How Good LLMs Are at Answering Bangla Medical Visual Questions? Dataset and Benchmarking

Rafid Ahmed[1], Intesar Tahmid[1], Mir Sazzat Hossain[2], Tasnimul Hossain Tomal[1], Md Fahim[1], Md Farhad Alam Bhuiyan[1]

[1]Penta Global Limited

[2]Center for Computational & Data Sciences, Independent University, Bangladesh

Abstract

Recent advancements in Large Language Models (LLMs) and Large Multimodal Models (LMMs) have enabled general-purpose systems to demonstrate promising capabilities in complex reasoning tasks, including those in the medical domain. Medical Visual Question Answering (MedVQA) has particularly benefited from these developments. However, despite Bangla being one of the most widely spoken languages globally, there exists no established MedVQA benchmark for it. To address this gap, we introduce a new Bangla Medical VQA dataset comprising clinically validated image–question–answer pairs, along with a comprehensive evaluation of current foundation models on this resource. Consistent with prior findings that report worse-than-random performance of current models on English MedVQA benchmarks, our analysis reveals that Bangla performance is substantially lower, reflecting the challenges inherent to low-resource languages. Even top-performing models such as Gemini and GPT-4.1 Mini fail to accurately answer specialized diagnostic questions, indicating severe limitations in fine-grained medical reasoning. Although certain open-source models, such as Gemma-3, occasionally outperform these models in general categories, they too struggle with clinically complex questions underscoring the urgent need for top-notch evaluation method.

1 Introduction

In recent years, foundation models such as Large Language Models (LLMs) (Achiam et al., 2023; Touvron et al., 2023; Anil et al., 2023) and Large Multimodal Models (LMMs) (Comanici et al., 2025; Li et al., 2023b; Liu et al., 2023; Chen et al., 2023) have attracted significant attention for their ability to process, generate, and reason over complex textual and visual inputs. These models produce human-like language and achieve high performance across a wide range of benchmarks. Their integration into the medical domain has already shown promising results, demonstrating potential for real-world clinical applications.

Applications such as automated radiology report generation (Sloan et al., 2024), clinical decision support, and interactive medical question answering highlight the capability of LLMs to assist both patients and healthcare professionals. Existing medical VQA benchmarks (Liu et al., 2021; He et al., 2020; Zhang et al., 2023), when combined with current foundation models, have reported encouraging results. However, despite being the seventh most spoken language in the world, with approximately 278 million speakers, Bangla has received very limited attention in the context of medical visual question answering. It is therefore crucial to evaluate the performance of LLMs in this domain for Bangla.

The effectiveness of medical AI systems is closely linked to the quality of the datasets on which they are trained (Gong et al., 2023). In high-resource languages such as English, medical VQA datasets have gradually evolved in complexity, progressing from simple factual questions to tasks requiring deeper clinical reasoning (Marino et al., 2019; Schwenk et al., 2022). In contrast, low-resource languages face a significant bottleneck: the absence of sufficiently comprehensive datasets capable of supporting even fundamental diagnostic question answering.

In the Bangla medical VQA domain, progress has been highly limited. To date, the only publicly available dataset is (Remon, 2025). However, this resource does not provide any information regarding the source of the images, nor does it describe the process of dataset creation or validation by medical experts. Furthermore, it suffers from annotation errors, lacks clinical oversight, and contains numerous irrelevant question–answer pairs, making it unsuitable as a reliable benchmark.

To address this gap, we propose a new Bangla

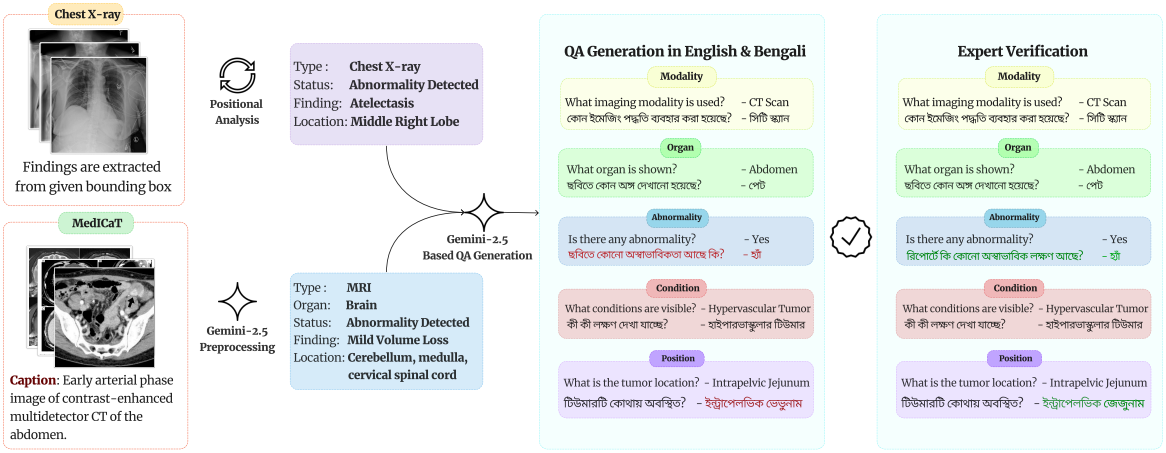


Figure 1: Workflow of the dataset curation process. Images and metadata were obtained from two widely used biomedical datasets, enabling the automatic generation of QA pairs, which were subsequently verified by domain experts.

medical Visual Question Answering (VQA) dataset built from images drawn from multiple medical domains (Subramanian et al., 2020; Wang et al., 2017), paired with diverse and clinically relevant questions. The dataset contains 1,374 unique image–caption pairs, with questions spanning five categories: modality, organ, abnormality, condition, and position. These categories can be grouped into two types: (i) generalized questions, which are relatively straightforward, and (ii) specialized questions, which require deeper medical understanding. Importantly, unlike the existing resource, our dataset was validated by two certified physicians to ensure both clinical accuracy and reliability.

