adv. pair	94.25	28.73	/	26.43	79.48
MiniGPT-v2	acc	1.14	41.37	0	12.64	5.12
	acc w.o. adv. pair	100	62.06	/	68.96	78.2
CheXagent	acc	4.59	27.58	50	10.34	2.56
	acc w.o. adv. pair	16.09	55.17	/	27.58	33.33
GPT-4V	acc	<u>81.39</u>	69.76	0	**33.73**	**25.97**
	acc w.o. adv. pair	81.39	81.39	/	39.53	42.85
Gemini Pro	acc	**87.2**	**77.9**	50	<u>22.09</u>	**25.97**
	acc w.o. adv. pair	94.18	93.02	/	65.11	50.64
	num	86	86	2	86	86

A.4 Abdominal CT Scan

Table 10: 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 <u>underlined</u>.

		General Question		Specialized Question		
		Modality	Organ	Abnormality	Condition/Finding	Position
Random Choice	acc	25	25	50	35.53	37.03
	acc w.o. adv. pair	50	50	/	40.86	41.88
LLaVA-v1-7B	acc	26.49	54.19	50	0.53	0
	acc w.o. adv. pair	100	100	/	94.54	99.83
LLaVA-v1.6-7B	acc	1.86	**82.82**	41.42	1.06	0.66
	acc w.o. adv. pair	99.86	96.4	/	70.17	87.76
LLaVA-Med	acc	5.05	45	30	15.44	5.28
	acc w.o. adv. pair	98	93.6	/	33.28	86.11
MiniGPT-v2	acc	0	37.15	48.57	6.12	2.14
	acc w.o. adv. pair	99.2	91.34	/	83.62	90.74
CheXagent	acc	25.03	38.21	<u>52.85</u>	15.57	6.61
	acc w.o. adv. pair	56.99	80.29	/	51.26	55.53
GPT-4V	acc	<u>95.72</u>	<u>72.72</u>	45.71	<u>27</u>	**23.25**
	acc w.o. adv. pair	96.39	92.11	/	40.9	34.71
Gemini Pro	acc	**98.31**	69.19	**65.71**	**28.79**	<u>20.39</u>
	acc w.o. adv. pair	99.2	96.8	/	61.86	58.2
	num	750	750	70	750	750

A.5 Brain MRI

Table 11: 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	25	25	50	36.7	36.64
	acc w.o. adv. pair	50	50	/	41.47	41.49
LLaVA-v1-7B	acc	1.23	32.86	50	0.53	0
	acc w.o. adv. pair	97.17	97.34	/	94.52	99.75
LLaVA-v1.6-7B	acc	17.49	88.51	28.57	0.53	0.48
	acc w.o. adv. pair	99.49	89.57	/	54.77	67.63
LLaVA-Med	acc	3	8.12	23.21	14.66	2.91
	acc w.o. adv. pair	90.45	44.34	/	22.08	89.29
MiniGPT-v2	acc	1.94	**96.64**	50	15.72	4.37
	acc w.o. adv. pair	99.82	100	/	66.78	88.8
CheXagent	acc	0	0	50	10.77	6.81
	acc w.o. adv. pair	1.41	0	/	16.07	23.66
GPT-4V	acc **96.99**	94.33	58.92	**36.1**	**27.8**	
	acc w.o. adv. pair	97.52	95.39	/	54.69	43.17
Gemini Pro	acc	95.22	94.87	**78.57**	35.51	19.7
	acc w.o. adv. pair	96.28	99.29	/	56.89	37.71
	num	566	566	56	566	566

A.6 Chest MRI

Table 12: 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	25	25	50	34.18	34.11
	acc w.o. adv. pair	50	50	/	39.84	39.88
LLaVA-v1-7B	acc	0	35	50	0	0
	acc w.o. adv. pair	80	92.5	/	90	100
LLaVA-v1.6-7B	acc	5	32.5	37.5	0	0
	acc w.o. adv. pair	90	50	/	50	76.47
LLaVA-Med	acc	5	45	12.5	12.5	5.88
	acc w.o. adv. pair	82.5	82.5	/	17.5	82.35
MiniGPT-v2	acc	0	35	50	10	8.82
	acc w.o. adv. pair	92.5	62.5	/	70	79.41
CheXagent	acc	0	**75**	50	15	0
	acc w.o. adv. pair	5	100	/	27.5	23.52
GPT-4V	acc	76.92	51.28	37.5	**25.64**	**18.18**
	acc w.o. adv. pair	76.92	53.84	/	30.76	30.3
Gemini Pro	acc	**87.5**	62.5	37.5	17.5	11.76
	acc w.o. adv. pair	87.5	82.5	/	55	44.11
	num	40	40	8	40	40

