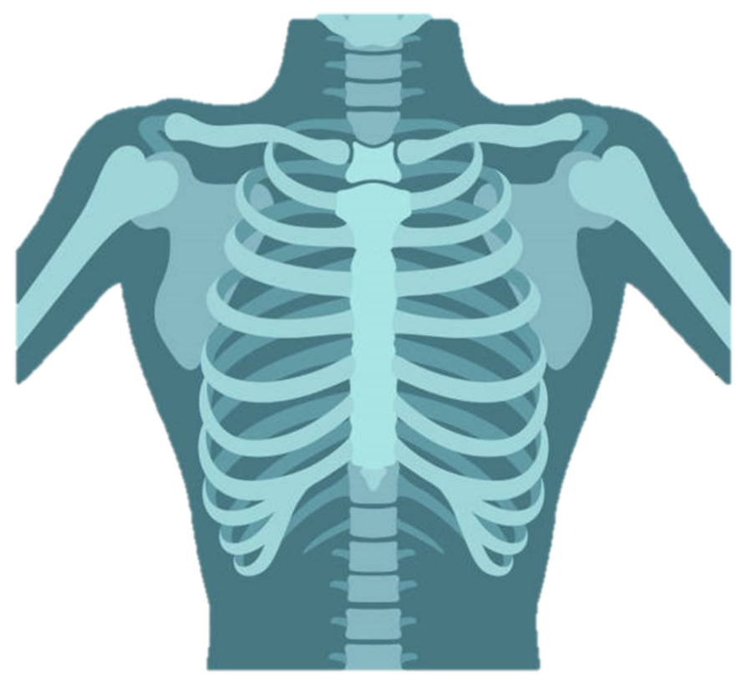
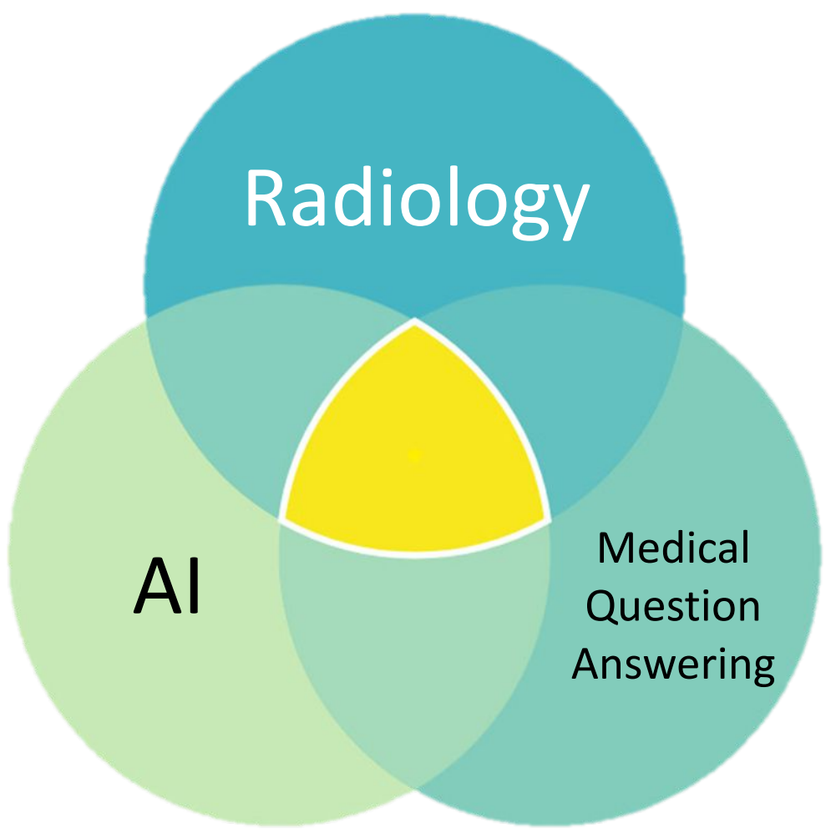
Vision Language Assistant for Medical Image Analysis and EHR Integration

Jain Arav, Mathur Amulya

Introduction



VLMs bridge visual and language understanding for applications in **Radiology** and **Pathology**. By analyzing both visual and textual data, they enhance diagnosis, treatment planning, and patient care. Integrated with Large Language Models (LLMs), they streamline clinical workflows, handling medical reports, insurance bills, and more for a comprehensive view of patient information.



