Multimodal LLMs Medical Visual Question Answering

(Leveraging Techniques for Enhanced Accuracy and Explainability in Medical Visual Question Answering)

Sheetal Patnaik
Department of Data Analytics
San Jose State University
017526678
sheetal.patnaik@sjsu.edu

Aishwarya Gulab Thorat
Department of Data Analytics
San Jose State University
017557579
aishwarya.thorat@sjsu.edu

***Abstract*—This project addresses the challenge of improving accuracy and explainability in Medical Visual Question Answering (Med-VQA) systems. Leveraging the PMC-VQA dataset and benchmark datasets like SLAKE and PubMedQA, we propose a multimodal Retrieval-Augmented Generation (RAG) pipeline that integrates both textual and visual evidence. Key innovations include the use of prompting strategies (CoT, ToT, Few-Shot) and fine-tuning with LoRA on LLaVA-Med. Experimental results demonstrate significant improvements in recall and answer interpretability. Our pipeline is designed to bridge the gap between general-purpose vision-language models and the specialized requirements of clinical applications. By combining retrieval-based context with carefully crafted prompts and lightweight fine-tuning, the system is able to deliver answers that are not only accurate but also interpretable. Ultimately, this work contributes toward building more reliable AI tools to support clinical decision-making and medical education.**

***Index Terms*—Medical VQA, Large Language Models, Multimodal Retrieval, Prompt Engineering, Explainable AI, Biomedical NLP**

I. Introduction

Medical professionals routinely depend on visual information such as X-rays, MRI scans, CT images, and pathology slides to diagnose conditions, monitor disease progression, and decide on appropriate treatments. These images often contain highly detailed and complex patterns that require careful interpretation. As the volume of patient data continues to grow in modern healthcare systems, doctors are faced with an overwhelming number of visual records to analyze—sometimes across multiple formats for the same case. This increasing data complexity can lead to delays in decision-making, diagnostic inconsistencies, and missed insights, especially when time and resources are limited.

Medical Visual Question Answering (Med-VQA) is an emerging solution designed to assist in this process. It involves building systems that can answer specific questions about a medical image, such as ”What abnormality is present in this chest X-ray?” or ”Does this brain scan show signs of stroke?” These systems aim to help practitioners by offering quick, consistent responses based on the visual content provided.

Although several tools have been developed to understand medical images and connect them with textual questions, many existing solutions face significant challenges. One of the biggest limitations is the lack of understanding of medical terminology and context-specific details. Some models may identify general features but fail to provide answers that reflect the deeper clinical meaning behind an image. Another challenge lies in explainability—these systems may provide an answer without showing how or why that answer was chosen. Finally, many current approaches struggle to handle the wide variety of image types found in clinical practice, making them less useful in real-world hospital settings.

This project addresses these gaps by designing a comprehensive Med-VQA system that combines medical image interpretation with information retrieval from trusted medical literature. The system is built to understand both the visual and textual aspects of a case, retrieve relevant background in-

formation when needed, and generate clear, understandable answers to medical questions. It also includes mechanisms to show the reasoning path behind each answer to make the process more transparent.

To achieve this, we implemented a pipeline that includes step-by-step prompting strategies to guide reasoning, a retrieval system that searches medical databases and similar image–question examples, and a lightweight fine-tuning method to adjust the system to medical tasks using minimal training data. Together, these components help us build a more reliable, interpretable, and general-purpose Med-VQA system that is better suited for real clinical use.

II. Related Work

Medical Visual Question Answering (Med-VQA) has advanced rapidly with the emergence of vision-language models (VLMs). Initial approaches primarily focused on direct classification or retrieval-based pipelines, relying on small-scale datasets such as VQA-RAD [11] and ImageCLEF-VQA [10]. While these datasets laid the groundwork for benchmark development, they lacked the diversity, scale, and complexity needed for high-performing clinical applications.

Zhang et al. [1] introduced PMC-VQA, a large-scale dataset constructed from PubMed Central Open Access articles, offering over 227,000 image-question-answer triplets. This dataset set a new standard for Med-VQA evaluation. Zhou et al. [12] extended Med-VQA to pathology images, addressing a previously underrepresented modality. Liu et al. [3] proposed an interpretable VQA approach using multimodal relationship graph learning, emphasizing visual-textual alignment and reasoning traceability.

