Adapting Lightweight Vision Language Models for Radiological Visual Question Answering

Aditya Shourya[1 [0009−0003−4094−5455]], Michel Dumontier[1,2 [0000−0003−4727−9435]], and Chang Sun[1,2 [0000−0001−8325−8848]]

[1] Department of Advanced Computing Sciences, Maastricht University, Netherlands
[2] Institute of Data Science, Maastricht University, Netherlands
{a.shourya, michel.dumontier, chang.sun}@maastrichtuniversity.nl

Abstract. Recent advancements in vision-language systems have improved the accuracy of Radiological Visual Question Answering (VQA) Models. However, some challenges remain across each stage of model development: limited expert-labeled images hinders data procurement at scale; the intricate and nuanced patterns of radiological images make modeling inherently difficult; and the lack of standard evaluation makes it difficult to identify cases where the model might be ill-conditioned. In this study, we fine-tune a lightweight 3B parameter vision-language model for Radiological VQA, demonstrating that small models, when appropriately tuned with curated data, can achieve robust performance across both open- and closed-ended questions. We propose a cost-effective training pipeline from synthetic question-answer pair generation to multi-stage fine-tuning on specialised radiological domain-targeted datasets (e.g., ROCO v2.0, MedPix v2.0). Our results show that despite operating at a fraction of the scale of state-of-the-art models such as LLaVA-Med, our model achieves promising performance given its small parameter size and the limited scale of training data. We introduce a lightweight saliency-based diagnostic tool that enables domain experts to inspect VQA model performance and identify ill-conditioned failure modes through saliency analysis. Project Link: `https://github.com/adishourya/MedM`

Keywords: Radiological Visual Question Answering · Vision-Language Models · Lightweight Models · Medical Imaging · Saliency Analysis Generative AI

1 Introduction

Vision-language models (VLMs) have made notable progress in general-domain tasks, such as crop anomaly detection[30] and intelligent video surveillance[41]. In the medical and healthcare domain, researchers have recently adapted VLMs to support medical visual question answering (VQA), with promising results from both academic initiatives [21, 34] and large-scale efforts [32, 17]. Alongside improvements in accuracy, recent VLMs have become increasingly accessible to small teams and individual researchers and practitioners to adapt off-the-shelf VLMs to domain-specific tasks through affordable fine-tuning. However, these

off-the-shelf VLMs still underperform on medical VQA tasks compared to general-domain VQA due to domain mismatch, limited data availability, and a lack of systematic evaluation and interpretability tools.

Developing robust medical VQA systems poses unique challenges. VLM models are trained on open-web datasets (like [5, 29]) that include general-domain data and struggle with the domain shift introduced by complex, multi-modality clinical inputs. Medical VQA tasks require not only visual understanding but also specialized reasoning grounded in clinical knowledge, which general-purpose VLMs typically lack. Moreover, the scarcity of large-scale, high-quality image–question–answer datasets in radiology limits the ability to fine-tune or evaluate these models systematically. In addition, the absence of standardized training pipelines and interpretability tools hampers both model development and clinical validation. Together, these challenges call for lightweight approaches that balance domain adaptation, performance analysis, and interpretability.

We address these challenges by adapting a lightweight VLM - 3B-parameter PaliGemma-mix-448 [3] for radiological VQA. Our approach combines a two-stage fine-tuning pipeline with parameter-efficient LoRA [12] adaptation, using a curated mixture of radiology datasets (SLAKE [20], PMC-VQA [46], ROCO v2.0 [28], MedPix 2.0 [33]). In the first stage of fine-tuning, we align the model's projection head with domain-specific anatomical vocabulary; in stage 2, we fine-tune the full model using enriched instruction-tuning data generated via a LLaMA-8B QA generation pipeline and annealing strategies to amplify high-quality supervision. To evaluate model performance, we introduce a saliency-based diagnostic tool that visualizes attention from image patches to response tokens and vice versa, enabling human experts to identify ill-conditioned outputs. Despite the model's small size, it achieves competitive accuracy on combined ROCO+MedPix VQA tasks, approaching the performance of much larger models like LLaVA-Med [17].

Our **key contributions** are as follows. First, we reassess model scaling trends in medical VQA by demonstrating that a compact 3B VLM, when appropriately fine-tuned, can achieve competitive performance on radiological VQA tasks, challenging the assumption that only large-scale models are capable of strong clinical reasoning. Second, we propose an end-to-end framework that spans dataset curation, synthetic QA pair generation, annealing-based enrichment, and a two-stage fine-tuning strategy. This pipeline enables medical domain specialization with minimal compute, serving as a practical guide for low-resource medical VLMs. Third, we develop a lightweight, attention-based interpretability tool to visualize cross-modal saliency between image regions and text outputs, supporting expert-driven auditing of model predictions. Finally, we empirically validate our model on both open- and closed-ended radiological QA tasks, highlighting that compact, interpretable models can be viable for domain-specific VQA applications.

2 Related Work

Our methodology builds upon recent development in medical VQA and text-based question-answering. Several studies have introduced comprehensive pipelines that span data collection, model training, and rigorous assessment, highlighting the evolving capabilities of a radiological VQA system. We now summarize key contributions from related works that have influenced our approach.

MedVInT-T(D,E) [45] presents a complete training and evaluation framework for medical VQA. Their approach involves fine-tuning a VLM model on an in-house curated synthetic dataset [46] using GPT-4 [26], which contains multiple choice questions to cover a variety of radiological images, and short fill-in-the-blanks style questions with the expectation that the resulting model also develops the capability of answering open-ended queries. The model, fine-tuned on public benchmarks, performs on par with existing radiological VQA systems. Additionally, they manually verify a sample of test set results to make the models robust against the current limitations of popular evaluation frameworks.

MedPaLM [32], introduces a comprehensive training and evaluation framework from scratch. They compile HealthSearchQA dataset [14] for answering both consumer- and professional-level text-based medical questions by sampling from existing medical QA datasets. They then fine-tune Flan-PaLM [7] on this dataset, achieving a new state-of-the-art model which is then evaluated by both professionals and laypersons on an extensive set of evaluation axes. Notably, their work exemplifies the design of human evaluation, incorporating assessments from both professionals and laypersons across a broad set of criteria.

