Multi-modal LLMs for Healthcare

Generative AI workshop
September 1rst, 2024

Agenda

جامعة نيويورك ابوظبي
NYU ABU DHABI

1. Overview of the Capstone Project
2. Key Milestones and Development Process
3. Lessons Learned and Common Pitfalls
4. Codebase Walkthrough
5. Closure: Glimpse on the programming assignment

جامعة نيويورك ابوظبي
NYU ABU DHABI

Overview of the Capstone Project

Project Team

جامعة نيويورك ابوظبي
NYU ABU DHABI

Aya El Mir, Computer Engineering, NYUAD Class' 2024

Lukelo Luoga, Computer Engineering, NYUAD Class' 2024

Under the support and guidance of: Professor Muhammad Shafique, Postdoctoral researcher Muhammad Abdullah, and PhD students Boyuan Chen and Minghao Shao.

Background

جامعة نيويورك ابوظبي
NYU ABU DHABI

What is an MLLM?

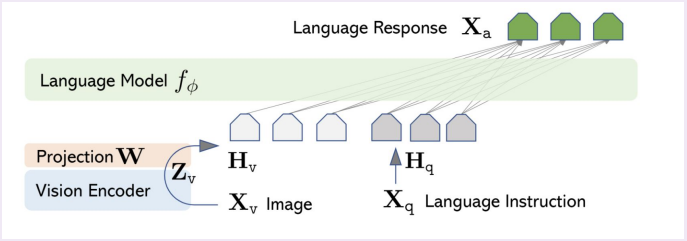
- **Multimodal Large Language Model (MLLM):** integrates multiple types of data, typically combining text and images, to perform complex tasks like medical image analysis, diagnosis, and medical question answering.
- Unlike standard Large Language Models (LLMs) that primarily process text, MLLMs are designed to understand and generate insights from multiple data modalities (e.g., text, images).

E.G:

- LLaVA: Large Language and Vision Assistant
- CLIP (Contrastive Language-Image Pre-Training)
- OpenFlamingo

How MLLMs Work?

- **Embeddings:** MLLMs create embeddings, which are dense vector representations of both text and image data. These embeddings allow the model to understand and relate information from different modalities.
- **Integration:** The model aligns the embeddings from text and images, enabling it to process and generate accurate responses to complex medical queries.



Defining the Problem

جامعة نيويورك ابوظبي
NYU ABU DHABI

Challenge in Low-Resource Environments:

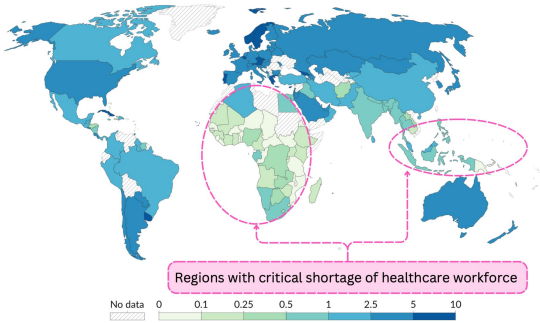
- Many low-resource countries, particularly in Africa, face a severe shortage of medical professionals, hindering effective healthcare delivery.
- AI technologies, such as MLLMs, have the potential to alleviate this burden by assisting in medical image analysis and diagnosis.

High Computational Demands:

- State-of-the-art MLLMs are incredibly GPU memory intensive, requiring powerful, high-end GPUs like NVIDIA A100 and V100 to operate effectively.
- These GPUs are often unavailable in low-resource environments due to their high cost and the lack of necessary infrastructure.

Barrier to Adoption:

- The excessive memory and computational requirements of current MLLMs prevent their deployment on consumer-grade GPUs, which are more accessible in low-resource settings.
- As a result, the regions most in need of these advanced AI tools are unable to benefit from their capabilities.



Global distribution of medical doctors per 1,000 people in 2021 compiled by *the World Bank* and visualized by *Our World in Data* [1]. The map illustrates significant disparities in medical personnel availability, with particularly low ratios in many African countries.

Our Solution

جامعة نيويورك ابوظبي
NYU ABU DHABI

A Potential solution could be optimizing MLLMs for Resource-Constrained Environments.

- Need to optimize MLLMs so that they can run efficiently on more affordable, consumer-grade hardware without sacrificing performance.
- This optimization is crucial for making AI-driven healthcare solutions accessible and impactful in resource-constrained environments.

