Final Project Proposed Topic

Extending MedVInT-TD to Microscopic Imaging: Domain-Specific Fine-Tuning for Medical Visual Question Answering

Class: Artificial Intelligence Course

Department of Computer Science

University of South Dakota

Group members:

Syed Ali Jaseem

Supervisor: Dr. Lina Chato

Fall 2025

Group Members Info			
No.	**Students' First & last name**	**CSC 457 or CSC 557**	**Section (U20, U15, or U19)**
1	**Syed Ali Jaseem**	**CSC 557**	**U20**
2			
3			
4			
5			

1. Objective

Medical Visual Question Answering (MedVQA) is an emerging field that combines medical image understanding with natural language reasoning, allowing AI systems to answer clinical questions based on medical images. It has the potential to support clinicians, improve patient communication, and advance medical education.

The paper "Development of a large-scale medical visual question-answering dataset" introduced PMC-VQA, a large dataset containing 227,000 question–answer pairs across 149,000 medical images. Alongside this dataset, the authors proposed two generative models, MedVInT-TD and MedVInT-TE, and fine-tuned them on public MedVQA benchmarks such as VQA-RAD, SLAKE, and ImageClef. Their models achieved state-of-the-art results, significantly outperforming prior approaches.

Building on this foundation, my project aims to further fine-tune the MedVInT-TD (open-ended "blank" variant) model on a custom dataset. The expectation is that leveraging domain-specific data on top of a strong pre-trained backbone will yield improved accuracy and robustness in generating free-text answers for medical visual questions. This fine-tuning will adapt the model to my dataset's unique characteristics, with the goal of achieving strong performance comparable to, or exceeding, existing benchmarks.