LLaVA-Med [17] curates PMC-15M dataset by sampling from PubMed Central [25] and prepares synthetically generated multi-turn instruction training data using GPT-4 [26]. The study trains the model for only 16 hours on 8xA100 GPUs [9], achieving state-of-the-art results in radiological visual question answering with a modest 8B LLM [39]. Their work demonstrates that individual researchers can achieve state-of-the-art performance even with a cost-effective training approach. [3]

3 Architecture

3.1 Model Design

Our vision-language model (VLM) builds on prior work [3, 1] and follows a multi-stage training pipeline (Figure 1). The training begins with the selection of an off-the-shelf vision-tower and an LLM, each demonstrating strong performance on their respective unimodal tasks, such as large-scale image classification for the vision-tower and natural language understanding and generation for the LLM. These components are then integrated and subjected to multimodal pretraining on a diverse set of tasks such as image captioning [42] and referring expression segmentation [15] to develop a broad understanding of visual concepts in the

[3] PMC-15M [17] remains unavailable to the public at the time of writing.

general domain. During the multimodal pretraining stage of the model, no weights are frozen in time, allowing all parameters to learn during backpropagation.

For domain adaptation such as radiological VQA, we conduct multi-stage fine-tuning on the selected off-the-shelf model using smaller but domain-specific datasets to adapt to specific tasks, mirroring methodologies in [32, 17, 45]. In our study, we employ *PaliGemma-mix-448* [3] as our base VLM. This choice is motivated by its transparent pretraining on a diverse and well-curated collection of open-web datasets [5, 29, 31, 35], in contrast to models with undisclosed training data [26]. This transparency enables a clearer understanding of the model's zero-shot (base) performance and would make it easier to compare after the base model is fine-tuned. The details about the main components of proposed VLM architecture are described below.

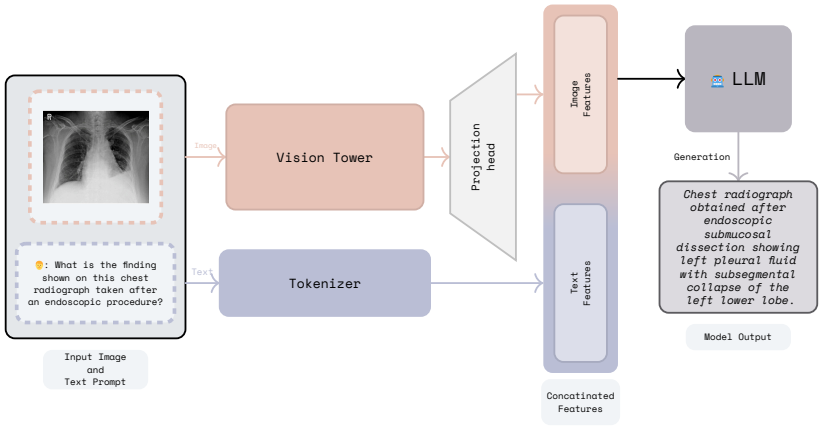


Fig. 1: Our and PaliGemma [3] Vision Language Model Architecture

Vision Tower: We employ a decoder-only SigLIP transformer [43] as the vision tower in our framework, which contains approximately 400M parameters. pretrained with a sigmoid contrastive loss and comprising 400M parameters. SigLIP is pretrained using a contrastive learning objective with a sigmoid loss, specifically to handle classification tasks involving a large number of labels where traditional cross-entropy loss becomes less effective [43]. The vision tower processes one or multiple input images by applying self-attention across image patches in a non-causal manner, generating image features that are independent of any accompanying text instruction.

Projection Head: A single linear layer aligns the output dimensionality of the vision tower with the token dimension of the language model's vocabulary, which is required for concatenation. While the projection can be implemented using multiple linear layers, the prior ablation study [3] found no significant advantage

to have more than one layer. Therefore, we use a single-layer projection in our VLM architecture.

Concatenation: The text prefix associated with the image is tokenized [16]) and concatenated with the projected image features from the vision tower. A special separator token is inserted between the image features and the tokenized text to delineate the two modalities. The resulting sequence is then padded or truncated as needed to match the input length of the language model.

LLM: The concatenated image and text features are passed to 2B-GEMMA LLM [38] as a single input. The model generates the first output token by jointly attending to both the visual features and the tokenized text prefix. Subsequent tokens are produced autoregressively, conditioned on the previously generated tokens along with the original multimodal input.

3.2 Diagnostic Design

To enhance interpretability and validate the clinical relevance of the proposed VQA, we analyze the model's attention mechanisms, which govern cross-modal interactions between image features and text tokens, inspired by [22]. We develop a diagnostic tool for saliency analysis aimed at aiding practicing radiologists during expert evaluation [37]. The interactions between text prefix, image features, and response tokens, which occur exclusively within the attention heads of the LLM, were analyzed with visualizations. Prior to the concatenation layer, there is no interaction between the text prefix and image features. Therefore, the attention heads of the LLM learn to selectively filter and attend to the relevant signals from both modalities to guide the generation process.

Although saliency is not the same as explainability [2], experts can often identify diagnostic indicators, as saliency is fundamentally tied to the learned weights of the model. For a self-attention-based model [40], this relation is easy to examine as self-attention operates by aggregating similarity scores between two learned representations for each tokens: queries and keys. These interactions determine how information is distributed across tokens which ultimately guides the generation process. We implemented the following two attention techniques.

Saliency via Raw Attention. Raw attention examines the interactions between queries and keys, which can be interpreted as measuring the affinity or relevance of a token of interest (query) with the rest of the tokens (keys), either within or across modalities. We compute attention weights between queries and keys to localize token-level contributions.

Saliency via Rollout Attention. In self-attention-based models, raw attention weights do not always provide meaningful insights as information propagates through multiple layers, embeddings become increasingly mixed. This is because self-attention does not inherently preserve token identity across layers; rather, it continuously blends representations from multiple input tokens. As a result individual token contributions become obscure, and raw attention weights fail to capture the original token relationship.[6]. We adopt rollout attention [6, 10], which recursively aggregates attention weights across layers while accounting for skip connections.

