Evaluating Large Vision Language Models on Bangla Medical Visual Question Answering

Rafid Ahmed[1,*], Intesar Tahmid[1,*], Mir Sazzat Hossain[1,2,*], Tasnimul Hossain Tomal[1], Md Mahir Jawad[1], Anam Borhan Uddin[1], Md Fahim[1,2,†], Md Farhad Alam Bhuiyan[1]

[1]*Penta Global Limited, Bangladesh*
[2]*Center for Computational & Data Sciences*
[*]Equal Contribution [†]Project Lead
Correspondence: {ahmedrafid023, intesar3006, fahimcse381, pdcsedu}@gmail.com

Abstract

Recent advancements in Large Language Models (LLMs) and Large Vision Language Models (LVLMs) have enabled general-purpose systems to demonstrate promising capabilities in complex reasoning tasks, including those in the medical domain. However, their evaluation has predominantly focused on high-resource languages, leaving low-resource contexts like Bangla underexplored. To address this gap, we introduce BanglaMedVQA, a multilingual Medical Visual Question Answering (VQA) dataset comprising clinically validated image–question–answer pairs, along with a comprehensive evaluation of current LVLMs on this resource. We rigorously evaluate nine state-of-the-art LVLMs using zero-shot, Chain-of-Thought (CoT), and LoRA fine-tuning strategies. Our results reveal a clear performance disparity: models perform well on generalized visual tasks but struggle with fine-grained diagnostic reasoning, achieving surprisingly low accuracy in specialized categories. While fine-tuning significantly improves overall accuracy, especially for Qwen2.5-VL and MedGemma 4B, limitations in specialized medical reasoning persist. Our work provides a foundation for future research in Bangla medical VQA. The code and dataset are available at https://ahmedrafid023.github.io/med-vqa-page/.

1 Introduction

In recent years, Large Language Models (LLMs), such as GPT (Achiam et al., 2023), LLaMA (Touvron et al., 2023), and PaLM (Anil et al., 2023), and Large Vision-Language Models (LVLMs), such as Gemini (Comanici et al., 2025), BLIP-2 (Li et al., 2023b), MiniGPT-v2 (Chen et al., 2023), and other multimodal architectures (Liu et al., 2023), have attracted significant attention for their ability to process, generate, and reason over complex multimodal input. These models produce human-like responses and achieve strong performance across a wide range of benchmarks. Their integration into the medical domain has already shown promising results, demonstrating substantial potential for real-world clinical applications.

Applications such as automated radiology report generation (Sloan et al., 2024), clinical decision support (Poulain et al., 2024), and interactive medical question answering (Kim et al., 2024) highlight the capability of LLMs to assist both patients and healthcare professionals. While LVLMs are increasingly being adopted across a wide range of medical applications, their evaluation in Bangla-language medical contexts has received limited attention. Medical visual question answering is commonly used to benchmark multimodal reasoning in the medical domain; however, existing benchmarks predominantly focus on English-language data (Liu et al., 2021; He et al., 2020; Zhang et al., 2023), leaving low-resource languages such as Bangla underexplored.

To systematically evaluate the capabilities of LVLMs in the Bangla medical domain, we adopt medical visual question answering as the evaluation paradigm, as it provides a standardized and interpretable framework for assessing multimodal reasoning. However, the lack of high-quality Bangla medical VQA datasets poses a major obstacle to such evaluation. To address this gap, we construct a new Bangla Medical VQA dataset using images drawn from multiple medical domains (Subramanian et al., 2020; Wang et al., 2017), paired with diverse and clinically relevant questions. The dataset contains 7,000 unique image–question 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.

We evaluate both open-source and closed-source

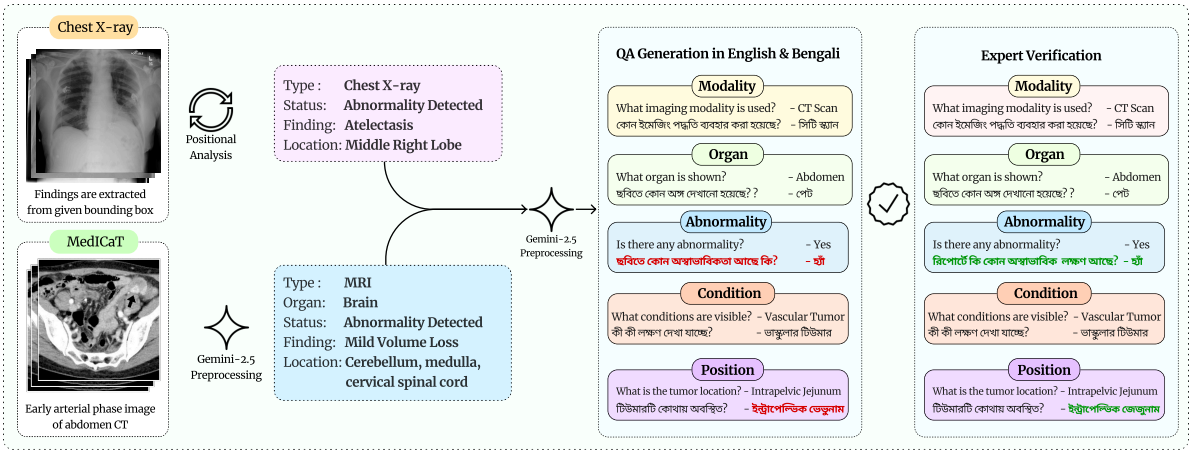


Figure 1: Workflow of the dataset curation process. Images and metadata were obtained from two widely used biomedical datasets, enabling the automatic generation of QA pairs, which were subsequently verified by domain experts.

LVLMs on our proposed dataset. Our experiments reveal that current LVLMs perform reasonably on generalized questions but struggle significantly on specialized diagnostic categories such as *Condition* and *Position*, often performing surprisingly low. This highlights critical gaps in the reliability of existing models for fine-grained medical reasoning. Previous studies have reported worse than random behavior on English medical VQA tasks (Yan et al., 2025); however, the problem is even more pronounced for Bangla, a low-resource language. Comparative experiments further confirm that closed-source LVLMs consistently outperform open-source models. Among the evaluated models, Gemini and Claude emerge as the best performers, achieving overall accuracies of 35.75% and 32.25%, respectively.