A.7 Spine MRI

Table 13: 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	25	25	50	31.51	31.52
	acc w.o. adv. pair	50	50	/	37.55	37.66
LLaVA-v1-7B	acc	0	32.09	50	50	0.3
	acc w.o. adv. pair	98.45	98.76	/	93.51	99.32
LLaVA-v1.6-7B	acc	3.08	86.72	27.77	0.3	0
	acc w.o. adv. pair	99.69	95.37	/	50.3	57.77
LLaVA-Med	acc	1.54	5.24	36.11	12.96	5.4
	acc w.o. adv. pair	88.27	23.45	/	17.96	77.7
MiniGPT-v2	acc	0.3	49.69	52.77	6.79	2.02
	acc w.o. adv. pair	100	84.56	/	80.86	87.83
CheXagent	acc	0	13.58	47.22	15.43	2.7
	acc w.o. adv. pair	1.54	27.77	/	22.53	25.33
GPT-4V	acc	96.28	**90.71**	55.55	22.6	**15.59**
	acc w.o. adv. pair	96.28	95.66	/	32.19	26.44
Gemini Pro	acc	**98.13**	88.81	**57.14**	**24.53**	14.91
	acc w.o. adv. pair	98.13	97.82	/	50	39.32
	num	332	332	35	332	332

A.8 Abdominal MRI

Table 14: 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	25	25	50	37.13	38.26
	acc w.o. adv. pair	50	50	/	41.81	42.23
LLaVA-v1-7B	acc	0	39.28	50	2.38	0
	acc w.o. adv. pair	96.42	100	/	90.47	100
LLaVA-v1.6-7B	acc	2.38	73.8	35.71	1.19	0
	acc w.o. adv. pair	98.8	92.85	/	63.09	80.3
LLaVA-Med	acc	2.38	47.61	50	14.28	9.09
	acc w.o. adv. pair	79.76	95.23	/	17.85	71.21
MiniGPT-v2	acc	0	36.9	50	8.33	4.54
	acc w.o. adv. pair	98.8	91.66	/	82.14	93.93
CheXagent	acc	0	26.19	50	11.9	10.6
	acc w.o. adv. pair	0	59.52	/	30.95	30.3
GPT-4V	acc	86.9	**75**	50	27.38	**25.75**
	acc w.o. adv. pair	86.9	86.9	/	50	42.42
Gemini Pro	acc	**89.28**	72.61	**85.71**	**28.57**	**25.75**
	acc w.o. adv. pair	91.66	94.04	/	67.85	56.06
	num	84	84	14	84	84

A.9 Brain X-ray

Table 15: 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	25	25	50	44.77	47.08
	acc w.o. adv. pair	50	50	/	46.51	48.33
LLaVA-v1-7B	acc	45.56	26.58	50	0	0
	acc w.o. adv. pair	100	72.15	/	96.2	100
LLaVA-v1.6-7B	acc	11.39	13.92	16.66	8.86	4.44
	acc w.o. adv. pair	94.93	20.25	/	72.15	91.11
LLaVA-Med	acc	8.86	8.86	0	20.25	4.44
	acc w.o. adv. pair	100	27.84	/	24.05	88.88
MiniGPT-v2	acc	18.98	**83.54**	50	18.98	17.77
	acc w.o. adv. pair	98.73	88.6	/	63.29	73.33
CheXagent	acc	<u>84.81</u>	0	50	12.65	8.88
	acc w.o. adv. pair	88.6	1.26	/	21.51	22.22
GPT-4V	acc	82.05	8.97	33.33	**43.58**	<u>22.72</u>
	acc w.o. adv. pair	83.33	8.97	/	50	36.36
Gemini Pro	acc	**89.87**	<u>51.89</u>	50	<u>31.64</u>	**31.11**
	acc w.o. adv. pair	93.67	62.02	/	50.63	46.66
	num	79	79	6	79	79