4 Datasets and Training Recipe

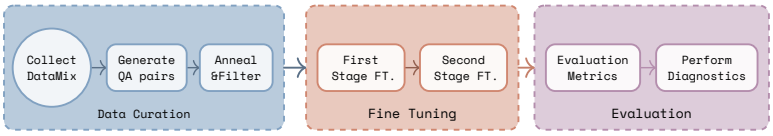


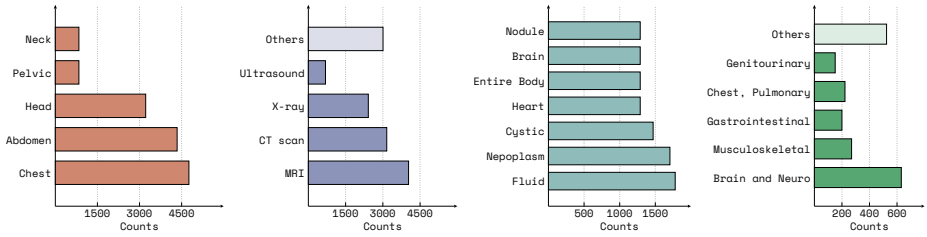
Fig. 2: Training Recipe Overview

The overall methodology of our training recipe is outlined in Figure 2. We begin by collecting publicly available radiological datasets and converting them into Visual Question-Answer (VQA) pairs [39]. The resulting dataset is then enriched and processed to ensure suitability for fine-tuning. Our fine-tuning approach uses a two-stage training strategy: the first stage focuses on learning foundational visual radiological concepts, while the second stage incorporates larger datasets to enhance the model's rigor and generalization capabilities.

To evaluate model performance, we measure classification accuracy on both open- and closed-ended questions, depending on the dataset composition. For generative responses from open-ended questions, we assess their factuality using GPT-4 [26] as an automated judge. We perform ablation studies across different stages of our data curation and finetuning methodology to quantify performance gains. In the absence of a medical expert, the authors of the paper conduct a diagnostic analysis on organ-level cases to identify model limitations.

4.1 Data Collections

Fine-tuning VLM requires not only substantial model capacity but also access to large, diverse, and semantically rich datasets. In our work, we combined four datasets that have been de-identified for privacy protection including SLAKE [20], PMC-VQA [46], ROCOv2 [28], and MedPix 2.0 [24]. The combination of them spans a wide range of pathology and radiological modalities (Figure 3, and concepts for open- and closed-ended questions.



(a) SLAKE: Organ semantic annotations (b) PMCVQA: modality distribution (c) ROCO2.0: Top7 UMLS concepts (d) MedPix2.0: Distribution on organs

Fig. 3: Distributions across four used datasets.

SLAKE contains ∼14,000 VQA pairs, annotated by practicing physicians. The dataset covers a wide range of anatomical regions and provides high-quality semantic annotations that are well-suited for evaluating radiological reasoning.

PMC-VQA is derived from PMC-CLIP [19] and includes ∼ 227,000 QA pairs. The questions are either multiple-choice or short fill-in-the-blank format. Its scale and diversity have been effectively leveraged in training models such as MedVInT-TE and MedVInT-TD. The dataset includes a diverse set of imaging modalities such as CT, MRI, ultrasound, and X-ray.

ROCOv2 contains ∼79,000 image-caption pairs from PubMed Central. Each caption provides a concise (∼ 20 word) description of the radiological images. Due to its breadth and structural consistency, ROCOv2 supports multiple tasks including image captioning, multi-label classification, and VLM pretraining.

MedPix 2.0 includes ∼ 12,000 curated cases from the MedPix database. Each case contains diagnostic images, detailed case descriptions, and relevant treatment information. The dataset is built using a semi-automated pipeline with manual validation to reduce label noise.

4.2 QA-Pairs Data Generation

Among our selected datasets, *SLAKE* and *PMC-VQA* natively provide image–QA pairs, while *ROCO v2.0* and *MedPix v2.0* contain image–caption pairs. Fine-tuning on image–QA triplets has been shown to be more effective than image–caption pairs for training VLMs on visual reasoning tasks [47]. Therefore, inspired by previous work [4, 47, 45, 17], we synthesize both open- and closed-ended QA pairs from image-caption pairs using LLaMA-8B [39]. LLaMA-8B was applied for its accessibility, inference efficiency, and reproducibility for other individual researchers. Importantly, its pretraining corpus contains limited medical content, allowing us to isolate and evaluate the performance of general-domain LLMs when applied to specialized medical tasks.

Medical VQA tasks demand not only visual understanding but also clinical reasoning, which general-purpose VLMs often lack. To address this, we prioritize datasets where questions are grounded in patient context and, where possible, linked to supporting medical literature. Figure 8 and 9 in the Appendix show the prompt templates to generate patient case-based and literature-based QA pairs from image-caption pairs. Synthetic QA generation introduces risks such as hallucinations or clinically irrelevant content. To ensure quality, we manually filter out noisy outputs and apply a form of dataset annealing to incrementally refine the corpus toward higher semantic and clinical relevance.

4.3 Annealing and Filtering

Annealing improves model performance by incrementally incorporating small, high-quality subsets into a larger training set. The objective is to improve the proportion of higher informative examples such as those rich in visual concepts and clinical reasoning within the overall dataset. By doing so, the model learns more reliable patterns that might otherwise be obscured by lower-quality data.

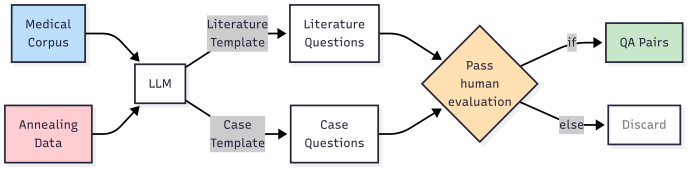


Fig. 4: Filtering and curation pipeline.