We also investigated whether fine-tuning could improve model performance. Our experiments show that fine-tuned models substantially improve overall reasoning ability, though they still exhibit limited capability on specialized diagnostic categories. For example, Qwen2.5-VL and MedGemma 4B-it become the top performers after fine-tuning, achieving overall accuracies of 58.25% and 57.50%, respectively, surpassing the performance of closed-source models. However, for specialized categories, the best-performing model, Qwen2.5-VL, only achieves 18.50% and 31.25% LAVE (Mañas et al., 2024) scores in the *Condition* and *Position* categories, respectively. These results indicate that, although fine-tuning improves general reasoning ability, substantial work remains to address fine-grained medical reasoning in domain-specific question categories. In particular, our paper makes the following contributions:

- We introduce a new Bangla Medical VQA dataset comprising diverse and clinically relevant QA pairs across multiple diagnostic categories for medical images and evaluate nine open and closed source VLMs on this benchmark.
- We analyze the impact of prompting strategies by systematically comparing zero-shot and Chain-of-Thought prompting and further perform LoRA fine-tuning on open-source models, demonstrating notable performance gains across evaluation settings.
- We find that while LVLMs perform well on generalized medical questions, they continue to struggle with specialized diagnostic queries, indicating that current models are not yet reliable for integration into real-world clinical decision-making.

2 Related Work

Visual Question Answering VQA has emerged as a challenging task at the intersection of vision and language. Antol et al. (2015) introduced a benchmark dataset designed to evaluate fine-grained visual understanding and commonsense reasoning. Shih et al. (2016) introduced a region-focused attention mechanism, aligning image regions with question embeddings. Shah et al. (2019) addressed knowledge-intensive scenarios with KVQA. More recently, MTVQA by Tang et al. (2025) introduced

a benchmark across nine languages to address the visual-textual misalignment in translation-based multilingual text-centric VQA datasets. LRCN by Han et al. (2025) mitigates information loss in deep VQA transformers by introducing a plug-and-play Layer-Residual Mechanism.

Medical Visual Question Answering Gu et al. (2024) proposed a novel architecture for medical VQA that generates answer-constrained latent prompts. Ultimately, it reported state-of-the-art gains on VQA-RAD (Narayanan et al., 2024), SLAKE (Liu et al., 2021). Yim et al. (2024) delivered a multilingual dermatology VQA dataset and benchmarks for consumer-generated images and free-text clinician responses. Xu et al. (2024) proposed a multi-level visual language model for MVQA, introducing a new instruction dataset (MLe-VQA), a feature alignment module, and an evaluation benchmark (MLe-Bench). Yu et al. (2025b) introduced adaptive region-level visual prompts and a hierarchical answer generator with PEFT techniques. Guo and Terzopoulos (2025) proposed a novel prompting strategy for Medical LVLMs that reduces hallucination and improves VQA performance.

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. Rafi et al. (2022) introduced the first human-annotated Bengali VQA dataset using images from VQA v2.0. Despite advancements in the general VQA, the medical domain for the Bangla language remains under-resourced and unexplored. Recently, (Ahmed et al., 2026) explores the effectiveness of LLMs in Bangla Medical domain.

3 Dataset Creation

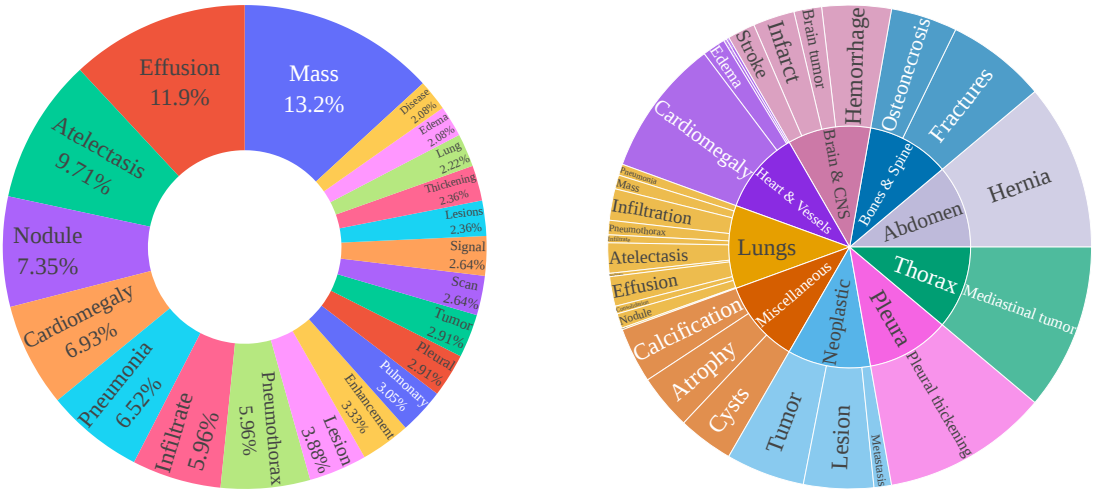
3.1 Data Collection

We constructed BanglaMedVQA by leveraging the proven scale and complementary strengths of two widely used and validated biomedical datasets: ChestX-ray8 (Wang et al., 2017) and MedICaT (Subramanian et al., 2020). The ChestX-ray8 dataset focuses exclusively on chest radiographs and provides structured condition metadata along with explicit positional annotations in the form of bounding boxes. To expand the dataset beyond chest X-rays and cover a broader range of medical imaging modalities and anatomical regions, we additionally incorporate MedICaT, which includes diverse clinical images such as radiology figures, medical illustrations, and pathology-related visuals, accompanied by detailed condition information and positional descriptions embedded within image caption pairs.

Previous research on medical VQA generation and benchmarking (Zhang et al., 2023; Li et al., 2023c; Yan et al., 2025) has demonstrated that few-shot generation and structured QA pipelines can yield clinically valid and semantically aligned question–answer pairs. Following this line of evidence, we adopted a similar methodological approach to maintain both the clinical integrity and linguistic coherence of our generated data. For each image, a standardized metadata record was created, including the imaging modality, anatomical region, and a list of detected clinical findings along with their positional descriptions. In the case of MedICaT, Gemini was used with few-shot prompting to extract information about abnormalities and their locations from captions. For ChestX-Ray8, a specialized positional reasoning module generated descriptive text from bounding box coordinates, as illustrated in Figure 1.