We systematically evaluate both open-source and closed-source LLMs on our proposed dataset. The best-performing models, Gemini and GPT, achieve overall accuracies of 55.08% and 38.60%, respectively. While results on generalized questions appear promising, performance on specialized diagnostic categories such as *Condition/Finding* and *Position* falls below random chance, revealing critical gaps in the reliability of current LMMs for fine-grained medical reasoning. Previous studies have reported worse-than-random behavior on English medical VQA tasks (Yan et al., 2024); however, the problem is even more pronounced for Bangla, a low-resource language. Comparative experiments further confirm that closed-source models consistently perform better in English than in Bangla.

Incorporating chain-of-thought reasoning and supplementing inputs with visual descriptions generated by GPT-4.1 Mini substantially improves performance, suggesting that limited visual understanding remains a key bottleneck. These results highlight the importance of augmenting LMMs with richer visual and linguistic context to enable more reliable medical VQA in low-resource languages.These findings highlight the urgent need for more robust Bangla medical datasets and stronger evaluation frameworks, despite the impressive progress of foundation models in general-domain benchmarks.

2 Related Work

2.1 Visual Question Answering

Visual Question Answering (VQA) has emerged as a challenging task at the intersection of vision and language. It was formally established by Antol et al. (2015), who introduced a benchmark dataset requiring fine-grained visual understanding and commonsense reasoning. Shih et al. (2016) further advanced the field by introducing a region-focused attention mechanism, aligning image regions with question embeddings to improve performance on queries requiring localized reasoning. However, such early models exploited dataset biases, which led Goyal et al. (2017) to propose VQA v2.0, a balanced dataset with complementary image-question pairs that force models to genuinely leverage visual content. Beyond visual grounding, Shah et al. (2019) addressed knowledge-intensive scenarios with KVQA, a dataset requiring external knowledge, particularly about named entities, to answer complex questions. Together, these works have progressively shaped VQA from simple recognition toward deeper reasoning.

2.2 Medical Visual Question Answering

Gu et al. (2024) proposes a Latent Prompt Assist architecture for medical VQA that generates answer-constrained latent prompts and fuses them with unimodal and multimodal features. It also integrates disease–organ priors to enhance clinical relevance, and ultimately, it reported state-of-the-art gains on VQA-RAD, SLAKE, and VQA-2019. Yim et al. (2024) delivers a multilingual dermatology VQA dataset and benchmarks for consumer-generated images and free-text clinician responses, emphasizing real-world noise, longer responses, and multilingual evaluation metrics. Xu et al. (2024) proposed a multi-level visual language model for MVQA, introducing a new multi-level instruction dataset (MLe-VQA), a feature alignment module, and an evaluation benchmark (MLe-Bench) to better capture recognition, details, diagnosis, knowledge, and reasoning. Gai et al. (2025) developed rationale-guided benchmarks (R-RAD, R-SLAKE, R-Path) and a framework that integrates medical decision-making rationales into MedVQA, thereby improving both interpretability and performance. Yu et al. (2025) introduced adaptive region-level visual prompts and a hierarchical answer generator with PEFT techniques to improve fine-grained localization and generative MedVQA.

2.3 Bangla Visual Questing Answering

Barua et al. (2024) introduced a regionally relevant Bangla VQA corpus named ChitroJera that emphasizes cultural and contextual relevance for Bangla speakers, filling important gaps left by predominantly English VQA resources. Bangla-Bayanno by Hasan et al. (2025) presents a large-scale, open-ended Bangla VQA benchmark (52.7K QA pairs over $\tilde{4}$.7K images) created via an LLM-assisted translation-refinement pipeline to produce fluent, context-preserving Bangla queries and answers. Rafi et al. (2022) introduced the first human-annotated Bengali VQA dataset using images from VQA v2.0 and proposed a deep learning-based top-down attention model to effectively balance visual and linguistic information in QA tasks.

3 Dataset Creation

3.1 Data Collection

The dataset was curated consisting of 1,374 unique image–caption pairs by combining samples from two widely used biomedical datasets: MedICaT (Subramanian et al., 2020) and ChestX-ray14 (Wang et al., 2017). To ensure consistency across sources, we created a unified metadata schema for each image, defined as follows:

$$D_i = \{\, mod_i,\ organ_i,\ \{(cond_j, pos_j)\}_{j=1}^{n_i} \,\} \tag{1}$$

Here, mod_i denotes the imaging modality, $organ_i$ represents the anatomical region under examination, and each pair $(cond_j, pos_j)$ corresponds to a medical finding and its associated spatial location.

For the MedICaT dataset, medical findings were extracted from the image captions. We employed the `gemini-2.5-flash` model to identify abnormality types and positional details, which were then incorporated into the metadata. For the ChestX-ray14 dataset, we utilized the provided bounding box annotations in combination with a positional reasoning module to derive precise location-aware metadata.