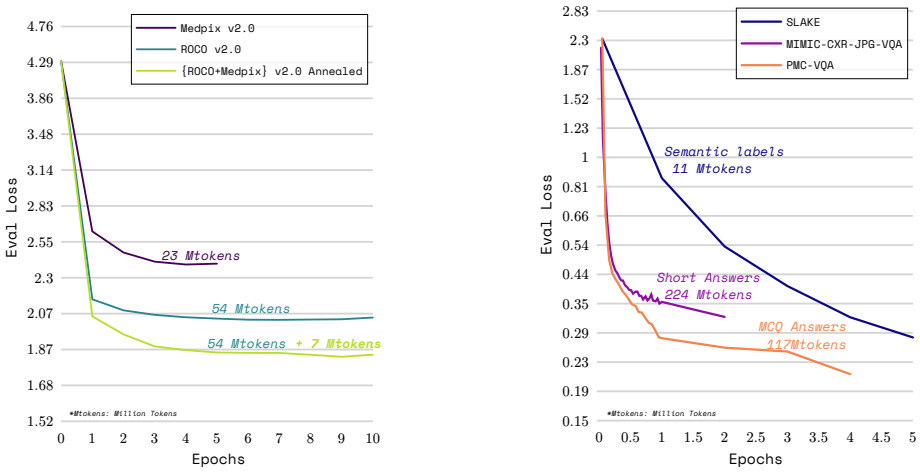
Evidence for annealing's effectiveness comes from [39], where LLaMA3-8B showed a 24% improvement on grade-school-level math questions GSM8K [8] and a 6.4% gain on competition-level math reasoning tasks [11]. Notably, the benefit diminished for larger models (e.g., LLaMA3-405B) [39], suggesting that small or mid-sized models, such as our 4B parameter VLM, are receptive to annealing.

This study uses the high-quality dataset *MedPix v2.0* [33] as the primary enrichment dataset for annealing *ROCO v2.0*. While *MedPix* is smaller in scale, it provides well-curated radiological cases and literature references, making it well-suited for improving domain-specific reasoning. A key component of effective annealing is systematic *filtering*, which ensures that only high-quality and domain-relevant data is incorporated into the dataset. Figure 4 outlines our data curation stratergy with annealing and filtering. The process begins with a medical corpus, filtered by the pathological relevance. Unlike conventional upsampling strategies that increase the variety of rare cases, our approach focuses on reinforcing the most common pathologies existing in our data mix to improve model generalization.

4.4 Two-stage Fine-Tuning

1st Stage: Off-the-shelf VLMs often exhibit inconsistent performance in recognizing anatomical structures occasionally producing incorrect generation when presented with slight variations in images. This inconsistency highlights the need for alignment between visual features and anatomical vocabulary. To address this, we adopt the SLAKE dataset [20] as a foundation for 1st stage fine-tuning. SLAKE offers well-annotated radiological visual concepts, making it particularly suitable for anatomical structure recognition. In this initial phase, we fine-tune only the projection head of the model while keeping all other parameters frozen. We train the projection layer for 5 epochs on SLAKE and use the resulting checkpoint as the initialization point for subsequent model training. Our method aligns with curriculum learning principles, emphasized in [36], starting with simpler radiological visual concepts followed by more diverse data.

2nd Stage: Using the checkpoint from the 1st stage as the model weight initialization, we fine-tune the model on larger and more diverse instruction sets - ROCO v2.0 [28], MedPix 2.0 [24], and PMC-VQA [23]. To perform parameter-efficient fine-tuning, we apply LoRA [12], a low-rank adaptation method, targeting the attention heads in both the vision tower and the language model. This allows us to significantly reduce computational and storage overhead, deviating from traditional fine-tuning methods that retain all or a large portion of model parameters.



(a) Eval Loss for open ended questions

(b) Eval loss for short ended questions

Fig. 5: Fine Tuning Evaluation Loss

5 Experiments and Evaluation

5.1 Experiment Setting

All experiments, including fine-tuning and evaluation, were conducted on a single NVIDIA H100 GPU. With adequate allocation, fine-tuning on the ROCO, MedPix, and combined ROCO+MedPix datasets each consumed approximately 2.1 TFLOP-days, PMC-VQA required about 4.2 TFLOP-days, and SLAKE completed in under 0.15 TFLOP-days.

5.2 Fine-Tuning Experiments

We first evaluated fine-tuning performance across instruction sets with varying token lengths and question formats. Evaluation loss curves over training epochs are shown in Figures 5a and 5b. For datasets where QA template is open-ended, we observe that the evaluation loss decreases approximately quadratically as the number of tokens in the instruction set increases (Figure 5a). However, this trend does not hold for datasets with close or short-ended QA templates, where the labels contain fewer tokens as the expected loss after a few training iterations becomes smaller and the loss plateaus earlier (Figure 5b).

We further analyzed **scaling behavior** using the empirical loss model:

$$\tilde{L}(X, D_f) = A \cdot \frac{1}{X^{\alpha}} \cdot \frac{1}{D_f^{\beta}} + E \tag{1}$$

where $\tilde{L}$ is evaluation loss, X is the fine-tuning parameters, D_f is the token size, and A, α, β, E are scaling exponents. Scaling properties for fine-tuning LLMs

are highly dependent on task type and data composition [44]. Consequently, the optimal fine-tuning strategies and scaling behavior can vary depending on the structure and semantics of the training data. We observed that scaling exponents ("A", α,"β", "E" in Equation 1) differ depending on the question-answer templates used across datasets. For example, ROCO v2.0 [28] and MedPix 2.0 [24] have open-ended instruction sets with an average label length of around 20 tokens. In this case, task dependence is less observable, and improvements in evaluation loss ($\tilde{L}$) tend to correlate more directly with data size (D_f).

In contrast, task dependence becomes more evident in close-ended QA, particularly when different templates are used for QA pairs (Figure 5b). While higher data volume generally leads to faster convergence, this trend breaks down when comparing MIMIC-CXR-JPG [13] and PMC-VQA [46]. Despite its smaller size, PMC-VQA yields greater learning gains in fewer epochs, likely due to the use of multiple-choice templates. These have a lower expected loss ($\tilde{L} = -\ln\left(\frac{1}{4}\right)$) than open-ended QA tasks, which typically involve more linguistic variation and semantic ambiguity.

These observations suggest that a single scaling law may not generalize across mixed-template datasets. As dataset mixtures grow, especially those combining open- and close-ended QA formats, it becomes increasingly difficult to preserve a consistent ratio of question types. Since each new addition may introduce variations in this ratio, it becomes challenging to predict the expected evaluation loss as the number of tokens in the instruction set grows. This variability complicates the application of scaling laws in Medical VQA, as the impact of additional training data is not uniform across different datasets and QA templates.