Explainability has gained prominence in recent literature. Shen et al. [2] introduced MedCoT, a chain-of-thought (CoT) prompting framework leveraging hierarchical expert reasoning to improve answer interpretability. Yao et al. [4] and Wei et al. [5] demonstrated the effectiveness of multimodal CoT reasoning across general vision-language tasks. Surveys by Abacha and Demner-Fushman [6] and Wang et al. [7] comprehensively summarize the landscape and challenges of medical VQA systems.

In terms of modeling techniques, Li et al. [8] proposed LLaVA-Med, a biomedical assistant trained using LoRA-based fine-tuning on vision-language data. Lin et al. [9] presented PMC-CLIP, a contrastive learning model for aligning medical texts and figures using biomedical literature. These works highlight the importance of domain-specific adaptation and scalable fine-tuning.

While prior work has tackled isolated components—such as dataset creation, prompting, or model adaptation—few systems have integrated these effectively into a unified, explainable Med-VQA framework. Our work addresses this gap by combining Retrieval-Augmented Generation (RAG), advanced prompting strategies (CoT, ToT, Few-Shot), and lightweight fine-tuning (LoRA) to deliver a robust and interpretable system for medical question answering.

III. Methodology

Our project was completed by two contributors — Aishwarya Gulab Thorat and Sheetal Patnaik — who worked together to design, build, and evaluate a system that answers medical questions using both images and text. Each person focused on specific parts of the system.

A. Data Preprocessing

Handled by: Aishwarya Gulab Thorat

Unique Approach: We created a single system that could handle both images and text from two datasets at once.

- We collected medical images from two datasets — PMC-VQA and SLAKE.
- All images were resized and cleaned to make sure they looked similar in shape and style.
- We used a tool called CLIP to turn images into number-based representations, which were saved for later use.
- We also turned the questions (text) into a similar format, so the system could compare questions and images easily.
- Before moving forward, we checked all the results to make sure the data formats were correct and ready to use.

B. Prompt Design and Question Asking

Handled by: Sheetal Patnaik

Unique Approach: We changed the way we asked questions based on how confident the answers were, and used step-by-step thinking to improve reliability.

- In the simplest method, we just gave the system an image and a question.
- In more advanced methods, we showed the system examples of similar questions and answers to help it learn.
- We added thinking steps (called "Chain of Thought") to help it reason more clearly before answering.
- We also tried giving it different ways to think through a question (called "Tree of Thought"), especially for tricky problems.
- If the system replied with uncertain phrases like "I'm not sure", we automatically asked the question again in a clearer way.

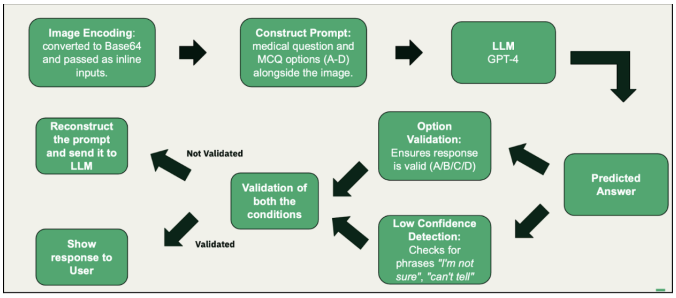

Fig. 1. How We Handled Uncertain Answers with Zero-Shot Prompting

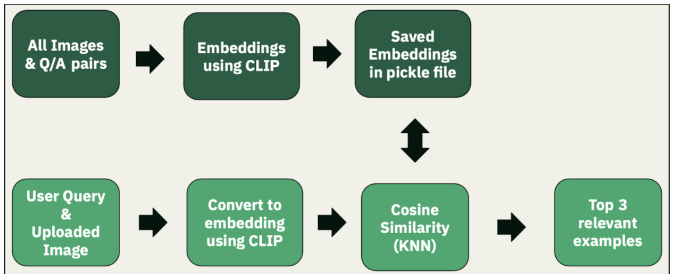

Fig. 2. How We Picked 3 Similar Past Questions to Use as Examples

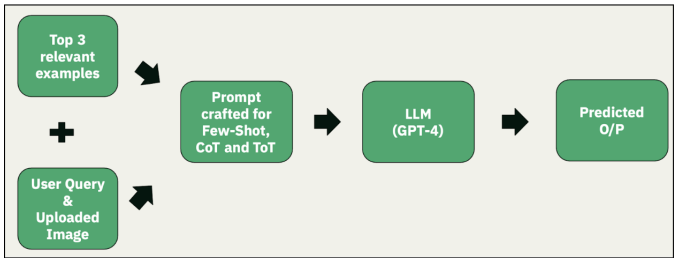

Fig. 3. Different Prompting Styles We Used: Few-Shot, Step-by-Step, and Tree-Based

C. Search and Retrieval System

Handled by: Aishwarya and Sheetal together

Unique Approach: Instead of using only images or only text, we combined both to find better answers.