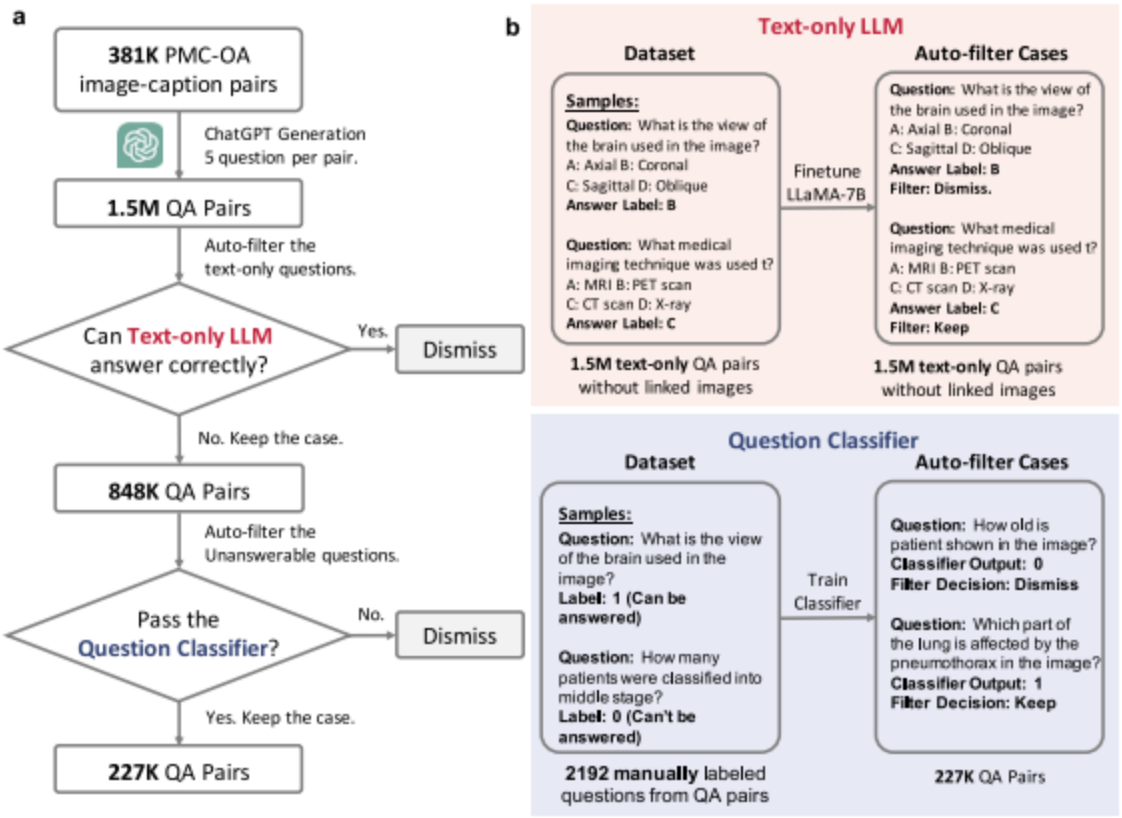
2. Data

The foundation of this work is built on the PMC-VQA dataset, which was introduced to address the lack of large-scale, multimodal datasets for Medical Visual Question Answering (MedVQA). Prior to this effort, most MedVQA datasets were relatively small in scale, limiting the ability to train high-performing generative models.

The authors constructed PMC-VQA by leveraging PMC-OA, a biomedical dataset containing 1.6 million image–text pairs collected from PubMed Central's Open Access subset, covering over 2.4 million scientific papers. The PMC-OA dataset was created in three stages:

1. Medical figure-caption collection
2. Subfigure separation
3. Subcaption separation and alignment

For PMC-VQA, the authors used a version containing 381K image–caption pairs taken from the first stage, preserving diversity and complexity by not performing subfigure auto-separation. Importantly, the dataset was constructed entirely from open-source resources available for research, ensuring compliance with ethical and regulatory standards.



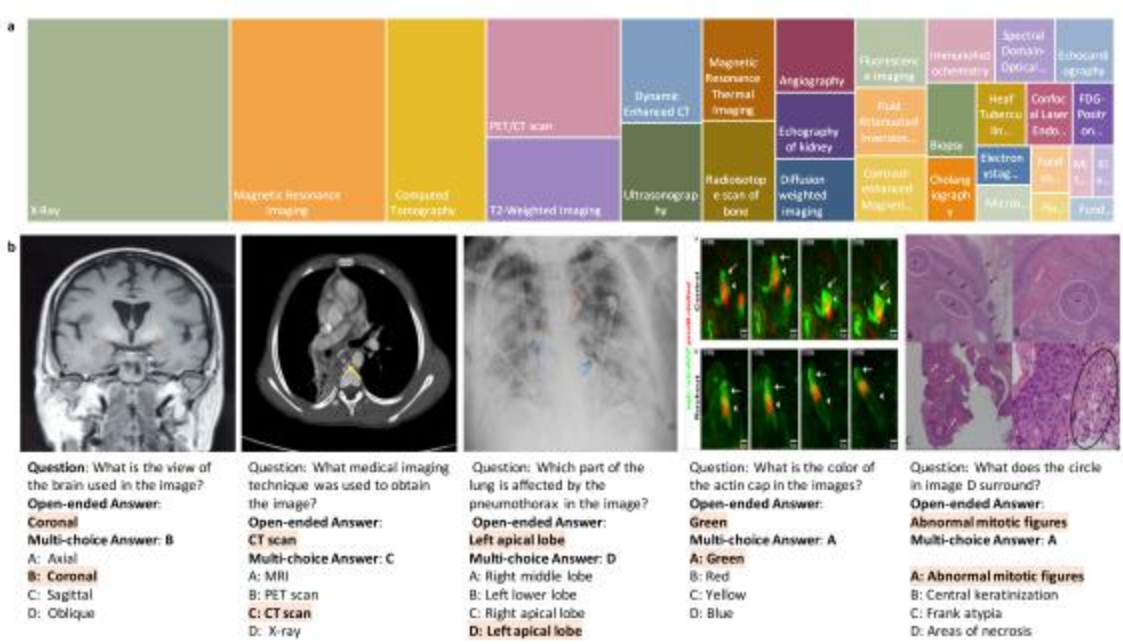
Question–Answer Generation

To automatically produce large-scale question–answer pairs, the image captions from PMC-OA were used as input to ChatGPT. The model was prompted to generate five Q&A pairs per caption, each with one correct answer and three distractors in multiple-choice format. This process produced 1.49 million Q&A pairs, averaging about 3.93 questions per image.