A.10 Chest X-ray

Table 16: 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	25	25	50	37.59	37.08
	acc w.o. adv. pair	50	50	/	42.15	41.87
LLaVA-v1-7B	acc	28.75	36.57	50	0.12	0.11
	acc w.o. adv. pair	99.9	99.83	/	68.52	99.96
LLaVA-v1.6-7B	acc	7.11	<u>83.97</u>	47.94	5.89	1.52
	acc w.o. adv. pair	99.51	97.53	/	53.94	94.02
LLaVA-Med	acc	6.25	39.77	40.24	26.28	6.14
	acc w.o. adv. pair	99	90.6	/	35.67	94.34
MiniGPT-v2	acc	4.93	**94.07**	50.05	18.78	11.94
	acc w.o. adv. pair	99.96	99.39	/	55.19	67.49
CheXagent	acc	53.68	39.64	**76.59**	**42.75**	8.9
	acc w.o. adv. pair	62.59	47.53	/	58.1	36.93
GPT-4V	acc	<u>91.53</u>	67.51	53.18	<u>39.35</u>	**21.35**
	acc w.o. adv. pair	91.88	94.09	/	46.18	34.38
Gemini Pro	acc	**98.07**	76.74	<u>61.29</u>	25.83	<u>15.31</u>
	acc w.o. adv. pair	99.29	99.67	/	57.83	66.42
	num	3120	3120	1948	3120	3120

A.11 Spine X-ray

Table 17: 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	25	25	50	30.95	30.99
	acc w.o. adv. pair	50	50	/	37.54	37.51
LLaVA-v1-7B	acc	44.55	45.04	50	0.49	0
	acc w.o. adv. pair	98.51	98.51	/	93.56	98.85
LLaVA-v1.6-7B	acc	4.45	**82.67**	33.33	1.48	0.57
	acc w.o. adv. pair	99.5	95.54	/	61.38	77.58
LLaVA-Med	acc	8.41	7.92	33.33	17.82	5.74
	acc w.o. adv. pair	97.02	28.21	/	25.24	78.16
MiniGPT-v2	acc	2.97	52.47	58.33	16.33	4.02
	acc w.o. adv. pair	99	75.24	/	61.38	69.54
CheXagent	acc	82.17	20.29	62.5	16.83	0.57
	acc w.o. adv. pair	90.09	40.59	/	30.69	31.6
GPT-4V	acc	85.57	72.13	47.82	**29.85**	18.49
	acc w.o. adv. pair	85.57	81.09	/	37.13	29.47
Gemini Pro	acc	**95.02**	70.14	**70.83**	17.91	**19.07**
	acc w.o. adv. pair	97.01	93.53	/	79.1	58.95
	num	201	201	24	201	201

A.12 Abdominal X-ray

Table 18: 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	25	25	50	36.55	37.46
	acc w.o. adv. pair	50	50	/	37.54	37.51
LLaVA-v1-7B	acc	53.87	53.01	50	2.15	0.56
	acc w.o. adv. pair	100	99.56	/	95.25	99.45
LLaVA-v1.6-7B	acc	5.17	56.46	46	6.46	1.12
	acc w.o. adv. pair	97.84	78.44	/	80.17	91.52
LLaVA-Med	acc	7.75	42.24	60	14.65	4.51
	acc w.o. adv. pair	98.27	91.37	/	23.7	89.83
MiniGPT-v2	acc	4.74	38.79	50	18.53	5.64
	acc w.o. adv. pair	99.13	86.63	/	56.03	67.23
CheXagent	acc	77.15	23.7	70	12.93	2.25
	acc w.o. adv. pair	90.51	46.55	/	38.79	37.85
GPT-4V	acc	84.84	50.21	60	**31.16**	**23.16**
	acc w.o. adv. pair	84.84	71.86	/	42.85	37.85
Gemini Pro	acc	**97.14**	**63.36**	**85**	27.15	19.2
	acc w.o. adv. pair	99.13	96.55	/	73.27	67.23
	num	232	232	20	232	232

B ProbMed

Table 19: Number of questions across each question type for each image. 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.

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