3.2 Question Generation from MetaData

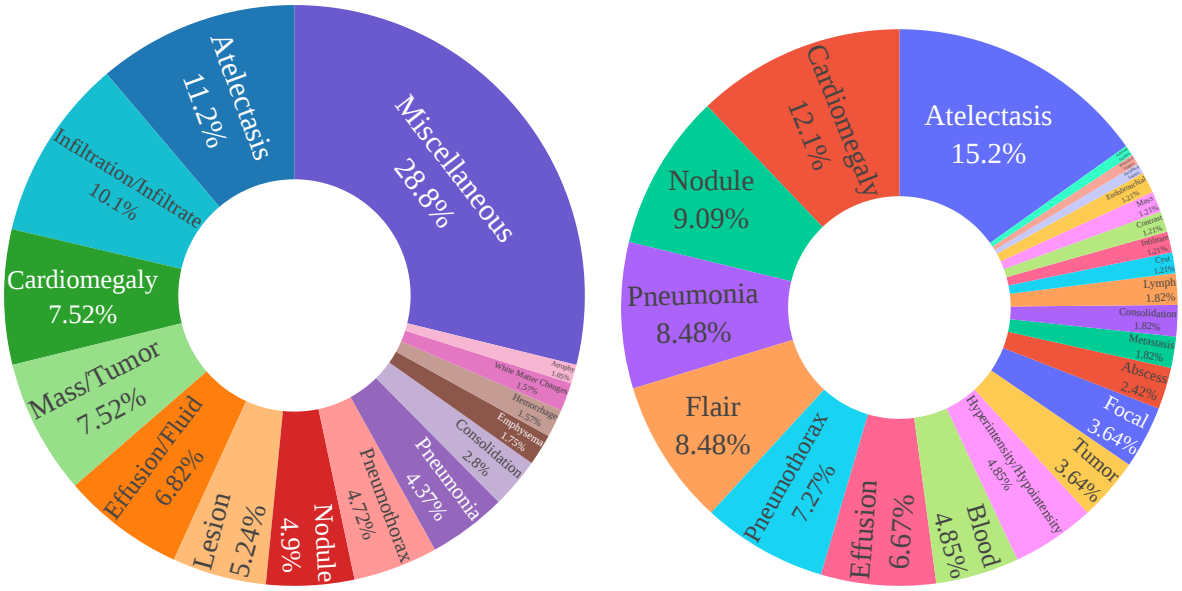
From the curated metadata, we generated English question–answer (QA) pairs for each image using the `gemini-2.5-flash` model through a few-shot prompting strategy. This approach enabled the model to produce clinically relevant and contextually accurate QA pairs aligned with the associated metadata. Subsequently, to support multilingual accessibility and facilitate research in non-English clinical contexts, we employed the `gemini-2.0-flash` model to translate the English QA pairs into Bangla. This two-stage pipeline ensured both linguistic diversity and consistency between the metadata and the QA pairs.

3.3 Question and Answer Verification

To evaluate the reliability of both the metadata and the corresponding question–answer pairs, we carried out an expert validation study. Two medical specialists independently examined a randomly selected subset of metadata entries together with their associated QA pairs. The experts, compensated on an hourly basis, verified the annotations with an observed accuracy of 95% for metadata and 98% for QA pairs.

4 Dataset Statistics

The dataset is balanced with respect to healthy and abnormal cases. Additionally, we generated 2,000 Visual Question Answering (VQA) instances.



(a) Distribution of Clinical Conditions in the Dataset.

(b) Distribution of Question Keywords in the Dataset.

Figure 2: Distributions of clinical conditions and question keywords in the curated dataset.

Each question is categorized into one of five types: *modality*, *organ*, *abnormality*, *condition*, and *position*, ensuring uniform coverage across these categories.

4.1 Question Statistics

The questions were generated using the Gemini-2.5-Flash model, with lengths varying between 3 and 24 words and an average length of approximately 6 words.

Analysis of question keywords reveals a strong emphasis on imaging interpretation, anatomical localization, and abnormality assessment.The most frequent terms include *abnormality*, *condition*, *identified*, *organ*, *technique*, *pneumonia*, *pneumothorax*, *atelectasis*, *cardiomegaly*, *metastasis*, *tumor*, *flare*, *consolidation*, *lymph*, *abscess*, *effusion*, *blood*, *nodule*, *cyst*, *infiltrate*, *contrast*, *hyperintensity / hypointense*, *mass*, *focal*, *endobronchial*, *peripheral*, *bronchial*, and *adrenal*, reflecting the central role of image understanding and diagnostic reasoning.

4.2 Answer Statistics

The answer distribution reflects the clinical focus areas of the dataset. With respect to imaging modalities, *X-ray* constitutes the majority of cases, followed by *CT* and *MRI*. This distribution highlights the dataset's emphasis on widely available and routinely used imaging techniques, particularly X-rays, while also incorporating advanced modalities such as CT and MRI.

Organ-wise, questions predominantly target the *Chest/Thorax*, followed by the *Brain*, *Abdomen*, *Spine*, *Pelvis*, and *Others*. In terms of clinical conditions, the dataset emphasizes normal cases, with labels as *No Finding*, encompassing variants such as "no abnormality detected" and "no condition identified." Beyond normal cases, the dataset covers a diverse range of abnormalities, including *Atelectasis*, *Infiltration/Infiltrate*, *Cardiomegaly*, *Mass/Tumor*, *Effusion/Fluid*, *Lesion*, *Nodule*, *Pneumothorax*, *Pneumonia*, *Consolidation*, *Emphysema*, *Hemorrhage*, *White Matter Changes*, and *Atrophy*, along with additional questions linked to rare or miscellaneous abnormalities.

This distribution demonstrates that while the dataset predominantly features normal cases, it also captures a wide spectrum of pathological findings, ensuring robustness for Medical VQA tasks.

5 Experiment Setup

5.1 Experimentation with Bangla Question

Zero-Shot Prompting. To assess the medical reasoning capabilities of large vision-language models (LVLMs), we adopt a *zero-shot prompting* approach. Each model receives a medical image I, a system prompt P, and a Bangla natural language question Q_{BAN}. Without any task-specific fine-tuning or example demonstrations, the model is expected to generate an open-ended response based solely on this input.

Models	Chest X-Ray			Medi-CAT			Overall (Avg.)		
	Acc	BScore	LAVE	Acc	BScore	LAVE	Acc	BScore	LAVE
Zero Shot Prompt on Bangla QA Pair									
Closed Source VLMs									
Gemini 2.5 Flash	37.00	82.75	**50.80**	34.50	81.14	**60.08**	35.75	81.95	**55.44**
GPT-4.1 Mini	15.00	72.93	37.70	16.00	73.51	39.50	15.50	73.22	38.60
Open Source VLMs									
Llama-3.2V 11B	8.50	73.43	17.50	11.00	81.01	33.50	9.75	77.22	25.50
Gemma-3 12B	29.50	72.62	39.70	39.50	76.70	53.37	34.50	74.66	46.54
Qwen2.5-VL 7B	21.00	75.36	32.85	28.50	75.53	38.30	24.75	75.45	35.58
LLaVA-1.5 7B	8.50	74.43	15.20	14.50	75.14	18.75	11.50	74.79	16.98
Open Source Medical VLMs									
Med-LLaVa 7B	7.00	41.87	12.70	6.00	26.21	30.72	6.50	34.04	21.71
Med-Gemma 4B	4.00	68.91	11.90	18.50	76.02	27.80	11.25	72.47	19.85
CoT Prompt on Bangla QA Pair									
Closed Source VLMs									
Gemini 2.5 Flash	31.00	83.10	**49.33**	34.00	82.13	**60.58**	32.50	82.62	**54.96**
GPT-4.1 Mini	11.50	69.70	36.15	16.00	72.91	35.70	13.75	71.31	35.93
Open Source VLMs									
Llama-3.2V 11B	13.00	73.68	30.17	21.50	82.94	47.18	17.25	78.31	38.68
Gemma-3 12B	31.00	74.26	41.00	40.50	78.72	56.33	35.75	76.49	48.67
Qwen2.5-VL 7B	20.00	74.52	31.85	29.00	75.51	37.15	24.50	75.02	34.50
LLaVA-1.5-7B	1.30	70.44	6.05	2.00	70.58	4.55	1.65	70.51	5.30
Open Source Medical VLMs									
Med-LLaVa 7B	2.60	15.93	14.35	3.27	31.79	32.31	2.94	23.86	23.33
Med-Gemma 4B	12.00	69.21	22.50	28.00	75.18	34.63	20.00	72.20	28.57

