CSE428 Final Report

Yushan (Shayla) Huang

June 2025

1 Abstract

This project explores domain specific adaptation of a large multi-modal vision-language model, LLaVA-NeXT, to improve its performance on biomedical visual quetsion answering (VQA) tasks. Although models like LLaVA-NeXT demonstrate impressive general-purpose capabilities, they often underperform in specialized domains such as medicine due to lack of domain-specific knowledge. To address this, this project fine-tunes the language components of LLaVA-NeXT using adapter-based methods (LoRA), allowing efficient training on a curated dataset of medical images from PubMed Central, paired with clinical captions and VQA annotations. Our fine-tuned model demonstrates a slight improvement of understanding of medical terminology compared to the base model. We also established an evaluation pipeline to compare performance. This work highlights the feasibility and importance of domain adaptation for multi-modal models in applications like medical AI.

2 Context - Related Work and Background

Recent advances in multi-modal large language models have made significant progress in tasks that requires understanding both visual and textual information.

LLaVA and LLaVA-NeXT

LLaVA (Large Language and Vision Assistant) combines a vision encoder wiith a language model using projection layers. LLaVA-NeXT, the latest in the series of LLaVA models, introduces improvement such as better multi-turn chat templates and performance enhancements like FlashAttention2. It is a strong general-purpose baseline, but not specifically tuned to handle the biomedical image and language.

LLaVA-Med

LLaVA-Med is a specialized adaptation of LLaVA for medical applications. It is trained on medical visual question answering (VQA) datasets derived from biomedical papers and clinical data (PMC-15M). The dataset used in this project is based on those provided in the LLaVA-Med benchmark, which includes annotated questions, captions and answers corresponding to medical images, which is valuable for domain adaptation and performance evaluation.

BiomedCLIP

BiomedCLIP is a vision-language foundation model pretrained on large-scale biomedical dataset using contrastive learning objective. BiomedCLIP provides high-quality image encoders that can be used for downstream biomedical visual tasks. In this project it was considered as a potential future integration to enhance the model's performance on domain-specific images.

While general-purpose Multi-modal LLMs like LLaVA-NeXT show impressive capabilities, their lack of biomedical specialization can result in underperformance on fine-grained tasks in healthcare. On the other hand, specialized models like BiomedCLIP focus primarily on representation learning rather than instruction following. This project aims to bridge the gap by fine-tuning LLaVA-NeXT model on biomedical VQA data, further improve the instruction-following capabilities in medical image-text knowledge.

3 Experimental Setup

3.1 Dataset

The dataset used for fine-tuning is based on the instruction dataset from LLaVA-Med, originally 10k image-caption-question-answer samples derived from PMC-15M, a large-scale biomedical image-text dataset. Each sample includes a medical image, a caption describing the image, a medical question, and a GPT-generated answer.

Before using it to fine-tune, I enhanced the dataset to integrate clinical reasoning. For each sample, I used the DeepSeek ai. model to generate step-by-step answers that include clinical reasoning to simulate how real doctors reason about medical images based on context and language. The final fine-tuning dataset consists of these reasoning-augmented answers, is supposed to teach the model to embed diagnostic thinking.

The evaluation was conducted using the LLaVA-Med test set, which includes 193 image-question pairs, also derived from PMC-15M. This set allows for evaluating how well the model performs to biomedical VQA.

3.2 Baseline

The baseline model is LLaVA-NeXT, a multi-modal model that integrates vision and language via pretrained transformers. I fine-tuned it using LoRA (Low=Rank Adaptation) applied only to the language model while keeping the image encoder frozen, to efficiently adapt the model to the biomedical domain without affecting the general visual understanding.

3.3 Methods

To adapt LLaVA-NeXT to the biomedical domain, I applied adapter-based fine tuning using LoRA. This method allows the language model to specialize in medical reasoning without retraining the entire network, which makes it much more computationally efficient and parameter-efficient. Specifically, LoRA adapters were inserted to the language model layers only, while the vision encoder remained frozen throughout training. In which case, there are 4,194,304 trainable parameters out of 8,358,955,008 total parameters, which is around 5%.