Objective

- Fine-tune the Large Language and Vision Assistant **(LLaVA)** model on a custom dataset, **EHRXQA**, and do a comparative analysis between the fine-tuned llava-13B model, the baseline llava-13B model and a baseline Llava-med model.

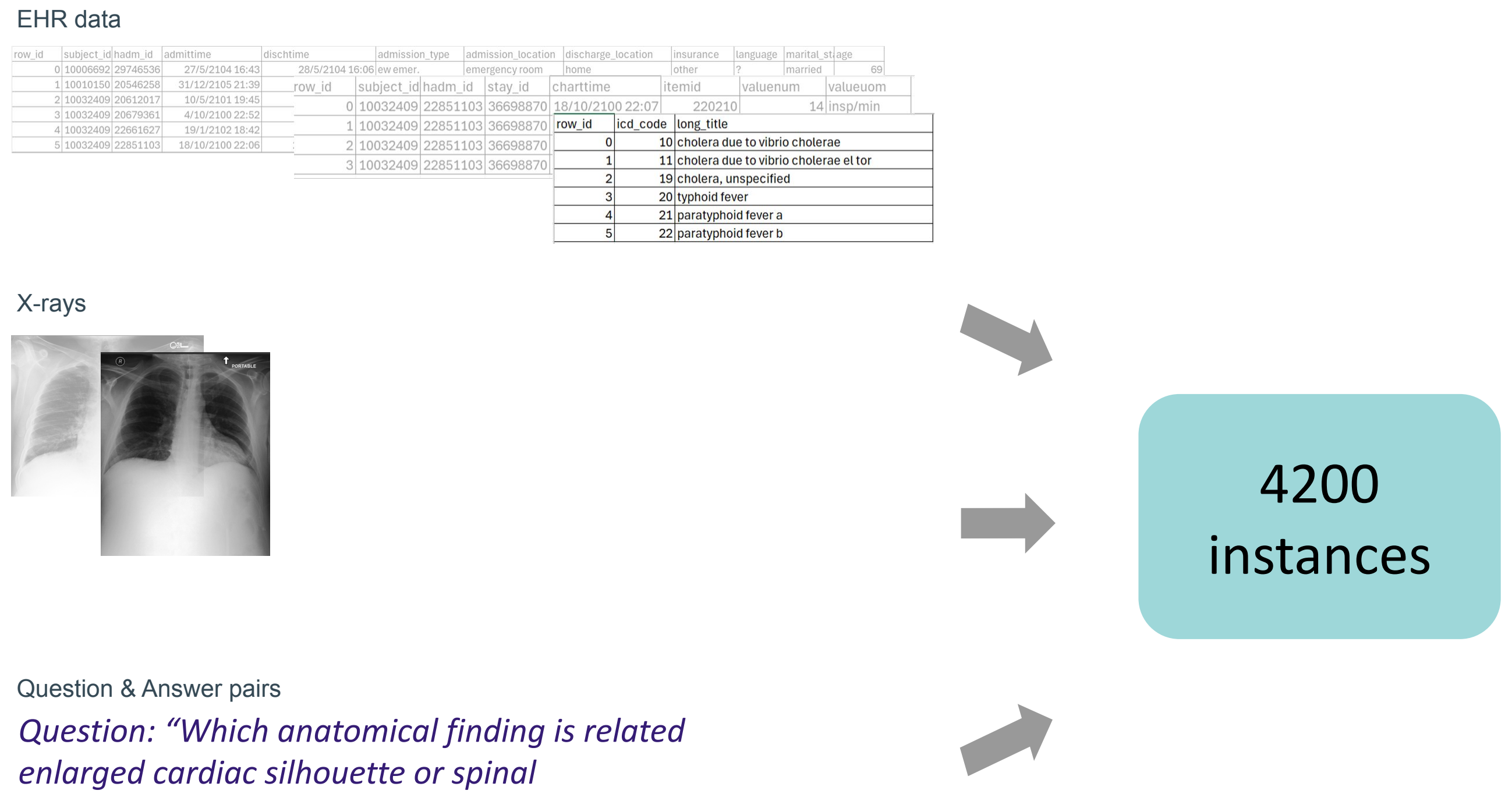
Dataset

The MIMIC-CXR QA dataset combines:

- **EHR tables (mimic-iv)**
- **Chest X-ray images**
- **Question and answer pairs**

Key fields of focus: patient_id, study_id (patient-specific X-ray study), questions, and answers.

The combined dataset consists of 4200 lung X-ray-related questions, designed to provide clinical insights of patient's health. The dataset was then split into train, validation, and testing for fine-tuning and evaluating the LLaVA model



EHR data

row_id	subject_id	hadm_id	admittime	dischtime	admission_type	admission_location	discharge_location	insurance	language	marital_st	age
0	10006692	29746536	27/5/2104 16:43	28/5/2104 16:06	ew emer.	emergency room	home	other	?	married	69
1	10010150	20546258	31/12/2105 21:39								
2	10032409	20612017	10/5/2101 19:45								
3	10032409	20679361	4/10/2100 22:52								
4	10032409	22661627	19/1/2102 18:42								
5	10032409	22851103	18/10/2100 22:06								

row_id	subject_id	hadm_id	stay_id	charttime	itemid	valuenum	valueuom
0	10032409	22851103	36698870	18/10/2100 22:07	220210	14	insp/min
1	10032409	22851103	36698870				
2	10032409	22851103	36698870				
3	10032409	22851103	36698870				

row_id	icd_code	long_title
0	10	cholera due to vibrio cholerae
1	11	cholera due to vibrio cholerae el tor
2	19	cholera, unspecified
3	20	typhoid fever
4	21	paratyphoid fever a
5	22	paratyphoid fever b

X-rays

4200 instances

Question & Answer pairs

Question: "Which anatomical finding is related enlarged cardiac silhouette or spinal degenerative changes?"
Answer: "enlarged cardiac silhouette"

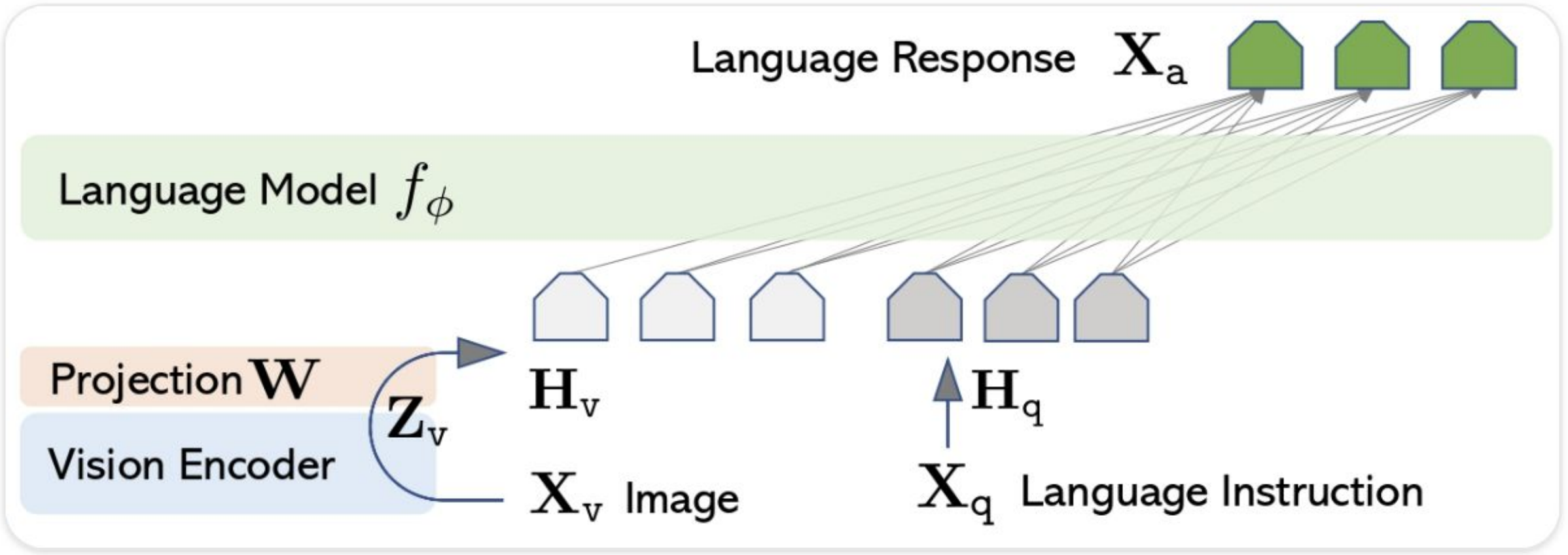
Table 1: Sample questions in EHRXQA, categorized by **modality-based** (*Image*, *Table*, *Image+Table*) and **patient-based** scope (*none*, *single*, *group*), illustrating our dataset's diversity and complexity.