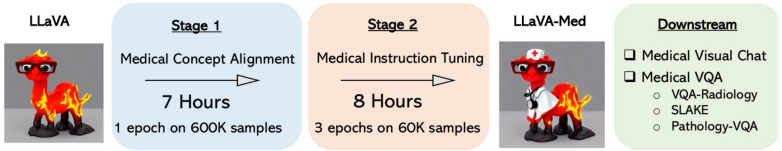
1. **Model Selection:** Chose **TinyLLaVA**, an open-source, compact model designed for efficient operation with fewer parameters.
2. **Dataset Utilized:** Leveraged the **PMC-15M dataset**, a large-scale biomedical dataset with 15 million image-text pairs, ensuring broad coverage of medical contexts.
3. **Fine-Tuning Process:** Conducted extensive fine-tuning of TinyLLaVA to adapt the model specifically to the biomedical domain, aligning it with medical tasks.
4. **Optimization Techniques:** Applied **post-training quantization** to further reduce the model's memory usage and computational requirements, making it suitable for deployment on consumer-grade GPUs.

Models and Dataset

جامعة نيويورك ابوظبي
NYU ABU DHABI

Inspiration Model: LLaVA-Med

LLaVA-Med (7B parameters) is a state-of-the-art Multimodal Large Language Model (MLLM) designed for medical applications, integrating image and text modalities for tasks like visual question answering (VQA) in healthcare.

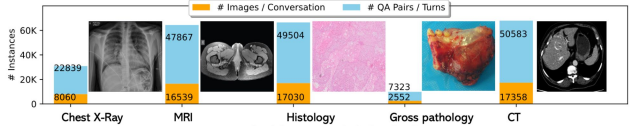


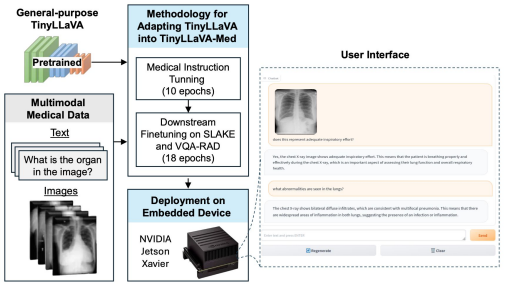
Base Model: TinyLLaVA

TinyLLaVA (1.5B parameters) was chosen as the base model due to its compact size and efficiency, making it a practical starting point for optimization.

Dataset: PMC-15M

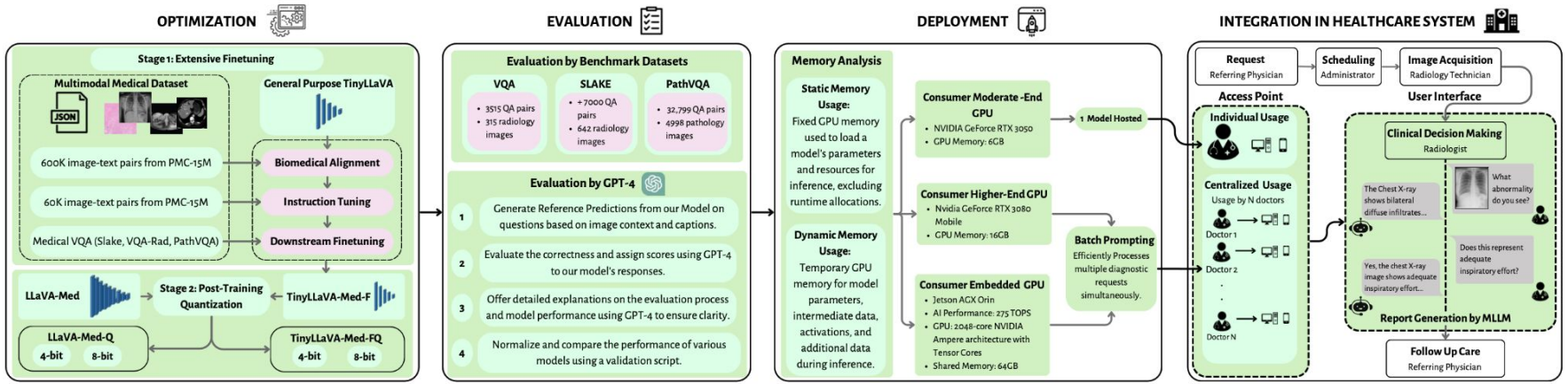
- Comprises 15 million high-quality biomedical image-text pairs extracted from PubMed Central (PMC) scientific publications.
- Chosen for its frequent use by state-of-the-art medical MLLMs due to its **open-source nature** and **high quality**.