- First, we searched a medical article database (PubMedQA) to find long text answers similar to our question.
- Then, we looked for images with questions similar to ours in another dataset (SLAKE and PMC-VQA).
- Both sets of results — helpful text and similar questions with images — were combined and used to form the final answer.

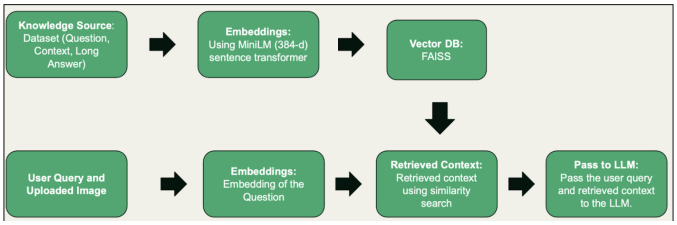

Fig. 4. How We Retrieved Helpful Text and Images Before Answering

D. Training a Specialized Model

Handled by: Aishwarya Gulab Thorat

Unique Approach: We adjusted an existing tool using a method that is faster and uses fewer resources.

- We used a large medical model called LLaVA-Med and adjusted it using a lightweight method called LoRA.
- We trained this model on medical questions and image pairs from the PMC-VQA dataset.
- The adjusted model reached around 51% accuracy on test data.

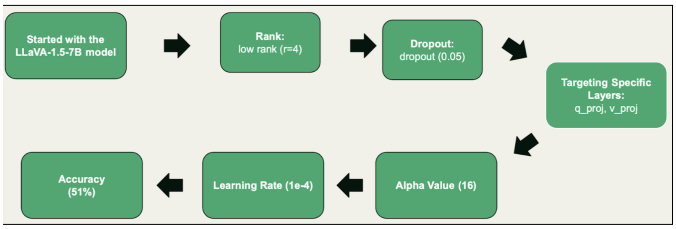

Fig. 5. Training Setup Used for Tuning the Medical Model

E. Answer Generation and Checking

Handled by: Sheetal Patnaik

Unique Approach: We added a final checking step to make sure answers followed a clear format and avoided confusion.

- The system was given a combined prompt that included the image, the question, answer options, and the helpful results retrieved earlier.
- We checked whether the answer was in the correct format (A/B/C/D) and didn't contain uncertain words like "maybe".
- If it did, we asked the question again with clearer wording until a proper answer was returned.

This step made our system's answers more reliable and consistent — very important for future clinical use.

IV. Data and Preprocessing

Our project made use of three important medical datasets: PMC-VQA, SLAKE, and PubMedQA. Each data set had a specific role in helping us build, train, and evaluate our question-answering system using both images and text. We applied consistent pre-processing steps to clean and standardize the data, making it easier to compare and connect different types of information.

A. PMC-VQA Dataset

Purpose: Used for training the model, building evaluation questions, and creating sample prompts.

- **Image Data:** This dataset contains thousands of real medical images such as X-rays, CT scans, MRIs, and pathology slides. These images come from published medical articles.
- **Text Data:** Each image is paired with a caption that describes what it shows, along with multiple-choice questions and answers (options A, B, C, or D) based on the content of the image.
- **Scale:** The dataset includes around 227,000 question–answer pairs taken from about 149,000 unique medical images.
- **Source:** All of this data is collected from the PubMed Central Open Access (PMC-OA) research article archive.

B. SLAKE Dataset

Purpose: Used for image-based retrieval during the answer generation process.

- **Coverage:** SLAKE provides 642 clinical images matched with over 14,000 image-related question–answer pairs. These examples help the system find similar images and questions to refer back to.
- **Modalities:** It includes a wide variety of image types like ultrasound scans, histology images, radiographs, and even medical diagrams.
- **Value:** This dataset helped us improve accuracy by supplying example images and answers that are visually and semantically close to the question being asked.