5.3 VQA Evaluation

Standard n-gram metrics such as BLEU [27] and ROUGE [18] offer limited insight into factual correctness, particularly in clinical VQA settings [32]. We report these scores in Table 5 in the Appendix, but propose and emphasize more robust evaluation methods below.

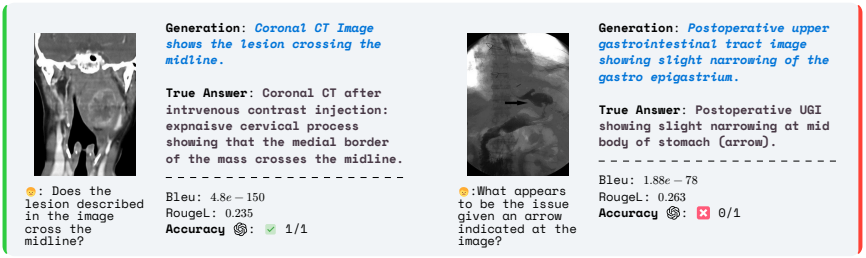


Fig. 6: LLM-based Evaluation Examples

Closed-ended QA Evaluation: For multiple-choice question answering (MCQA) such as PMC-VQA [46], we measure model accuracy across five stochastic generations per test instance. Inspired by [45], we define a prediction as *non-robust* if the model produces different answers in three or more out of five inferences. In such cases, we penalize the accuracy by one point to account for uncertainty and instability in the output.

Open-Ended QA Evaluation: For open-ended question that demands clinical reasoning, we employ LLM-based evaluation. We design a prompt template (Figure 10 in the Appendix) and use GPT-4.0 [26] to judge each generated answer based on factual correctness. Examples are presented in Figure 6.

Table 1 compares accuracy across four datasets, evaluating the effect of a two-stage fine-tuning approach. Results are reported as the mean accuracy ± standard deviation over five inference runs, with LLaVA-Med serving as a high-capacity baseline. On SLAKE (closed-ended QA), two-stage fine-tuning achieves 79% accuracy, highlighting strong gains even without large model capacity. For PMC-VQA, ROCO, and the ROCO+MedPix annealing set, two-stage fine-tuning consistently outperforms single-stage fine-tuning, demonstrating its effectiveness across different QA formats. Although the accuracy gains of 2-stage fine-tuning are slight, it accelerated convergence, reducing the number of epochs needed to reach target evaluation loss. Finally, comparing ROCO to the ROCO+MedPix annealing set shows clear performance gains from annealing, even with small data volumes. These results indicate that modest instruction set annealing offers a cost-effective way to improve generalization and robustness, with potentials for further gains using larger annealing sets.

Dataset	**w/o Stage 1**	**with Stage 1**	**LLaVA-Med**
SLAKE (Closed)	–	79.00 ± 2.75	86.50 ± 1.60
PMC-VQA (Short)	32.22 ± 2.23	33.15 ± 3.15	58.44 ± 2.53
ROCO v2.0 (Open)	32.25 ± 3.60	34.00 ± 3.95	56.56 ± 3.57
ROCO + MedPix (Annealing)	39.53 ± 3.22	41.48 ± 3.90	56.63 ± 3.22

Table 1: Accuracy (%) without and with Stage 1 fine-tuning across datasets. Means ± standard deviations are reported across five inference runs (each on a sample of 200).

5.4 Manual Verification via Saliency Diagnostic

Inspired from previous work [45, 32], we conduct manual verification of model-generated responses on test samples, incorporating saliency diagnostic wherever possible for the authors of the study. As discussed, we applied raw attention and attention rollout methods for saliency analysis (Figure 7). In the case of response-to-image saliency, we select a response token (e.g., "narrowing") as the query and visualize the average saliency over the input image (used as keys). Conversely, we examine image-to-response saliency, where we select a specific image patch (e.g., the blue arrow) as the query and plot the resulting saliency over the response tokens based on their key representations. Compared to raw attention, we found

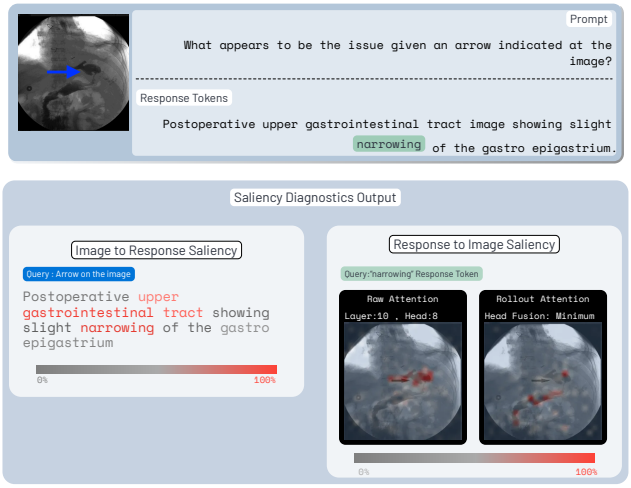


Fig. 7: An example saliency analysis with Raw Attention and Attention Rollout for a patient with a Post-Operative UGI with slight narrowing at mid-body of stomach.

the resulting saliency of attention rollout highlights more abstract features such as the passage of the gastrointestinal tract that are semantically relevant to the given example. More details and examples are presented in Appendix B.

Using the saliency tool, we evaluate the factuality of the generated responses from our model against the corresponding ground-truth labels. Furthermore, we report a broad per-class accuracy (Table 2) at the organ level to highlight variability in model performance, as certain anatomical regions exhibit greater nuance and complexity than others.

Organ-level Pathologies	**Accuracy (%)**
Chest	15/50 (19/50)
Gastrointestinal	28/50 (32/50)
Musculoskeletal	39/50 (41/50)
Brain and Neuro	14/50 (22/50)

Table 2: Manual Verification over a single inference (LLava-Med [17] as a baseline).

6 Conclusion

This study shows that a compact 3B VLM, when fine-tuned with an end-to-end pipeline, can achieve strong performance on radiological VQA tasks. Our framework, including synthetic QA generation, instruction annealing, and two-stage fine-tuning, enables low-resource specialization for medical VLMs. We further introduce a lightweight saliency tool for cross-modal interpretability and validate our approach on both open- and closed-ended QA pairs.