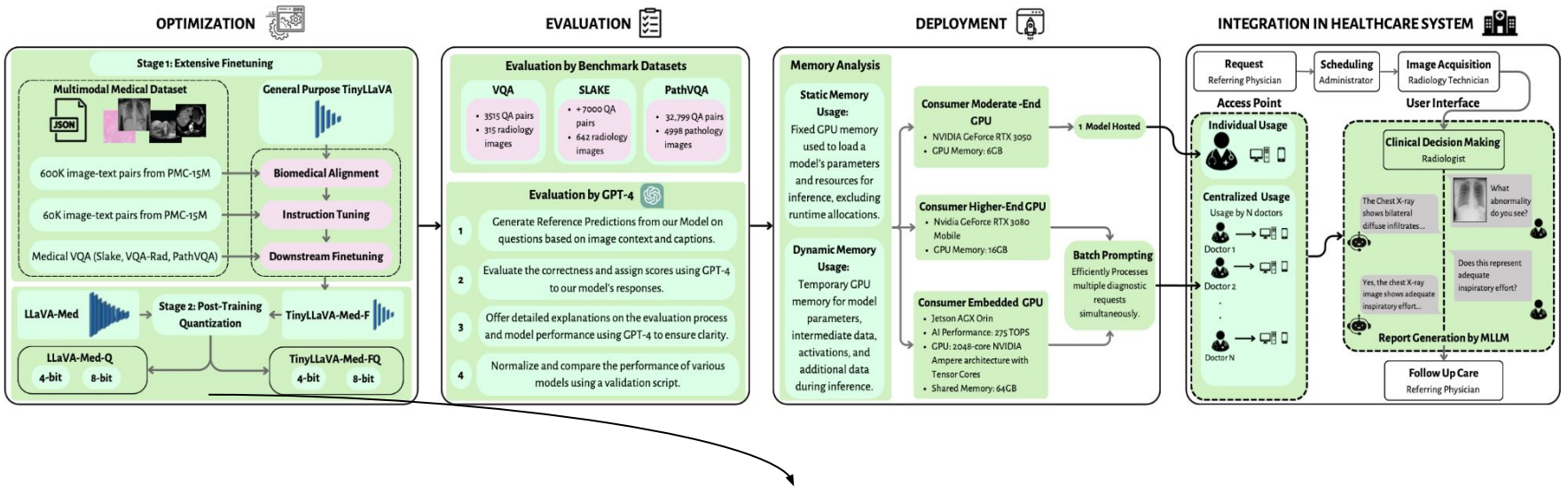
Framework

جامعة نيويورك ابوظبي
NYU ABU DHABI



Step 1: Optimization

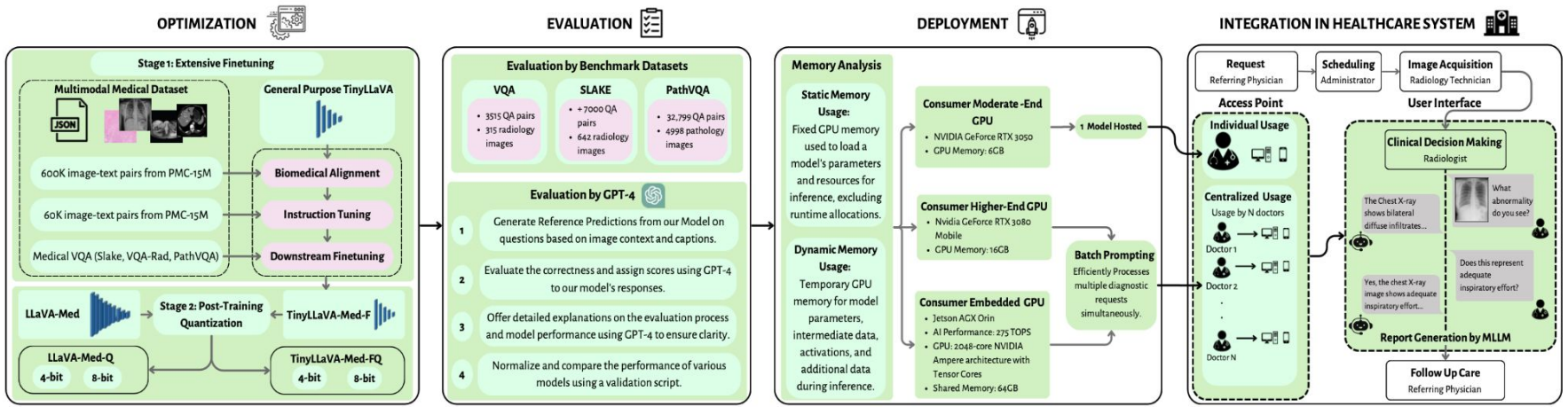
جامعة نيويورك ابوظبي
NYU ABU DHABI



- **Extensive Fine-Tuning:** Adapted the TinyLLaVA model to the biomedical domain using the PMC-15M dataset, focusing on aligning image and text embeddings for medical applications.
- **Post-Training Quantization:** Applied 4-bit and 8-bit quantization to reduce the model's memory footprint, making it suitable for deployment on consumer-grade GPUs.

Step 2: Evaluation

جامعة نيويورك ابوظبي
NYU ABU DHABI



- **Benchmark Dataset Testing:** Evaluated the optimized models on multiple medical VQA datasets (VQA-RAD, SLAKE, PathVQA) to assess their accuracy in handling both open and closed-ended questions.
- **GPT-4 Comparative Analysis:** Conducted a comparative evaluation using GPT-4, assessing the models' performance in medical conversations and their ability to handle domain-specific tasks.

Results

جامعة نيويورك ابوظبي
NYU ABU DHABI

TABLE I

COMPARATIVE PERFORMANCE ANALYSIS OF VARIOUS MODELS ON MEDICAL VISUAL QUESTION ANSWERING DATASETS. THIS TABLE DISPLAYS THE ACCURACY PERCENTAGES FOR BOTH OPEN AND CLOSED QUESTION TYPES ACROSS THREE DATASETS: VQA-RAD, SLAKE, AND PATHVQA. IT INCLUDES RESULTS FROM SUPERVISED FINETUNING EXPERIMENTS ALONGSIDE ZERO-SHOT EVALUATIONS, HIGHLIGHTING THE EFFECTIVENESS OF EACH MODEL UNDER DIFFERENT TRAINING CONDITIONS.

Method	VQA-RAD		SLAKE		PathVQA	