3.2 Question Generation

Prior studies have shown that Gemini models demonstrate stronger performance and reliability in medical reasoning and multimodal clinical tasks compared to other models (Yu et al., 2025a; Yan et al., 2025). Motivated by these findings, we selected Gemini as the backbone model for VQA generation. From the curated metadata and images, we then generated English VQA pairs using the `gemini-2.5-flash` model with a few-shot prompting strategy. Subsequently, to support multilingual accessibility and facilitate research in Bangla clinical contexts, we employed the `gemini-2.5-flash` model to translate the English QA pairs into Bangla. This two-stage pipeline ensures both linguistic diversity and semantic consistency between the metadata and the generated VQA pairs.



(a) Distribution of Question Keyword in the Dataset.

(b) Distribution of Conditions in the Dataset.

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

3.3 Question and Answer Verification

The translated QA pairs may exhibit linguistic inconsistencies or positional mismatches between textual descriptions and visual regions. To rigorously evaluate the reliability of the generated Bangla QA pairs, we carried out an expert validation study on the 1000-sample subset. Two medical specialists were compensated on an hourly basis to independently examine and verify the annotations. This focused review covered two high-risk areas.

1. QA Validation and Translation Check The specialists verified the clinical correctness and visual grounding of the original English VQA pairs, ensuring medical validity and image consistency. The corresponding Bangla translations were then reviewed for accurate medical terminology and structural decoding errors common in low-resource clinical translation.

2. Anatomical Position Verification: The specialists verified whether the LLM-inferred positional information correctly matched the pathological regions visible in the images.

The average acceptance rate across both validators and both quality checks demonstrated an overall accuracy of 97%. Based on this high level of verification and the strong inter-rater agreement, the medical specialists provided a favorable verdict, confirming the robustness and clinical reliability of the proposed LLM-driven generation method. Following this successful validation, we proceeded to generate the remaining portion of the dataset using the identical two-stage pipeline, culminating in the final set of 7,000 unique VQA pairs.

QA Pair Validation		
Samples used	1000	
Annotator	**Acc. (#)**	**Acc. (%)**
Annotator 1	972	97.20
Annotator 2	976	97.60
Inter-Rater Reliability (Cohen's Kappa)		
Samples used	1000	
Cohen's κ coefficient	0.89	

Table 1: Verification statistics of the curated dataset. Accuracy refers to the correctness of QA pairs validated by medical specialists, and Cohen's κ measures inter-rater reliability.

4 BanglaMedVQA Data Analysis

4.1 Data Statistics

The proposed dataset comprises a total of 7,000 VQA instances. The data are sourced equally from two prominent biomedical datasets, 3,500 instances from ChestX-ray8 and 3,500 from MedICaT. The dataset is balanced with respect to healthy and abnormal cases and is uniformly distributed across five question categories: modality, organ, abnormality, condition, and position, each representing approximately 20% of the total questions. Analysis of question keywords, as shown in Figure 2a, illustrates a strong focus on imaging interpretation, anatomical localization, and abnormality assessment. The most frequent terms include *abnormality*, *condition*, *organ*, *technique*, and specific findings such as *mass*, *effusion*, *atelectasis*, *cardiomegaly*, *nodule*, *pneumonia*, and *infiltrate*, reflecting the dataset's emphasis on clinical image

understanding and diagnostic reasoning.

4.2 Clinical Conditions

The distribution of clinical conditions in the dataset is summarized in Figure 2b. Organ-wise, the majority of conditions target the *Lungs*, followed by *Heart & Vessels*, *Brain & CNS*, *Pleura*, *Bones & Spine*, *Thorax*, *Abdomen*, and *Neoplastic* cases. Within each organ/system, the dataset covers a wide range of conditions, with more frequent labels including *Infiltration*, *Atelectasis*, *Effusion*, *Cardiomegaly*, *Nodule*, *Pneumothorax*, *Mass*, *Pneumonia*, *Consolidation*, *Pleural thickening*, *Edema*, *Emphysema*, *Tumor*, and *Lesion*, as well as rarer conditions such as *Fibrosis*, *Hemorrhage*, *Calcification*, *Fractures*, *Cysts*, *Metastasis*, and *Stroke*.

5 Experiment Design

5.1 Experimentation with Bangla

Zero-Shot Prompting. To assess the medical reasoning capabilities of 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.

Chain-of-Thought (CoT) Prompting. For 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.

LoRA Finetuning. We also experiment with LoRA (Hu et al., 2022) finetuning approach to see the impact of further visual and textual alignment will help the the performance. For the pretrained model ϕ_M with frozen weights $\mathbf{W}$, LoRA fine-tunes via a low-rank update $\Delta\mathbf{W} = \mathbf{A}\mathbf{B}^T$, where $\mathbf{A} \in \mathbb{R}^{d\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times d}$, with $r \ll d$. Only $\mathbf{A}$ and $\mathbf{B}$ are trainable, reducing parameter count. The adapted weights for inference are: $\mathbf{W}' = \mathbf{W} + \mathbf{A}\mathbf{B}^T$. We input ϕ_M with I, and Q_{BAN}, and train ϕ_M using the loss computed from A_{BAN}.

5.2 Experimentation with English

To investigate the effect of language variation, we further extend our experiments to include English-language questions. Specifically, since the Bangla question–answer pairs were originally translated from English, we utilize their corresponding English versions Q_{ENG} and A_{ENG} for each medical image I in this experiment to ensure consistency and enable cross-linguistic comparison. 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 all the experiments. The details about the model & its configuration and evaluation metrics are provided on Appendix A.

6 Result and Analysis

Results on Bangla and English QA pairs are presented in Tables 2, 3 and Figure 3 respectively.

Bangla QA Pairs. Overall, closed-source VLMs substantially outperform both open-source and medical-domain counterparts under zero-shot and chain-of-thought prompting. Gemini 2.5 Flash achieves the highest LAVE scores with an average of 55.44% under zero-shot prompting. Claude Sonnet-4 follows closely, achieving a LAVE score of 45.28%. 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 an average of 46.54% 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, with accuracies frequently dropping below 10% in Bangla.

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. Notably, medical-domain VLMs show poor results, with Med-LLaVA exhibiting near-zero accuracy on both Chest X-Ray and MedICaT.