Automatic and Manual Filtering

Since some caption-derived questions could be answered with text alone (without the image), a LLaMA-7B language-only model was fine-tuned to identify and remove such cases. If the model could consistently answer a question correctly, that question was excluded. This filtering step reduced the dataset to 848,433 pairs that truly required visual grounding.

Next, the authors addressed questions that relied too heavily on caption information rather than the image itself. A binary classifier was trained on a manually annotated set of 2,192 Q&A pairs to distinguish between image-answerable and non-image-answerable cases. With this classifier achieving 81.77% accuracy, the dataset was further refined, resulting in the final PMC-VQA dataset of 226,946 Q&A pairs across 149,075 images.



Dataset Splits and Evaluation Sets
From this refined dataset, the authors created evaluation sets:

- PMC-VQA-test-initial: 50,000 image–question pairs with no image overlap between training and test sets.
- PMC-VQA-test: A high-quality test set of 2,000 pairs manually verified for relevance, difficulty of distractors, and image quality.

Relevance to This Project
The PMC-VQA dataset, being the largest and most diverse MedVQA dataset to date, provided the foundation for training the MedVInT-TD and MedVInT-TE models, which were then fine-tuned on multiple public benchmarks (VQA-RAD, SLAKE, ImageClef). These models consistently achieved state-of-the-art performance across open- and close-ended tasks, demonstrating the effectiveness of large-scale pretraining followed by benchmark-specific fine-tuning.

Because MedVInT-TD (blank variant) has already been exposed to vast multimodal data and shown strong generalization on multiple benchmarks, I expect that further fine-tuning it on my own dataset of cell images will yield strong performance. The combination of large-scale pretraining and domain-specific adaptation makes this model a powerful candidate for achieving accurate, context-aware answers in my project.

PMC-VQA Dataset

3. Hypothesis and expectations

The central hypothesis of this project is that the MedVInT-TD (blank variant) model, after fine-tuning on my custom dataset of microscopic cell images, will be able to detect and localize unhealthy cells and provide accurate free-text responses to clinically relevant questions.

The original PMC-VQA dataset exposed MedVInT to a wide range of medical modalities (radiology, pathology, microscopy, and others), enabling it to learn strong multimodal reasoning skills. Although its pretraining was not explicitly focused on cellular-level analysis, the model's architecture and generative nature position it well for transfer learning into this finer-grained biomedical context.

In my dataset, each image contains multiple cells, and the associated questions will emphasize both diagnosis ("Are there unhealthy cells present?") and localization ("Where are the unhealthy cells located?"). This extends the MedVQA paradigm from global image-level interpretation to spatially-aware, localized reasoning.

I expect that fine-tuning will lead to the following outcomes:

1. **Accurate Abnormality Detection**

   The model should learn to distinguish between healthy and unhealthy cells based on subtle visual cues, such as morphology, staining intensity, or structural irregularities.

   Given a binary-type question ("Are there unhealthy cells present?"), the model should reliably provide correct yes/no answers with minimal hallucination.

2. **Localization in Natural Language**

   Beyond binary classification, the model should generate free-text answers that indicate location of abnormalities, e.g., "Yes, unhealthy cells are clustered in the upper left region of the image."

   This requires the model to map visual features to interpretable spatial descriptors, a step toward bridging detection with natural language explanation.

3. **Generalization Across Samples**

   By fine-tuning on cell-level images, the model is expected to generalize to unseen samples within the same domain, maintaining robustness across variations in cell density, lighting, or imaging quality.

   This will validate that pretraining on large-scale PMC-VQA, followed by domain-specific adaptation, allows the model to transfer knowledge effectively from radiology/pathology contexts to cell microscopy tasks.