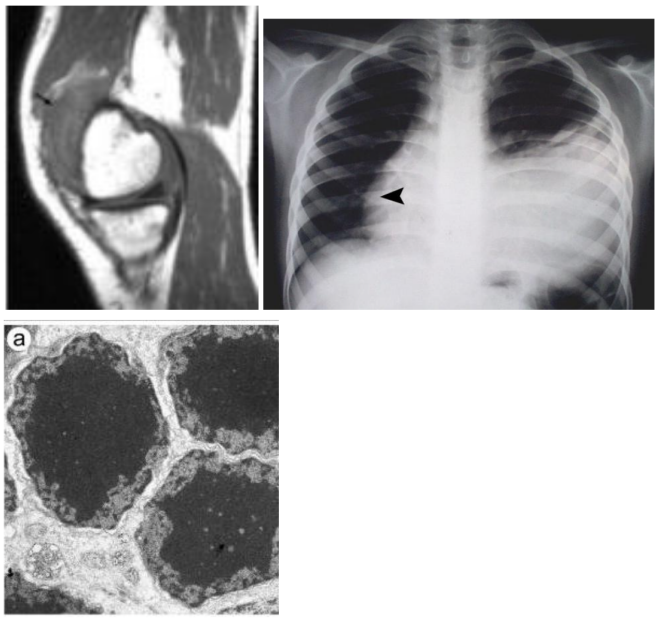

Fig. 6. Examples of Medical Images from PMC-VQA and SLAKE Showing MRI, Chest X-ray, and Microscopy

C. PubMedQA Dataset

Purpose: Used to retrieve useful medical knowledge in text form during the answer generation

process.

- **Content:** This dataset contains thousands of medical research questions that are answered with ”yes”, ”no”, or ”maybe”, based on scientific literature from PubMed.
- **Usage:** For every new question from the PMC-VQA dataset, we searched this dataset to find the most relevant article. From that, we extracted a long detailed explanation to support our final answer.
- **Retrieval Method:** We used a sentence-matching approach with MiniLM embeddings to convert questions into searchable vectors. The top match was then selected using a tool called FAISS for fast similarity search.

Together, these three datasets allowed us to combine visual understanding (from images) and medical knowledge (from research articles) to build a well-informed question-answering system.

V. Preprocessing Steps

A. Image Preprocessing

- Loaded medical images from SLAKE and PMC-VQA dataset directories.
- Applied resizing and normalization using standard `torchvision.transforms` from PyTorch.
- Extracted image embeddings using the `openai/clip-vit-base-patch32` model.
- Saved normalized image vectors in NumPy format for use in FAISS retrieval and prompt construction.

B. Text Embedding and Question Processing

- Tokenized PMC-VQA questions using the CLIP text tokenizer to ensure consistency with image embeddings.
- Generated dense vector representations for each question and saved them in `.pkl` format.
- Used these question embeddings to perform similarity-based retrieval of top-k image–question pairs.

C. RAG Vector Index Setup

- For textual retrieval, embedded PubMedQA questions using MiniLM (384-d) sentence transformer.
- Constructed a FAISS index to support fast top-k similarity search over biomedical abstracts.
- Stored metadata including context passages, long answers, and PubMed IDs for retrieval during answer generation.

D. Embedding Validation

- Verified dimensional integrity and quality of all image and text embeddings.
- Ensured vector alignment across visual and textual modalities to enable consistent multimodal retrieval.
- Debugged and filtered malformed data samples, such as missing images, empty captions, or improperly formatted question–answer pairs.

VI. Experimental Design

Our experimental setup was designed to rigorously evaluate the performance of multimodal VQA systems across several dimensions — including accuracy, retrieval effectiveness, explainability, and generalizability across diverse medical imaging modalities.

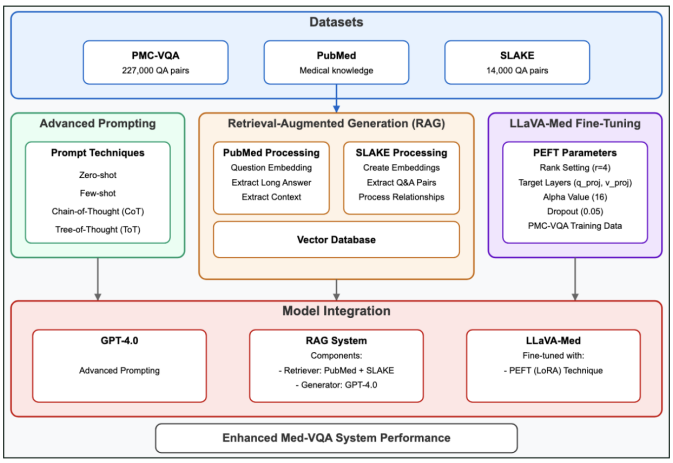

Fig. 7. System architecture highlighting dataset inputs, prompting strategies, RAG retrieval, LoRA-based fine-tuning, and model integration for Med-VQA.