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-

Models	Chest X-Ray			MedICaT			Overall (Avg.)		
	Acc	BScore	LAVE	Acc	BScore	LAVE	Acc	BScore	LAVE
Zero Shot Prompt on Bangla QA Pairs									
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
Claude Sonnet 4	34.50	80.80	43.08	29.00	80.22	47.47	31.75	80.51	45.28
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 Pairs									
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
Claude Sonnet 4	33.50	81.32	43.63	31.00	81.95	51.60	32.25	81.64	47.62
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
LoRA Fine-Tuning on Bangla QA Pairs									
Open Source VLMs									
Llama-3.2V 11B	47.00	92.77	70.70	48.00	89.00	62.78	47.50	90.89	66.74
Gemma-3 12B	44.00	92.13	67.45	48.50	88.57	62.60	46.25	90.35	65.03
Qwen2.5-VL 7B	47.00	92.64	69.75	**51.50**	89.08	**64.60**	49.25	90.86	67.18
LLaVA-1.5-7B	42.50	91.55	64.65	46.60	87.43	58.22	44.55	89.49	61.44
Open Source Medical VLMs									
Med-LLaVa 7B	43.00	91.68	65.38	47.50	87.23	58.57	45.25	89.46	61.98
Med-Gemma 4B	**52.00**	**94.06**	**76.55**	50.00	**89.56**	63.35	**51.00**	**91.81**	**69.95**

Table 2: Model Benchmarking for **Bangla** with average scores across ChestX-Ray and MedICaT datasets.Blue highlights the highest-performing model in the Benchmark, while Cyan marks the best-performing models for each metric across the model types.

3.2V 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, though the gains are more modest. In contrast, medical VLMs such as Med-LLaVA generally show negligible gains. Med-Gemma, however, demonstrates substantial improvements under CoT prompting.

Impact of LoRA Fine-Tuning. LoRA fine-tuning brings a clear performance boost across both Bangla and English QA pairs, surpassing all zero-shot and CoT settings. In the Bangla VQA setup (Table 2), open-source models such as Qwen2.5-VL and Med-Gemma reach overall LAVE scores of 67.18% and 69.95%, showing major gains over their earlier results. Accuracy also rises notably, with most models scoring above 45%. Models pre-trained on medical data like Med-Gemma and Med-LLaVA, which struggled before, improve sharply after fine-tuning, showing LoRA's ability to bridge domain and language gaps through lightweight adaptation. A similar pattern appears in English (Table 3), where fine-tuned models reach or even match closed-source performance. Llama-3.2V and Med-Gemma achieve LAVE scores above 75%, while Qwen2.5-VL and Gemma-3 also perform strongly across both datasets. Overall, LoRA fine-tuning helps open-source and medical VLMs improve their performance, reducing the gap with

Models	ChestX-Ray			MedICaT			Overall (Avg.)		
	Acc	BScore	LAVE	Acc	BScore	LAVE	Acc	BScore	LAVE
Zero Shot Prompt on English QA Pairs									
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
Claude Sonnet 4	41.00	66.56	51.97	47.00	70.48	57.40	44.00	68.52	54.69
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 Pairs									
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
Claude Sonnet 4	38.50	65.25	48.65	47.50	70.67	57.70	43.00	67.96	53.18
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
LoRA Fine-Tuning on English QA Pairs									
Open Source VLMs									
Llama-3.2V 11B	**68.00**	**88.50**	**82.53**	64.00	**78.42**	**70.37**	66.00	**83.46**	**76.45**
Gemma-3 12B	60.00	85.40	73.18	**65.00**	76.59	67.61	62.50	80.99	70.40
Qwen2.5-VL 7B	62.50	85.13	74.15	**65.00**	76.21	68.45	63.75	80.67	71.30
LLaVA-1.5-7B	63.50	84.67	74.93	59.50	70.47	61.15	61.50	77.57	68.04
Open Source Medical VLMs									
Med-LLaVa 7B	62.00	84.22	71.88	60.00	70.83	61.00	61.00	77.53	66.44
Med-Gemma 4B	**68.00**	88.21	82.32	**65.00**	76.21	69.65	**66.50**	82.21	75.99

Table 3: Model Benchmarking for **English** with average scores across ChestX-Ray and MedICaT datasets. **Blue** highlights the highest-performing model in the Benchmark, while **Cyan** marks the best-performing models for each metric across the model types.

closed-source models in medical QA tasks.

Performance Across Question Categories. We further analyze model behavior across different question types, as shown in Figures 4–5 and Tables 4–5(in Appendix B). The questions are grouped into *general* (Modality, Organ) and *specialized* (Abnormality, Condition, Position) categories.

In the zero-shot setting, a consistent pattern emerges across both datasets and languages. Among closed-source models, Gemini 2.5 achieves the strongest performance, with average accuracies of 87.5% on Modality and 80.8% on Abnormality across the dataset. GPT-4.1 Mini, Claude-4 show similar gaps, while open-source models like Gemma-3 and Qwen2.5-VL perform competitively on general but poorly on specialized questions. Medical VLMs such as Med-LLaVA and Med-Gemma show limited multilingual adaptation. Under CoT prompting, moderate gains appear in specialized reasoning, but the overall pattern remains unchanged. Gemini 2.5 and Gemma-3 improve slightly, while medical VLMs got little gain.

LoRA fine-tuning leads to substantial and consistent improvements across all categories and languages. When averaged across MedICaT and ChestX-Ray8, Abnormality accuracy exceeds 90% in Bangla, while Condition improves to approxi-

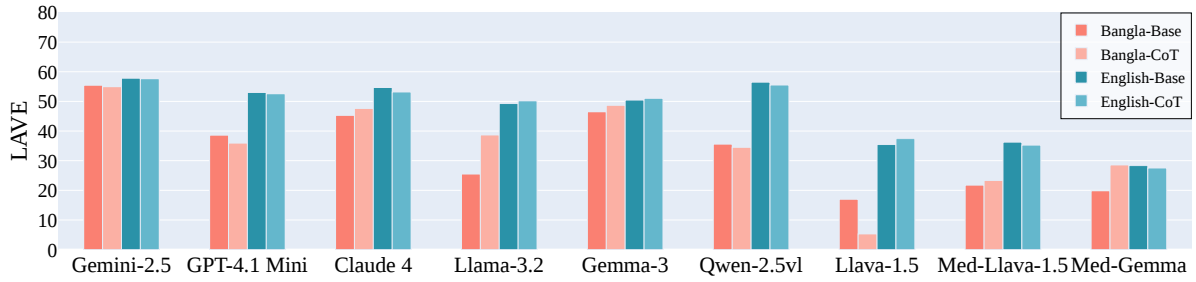
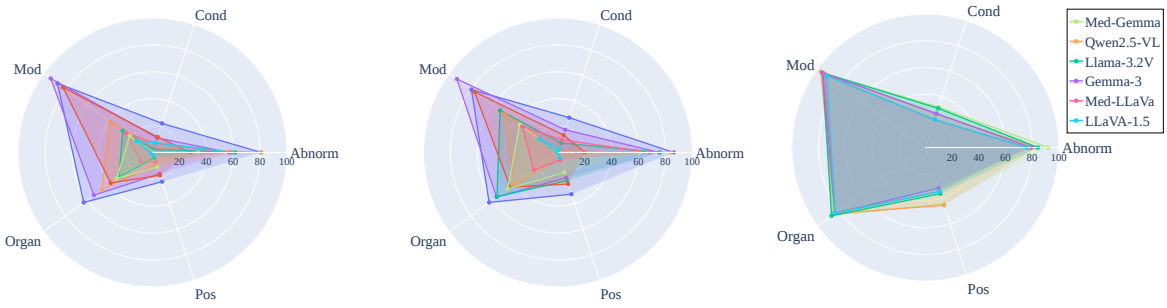