Table 1: Model Benchmarking for **Bangla** with average scores across Chest X-Ray and Medi-CAT datasets.

Chain-of-Thought (CoT) Prompting. For *Chain-of-Thought (CoT)* prompting, we explicitly guide the model to reason through the problem before producing an answer. This is done by appending the phrase ***"Let's think step by step"*** to the system prompt P, encouraging multi-step reasoning and intermediate inference prior to answer generation.

5.2 Experimentation with English

To investigate the effect of language variation, we extend our experiments to include English-language questions. Specifically, for each Bangla question Q_{BAN} and its corresponding answer A_{BAN} associated with a medical image I, we generate an English translation, denoted as Q_{ENG} and A_{ENG}, respectively. `gemini-2.0-flash` model is used for this translation purpose.

For the English-based experiments, we maintain the same system prompt P and input the English question Q_{ENG} in place of the original Bangla question Q_{BAN}. This setup is used for both zero-shot and CoT prompting scenarios, allowing us to directly compare model performance across languages under identical conditions.

5.3 Experimented Models

We conduct a systematic evaluation of eight LVLMs, spanning closed-source, open-source, and domain-specific medical variants, on our dataset. Specifically, we benchmark two proprietary models, Gemini 2.5 Flash (Comanici et al., 2025) and GPT-4.1 Mini (OpenAI, 2023), alongside a suite of open-source general-purpose models including LLaMA-3.2V (Dubey et al., 2024), Gemma-3 (Team et al., 2025), Qwen2.5-VL (Xu et al., 2025), and LLaVA-1.5 (Liu et al., 2024). In addition, we assessed two medical-domain LVLMs: Med-LLaVA (Li et al., 2023a) and Med-Gemma (Sellergren et al., 2025).

5.4 Evaluation Metrics

Performance of the models is reported using three complementary metrics: Accuracy (Acc), BERTScore (BScore), and LAVE (LLM-Assisted VQA Evaluation). Accuracy measures the proportion of exact matches between predicted and ground-truth answers. BERTScore evaluates the semantic similarity between predicted and reference answers using contextual embeddings.LAVE (LLM-Assisted VQA Evaluation) (Mañas et al.,

Models	Chest X-Ray			Medi-CAT			Overall (Avg.)		
	Acc	BScore	LAVE	Acc	BScore	LAVE	Acc	BScore	LAVE
Zero Shot Prompt on English QA Pair									
Closed Source VLMs									
Gemini 2.5 Flash	40.50	67.04	52.47	49.50	73.55	**63.15**	45.00	70.30	**57.81**
GPT-4.1 Mini	41.00	67.23	54.70	33.00	63.77	51.31	37.00	65.50	53.01
Open Source VLMs									
Llama-3.2V 11B	10.50	59.92	45.30	11.00	66.00	53.25	10.75	62.96	49.28
Gemma-3 12B	19.50	56.05	43.30	46.50	67.14	57.60	33.00	61.60	50.45
Qwen2.5-VL 7B	46.50	69.53	**55.85**	50.00	69.44	57.10	48.25	69.49	56.48
LLaVA-1.5 7B	17.00	45.94	23.90	40.50	59.14	47.00	28.75	52.54	35.45
Open Source Medical VLMs									
Med-LLaVa 7B	1.80	37.01	32.75	2.17	31.21	39.75	1.99	34.11	36.25
Med-Gemma 4B	13.00	40.32	16.70	36.00	53.84	40.00	24.50	47.08	28.35
CoT Prompt on English QA Pair									
Closed Source VLMs									
Gemini 2.5 Flash	40.50	67.17	53.75	48.50	72.32	**61.55**	44.50	69.75	**57.65**
GPT-4.1 Mini	33.50	62.83	46.85	46.00	68.88	58.31	39.75	65.86	52.58
Open Source VLMs									
Llama-3.2V 11B	29.00	59.21	50.12	45.50	67.68	50.25	37.25	63.45	50.19
Gemma-3 12B	38.00	66.43	46.55	45.00	67.31	55.50	41.50	66.87	51.03
Qwen2.5-VL 7B	48.00	69.52	**56.55**	46.00	67.82	54.50	47.00	68.67	55.53
LLaVA-1.5-7B	20.50	47.53	27.10	42.00	60.61	47.80	31.25	54.07	37.45
Open Source Medical VLMs									
Med-LLaVa 7B	2.10	37.56	31.25	2.35	13.89	39.25	2.23	25.73	35.25
Med-Gemma 4B	11.50	41.30	15.93	35.50	53.59	39.15	23.50	47.45	27.54

Table 2: Model Benchmarking for **English** with average scores across Chest X-Ray and Medi-CAT datasets.

2024) metric designed to provide a more reliable evaluation of model performance in VQA tasks.

6 Result and Analysis

Results on Bangla and English QA pairs are presented in Tables 1 and 2, respectively.