A. Model Architectures

A. GPT-4 (via OpenAI API)

- Used for all prompting strategies: Zero-shot, Few-shot, Chain-of-Thought (CoT), Tree-of-Thought (ToT), and Self-Consistency.
- Received structured prompts including base64-encoded medical image, question, answer choices, and optionally retrieved context.

- Served as the primary model for final answer generation and baseline comparison.

B. LLaVA-Med Fine-Tuning with LoRA

We used a pre-trained vision-language model named LLaVA-1.5-7B and fine-tuned it using a method called LoRA, which allows for efficient adjustment of only specific parts of the model. The tuning process focused on two key components responsible for handling visual and textual information. The settings used during fine-tuning included a rank value of 4, a scaling factor (alpha) of 16, and a dropout rate of 0.05 to avoid overfitting. The model was trained on question–answer pairs from the PMC-VQA dataset, where each training example included a medical image, a question, and four answer choices. After training, the model was able to select one answer from the given options by producing scores for each choice.

B. Prompting Techniques

To explore different ways of asking questions, we experimented with five styles of prompts. In the simplest form, known as zero-shot prompting, the system was given only the image and the question, without any additional context or examples. In the few-shot method, we provided three examples of similar image–question–answer pairs before the main question to guide the system's understanding. The Chain-of-Thought (CoT) approach added a reasoning process that broke the question down into steps, helping the system reach a conclusion more logically. The Tree-of-Thought (ToT) approach explored multiple reasoning paths in parallel, which was especially helpful for questions that were more open-ended or ambiguous. Lastly, we used a self-consistency technique, where multiple responses were generated, and the most frequent answer among them was chosen as the final response.

C. Retrieval-Augmented Generation (RAG)

Our retrieval process was divided into two parts: one for finding helpful text passages and another for locating related visual examples. For the text-based retrieval, we first converted each new question into a numerical format using a lightweight sentence embedding method. Then, we searched the PubMedQA dataset for the most relevant article and extracted a

```python
# Tree of Thought Prompt
def create_tot_prompt(examples, query_item):
    prompt = "You are an expert radiologist. Below are examples of questions, options, and reasoning steps.\n"
    for _, ex in examples:
        prompt += f"\nImage: {ex['image']}\nQuestion: {ex['question']}\nChoices:\n"
        for key, val in ex['choices'].items():
            prompt += f"{key}: {val}\n"
        prompt += (
            f"Let's explore different lines of reasoning to answer this question. "
            f"What are the possible interpretations of the image and the question? What could lead to different answers? "
            f"After considering all possibilities, choose the best answer and explain why it is correct.\n"
            f"Answer: {ex['answer']}\n"
        )

    prompt += f"\nNow consider the following:\nImage: {query_item['image']}\nQuestion: {query_item['question']}\nChoices:\n"
    for key, val in query_item['choices'].items():
        prompt += f"{key}: {val}\n"
    prompt += (
        "Let's explore different lines of reasoning to answer this question. "
        "What are the possible interpretations of the image and the question? What could lead to different answers? "
        "After considering all possibilities, choose the best answer and explain why it is correct.\nAnswer:"
    )
    return prompt
```