Figure 3: LAVE score comparison across all models on the full dataset under four evaluation settings: Bangla (Base), Bangla (Chain-of-Thought), English (Base), and English (Chain-of-Thought).



(a) Zero-Shot (b) CoT Prompt (c) LoRA Fine Tuning

Figure 4: Comparison of LAVE scores across categorical question types on the overall dataset for Bangla QA pair. The three plots depict performance for (a) the base model, (b) the model with chain-of-thought reasoning, and (c) the model with LoRA fine-tuning.Here *Mod* refers to *Modality*, *Abnorm* refers to *Abnormality*,*Cond* referes to *Condition*, and *Pos* refers to *Position*.

mately 30%, with similar trends observed in English. Med-Gemma 4B and Qwen2.5-VL 7B achieve the most balanced results (e.g., 95.6% and 98.4% on Modality, 45.4% and 54.6% on Position). The radar plots in Figures 4c and 5c clearly show these consistent category-wise gains. Nevertheless, even in this setting, performance on Condition and Position remains markedly lower than on general categories, underscoring the persistent difficulty of diagnostic reasoning.

Cross-Lingual Observations. Comparing Bangla and English benchmarks (Figure 3), we find that performance in Bangla is generally weaker, particularly for open-source and domain-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 VLMs, we observed that while top-performing models such as Gemini, GPT-4.1 Mini, and Claude Sonnet-4 achieve reasonable performance on general questions, they struggle severely on specialized diagnostic tasks. Open-source models, including Gemma-3, occasionally outperform closed-source models on general questions but similarly fail on fine-grained medical queries. Incorporating LoRA fine-tuning provides substantial improvements, enabling models such as Qwen2.5-VL and Med-Gemma 4B-it to surpass closed-source baselines. Despite these gains, performance on domain-

specific categories such as Condition and Position remains limited, indicating that current VLMs still lack fine-grained medical reasoning capabilities. The BanglaMedVQA dataset thus serves as a crucial benchmark to drive future progress in this area.

Limitations

Despite the contributions of this work, several limitations remain. Although overall model performance improves, accuracy on fine-grained and specialized categories such as *Condition* and *Position* remains limited. In addition, due to resource and scalability constraints, it was not feasible to conduct expert validation for all 7,000 VQA pairs. Finally, while the dataset spans multiple medical domains, further expansion in image modalities and more extensive expert-verified annotations could strengthen its clinical breadth and robustness.

To build effective real-world healthcare AI systems, it is important to account for widely used text forms, particularly transliterated content in low-resource languages such as Bangla (Fahim et al., 2024; Haider et al., 2025). Addressing this challenge presents a valuable direction for future research. Additionally, (Ahmed et al., 2024) explores various strategies for enhancing LLM performance on transliterated text. It would be worthwhile to investigate whether these techniques are equally effective within the medical domain. Also current work can easily be extended to investigate the model performance of Bangla dialects following previous work by (Jawad et al., 2025).

Ethical Statement

This work involves the creation and use of a medical visual question-answering dataset in Bangla. All images and associated annotations were sourced from publicly available datasets or generated with appropriate institutional permissions. Annotations and dataset curation were performed by qualified medical experts to ensure accuracy and minimize the risk of misinformation. We acknowledge that models trained on this dataset are intended for research purposes only and should not be used for clinical decision-making. Patient data privacy has been maintained throughout, and no personally identifiable information is included in the dataset. We encourage responsible use of both the dataset and models, and recommend thorough evaluation before any deployment in real-world medical contexts. Additionally, LLMs were utilized during the drafting process to assist with grammatical refinement and structural clarity; however, all technical content and final interpretations were produced and verified by the authors.

References

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

Fahim Ahmed, Md Fahim, Md Ashraful Amin, Amin Ahsan Ali, and AKM Rahman. 2024. Improving the performance of transformer-based models over classical baselines in multiple transliterated languages. In *ECAI 2024*, pages 4043–4050. IOS Press.

Rafid Ahmed, Intesar Tahmid, Mir Sazzat Hossain, Tasnimul Hossain Tomal, Md Fahim, and Md Farhad Alam Bhuiyan. 2026. How good llms are at answering bangla medical visual questions? dataset and benchmarking. In *AAAI Bridge Program on AI for Medicine and Healthcare*, pages 1–14. PMLR.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, and 1 others. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

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, and 1 others. 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, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Md Fahim. 2023. Aambela at blp-2023 task 2: Enhancing banglabert performance for bangla sentiment analysis task with in task pretraining and adversarial weight perturbation. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 317–323.

Md Fahim, Fariha Tanjim Shifat, Fabiha Haider, Deeparghya Dutta Barua, MD Sakib Ul Rahman Sourove, Md Farhan Ishmam, and Md Farhad Alam Bhuiyan. 2024. Banglatlit: A benchmark dataset for back-transliteration of romanized bangla. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14656–14672.

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.

Danfeng Guo and Demetri Terzopoulos. 2025. Prompting medical large vision-language models to diagnose pathologies by visual question answering. *Machine Learning for Biomedical Imaging*, 3(March 2024):59–71.

Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Md Sakib Ul Rahman Sourove, Deeparghya Dutta Barua, Md Fahim, and Md Farhad Alam Bhuiyan. 2025. Banth: A multi-label hate speech detection dataset for transliterated bangla. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7217–7236.

Dezhi Han, Jingya Shi, Jiahao Zhao, Huafeng Wu, Yachao Zhou, Ling-Huey Li, Muhammad Khurram Khan, and Kuan-Ching Li. 2025. Lrcn: Layer-residual co-attention networks for visual question answering. *Expert Systems with Applications*, 263:125658.