Our study has several limitations that highlight directions for future work. First, our ablation analysis focused primarily on LLM scaling, with a limited investigation into the vision encoder. Second, saliency analysis was conducted without expert involvement, limiting interpretability to broad organ-level patterns. Future work should involve clinicians to evaluate fine-grained anatomical and pathological relevance. Last, our evaluation framework focused only on single-turn QA, whereas real-world clinical workflows involve multi-turn interactions. Expanding the evaluation to multi-turn dialogues would offer a more comprehensive assessment of model reasoning and consistency.

Code and Data Availability

The code is publicly available at: `https://github.com/adishourya/MedM`. The dataset derived from MedPix v2.0 [33] and used for our annealing experiments can be accessed at: `https://huggingface.co/datasets/adishourya/MEDPIX-ShortQA`

Synthetically generated question-answer pairs based on the ROCO V2.0 dataset [28] are available at the following locations: *Training split*: `https://huggingface.co/datasets/adishourya/ROCO-QA-Train`; *Validation and test splits*: `https://huggingface.co/datasets/adishourya/ROCO-QA`

Bibliography

[1] Abdin, M.I., Ade Jacobs, S., Awan, A.A., Aneja, J., Awadallah, A., Hassan Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Mendes, C.C.T., Chen, W., Chaudhary, V., Chopra, P., Giorno, A.D., de Rosa, G., Dixon, M., Eldan, R., Iter, D., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R.J., Huynh, J., Javaheripi, M., Jin, X., Kauffmann, P., Karampatziakis, N., Kim, D., Khademi, M., Kurilenko, L., Lee, J.R., Lee, Y.T., Li, Y., Liang, C., Liu, W., Lin, X.E., Lin, Z., Madan, P., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Rosset, C., Roy, S., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Song, X., Ruwase, O., Wang, X., Ward, R., Wang, G., Witte, P., Wyatt, M., Xu, C., Xu, J., Xu, W., Yadav, S., Yang, F., Yang, Z., Yu, D., Zhang, C., Zhang, C., Zhang, J., Zhang, L.L., Zhang, Y., Zhang, Y., Zhou, X.: Phi-3 technical report: A highly capable language model locally on your phone. Tech. Rep. MSR-TR-2024-12, Microsoft (August 2024), https://www.microsoft.com/en-us/research/publication/phi-3-technical-report-a-highly-capable-language-model-locally-on-your-phone/

[2] Bertrand, A., Pearce, A., Thain, N.: Searching for unintended biases with saliency. PAIR Explorables (2022), https://pair.withgoogle.com/explorables/saliency/

[3] Beyer*, L., Steiner*, A., Pinto*, A.S., Kolesnikov*, A., Wang*, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., Unterthiner, T., Keysers, D., Koppula, S., Liu, F., Grycner, A., Gritsenko, A., Houlsby, N., Kumar, M., Rong, K., Eisenschlos, J., Kabra, R., Bauer, M., Bošnjak, M., Chen, X., Minderer, M., Voigtlaender, P., Bica, I., Balazevic, I., Puigcerver, J., Papalampidi, P., Henaff, O., Xiong, X., Soricut, R., Harmsen, J., Zhai*, X.: PaliGemma: A versatile 3B VLM for transfer. arXiv preprint arXiv:2407.07726 (2024)

[4] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shinn, N., Ziegler, D., Wu, J., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)

[5] Changpinyo, S., Kukliansy, D., Szpektor, I., Chen, X., Ding, N., Soricut, R.: All you may need for VQA are image captions. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1947–1963. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.naacl-main.142, https://aclanthology.org/2022.naacl-main.142/

[6] Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: 2021 IEEE/CVF Conference on Computer Vision and

Pattern Recognition (CVPR). pp. 782–791 (June 2021). https://doi.org/10.1109/CVPR46437.2021.00084

[7] Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models (2022). https://doi.org/10.48550/ARXIV.2210.11416, https://arxiv.org/abs/2210.11416

[8] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., Schulman, J.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021)

[9] Corporation, N.: Nvidia a100 tensor core gpu (2020), https://www.nvidia.com/en-us/data-center/a100/, accessed: 2025-03-30

[10] Gildenblat, J.: Explainability for vision transformers. https://github.com/jacobgil/vit-explain (2021), accessed: 2025-02-03

[11] Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., Steinhardt, J.: Measuring mathematical problem solving with the math dataset. NeurIPS (2021)

[12] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685

[13] Johnson, A.E.W., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., ying Deng, C., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs (2019), https://arxiv.org/abs/1901.07042

[14] katielink: Healthsearchqa. https://huggingface.co/datasets/katielink/healthsearchqa (2023), accessed: 2025-05-01

[15] Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.L.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP (2014)

[16] Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing (2018), https://arxiv.org/abs/1808.06226

[17] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day (2023), https://arxiv.org/abs/2306.00890

[18] Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://www.aclweb.org/anthology/W04-1013

[19] Lin, W., et al.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: Greenspan, H. (ed.) Medical Image Computing and Computer Assisted Intervention, pp. 525–536. Springer (2023)

[20] Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering (2021), https://arxiv.org/abs/2102.09542