6.1 Bangla QA Pairs

Overall, closed-source VLMs substantially outperform both open-source and medical-domain counterparts under zero-shot and chain-of-thought (CoT) prompting. Gemini 2.5 Flash achieves the highest LAVE scores in both Chest X-Ray and MedICaT, with 50.80% and 60.08% respectively under zero-shot prompting. In contrast, GPT-4.1 Mini performs significantly worse, with accuracy between 15–16% and weaker LAVE alignment. Among open-source general-purpose VLMs, Gemma-3 12B achieves the strongest performance, reaching 39.70% (Chest X-Ray) and 53.37% (MedICaT) in zero-shot settings, with further gains under CoT prompting. Medical-specific VLMs such as Med-LLaVA and Med-Gemma underperform relative to general-purpose models, likely due to limited multilingual adaptation, with accuracies frequently dropping below 10% in Bangla.

6.2 English QA Pairs

In English, the performance gap between closed-source and open-source VLMs is narrower. Both Gemini 2.5 Flash and Qwen2.5-VL 7B yield competitive results, with Qwen2.5-VL achieving the best overall LAVE on Chest X-Ray (55.85%) and Gemini 2.5 Flash leading on MedICaT (63.15%) under zero-shot prompting. GPT-4.1 Mini, while weak in Bangla, demonstrates stronger alignment in English, with accuracy exceeding 40% on Chest X-Ray. Interestingly, the open-source Gemma-3 12B performs competitively, surpassing GPT-4.1 Mini on MedICaT. Medical-domain VLMs again fail to generalize, with Med-LLaVA yielding near-zero accuracy across both datasets.

6.3 Impact of Chain-of-Thought Prompting

We observe a consistent trend where CoT prompting improves the performance of most models, particularly open-source VLMs. For instance, Llama-3.2V 11B shows a notable increase in accuracy and LAVE, from 17.50% to 30.17% in Bangla and from 45.30% to 50.12% in English. Closed-source Gemini 2.5 Flash also benefits, particularly in Bangla,

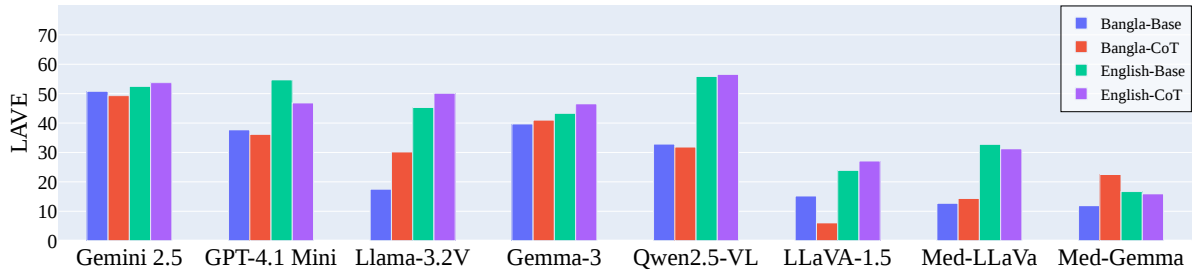


Figure 3: LAVE score comparison of the Chest X-Ray dataset for different models under four different settings: vanilla (baseline performance) for Bangla, chain-of-thought (CoT) reasoning for Bangla, vanilla (baseline performance) for English, chain-of-thought (CoT) reasoning for English.

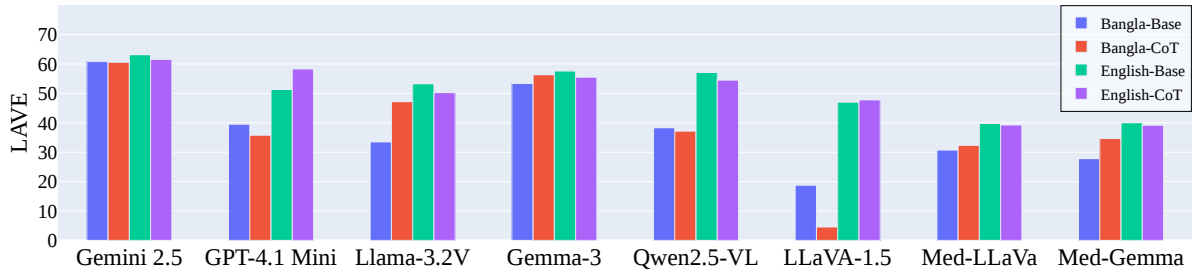


Figure 4: LAVE score comparison of the MedICat dataset for different models under four different settings: vanilla (baseline performance) for Bangla, chain-of-thought (CoT) reasoning for Bangla, vanilla (baseline performance) for English, chain-of-thought (CoT) reasoning for English.

though the gains are more modest. In contrast, medical VLMs such as Med-LLaVA generally show negligible gains, reflecting their limited reasoning in multilingual contexts. Med-Gemma, however, demonstrates substantial improvements under CoT prompting, highlighting its potential when guided by reasoning-based prompts.

6.4 Performance Across Specialized Questions

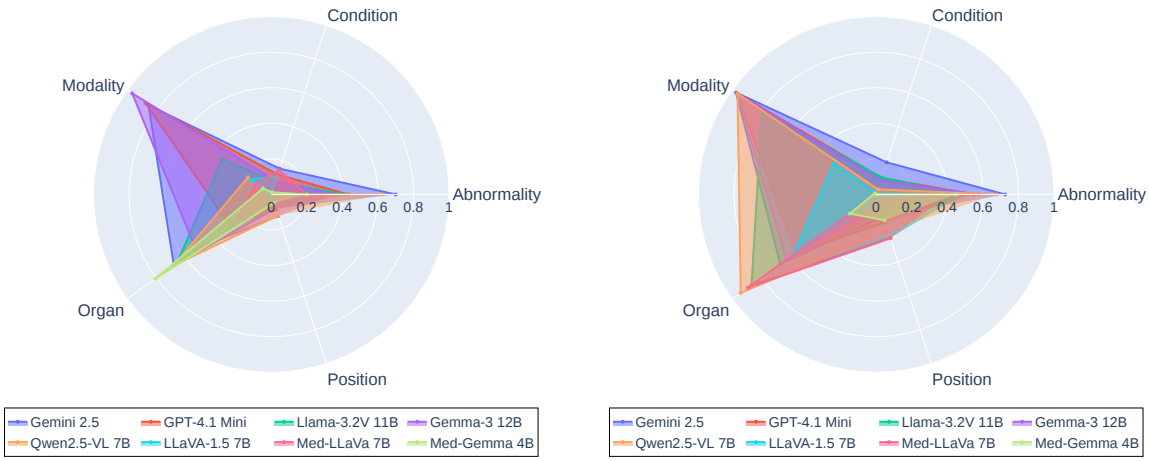
We further evaluate model performance across different question categories. The results are presented in Figures 5a–6b and Tables 3 and 4.The questions can be broadly divided into two categories: *general* questions, which involve tasks such as identifying image modality or organ type, and *specialized* questions, which require deeper medical knowledge to assess conditions and anatomical positions.