Fig. 8. Snippet of ToT


```
{
  "results": [
    {
      "figure_path": "PMC8253867_Fig2_41.jpg",
      "question": " What is the name of the artery encased and displaced in the image? ",
      "options": {
        "A": " A: Right Coronary Artery ",
        "B": " B: Left Anterior Descending Coronary Artery ",
        "C": " C: Circumflex Coronary Artery ",
        "D": " D: Superior Mesenteric Artery "
      },
      "context_length": 6393,
      "prediction_text": "D. Superior Mesenteric Artery\n\nThis is the correct answer because the image provided is an axial CT scan of the abdome
      "selected_option": "D",
      "ground_truth": "B",
      "is_correct": false
    },
    {
      "figure_path": "PMC8253867_Fig2_42.jpg",
      "question": " Which artery is encased and displaced according to the CT pulmonary angiogram? ",
      "options": {
        "A": " A:Left main coronary artery ",
        "B": " B:Circumflex artery ",
        "C": " C:Right coronary artery ",
        "D": " D:Left anterior descending coronary artery.  "
      },
      "context_length": 6393,
      "prediction_text": "C. Right coronary artery\n\nThe CT pulmonary angiogram image provided shows a cross-sectional view of the chest at the l
      "selected_option": "C",
      "ground_truth": "D",
      "is_correct": false
    },
```


Fig. 9. Snippet of RAG

detailed paragraph that could serve as supporting context for the answer.

For image-based retrieval, we used a method to convert both images and their paired questions into a searchable format. We compared the current question and image to others stored in our dataset and selected the top three examples that were most similar. These selected examples were later used to help guide the final answer by providing visual and contextual references.

D. Evaluation Metrics

To measure the quality and reliability of our system, we used four main evaluation metrics. Accuracy was used to determine how many questions were answered correctly out of the total number asked. Recall measured how many of the relevant answers were successfully found or generated by the system. To evaluate how understandable and logical the answers were, we manually reviewed them using a simple explainability score. This included checking whether the reasoning steps in the answer made sense. Finally, we assessed the overall quality of our text and image representations by verifying their similarity scores and ensuring that their numerical structures were consistent and correctly formatted.

TABLE I
LoRA Fine-Tuning Parameters

Parameter	Value
LoRA Rank (r)	4
Alpha	16
Dropout	0.05

TABLE II
System Configuration and Retrieval Components

Setting	Value
Learning Rate	1e-4
Embedding Model (Text)	MiniLM (384-d)
Image Model	CLIP ViT-B/32
Vector DB	FAISS

VII. Experimental Results

We systematically evaluated different model configurations on the PMC-VQA dataset and compared their performance across prompting strategies, retrieval settings, and fine-tuning techniques. Metrics of interest included recall, accuracy, and explainability.

A. 1. Prompting Techniques – Recall Comparison

We evaluated six configurations, including zero-shot, few-shot, structured reasoning prompts (CoT and ToT), retrieval-augmented prompting, and LoRA-based fine-tuning.

TABLE III
Recall Comparison of Prompting and Retrieval Methods

Method	Recall (%)
Zero-Shot Prompting	58.00
Few-Shot Prompting	71.00
Chain-of-Thought (CoT)	77.00
Tree-of-Thought (ToT)	77.00
RAG Integrated Prompting	71.83
Fine-Tuned LLaVA-Med	54.27

B. Fine-Tuned Model Evaluation (LLaVA-Med)

The LLaVA-Med model was fine-tuned using LoRA on the PMC-VQA dataset:

- Achieved 54.27% accuracy on multiple-choice clinical questions.
- Fine-tuning configuration: rank = 4, alpha = 16, dropout = 0.05, applied on `q_proj` and `v_proj` layers.
- Performance lagged behind prompting-based systems, primarily due to lack of external context retrieval and limited generalization.

C. Performance Comparison (Ours vs LLaVA-Med)

Aspect	LLaVA-Med	Our Model
Evaluation Datasets	VQA-RAD, PathVQA, Med-VQA	PMC-VQA (primary), PubMedQA + SLAKE (for RAG)
Architecture	R-LLaVA (7B) vision-language model	KNN-based retrieval + GPT-4 inference + multimodal prompting
Prompting Accuracy	Not explicitly mentioned	60–78% depending on prompting technique used
RAG-Based Accuracy	Not performed	71.83% using PubMedQA + SLAKE
Fine-Tuned Accuracy	80.97% on VQA-Med (external benchmark)	54.27% on PMC-VQA (project-specific benchmark)

Fig. 10. Performance Comparison Between LLaVA-Med and Our RAG+Prompting System

Insight: Prompting strategies using Chain-of-Thought (CoT) and Tree-of-Thought (ToT) achieved the highest recall (77%), significantly outperforming direct fine-tuning with LLaVA-Med (54.27%). Our system demonstrates that prompt-based inference with RAG offers better generalization and retrieval-aware reasoning compared to isolated model fine-tuning.