modality-based	patient-based		Sample question
Image	*single*	1-*image*	Given the last study of patient 15439, which anatomical finding is associated with the right lower lung zone, pneumothorax or vascular redistribution?
		2-*image*	Enumerate all diseases that are newly detected based on the last study of patient 19290 in 2103 compared to the previous study.
		N-*image*	How many times has the chest X-ray of patient 18489 shown linear/patchy atelectasis in the left lung on the current hospital visit?
	group		Count the number of patients whose chest X-ray studies this year showed any abnormalities in the mediastinum.
Table	*none*		What's the cost of a drug named lopinavir-ritonavir?
	single		Did patient 16164 receive any magnesium lab tests last year?
	group		What was the top three diagnosis that had the highest two year mortality rate?
Image+Table	*single*		Did a chest X-ray study for patient 15110 reveal any anatomical findings within 2 month after the prescription of hydralazine since 2102?
	group		Provide the ids of patients in the 20s whose chest X-ray showed low lung volumes in the right lung this month.

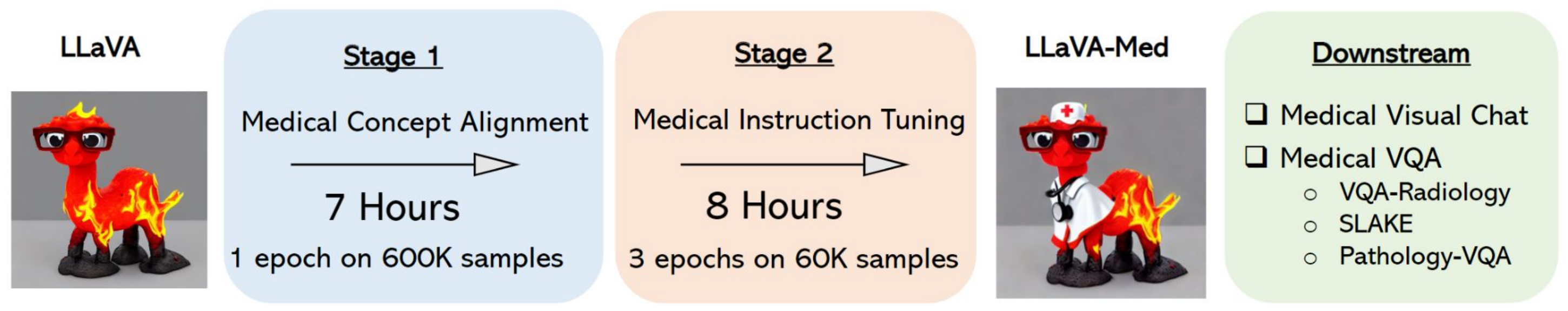
LLaVA and LLaVA-Med

LLaVA represents a novel end-to-end trained large multimodal model that combines a vision encoder and Vicuna for general-purpose visual and language understanding, achieving impressive chat capabilities.