[21] Markus Zhang, V.C.: Babydoctor. https://github.com/photomz/BabyDoctor (2023), gitHub
[22] Mondal, A.K., Bhattacharjee, A., Singla, P., Prathosh, A.P.: xvitcos: Explainable vision transformer based covid-19 screening using radiography. IEEE Journal of Translational Engineering in Health and Medicine **10**, 1–10 (2022). https://doi.org/10.1109/JTEHM.2021.3134096
[23] National Library of Medicine: PMC Open Access Subset. Online (2003), bethesda (MD): National Library of Medicine. [cited YEAR MONTH DAY]. Available from: https://pmc.ncbi.nlm.nih.gov/tools/openftlist/
[24] National Library of Medicine: MedPix: Free Online Medical Image Database. Online (2024), available at: https://medpix.nlm.nih.gov/home [Accessed YEAR MONTH DAY]
[25] National Library of Medicine: PubMed Central (PMC) (2024), https://www.ncbi.nlm.nih.gov/pmc/, accessed: 2024-02-08
[26] OpenAI: ChatGPT: A Large Language Model. Online (2024), available at: https://openai.com/chatgpt [Accessed YEAR MONTH DAY]
[27] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics (2002). https://doi.org/10.3115/1073083.1073135
[28] Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology objects in context (roco): A multimodal image dataset. Tech. rep., University of Applied Sciences and Arts Dortmund, TU Dortmund University, University of Duisburg-Essen (2018), https://labels.tue-image.nl/wp-content/uploads/2018/09/AM-04.pdf, accessed: 2024-11-02
[29] Piergiovanni, A., Kuo, W., Angelova, A.: Pre-training image-language transformers for open-vocabulary tasks (2022), https://arxiv.org/abs/2209.04372
[30] Sajid, H.: Ai agriculture: Boost yields with yolo11. https://www.ultralytics.com/blog/the-changing-landscape-of-ai-in-agriculture (2024), accessed: 2025-04-30
[31] Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-1238, https://aclanthology.org/P18-1238/
[32] Singhal, K., Azizi, S., Tu, T., et al.: Large language models encode clinical knowledge. Nature **620**, 172–180 (2023). https://doi.org/10.1038/s41586-023-06291-2, https://doi.org/10.1038/s41586-023-06291-2
[33] Siragusa, I., Contino, S., Ciura, M.L., Alicata, R., Pirrone, R.: Medpix 2.0: A comprehensive multimodal biomedical dataset for advanced ai applications (2024), https://arxiv.org/abs/2407.02994

[34] van Sonsbeek, T., Derakhshani, M.M., Najdenkoska, I., Snoek, C.G.M., Worring, M.: Open-ended medical visual question answering through prefix tuning of language models (2023), https://arxiv.org/abs/2303.05977

[35] Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M.: Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2443–2449. SIGIR '21, ACM (Jul 2021). https://doi.org/10.1145/3404835.3463257, http://dx.doi.org/10.1145/3404835.3463257

[36] Srinivasan, T., Ren, X., Thomason, J.: Curriculum learning for data-efficient vision-language alignment (2022), https://arxiv.org/abs/2207.14525

[37] Stan, G.B.M., Aflalo, E., Rohekar, R.Y., Bhiwandiwalla, A., Tseng, S.Y., Olson, M.L., Gurwicz, Y., Wu, C., Duan, N., Lal, V.: Lvlm-interpret: An interpretability tool for large vision-language models (2024), https://arxiv.org/abs/2404.03118

[38] Team, G., Riviere, M., Pathak, S., Sessa, P.G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C.L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.B., Neyshabur, B., Bachem, O., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C., Choquette-Choo, C.A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozińska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucińska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J.P., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J., Gordon, J., Lipschultz, J., Newlan, J., yeong Ji, J., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L.L., Usui, L., Sifre, L., Heuermann, L., Lago, L., McNealus, L., Soares, L.B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Görner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R.A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S.M., Cogan, S., Perrin, S., Arnold, S.M.R., Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kocisky, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A.,

Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R., Andreev, A.: Gemma 2: Improving open language models at a practical size (2024), https://arxiv.org/abs/2408.00118

[39] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023), https://arxiv.org/abs/2302.13971

[40] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023), https://arxiv.org/abs/1706.03762

[41] Verkada Inc.: Video alarms for 24/7 security & monitoring. https://www.verkada.com/alarms/video-alarms/, accessed: 2025-04-30

[42] Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2015)

[43] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training (2023), https://arxiv.org/abs/2303.15343

[44] Zhang, B., Liu, Z., Cherry, C., Firat, O.: When scaling meets llm finetuning: The effect of data, model and finetuning method. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024), https://openreview.net/forum?id=5HCnKDeTws

[45] Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Pmc-vqa: Visual instruction tuning for medical visual question answering (2024), https://arxiv.org/abs/2305.10415

[46] Zhang, X., et al.: Pmc-vqa dataset (2023), https://huggingface.co/datasets/xmcmic/PMC-VQA, accessed: 2025-02-08

[47] Zhu, D., Chen, J., Haydarov, K., Shen, X., Zhang, W., Elhoseiny, M.: Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions (2023), https://arxiv.org/abs/2303.06594

Appendix

A Generating Question Answer Pairs

```
def generate_qapairs(caption):
  # Construct the prompt for ollama for mcq based questions
  prompt = f"""
  Based on the following medical image captions generate appropriate and insightful
question for the caption. Treat this caption as the ground truth to generate your
question: {caption}
  """
  response = ollama.chat(model='llama3.1',
      messages=[ {
          'role': 'user',
          'content': prompt } ])

  # Return the generated text from the response
  return response['message']['content'].strip()
```

Fig. 8: Generate Case Based Questions Prompt

```
def generate_qapairs_lit(caption,case_topic):
  prompt = f"""
  Generate a relevant,short and insightful literature-based question for the given
case topic. Treat the caption as ground truth: {caption}, with associated
literature: {case_topic}.
  """
  response = ollama.chat(model='llama3.1',
      messages=[ {
          'role': 'user',
          'content': prompt } ])

  # Return the generated text from the response
  return response['message']['content'].strip()

"""
```

Fig. 9: Generate Literature Based Questions Prompt

B Evaluation and Saliency Diagnostics

```python
def evaluate_generation(generation,ground):
  prompt = f"""
 Assign a score of 0 if the generated response is not both factually correct and
aligned in meaning with the ground truth else reward 1.
 Generation : {generation}
 Ground Truth : {ground}
  """
  response = ollama.chat(model='gpt-4',
      messages=[ {
          'role': 'user',
          'content': prompt } ])

  # Return the generated text from the response
  return response['message']['content'].strip()
```

Fig. 10: Evaluation prompt for GPT-4 as a judge.

Training Args	**Value**
learning_rate	1×10^{-5}
lr_schedule	constant
label_smoothing	0.0
weight_decay	0.0
fp16	True
gradient_accumulation	16
batch_size	6

Table 3: First and second stage training hyperparameters.