While closed-source models such as Gemini-2.5 and GPT-4.1 Mini outperform other models and achieve relatively strong results on general questions, their performance on specialized questions remains severely limited. For Bangla, under chain-of-thought reasoning across the entire dataset, Gemini-2.5 achieves the highest average specialized LAVE score, but only 23.25% on *Condition* and 24.50% on *Position*. GPT-4.1 Mini performs even worse, with scores of 13% and 19.50% on these categories, respectively. These results indicate that even the strongest models perform close to random guessing on specialized diagnostic tasks.

Among open-source models, Gemma-3 consistently outperforms other open-source counterparts across all categories and occasionally surpasses closed-source models such as GPT on general questions. However, it struggles significantly on specialized questions. Furthermore, the performance of open-source medical VLMs on these specialized categories is alarmingly low, highlighting a substantial gap in their ability to support real-world medical diagnosis.

6.5 Cross-Lingual Observations.

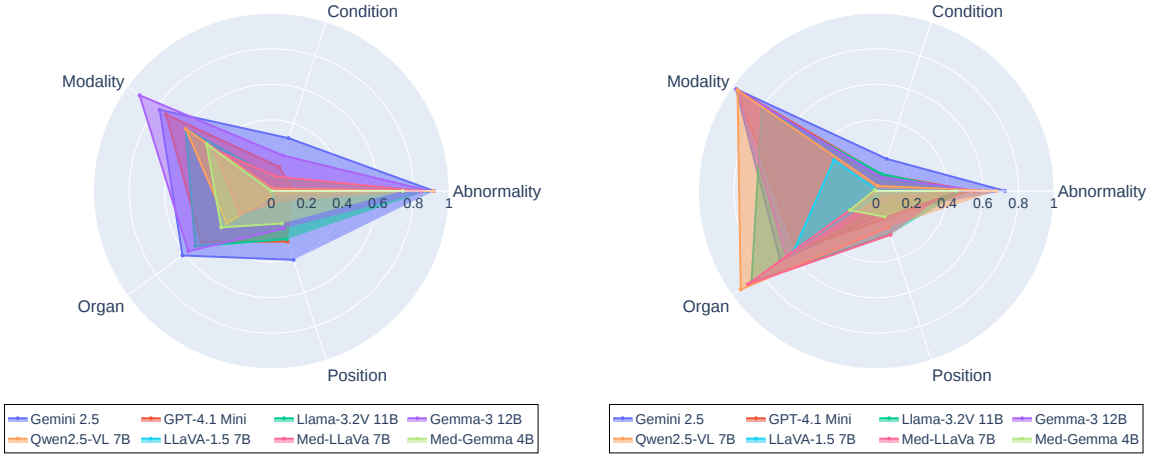
Comparing Bangla and English benchmarks, we find that performance in Bangla is generally weaker, particularly for open-source and domain-



(a) CoT prompt on Bangla QA pair

(b) CoT prompt on English QA pair

Figure 5: LAVE score comparison on the Chest X-Ray dataset across categorical question types with chain-of-thought reasoning in Bangla and English.



(a) CoT prompt on Bangla QA pair

(b) CoT prompt on English QA pair

Figure 6: LAVE score comparison on the MedICat dataset across categorical question types with chain-of-thought reasoning in Bangla and English

specific VLMs. Closed-source VLMs like Gemini exhibit stronger multilingual transfer, maintaining relatively robust scores in Bangla. This suggests that large-scale multilingual pretraining and broader instruction tuning are critical for achieving cross-lingual generalization in medical VQA.Our findings highlight three key insights: (1) closed-source VLMs maintain superior robustness and generalization across languages, (2) CoT prompting consistently enhances reasoning and alignment, and (3) medical-domain VLMs, while specialized, show limited cross-lingual capability and require further adaptation for multilingual medical tasks.

7 Conclusion

We are the first to propose Medical Visual Question Answering for Bangla addressing a significant gap in low-resource language evaluation for multimodal AI systems.Through systematic evaluation of both open-source and closed-source LLMs and LMMs, we observed that while top-performing models such as Gemini and GPT-4.1 Mini achieve reasonable performance on general questions, they struggle severely on specialized diagnostic tasks, often performing close to random. Open-source models, including Gemma-3, occasionally outperform closed-source models on general questions but also fail on specialized medical queries. Incorporating chain-of-thought prompting provides moderate improvements, highlighting insufficient visual understanding as a key limitation.

Limitations

Despite the contributions of this work, several limitations remain. While our evaluation provides valuable insights into the performance of existing models, we were unable to improve diagnostic accuracy through model fine-tuning. Fine-tuning large multimodal models on our Bangla MedVQA dataset could potentially yield significant performance gains; however, due to computational and resource constraints, this was not feasible within the scope of this study.

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*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Deeparghya Dutta Barua, Md Sakib Ul Rahman Sourove, Md Farhan Ishmam, Fabiha Haider, Fariha Tanjim Shifat, Md Fahim, and Md. Farhad Alam. 2024. Chitrojera: A regionally relevant visual question answering dataset for bangla. *CoRR*, abs/2410.14991.