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*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Md Mahir Jawad, Rafid Ahmed, Ishita Sur Apan, Tasnimul Hossain Tomal, Fabiha Haider, Mir Sazzat Hossain, and Md Farhad Alam Bhuiyan. 2025. Benchmarking large language models on bangla dialect translation and dialectal sentiment analysis. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*, pages 322–337.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. Medexqa: Medical question answering benchmark with multiple explanations. *arXiv preprint arXiv:2406.06331*.

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.

Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. 2023c. Self-supervised vision-language pretraining for medial visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

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.

Abhishek Narayanan, Rushabh Musthyala, Rahul Sankar, Anirudh Prasad Nistala, Pranav Singh, and Jacopo Cirrone. 2024. Free form medical visual question answering in radiology. *arXiv preprint arXiv:2401.13081*.

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

Raphael Poulain, Hamed Fayyaz, and Rahmatollah Beheshti. 2024. Bias patterns in the application of llms for clinical decision support: A comprehensive study. *arXiv preprint arXiv:2404.15149*.

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.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 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.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, An-Lan Wang, Chunhui Lin, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. 2025. MTVQA: Benchmarking multilingual text-centric visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7748–7763, Vienna, Austria. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 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, and 1 others. 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, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. 2025. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical VQA. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19188–19205, Vienna, Austria. Association for Computational Linguistics.

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.

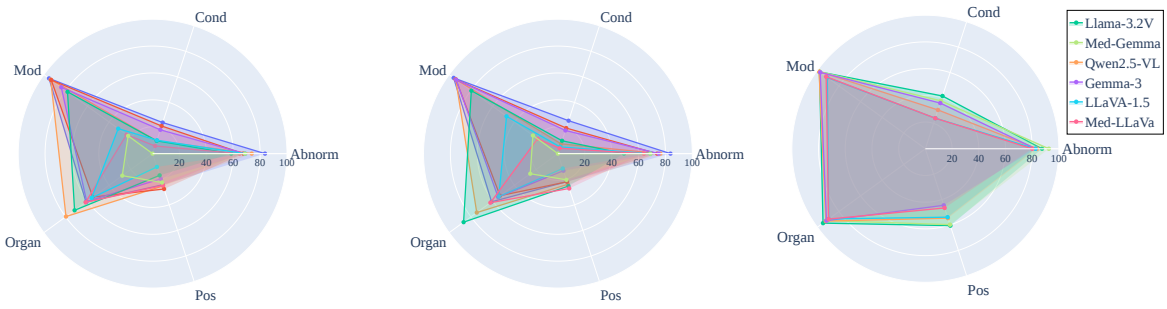
Suhao Yu, Haojin Wang, Juncheng Wu, Cihang Xie, and Yuyin Zhou. 2025a. Medframeqa: A multi-image medical vqa benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*.

Ting Yu, Zixuan Tong, Jun Yu, and Ke Zhang. 2025b. 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 Experimented Setup

Experimented Models. We conduct a systematic evaluation of nine VLMs, 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,



(a) Zero Shot (b) CoT Prompt (c) LoRA Fine Tuning

Figure 5: Comparison of LAVE scores across categorical question types on the overall dataset for English QA pair. The three plots depict performance for (a) the base model, (b) the model with chain-of-thought reasoning, and (c) the model with LoRA fine-tuning.Here *Mod* refers to *Modality*, *Abnorm* refers to *Abnormality*,*Cond* referes to *Condition*, and *Pos* refers to *Position*.

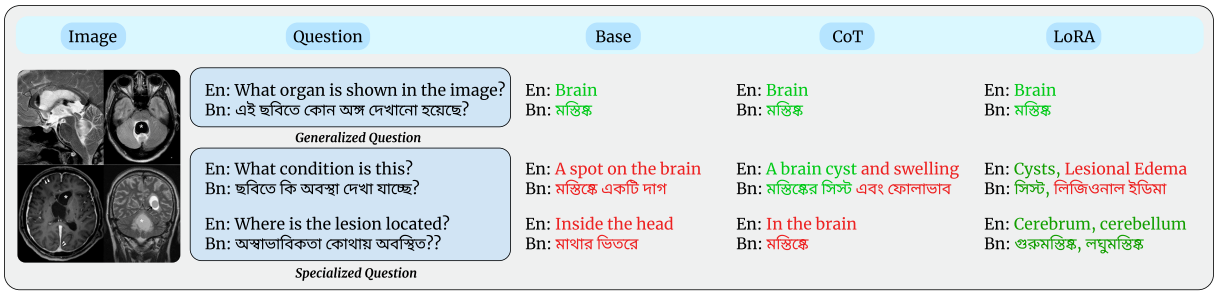


Figure 6: Error Analysis of Medical Visual Question Answering (MedVQA) pairs from the proposed dataset, showcasing aligned English–Bangla questions and clinically validated answers across categories such as condition and abnormality localization.

2023), Claude Sonnet 4 (Anthropic, 2024), 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 VLMs: Med-LLaVA (Li et al., 2023a) and Med-Gemma (Sellergren et al., 2025).

Model Configuration. While zero-shot and CoT prompting were evaluated on both closed and open-source LVLMs, for LoRA fine-tuning, we used the open-source models due to their modifiable architectures and permissible licensing. We fine-tuned the models using LoRA on the MedICaT and the ChestX-Ray8 dataset. The adapter was applied to all linear layers (`lora_target:all`) with a rank of 16 and an alpha value of 32. The model was trained for 2 epochs with a global batch size of 16, using the AdamW optimizer and a linear learning rate scheduler with a warm-up ratio of 0.03 and a learning rate of $2\text{x}10^{-5}$ following previous ablation from (Fahim, 2023). The maximum sequence length was set to 512 tokens. We used an NVIDIA A100 GPU with 80GB of VRAM.

Evaluation Metrics. Performance of the models is reported using three complementary metrics: Accuracy (Acc), BERTScore (BScore), and LLM-Assisted VQA Evaluation (LAVE) (Mañas et al., 2024). 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. The LAVE metric is designed to provide a more reliable evaluation by employing a large language model as an automatic judge. In our experiments, we used `GPT-4.1-mini` as the judging model to compute LAVE scores.