	Open	Closed	Open	Closed	Open	Closed
Supervised finetuning results with our own experiment runs(MLLM Based Methods)						
LLAVA	50.00	65.07	78.18	63.22	7.74	63.2
LLAVA-Med (LLama7B)	61.52	84.19	**85.34**	85.34	37.95	91.21
LLAVA-Med (Vicuna7B)	**64.39**	81.98	84.71	83.17	38.87	91.65
Med-Moe (Phi2:3.6B)	58.55	**82.72**	85.06	**85.58**	34.74	**91.98**
Med-Moe (StableLM:2.0B)	50.08	80.07	83.16	83.41	33.79	91.30
TinyLLaVA-Med-F (TinyLLaVA-1.5B)	50.6	81.25	**85.34**	85.43	**39.25**	90.56
Zero-shot results						
LLAVA-Med (LLama7B)	36.23	60.16	41.72	47.6	**10.86**	59.75
LLAVA-MED (Mistral7B)	**36.79**	65.44	42.83	60.82	10.04	69.04
LLAVA-MED-Q8 (Mistral7B)	32.98	**68.01**	43.92	**64.18**	10.11	**69.45**
LLAVA-MED-Q4 (Mistral7B)	29.82	62.87	**43.98**	62.50	9.85	69.15
Med-Moe (Phi2:3.6B)	36.73	61.75	43.93	56.97	6.94	66.46
Med-Moe (StableLM:2.0B)	28.02	66.91	40.63	52.64	9.40	69.09
TinyLLaVA-Med-F (1.5B)	29.89	**68.01**	36.43	58.46	10.53	53.52
TinyLLaVA-Med-FQ8 (TinyLLaVA-1.5B)	31.17	65.07	36.16	57.45	10.33	53.44
TinyLLaVA-Med-FQ4 (TinyLLaVA-1.5B)	34.58	63.24	34.66	62.26	10.06	54.11
Representative and SoTA methods with numbers reported in the literature (Non-MLLM Based Methods)						
VL Encoder-Decoder [27]	71.49	82.47			71.49	85.61
Q2ATransformer [28]	79.19	81.20			54.85	88.85
Prefix T. Medical LM [29]			84.30	82.01	40.00	87.00
PubMedCLIP [11]	60.10	80.00	78.40	82.50		
BiomedCLIP [12]	67.60	79.80	82.05	89.70		
M2I2 [30]	66.50	83.50	74.70	91.1	36.30	88.00