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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. 2025. MedThink: A rationale-guided framework for explaining medical visual question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7438–7450, Albuquerque, New Mexico. Association for Computational Linguistics.

Youdi Gong, Guangzhen Liu, Yunzhi Xue, Rui Li, and Lingzhong Meng. 2023. A survey on dataset quality in machine learning. *Information and Software Technology*, 162:107268.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tiancheng Gu, Kaicheng Yang, Dongnan Liu, and Weidong Cai. 2024. Lapa: Latent prompt assist model for medical visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4971–4980.

Mohammed Rakibul Hasan, Rafi Majid, and Ahanaf Tahmid. 2025. Bangla-bayanno: A 52k-pair bengali visual question answering dataset with llm-assisted translation refinement. *Preprint*, arXiv:2508.19887.

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

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. 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. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4171–4179.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3199.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5):1.

Mahamudul Hasan Rafi, Shifat Islam, S. M. Hasan Imtiaz Labib, SM Sajid Hasan, Faisal Muhammad Shah, and Sifat Ahmed. 2022. A deep learning-based bengali visual question answering system. In *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pages 114–119.

S. H. Remon. 2025. Med vqa bn overall. https://huggingface.co/datasets/SHRemon97/Med-VQA_Bn_Overall.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8876–8884.

Kevin J. Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to look: Focus regions for visual question answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4613–4621.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. 2024. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, 18:368–387.

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.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv 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.

Dexuan Xu, Yanyuan Chen, Jieyi Wang, Yue Huang, Hanpin Wang, Zhi Jin, Hongxing Wang, Weihua Yue, Jing He, Hang Li, and Yu Huang. 2024. MLeVLM: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4977–4997, Bangkok, Thailand. Association for Computational Linguistics.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. 2024. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical vqa. *arXiv preprint arXiv:2405.20421*.

Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2024. DermaVQA: A Multilingual Visual Question Answering Dataset for Dermatology . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15005. Springer Nature Switzerland.

Ting Yu, Zixuan Tong, Jun Yu, and Ke Zhang. 2025. Fine-grained adaptive visual prompt for generative medical visual question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9):9662–9670.

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. *arXiv preprint arXiv:2305.10415*.

A Appendix

Detailed model performance breakdown:

Models	Generalized Question		Specialized Question			Overall
	Modality	Organ	Abnormality	Condition	Position	
Base Prompt on Bangla Chest X-Ray QA Pair						
Closed Source VLMs						
Gemini 2.5	95.75	**65.00**	70.00	**15.75**	7.50	**50.80**
GPT-4.1 Mini	85.25	29.75	47.50	11.00	15.00	37.70
Open Source VLMs						
Llama-3.2V 11B	11.00	29.50	40.00	3.00	1.75	17.50
Gemma-3 12B	**97.50**	50.25	30.00	8.00	12.75	39.70
Qwen2.5-VL 7B	13.00	63.50	**72.50**	2.85	**15.25**	32.85
LLaVA-1.5 7B	13.50	2.00	42.50	14.75	3.25	15.20
Open Source Medical VLMs						
Med-LLaVa 7B	6.75	6.75	39.75	8.00	2.25	12.70
Med-Gemma 4B	6.00	50.00	0.00	0.50	3.00	11.90
Base Prompt on English Chest X-Ray QA Pair						
Closed Source VLMs						
Gemini 2.5	**97.00**	65.25	72.25	18.58	9.25	**52.47**
GPT-4.1 Mini	97.50	60.00	**75.00**	**21.75**	19.25	54.70
Open Source VLMs						
Llama-3.2V 11B	67.25	67.75	67.50	10.00	14.00	45.30
Gemma-3 12B	76.25	66.25	47.50	11.50	14.00	43.30
Qwen2.5-VL 7B	97.00	**94.25**	60.00	2.25	**25.75**	55.85
LLaVA-1.5 7B	8.25	59.50	45.00	4.75	2.00	23.90
Open Source Medical VLMs						
Med-LLaVa 7B	3.00	88.25	40.25	8.75	23.50	32.75
Med-Gemma 4B	0.00	23.25	45.00	0.00	15.25	16.70
CoT Prompt on Bangla Chest X-Ray QA Pair						
Closed Source VLMs						
Gemini 2.5	81.51	**64.75**	**80.75**	**23.25**	**24.50**	**49.33**
GPT-4.1 Mini	80.38	39.00	27.75	13.00	19.50	36.15
Open Source VLMs						
Llama-3.2V 11B	46.81	59.19	64.25	7.39	15.75	30.17
Gemma-3 12B	**94.19**	55.13	62.00	14.50	17.50	41.00
Qwen2.5-VL 7B	37.88	49.50	76.88	0.75	7.50	31.85
LLaVA-1.5 7B	16.25	0.38	0.00	5.50	4.38	6.05
Open Source Medical VLMs						
Med-LLaVa 7B	25.38	22.44	52.13	11.38	5.35	14.35
Med-Gemma 4B	25.38	57.69	47.00	0.63	12.13	22.50
CoT Prompt on English Chest X-Ray QA Pair						
Closed Source VLMs						
Gemini 2.5	**97.50**	66.75	**72.50**	**19.00**	13.00	53.75
GPT-4.1 Mini	97.00	56.00	55.00	9.75	16.50	46.85
Open Source VLMs						
Llama-3.2V 11B	79.50	86.75	48.75	10.00	**25.63**	50.12
Gemma-3 12B	97.00	62.50	57.50	8.25	7.50	46.55
Qwen2.5-VL 7B	96.50	**94.25**	67.50	3.00	21.50	**56.55**
LLaVA-1.5 7B	29.50	57.50	45.00	0.75	2.75	27.10
Open Source Medical VLMs						
Med-LLaVa 7B	0.00	89.25	40.75	0.25	26.00	31.25
Med-Gemma 4B	1.00	18.38	45.00	0.00	15.25	15.93

Table 3: Model performance for different categorical question on Chest X-ray Dataset