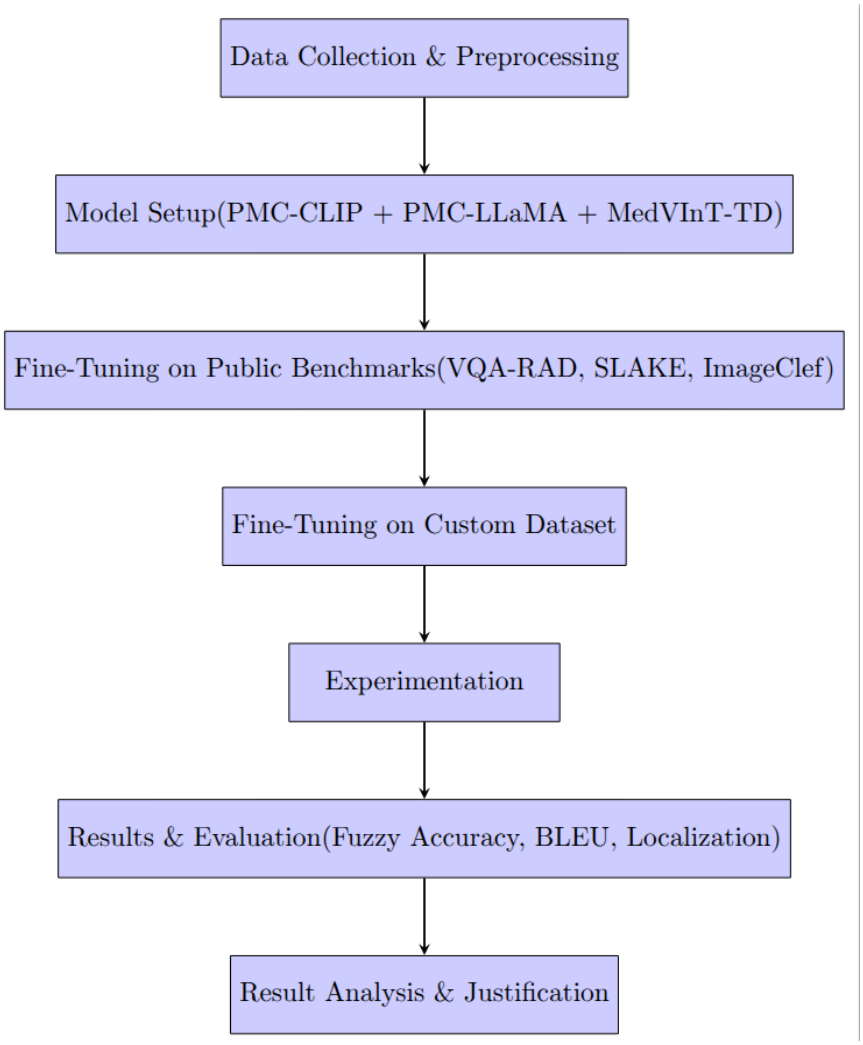
4. **Improved Biomedical Applicability**

   If successful, this approach demonstrates the feasibility of using generative MedVQA models in microscopic diagnostic workflows, where both presence and location of abnormalities are crucial.

The ability to generate interpretable explanations in natural language can make the system more useful for medical research and educational purposes.

Overall, I expect that fine-tuning MedVInT-TD on this dataset will yield a model capable of producing accurate, spatially-aware, and clinically relevant answers, going beyond global image understanding and pushing MedVQA into more granular diagnostic applications.

4. Project Process and Design
can be flowchart or graph to show steps starting from data processing to result justification



References

Main

1. Paper: Development of a large-scale medical visual question-answering dataset
2. MedVInT_TD blank (open-ended questions) Model Trained on PMC-VQA Dataset
3. PMC-CLIP: Contrastive Language-Image Pre-training using Biomedical Documents (Trained on PMC-OA Dataset)
4. PMC-LLAMA: LLaMA-7b finetuned on the PMC papers in S2ORC dataset.
5. PMC-VQA Dataset
6. VQA-RAD Dataset
7. SLAKE Dataset
8. ImageClef Dataset

Additional

9. https://arxiv.org/pdf/2303.07240
10. https://pmc.ncbi.nlm.nih.gov/tools/openftlist/
11. https://medmnist.com/