TABLE II

GPT-4 EVALUATION OF MODELS ON BIOMEDICAL MULTIMODAL CONVERSATION. THIS TABLE DISPLAYS THE PERFORMANCE ACROSS CONVERSATION AND DESCRIPTION QUESTION TYPES, SHOWING THE MODELS' PROFICIENCY IN HANDLING SPECIFIC MEDICAL DOMAINS. THE OVERALL SCORE REFLECTS THE AVERAGE CAPABILITY ACROSS ALL TESTED SCENARIOS.

Model	Conversation	Description	Chest X-Ray	MRI	Histology	Gross	CT Scan	Overall
LLAVA-MED (Mistral7b)	59.57	52.59	64.04	48.82	63.68	54.31	56.89	57.77
LLAVA-MED-Q8 (Mistral7b)	60.03	50.23	61.71	48.52	63.21	58.20	55.22	57.49
LLAVA-Med-Q4 (Mistral7b)	58.65	48.94	61.00	47.96	63.31	53.36	53.88	56.14
Med-Moe (Phi2:3.6B)	55.49	43.79	60.37	46.68	55.91	47.11	51.40	52.46
Med-Moe (StableLM:2.0B)	52.99	40.81	56.44	44.29	54.03	50.37	43.91	49.83
TinyLLaVA-Med-F (TinyLLaVA-1.5B)	52.92	41.04	63.85	40.70	51.43	52.02	41.97	49.84
TinyLLaVA-Med-FQ8 (TinyLLaVA-1.5B)	53.80	39.89	63.13	42.09	54.96	46.55	43.80	50.20
TinyLLaVA-Med-FQ4 (TinyLLaVA-1.5B)	51.60	38.07	59.42	41.94	49.43	49.93	40.42	48.09

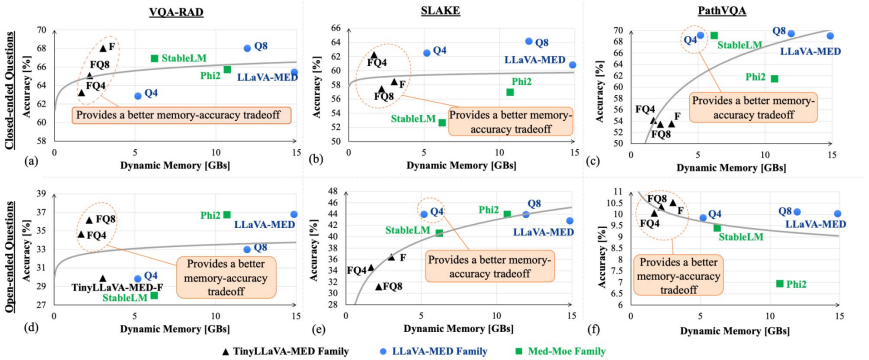


Fig. 4. Comparative analysis of memory-accuracy tradeoffs across three Visual Question Answering (VQA) datasets: VQA-RAD, SLAKE, and PathVOA. Notations are as follows: "FQ4" denotes a fine-tuned and quantized 4-bit version; "FQ8" refers to a fine-tuned and quantized 8-bit version; "Q4" signifies a quantized 4-bit version without fine-tuning; "Q8" indicates a quantized 8-bit version without fine-tuning. These plots demonstrate our optimized models maintain accuracy with minimal memory usage