Models	Generalized Question		Specialized Question			Overall
	Modality	Organ	Abnormality	Condition	Position	
Base Prompt on Bangla MediCat QA Pair						
Closed Source VLMs						
Gemini 2.5	79.25	**61.63**	**91.50**	**29.75**	**38.25**	**60.08**
GPT-4.1 Mini	79.63	47.38	8.50	23.75	38.25	39.50
Open Source VLMs						
Llama-3.2V 11B	44.00	32.25	83.00	1.75	6.50	33.50
Gemma-3 12B	**89.88**	57.75	83.75	14.75	20.75	53.70
Qwen2.5-VL 7B	64.50	30.00	89.00	4.25	3.75	38.30
LLaVA-1.5 7B	15.75	0.00	76.75	0.50	0.75	18.75
Open Source Medical VLMs						
Med-LLaVa 7B	42.25	15.38	76.38	6.38	13.25	30.72
Med-Gemma 4B	36.13	16.25	67.13	0.00	19.50	27.80
Base Prompt on English MediCat QA Pair						
Closed Source VLMs						
Gemini 2.5	**93.50**	57.25	**95.00**	**30.50**	**39.50**	**63.15**
GPT-4.1 Mini	89.00	50.00	60.00	21.50	36.08	51.31
Open Source VLMs						
Llama-3.2V 11B	89.25	55.25	86.75	15.00	20.00	53.25
Gemma-3 12B	91.75	55.00	90.00	25.75	25.50	57.60
Qwen2.5-VL 7B	90.50	**65.00**	87.50	17.75	24.75	57.10
LLaVA-1.5 7B	55.00	52.50	92.50	16.25	18.75	47.00
Open Source Medical VLMs						
Med-LLaVa 7B	45.00	35.00	89.25	3.25	26.25	39.75
Med-Gemma 4B	45.00	32.50	92.50	0.00	30.00	40.00
CoT Prompt on Bangla MediCat QA Pair						
Closed Source VLMs						
Gemini 2.5	77.63	**61.75**	**91.50**	**31.25**	**40.75**	**60.58**
GPT-4.1 Mini	73.25	48.75	12.25	14.25	30.00	35.70
Open Source VLMs						
Llama-3.2V 11B	59.63	52.88	87.50	7.40	28.50	47.18
Gemma-3 12B	**91.63**	57.75	89.00	21.00	22.25	56.33
Qwen2.5-VL 7B	59.50	31.00	91.25	1.50	2.50	37.15
LLaVA-1.5 7B	18.50	0.00	0.00	0.25	4.00	4.55
Open Source Medical VLMs						
Med-LLaVa 7B	40.25	21.88	84.25	8.50	6.70	32.31
Med-Gemma 4B	45.00	34.88	74.00	0.00	19.25	34.63
CoT Prompt on English MediCat QA Pair						
Closed Source VLMs						
Gemini 2.5	**94.25**	55.25	**95.00**	**32.50**	**30.75**	**61.55**
GPT-4.1 Mini	90.75	50.50	92.50	30.25	27.58	58.31
Open Source VLMs						
Llama-3.2V 11B	79.75	**86.75**	49.50	10.13	25.13	50.25
Gemma-3 12B	90.25	47.50	92.50	28.00	19.25	55.50
Qwen2.5-VL 7B	91.50	55.00	85.00	12.25	26.50	54.50
LLaVA-1.5 7B	65.00	51.25	92.50	9.75	20.50	47.80
Open Source Medical VLMs						
Med-LLaVa 7B	37.50	35.00	87.00	8.25	28.50	39.25
Med-Gemma 4B	45.00	32.50	92.50	0.00	25.75	39.15

Table 4: Model performance for different categorical question on MedICaT dataset

B Used Prompts in the Paper

Prompt Used for LAVE Evaluation

Prompt for LAVE evaluation

Acts as a judge to compute LAVE scores for reference/prediction pairs.
Compare each prediction with its reference. Output **ONLY** a JSON list of floats (0 to 1). No extra text.

"You are a strict evaluator."
"Return ONLY a valid JSON array of floats between 0 and 1."
"Each float represents similarity between reference and prediction."
"If a reference answer is missing, treat it as 'No'."
"Do NOT return any text outside the JSON."

ZeroShot Prompt

ZeroShot Prompt for Bangla Med-VQA

You are a careful, clinically grounded medical vision-language assistant.
Task: You will be given a medical image and a single question about that image (e.g., modality, organ, abnormality, condition, or the position of the condition). Your job is to look at the image and provide the exact, minimal answer to the question.

CRITICAL: You must respond with **ONLY** the exact answer. No explanations, no sentences, no extra words.

STRICT OUTPUT RULES:

- ONE word or short phrase only
- NO sentences or explanations
- NO "This is..." or "The image shows..."
- NO punctuation unless part of the answer

Chain-of-Thought Prompt

CoT Prompt for Bangla Med-VQA

You are an expert medical vision-language assistant.
Task: You will be given a medical image and a single question about that image (e.g., modality, organ, abnormality, or specific finding). Your job is to think and reason step by step internally using the process below, then provide only the final answer without showing your reasoning.

Step-by-step internal reasoning (do not output this):

1. **Image Type Identification:**
    - Identify the imaging modality (X-ray, CT, MRI, ultrasound, etc.)
    - Note the anatomical region or body part being examined
2. **Visual Analysis:**
    - Observe anatomical structures and overall appearance
    - Check for abnormalities, lesions, devices, or unusual findings
    - Consider image quality, positioning, and technical factors
3. **Clinical Context Assessment:**
    - Recall the expected normal appearance for this view
    - Identify deviations from normal and their significance
4. **Question-Specific Reasoning:**
    - Link the visual findings directly to the question asked
    - Consider differential diagnoses only if needed to answer
5. **Evidence-Based Conclusion:**
    - Decide the most accurate answer supported by the image
    - Acknowledge uncertainty if evidence is insufficient

Output instructions (what to output):
CRITICAL: You must respond with **ONLY** the exact answer. No explanations, no sentences, no extra words.

STRICT OUTPUT RULES:

- ONE word or short phrase only
- NO sentences or explanations
- NO "This is..." or "The image shows..."
- NO punctuation unless part of the answer

Remember: Respond with ONLY the answer, nothing else.