B Further Experiment Results

B.1 Error Analysis

Our qualitative analysis shows that while the models generally handle straightforward questions such as identifying the imaging modality or the organ involved, they struggle with more specialized diagnostic reasoning. Errors are most frequent in questions requiring precise localization or interpretation of pathological findings, where models

often confuse nearby anatomical regions or misclassify related conditions such as cysts and edema. In several cases, responses were partially correct, capturing general abnormalities but missing more precise clinical details. These limitations highlight gaps in visual–textual grounding and domain understanding. Nevertheless, fine-tuning with LoRA techniques noticeably improves accuracy in such complex cases.

B.2 Question Category-Wise Performance Comparison

Tables 4 and 5 present category-wise results for all evaluated models under three experimental settings, Base, Chain-of-Thought (CoT), and LoRA fine-tuning, on both the *ChestX-Ray8* and *MedICaT* datasets. The questions are divided into *general* categories (*Modality*, *Organ*) and *specialized* categories (*Abnormality*, *Condition*, *Position*).

In the Bangla benchmark, closed-source models dominate under the base prompt. `Gemini-2.5` achieves an average of 87.5% on *Modality* and 80.75% on *Abnormality*, substantially outperforming open-source models in general questions. However, accuracy falls sharply for specialized categories, where *Condition* and *Position* drop to 22.75% and 22.87%, respectively. Open-source models such as `Gemma-3` (93.69% on *Modality*) and `Qwen2.5-VL` (80.75% on *Abnormality*) show partial strength in general categories but remain limited in reasoning-heavy ones.

Under CoT prompting, closed-source models exhibit modest improvements across all categories. `Gemini-2.5` increases to 86.12% for *Abnormality*, 27.25% for *Condition*, and 32.62% for *Position*, suggesting that stepwise reasoning benefits certain visual–clinical inferences. Open-source models also gain from CoT, with `Gemma-3` improving to 75.50% on *Abnormality* and `Qwen2.5-VL` reaching 84.06% on the same category. Nonetheless, specialized reasoning in Bangla remains consistently weaker than in English.

LoRA fine-tuning produces the most substantial gains across categories and model families. For Bangla, fine-tuned models reach benchmark highs: `Med-Gemma` attains 95.50% on *Modality*, 92.00% on *Abnormality*, and 31.87% on *Condition*, while `LLaMA-3.2V` records 95.06% on *Modality* and 36.31% on *Position*. `Qwen2.5-VL` performs competitively with 96.00–97.50% in general categories and 45.45% in *Position*. These results indicate that parameter-efficient LoRA fine-tuning effectively enhances both visual and linguistic alignment for Bangla MedVQA.

In the English benchmark, overall accuracy is higher and more stable across categories. Under the base prompt, `Gemini-2.5` achieves 95.25% in *Modality* and 83.62% in *Abnormality*, while open-source `Qwen2.5-VL` reaches an average of 93.75% and 94.25% in *Modality* and *Organ*, respectively. CoT prompting provides small but consistent boosts, particularly for `Gemini 2.5`, which increases to 95.87% in *Modality* and 25.75% in *Condition*.

With LoRA fine-tuning, the English benchmark achieves its highest results. `LLaMA-3.2V` obtains the top *Organ* (94.75%) and *Position* (60.53%) accuracies, while `Med-Gemma` peaks at 92.50% on *Abnormality* and 37.75% on *Condition*. `Qwen2.5-VL` matches the highest *Modality* score (98.37%) and achieves 83.75% on *Abnormality*. These results confirm that LoRA adaptation consistently improves specialized diagnostic understanding while maintaining strong general performance.

C Used Prompts in the Paper

VQA Creation Prompt

Structured Prompt for Med-VQA Generation

You are a clinically grounded medical vision-language assistant. Your task is to create structured question–answer (QA) pairs from a given medical image and its metadata.

We Provide:

- A medical image
- Structured metadata describing that image

Metadata Example: `{ "modality": "X-ray", "organ": "Lung", "abnormality": ["Nodule"], "condition": ["Malignancy"], "position": ["Upper lobe of the right lung"] }`

Task Instructions:

1. For each category — modality, organ, abnormality, condition — generate one structured question. Keep each answer as short as possible.
2. For the position category, generate one or more questions depending on the data in the metadata.
3. For the condition category:
   - Generate a single question such as: "What specific conditions are identified here?"
   - Include all conditions listed in the metadata in the answer, along with their positions if available.
4. For position, generate QA about the anatomical location of each disease.
5. Derive all answers only from the metadata and the image.
6. Do not invent findings that are not present in the metadata.

Output Format:

Return JSON in the following structure: `{ "image_id": "IMG_001", "qa_pairs": [ { "category": "modality", "question": "What is the imaging modality?", "answer": "X-ray" }, { "category": "organ", "question": "Which organ is the focus of this image?", "answer": "Lung" }, . . . ] }`

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.