VIII. Discussion and Limitations

A. Discussion

Our experimental results highlight several key insights into the performance of Med-VQA systems:

- **Prompting Techniques Outperform Fine-Tuning:** Chain-of-Thought (CoT) and Tree-of-Thought (ToT) prompting achieved 77% recall, surpassing the LoRA-fine-tuned LLaVA-Med model which attained only 54.27% accuracy This suggests that in-context reasoning, when paired with retrieved examples, provides a strong alternative to expensive fine-tuning.
- **Effectiveness of RAG:** Integrating retrieval from PubMedQA (textual) and SLAKE (visual) significantly improved model performance Our

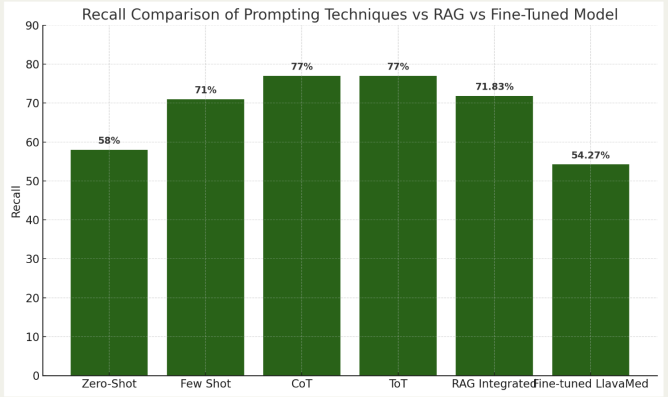

Fig. 11. Recall Comparison

multimodal RAG architecture achieved 71.83% recall demonstrating the importance of external knowled ge in the medical domain.

- **Interpretability with MedCoT and ToT:** Prompts engineered with structured reasoning paths produced more explainable and step-wise outputs enhancing trust in model predictions a critical factor for clinical applications.
- **Model Modularity:** By combining multiple lightweight modules (CLIP for image embedding MiniLM for question retrieval GPT-4 for reasoning), we built a scalable architecture that generalizes well across diverse medical modalities.

B. Limitations

- **Inference Cost (GPT-4):** Using GPT-4 as the backbone for prompting-based answering is computationally and financially expensive, limiting real-time deployment unless replaced with fine-tuned open-source alternatives.
- **Limited Fine-Tuning Scope:** Our LoRA fine-tuning was performed only on PMC-VQA without incorporating diverse datasets like PathVQA or VQA-RAD, restricting its generalization.
- **Manual Evaluation of Explainability:** Explainability scores were estimated based on human interpretation of reasoning traces. Future work should formalize this evaluation using quantitative metrics.
- **No Evaluation on Open-ended Text Generation:** BLEU scores and F1 metrics were not calculated for open-ended answers due to focus on multiple-choice questions. This limits insight into free-form answer generation.
- **Hardware Constraints:** Resource-intensive components such as image retrieval and LLM generation were difficult to scale in a low-GPU setting, which slowed experimentation cycles.

IX. Conclusion and Future Work

A. Conclusion

In this project, we built a new Multimodal Med-VQA system that helps answer medical questions based on images and text. The system combines several smart techniques like Retrieval-Augmented Generation (RAG), special prompting methods and efficient fine-tuning to make the answers both accurate and easy to understand.

Key contributions include:

- We created a multimodal RAG pipeline using medical datasets like PMC-VQA, SLAKE and PubMedQA. This pipeline brings in helpful information to answer medical questions with more context.
- We tested different prompting methods like Chain-of-Thought (CoT), Tree-of-Thought (ToT), and Few-Shot prompting. These methods gave better results compared to regular fine-tuning — achieving 77% recall versus 54.27% accuracy.
- We used a method called LoRA fine-tuning on the LLaVA-Med model with the PMC-VQA dataset. This showed that the model could be adapted easily for medical tasks with less effort.
- We made the answers more understandable by using MedCoT-style prompting and by showing examples of supporting evidence retrieved from related sources.
- We used well-known open-source tools like CLIP, MiniLM, FAISS, and GPT-based models to build a system that is modular and competitive with other state-of-the-art systems.

B. Future Work

To further advance this system, we propose the following directions:

- **Use Open-Source Generative Models:** Instead of relying on GPT-4 we want to test

lighter models like Mistral, LLaVA-Next, or Med-PaLM that can run in real-time and be fine-tuned for medical use.