Large Language and Vision Assistant for BioMedicine **(LLaVA-Med)** is a LLaVA model which is trained on the **PMC-15M dataset.** LLaVA-Med exhibits excellent multimodal conversational capability and can follow open-ended instruction to assist with inquiries about a biomedical image.



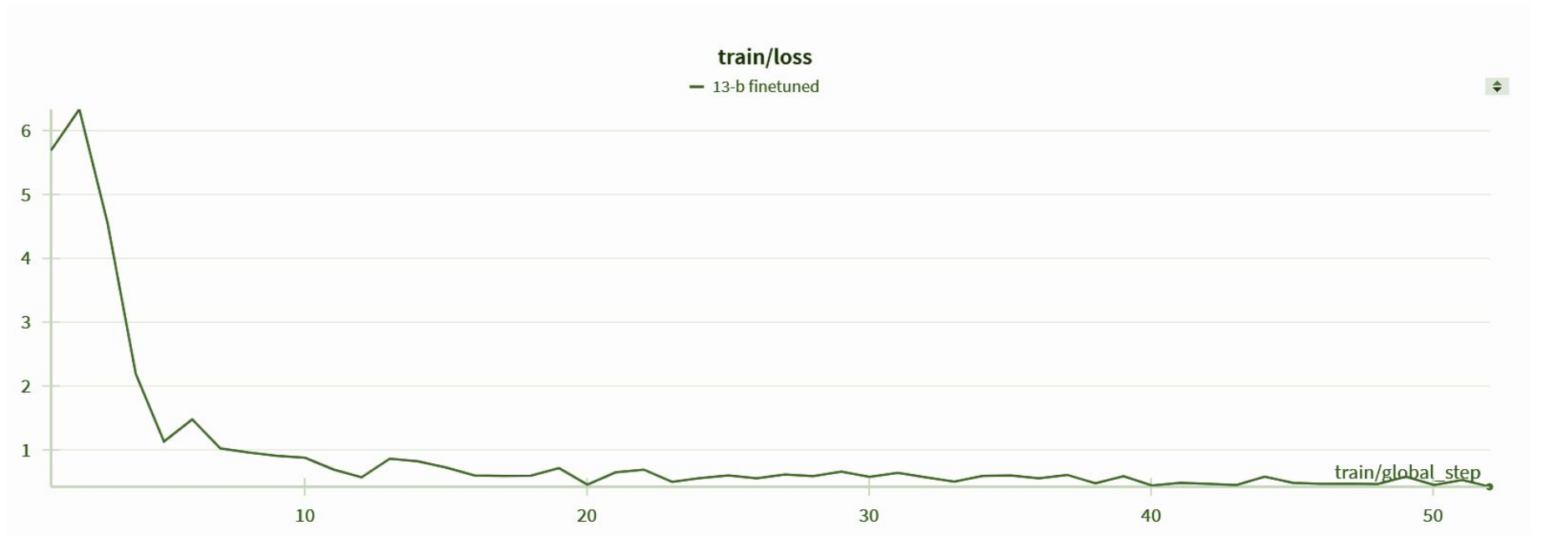
LLaVa architecture. Taken from the original paper.



LLaVA-Med was initialized with the general-domain LLaVA and then continuously trained in a curriculum learning fashion (first biomedical concept alignment then full-blown instructiontuning). We evaluated LLaVA-Med on standard visual conversation and question answering tasks. Taken from the original paper.