The primary innovation in the project lies in the clinical reasoning enhancement applied to the dataset. Rather than using the original GPT-generated answers from the LLaVA-Med instruction dataset, I prompted the DeepSeek-LLM-7B-base model to generate step-by-step reasoned answers for each question based on the caption, question, and the original answer, excluding the image itself. This emulates the kind of logical thinking a clinician might use when interpreting medical images, even only with text.

The fine-tuning process ran for 20 epochs with a relatively small dataset (600 examples), which was sufficient to fit and generalize to structured reasoning in the medical context. Padding tokens were set explicitly, and all model outputs were generated with deterministic decoding (temperature set to 0) to encourage consistency during evaluation.

This setup differentiates from the baseline LLaVA-NeXT model, which lacks domain specific reasoning and is not fine-tuned on biomedical tasks. By integrating clinical step-wise logic into the training process while maintaining efficient adaptation, this method links the general-purpose vision-language models with speciaalized medical VQA system.

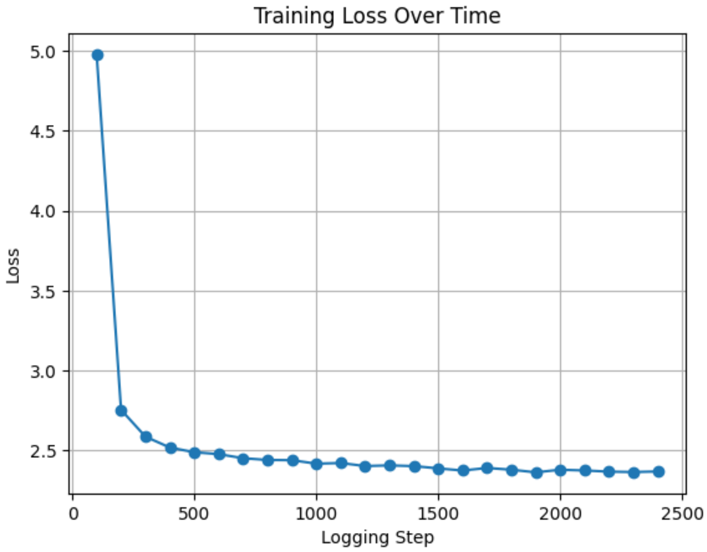

Figure 1: Fine-tuning Training Loss

Evaluation methods

To evaluate the effectiveness of the fine-tuned model, I used a comparative, LLM-based evaluation framework. Each model's answer, both the base LLaVA-NeXT and the fine-tuned model) was compared against the ground truth answer, which was originally generated by GPT-4 in the LLaVA-Med dataset.

The evaluation was done using Qwen1.5-7B-Chat, a general-purpose model. For each question-image pair, the prompt provided the image caption, user question, and two assistant responses (from the model and the ground truth answer). The prompt requested Qwen to evaluate the helpfulness, relevance, accuracy, and level of detail of each answer and assign a numerical score from 1 to 10.

Prompt: *We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above. The user asks the question on observing an image. For your reference, the visual content in the image is represented with caption describing the same image. Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance. Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. Pleas only include the two scores in the output, without any additional text or explanation.*

The scores are used to compute the average rating; the average rating for each question type (conversation and detailed description); and the average rating for each category (Chest X-ray, MRI, Histology, Gross, and CT-scan), as well as the relative score compared to the ground truth, per model. This setup enables a a more human-aligned and quantitative evaluation for tasks with no binary answer like VQA.

4 Result

We used QWEN to examine the performance of both model, compared to the ground truth GPT-4 generated answer. Overall, the fine-tuned model demonstrates comparable performance to the base model, with an overall score of 6.92 versus base model's 7.01. Relative to the GPT-4 ground truth, the base model achieves 79.02% while the fine-tuned model achieves 78.02%. From the distribution, two model's performances are very similar, except the fine-tuned model has more samples around 80% relative to the ground truth. This minor drop suggests that the adapter-based fine-tuning still preserves performance of the base model.