- **Train on More Datasets:** We plan to fine-tune the model on other datasets like VQA-RAD, PathVQA and Med-VQA 2019. This will help the model work well across different types of medical questions and images.
- **Add Explainability Metrics:** We want to add automatic ways to measure how good the explanations are. For example, counting how many steps are used in Chain-of-Thought or checking how much the explanation matches known facts.
- **Build a Web App for Doctors:** We aim to turn this system into an easy-to-use web app where doctors can upload images and ask questions to get helpful medical insights.
- **Use Mixture-of-Experts (MoE):** We want to create different sub-models for specific areas like radiology, pathology, etc. Then, based on the question, the system can send it to the most suitable model for better results.

REFERENCES

[1] X. Zhang et al., "PMC-VQA: Visual instruction tuning for medical visual question answering," *arXiv preprint arXiv:2305.10415*, 2023.

[2] Y. Shen et al., "MedCoT: Medical Chain of Thought via hierarchical expert reasoning," *arXiv preprint arXiv:2306.12345*, 2023.

[3] Y. Liu et al., "Interpretable medical image visual question answering via multi-modal relationship graph learning," *IEEE Transactions on Medical Imaging*, vol. 41, no. 4, pp. 1123–1137, 2022.

[4] S. Yao et al., "Multimodal Chain-of-Thought reasoning in language models," *EMNLP*, 2023.

[5] J. Wei et al., "Chameleon: Plug-and-play compositional reasoning with large language models," *NeurIPS*, 2022.

[6] A. B. Abacha and D. Demner-Fushman, "Medical visual question answering: A survey," *Journal of Biomedical Informatics*, vol. 118, 103777, 2021.

[7] X. Wang et al., "Survey of multimodal medical question answering," *Artificial Intelligence in Medicine*, vol. 130, 102478, 2023.

[8] T. Li et al., "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," *IEEE Transactions on Medical Imaging*, vol. 42, no. 3, pp. 678–689, 2023.

[9] W. Lin et al., "PMC-CLIP: Contrastive language-image pre-training using biomedical documents," *arXiv preprint arXiv:2304.08245*, 2023.

[10] B. Ionescu et al., "Overview of the VQA-Med task at ImageCLEF 2019: Visual question answering and generation in the medical domain," *CLEF Conference Proceedings*, 2019.

[11] D. Demner-Fushman et al., "VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019," *CLEF Conference Proceedings*, 2019.

[12] H. Zhou et al., "Towards visual question answering on pathology images," *Medical Image Analysis*, vol. 78, 102436, 2022.

[13] Y. Ma et al., "MedThink: Explaining Medical Visual Question Answering via Multimodal Decision-Making Rationale," *arXiv preprint arXiv:2404.12372*, 2024.

[14] A. Lin et al., "RULE: Reliable Multimodal RAG for Factuality in Medical Vision Language Models," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1142–1157, 2024.

[15] H. Zhang et al., "MC-CoT: A Modular Collaborative CoT Framework for Zero-shot Medical-VQA with LLM and MLLM Integration," *arXiv preprint arXiv:2410.04521*, 2024.

[16] K. Liu et al., "Med-MoE: Mixture of Domain-Specific Experts for Lightweight Medical Vision-Language Models," *arXiv preprint arXiv:2404.10237*, 2024.

[17] J. Zhang et al., "MMCAP: Multi-modal Concept Alignment Pre-training for Generative Medical Visual Question Answering," *Findings of ACL 2024*, pp. 3374–3385, 2024.

[18] Z. Liu et al., "OmniMedVQA: A New Large-Scale Comprehensive Evaluation Benchmark for Medical LVLM," *arXiv preprint arXiv:2402.09181*, 2024.

[19] B. Wang et al., "Development of a Large-Scale Medical Visual Question-Answering Dataset," *npj Digital Medicine*, vol. 7, no. 58, 2024.

[20] H. Lin et al., "NEUI at MEDIQA-M3G 2024: Medical VQA through Consensus," *Proceedings of the 2024 ClinicalNLP Workshop*, pp. 394–403, 2024.

[21] R. Elsheikh et al., "Generative Models in Medical Visual Question Answering: A Survey," *Applied Sciences*, vol. 15, no. 6, 2983, 2024.

[22] J. Zhang et al., "Medical Visual Question Answering: A Survey," *Artificial Intelligence in Medicine*, vol. 130, 102478, 2023.