Models	Chest X-Ray					MedICaT				
	Generalized Question		Specialized Question			Generalized Question		Specialized Question		
	Mod	Organ	Abnorm	Cond	Pos	Mod	Organ	Abnorm	Cond	Pos
Base Prompt on Bangla QA Pairs										
Closed Source VLMs										
Gemini 2.5	95.75	65.00	70.00	15.75	7.50	79.25	61.63	**91.50**	29.75	38.25
GPT-4.1 Mini	85.25	29.75	47.50	11.00	15.00	79.63	47.38	8.50	23.75	38.25
Claude Sonnet 4	**98.25**	36.37	60.00	12.00	8.75	69.25	45.00	84.12	16.75	22.25
Open Source VLMs										
Llama-3.2V 11B	11.00	29.50	40.00	3.00	1.75	44.00	32.25	83.00	1.75	6.50
Gemma-3 12B	97.50	50.25	30.00	8.00	12.75	89.88	57.75	83.75	14.75	20.75
Qwen2.5-VL 7B	13.00	63.50	72.50	2.85	15.25	64.50	30.00	89.00	4.25	3.75
LLaVA-1.5 7B	13.50	2.00	42.50	14.75	3.25	15.75	0.00	76.75	0.50	0.75
Open Source Medical VLMs										
Med-LLaVa 7B	6.75	6.75	39.75	8.00	2.25	42.25	15.38	76.38	6.38	13.25
Med-Gemma 4B	6.00	50.00	0.00	0.50	3.00	36.13	16.25	67.13	0.00	19.50
CoT Prompt on Bangla QA Pairs										
Closed Source VLMs										
Gemini 2.5	81.51	64.75	80.75	23.25	24.50	77.63	61.75	**91.50**	**31.25**	**40.75**
GPT-4.1 Mini	80.38	39.00	27.75	13.00	19.50	73.25	48.75	12.25	14.25	30.00
Claude Sonnet 4	98.00	43.12	57.50	9.75	9.75	79.25	47.50	88.75	20.00	22.50
Open Source VLMs										
Llama-3.2V 11B	46.81	59.19	64.25	7.39	15.75	59.63	52.88	87.50	7.40	28.50
Gemma-3 12B	94.19	55.13	62.00	14.50	17.50	91.63	57.75	89.00	21.00	22.25
Qwen2.5-VL 7B	37.88	49.50	76.88	0.75	7.50	59.50	31.00	91.25	1.50	2.50
LLaVA-1.5 7B	16.25	0.38	0.00	5.50	4.38	18.50	0.00	0.00	0.25	4.00
Open Source Medical VLMs										
Med-LLaVa 7B	25.38	22.44	52.13	11.38	5.35	40.25	21.88	84.25	8.50	6.70
Med-Gemma 4B	25.38	57.69	47.00	0.63	12.13	45.00	34.88	74.00	0.00	19.25
LoRA Fine-Tuning on Bangla QA Pairs										
Open Source VLMs										
Llama-3.2V 11B	97.50	82.00	77.50	42.62	53.87	92.62	**91.75**	**91.50**	19.25	18.75
Gemma-3 12B	96.75	83.00	67.50	36.62	36.62	93.25	84.37	91.25	17.00	27.12
Qwen2.5-VL 7B	96.00	83.75	75.00	34.37	59.62	**97.50**	84.37	91.37	18.50	31.25
LLaVA-1.5-7B	96.25	82.62	60.00	33.50	50.87	86.00	84.37	**91.50**	11.00	18.25
Open Source Medical VLMs										
Med-LLaVa 7B	96.00	82.25	62.25	33.75	52.37	90.37	84.50	91.25	10.00	16.75
Med-Gemma 4B	97.50	**84.62**	**92.50**	**46.50**	**61.62**	93.62	86.87	**91.50**	17.25	27.50

Table 4: Model performance for different categorical question on Chest X-ray and MedICat Dataset for Bangla. Here *Mod* refers to *Modality*, *Abnorm* refers to *Abnormality*,*Cond* referes to *Condition*, and *Pos* refers to *Position*. Here **Blue** highlights the highest-performing model in the Benchmark, while **Cyan** marks the best-performing models for each metric across the model types.

Models	Chest X-Ray					MedICaT				
	Generalized Question		Specialized Question			Generalized Question		Specialized Question		
	Mod	Organ	Abnorm	Cond	Pos	Mod	Organ	Abnorm	Cond	Pos
Base Prompt on English QA Pairs										
Closed Source VLMs										
Gemini 2.5	97.00	65.25	72.25	18.58	9.25	93.50	57.25	**95.00**	30.50	39.50
GPT-4.1 Mini	97.50	60.00	75.00	21.75	19.25	89.00	50.00	60.00	21.50	36.08
Claude Sonnet 4	97.50	63.25	70.00	13.32	15.75	91.50	51.25	92.50	24.50	27.25
Open Source VLMs										
Llama-3.2V 11B	67.25	67.75	67.50	10.00	14.00	89.25	55.25	86.75	15.00	20.00
Gemma-3 12B	76.25	66.25	47.50	11.50	14.00	91.75	55.00	90.00	25.75	25.50
Qwen2.5-VL 7B	97.00	94.25	60.00	2.25	25.75	90.50	65.00	87.50	17.75	24.75
LLaVA-1.5 7B	8.25	59.50	45.00	4.75	2.00	55.00	52.50	92.50	16.25	18.75
Open Source Medical VLMs										
Med-LLaVa 7B	3.00	88.25	40.25	8.75	23.50	45.00	35.00	89.25	3.25	26.25
Med-Gemma 4B	1.50	23.25	45.00	0.00	15.25	45.00	32.50	92.50	0.00	30.00
CoT Prompt on English QA Pairs										
Closed Source VLMs										
Gemini 2.5	97.50	66.75	72.50	19.00	13.00	94.25	55.25	**95.00**	**32.50**	30.75
GPT-4.1 Mini	97.00	56.00	55.00	9.75	16.50	90.75	50.50	92.50	30.25	27.58
Claude Sonnet 4	**97.75**	62.00	57.50	12.25	13.75	93.00	51.00	95.50	24.00	25.50
Open Source VLMs										
Llama-3.2V 11B	79.50	86.75	48.75	10.00	25.63	79.75	86.75	49.50	10.13	25.13
Gemma-3 12B	97.00	62.50	57.50	8.25	7.50	90.25	47.50	92.50	28.00	19.25
Qwen2.5-VL 7B	96.50	94.25	67.50	3.00	21.50	91.50	55.00	85.00	12.25	26.50
LLaVA-1.5 7B	29.50	57.50	45.00	0.75	2.75	65.00	51.25	92.50	9.75	20.50
Open Source Medical VLMs										
Med-LLaVa 7B	1.75	89.25	40.75	0.25	26.00	37.50	35.00	87.00	8.25	28.50
Med-Gemma 4B	1.00	18.38	45.00	0.00	15.25	45.00	32.50	92.50	0.00	25.75
LoRA Fine-Tuning on English QA Pairs										
Open Source VLMs										
Llama-3.2V 11B	96.00	**97.00**	89.75	50.87	**79.00**	**100.00**	**92.50**	85.00	32.23	42.07
Gemma-3 12B	97.25	95.62	72.50	43.25	57.25	97.50	87.50	92.50	28.75	31.82
Qwen2.5-VL 7B	96.75	94.37	75.00	38.37	66.25	**100.00**	90.00	92.50	22.75	**43.00**
LLaVA-1.5-7B	96.50	95.50	72.50	35.62	74.50	85.00	82.50	92.50	12.50	33.25
Open Source Medical VLMs										
Med-LLaVa 7B	96.25	94.25	72.50	33.87	62.50	87.50	85.00	87.50	14.25	30.75
Med-Gemma 4B	96.75	93.12	**92.50**	**52.75**	76.50	**100.00**	90.00	92.50	22.75	**43.00**

Table 5: Model performance for different categorical questions on the Chest X-ray and MedICat Dataset for English. Here *Mod* refers to *Modality*, *Abnorm* refers to *Abnormality*,*Cond* referes to *Condition*, and *Pos* refers to *Position*. Here **Blue** highlights the highest-performing model in the Benchmark, while **Cyan** marks the best-performing models for each metric across the model types.