Category	GPT-4 Score	Base Model Score	Fine-tuned Score	Base Relative (%)	Fine-tuned Relative (%)
Conversation	~8.85	7.10	6.98	80.37	79.17
Detailed Description	~8.93	6.72	6.76	75.18	75.79
Chest X-ray	~8.91	7.22	7.27	80.88	81.70
MRI	~8.83	7.03	6.89	79.48	78.13
Histology	~8.98	7.05	7.00	78.04	78.13
Gross	~8.74	6.88	6.76	79.12	78.83
CT Scan	~8.86	6.85	6.67	77.88	75.04
Overall	**8.89**	**7.01**	**6.92**	**79.03**	**78.30**

Table 1: Comparison of GPT-4, Base Model, and Fine-tuned Model Scores Across Categories

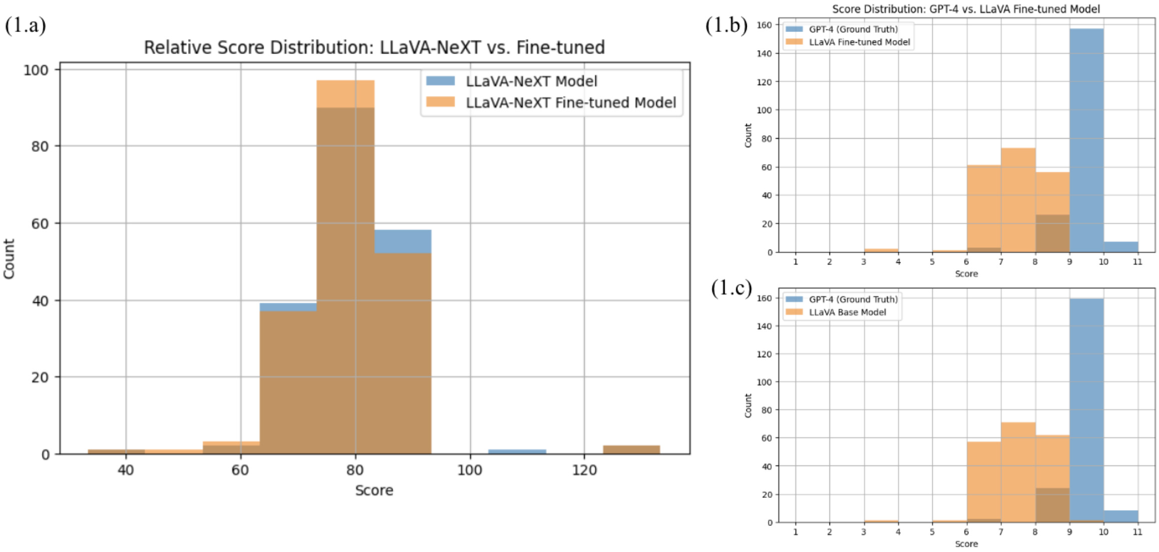

Figure 2: **1.a** shows the histogram for the relative score of both LLaVA-NeXT baseline model and the fine-tuned model, relative to the ground truth. (Calculated as score / ground truth score). **1.b** and **1.c** shows the distribution of absolute score, out of 10, of each model and the Ground truth.

By the type of each question, if it is in conversation type, or asking for detailed description. The performances on conversation are very close, with LLaVA-NeXT model slightly outperforms, while the fine-tuned model performs better in the detailed description tasks. The overall distribution of both models are almost identical.

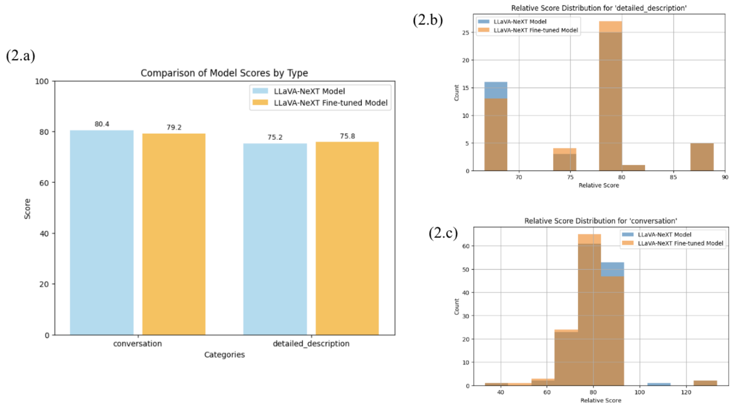

Figure 3: **2.a** shows the relative scores of each model (LLaVA-NeXT and fine-tuned) categorized by type (conversation and detailed description). **2.b** is the distribution of the relative scores of each model if the sample is in "conversation type.**2.c** is the distribution of the relative scores of each model if the sample is in "detailed description" type.