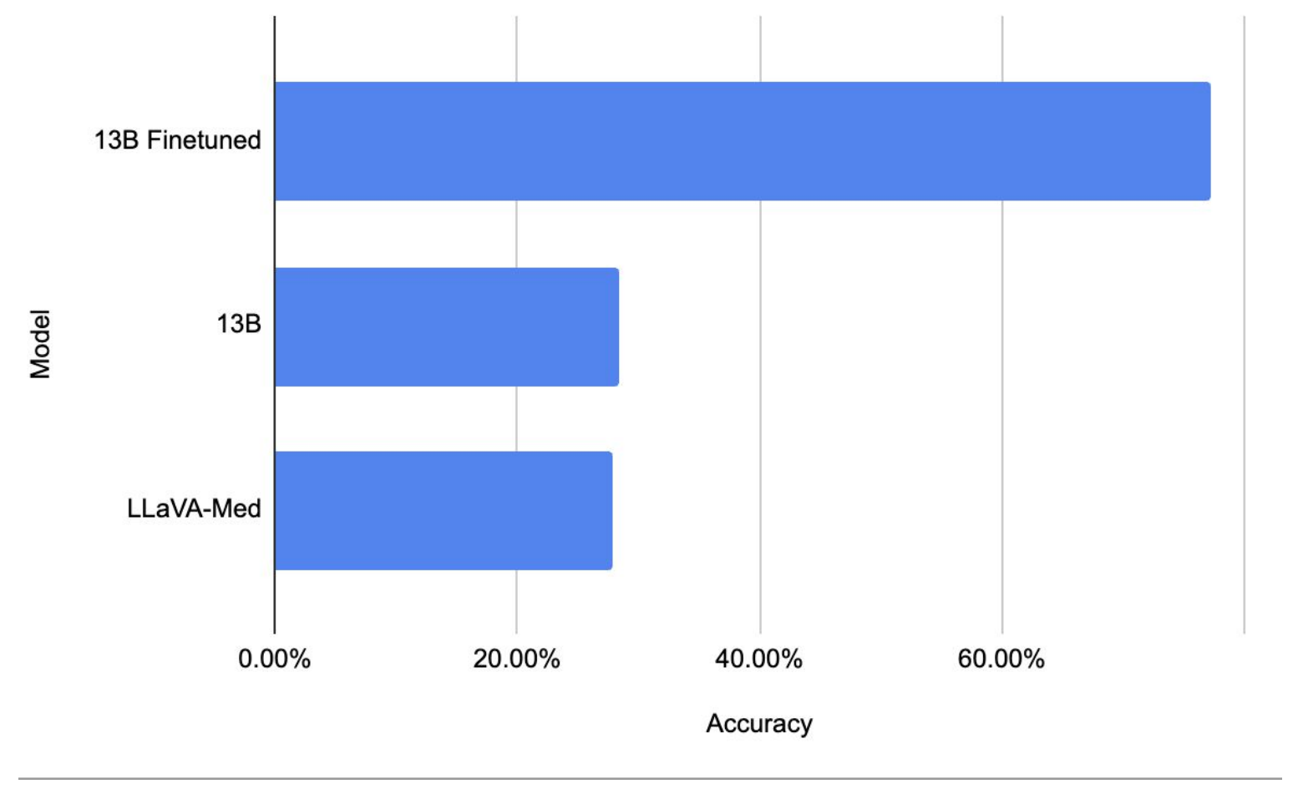
Observations

Training loss for LLaVA-13b-EHRXQA finetuning



Ground truth answers (Validation data)
Evaluation
Model-generated answers

Accuracy (%) = Number of correct answers / Total number of answers x 100



- Higher accuracy ≠ Better answers
- There seems to be a dataset bias, leading to the fine tuned model overfitting on the dataset quirks.
- Qualitative review crucial for VQA assessment

Future Work

- Curate high-quality VQA dataset with reliable/detailed ground truths, ensuring diverse and representative samples
- Expand the modality to include images, language and tabular data.

References

Liu, H., Li, C., Wu, Q. and Lee, Y.J., 2024. Visual instruction tuning. Advances in neural information processing systems, 36.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H. and Gao, J., 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems, 36.

Bae, S., Kyung, D., Ryu, J., Cho, E., Lee, G., Kweon, S., Oh, J., Ji, L., Chang, E., Kim, T. and Choi, E., 2024. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. Advances in Neural Information Processing Systems, 36.

Code: https://github.com/AmulyaMat/LLaVA_ChestXRay