B.1 Saliency Diagnostic Examples

Note that we examine attention across modalities (Image to Response and Response to Image) as given by Equation 2

query from selected response token — query from image patch

$$Attention = \text{softmax}\left(\frac{Q\ K^{\top}}{\sqrt{d_k}}\right) V \quad \text{or} \quad \text{softmax}\left(\frac{Q\ K^{\top}}{\sqrt{d_k}}\right) V \qquad (2)$$

keys from image patches — keys from response

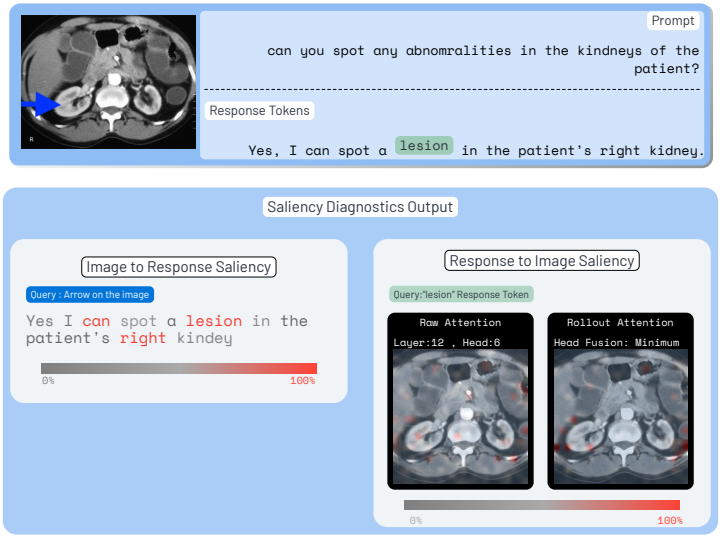


Fig. 11: A Patient suffering from lesion on their right kidney [Notice High Rollout Saliency in the right kidney of the patient]

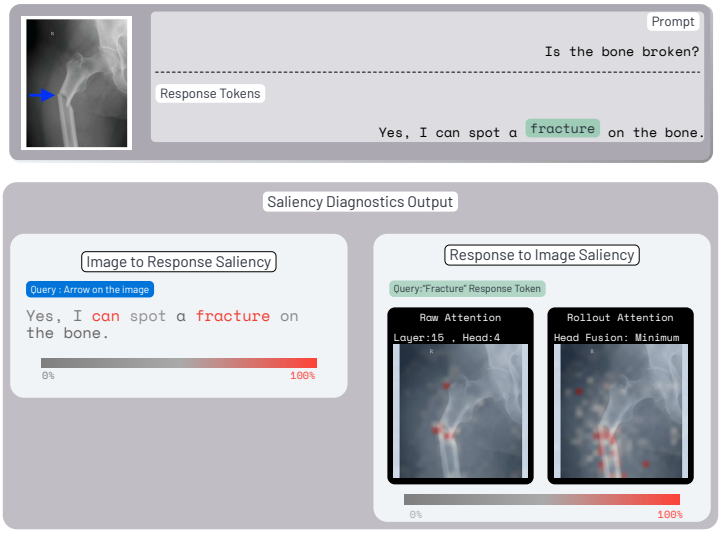


Fig. 12: An X-ray of a patient with Bone Fracture[High Saliency on the fractured region]

Metrics	MedPix v2.0	ROCO v2.0	ROCO + MedPix v2.0	PMC-VQA
ROUGE-S	0.311 ± 0.255	0.325 ± 0.132	0.334 ± 0.122	–
ROUGE-M	0.167 ± 0.082	0.179 ± 0.124	0.181 ± 0.124	–
ROUGE-L	0.308 ± 0.125	0.278 ± 0.120	0.304 ± 0.180	–
BLEU	0.055 ± 0.111	0.059 ± 0.090	0.077 ± 0.065	–
Accuracy	34/200 *(82/200)*	63/200 *(113/200)*	71/200 *(113/200)*	0.3002

Table 4: Results with no first-stage fine-tuning ($\pm$ 1 standard deviation). *LLava-Med* [17] used as baseline.

Metrics	SLAKE (Stage 1)	MedPix v2.0 (Stage 2)	ROCO v2.0 (Stage 2)	ROCO + MedPix v2.0 (Stage 2)	PMC-VQA (Stage 2)
ROUGE-S	–	0.322 ± 0.240	0.325 ± 0.182	0.380 ± 0.077	–
ROUGE-M	–	0.165 ± 0.088	0.196 ± 0.122	0.219 ± 0.111	–
ROUGE-L	–	0.318 ± 0.125	0.266 ± 0.080	0.412 ± 0.109	–
BLEU	–	0.024 ± 0.212	0.008 ± 0.121	0.430 ± 0.086	–
Accuracy	0.26 (*0.59*)	33/200 (*82/200*)	68/200 (*113/200*)	74/200 (*113/200*)	0.31046

Table 5: Results $\pm$ 1 standard deviation

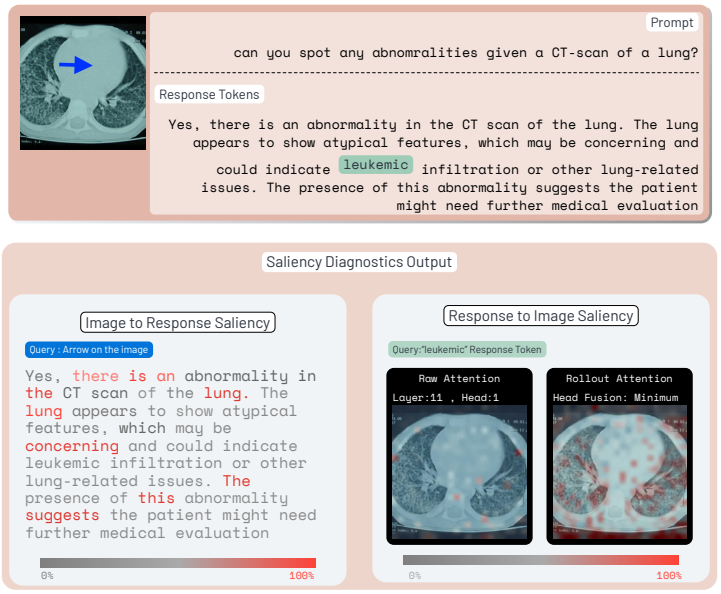


Fig. 13: A patient suffering from Leukemia [In such examples the authors of the study refrain from performing saliency diagnostics]