By the category of each question, we can see the Chest X-ray and histology slightly outperforms the base model, indicating improved interpretability in image-related and scenarios that requires heavy explanation. This aligns with the goal of enhancing clinical reasoning. On the other hand, the performance in CT scan and MRI drops, which reflect the domain-specific challenges in generalization.

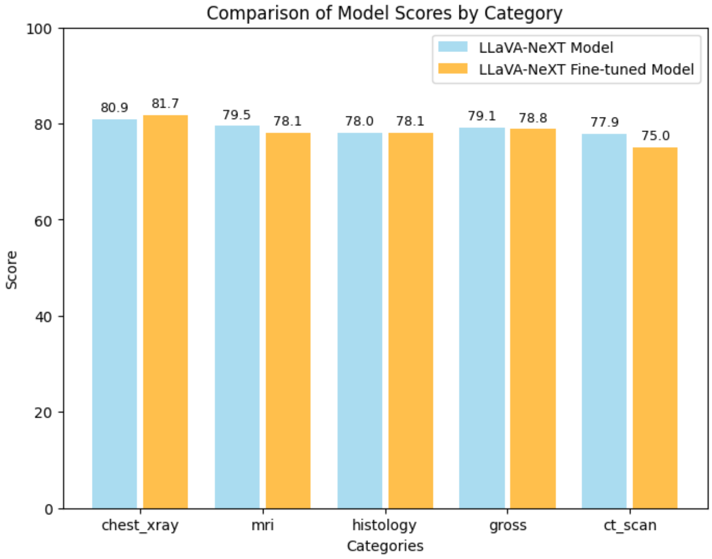

Figure 4: This bar chart shows the relative scores of each model (LLaVA-NeXT and fine-tuned) categorized by category (Chest X-ray, MRI, Histology, Gross, and CT-scan)

5 Discussion

The results showed trade-offs in integrating clinical reasoning into a vision-language model through lightweight adapter-based fine-tuning. Although the overall performance of the fine-tuned model is slightly lower than the base model (78.30% vs. 79.03% relative to GPT-4), the detailed comparison across categories shows that fine-tuning did not decrease the model's general performance and even offered improvement in areas that requires more reasoning.

Specifically, the fine-tuned model shows improvement in chest X-ray and detailed description, categories that typically require step-by-step logic, clinical observations, and narrative construction. These improvements suggest that the reasoning added via DeepSeek-generated explanations enhanced the model's ability to mimic clinical thought processes. This aligns with the goal of "thinking" more like clinicians.

However, the decline in scores of CT scan, MRI, and gross pathology may due to several factoes:

- They may involve more domain-specific knowledge that was not fully captured by the text-only reasoning added during fine-tuning.
- The size of the fine-tuning dataset (∼600 samples) may be insufficient for the model to generalize across all categories.
- The evaluation is fully relied on the captions without direct access to the image itself, which may limit the full potential of reasoning.

In terms of evaluation method, like any LLM-based evaluation, using Qwen may introduce variance or subtle biases. Yet this is one of the most interpretable and consistent way to quantify the performance on VQA.

Overall, the results shows the feasibility of using adapter-based fine-tuning to integrate with domain-specific reasoning without significant loss of general performance. With a larger dataset and integration of image features, the model has the potential to be more accurate and explainable in clinical VQA tasks.

6 Future work

In future work, I plan to incorporate image encoder fine-tuning to more directly adapt visual understanding to biomedical domains, also experimenting with more detailed reasoning chains during training, and exploring alignment techniques. More broadly, I believe the intersection of vision-language models and clinical reasoning has the potential to significantly improve AI-assisted diagnosis and education in healthcare.