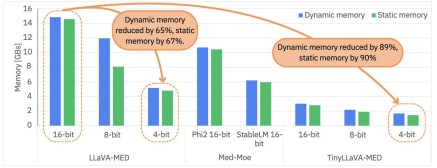
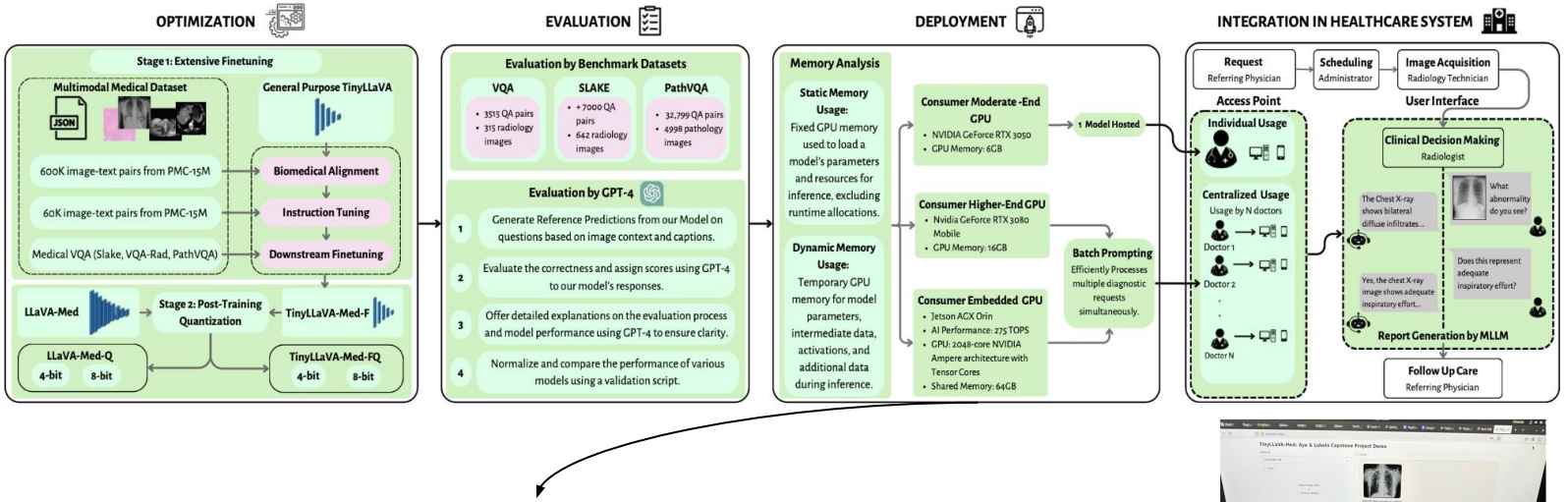


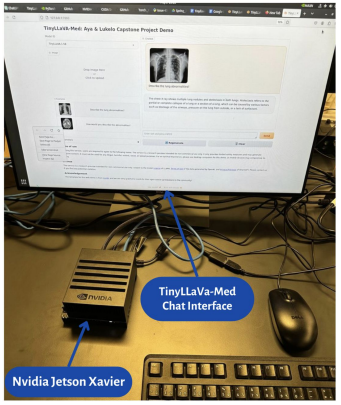
Fig. 3. Comparison of dynamic and static memory consumption across our models and other State-of-art MLLMs.

Step 3: Deployment

جامعة نيويورك ابوظبي
NYU ABU DHABI

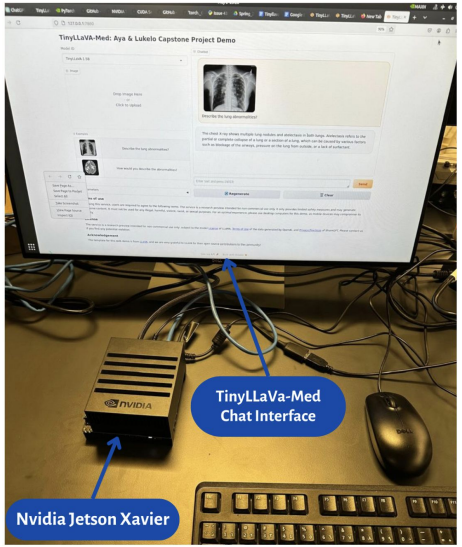


- **Consumer-Grade GPU Implementation:** Deployed the models across various consumer-grade GPUs, such as the NVIDIA GeForce RTX 3050, to ensure they function efficiently in resource-limited settings.
- **Memory Consumption Analysis:** Performed a detailed analysis of static and dynamic memory usage during model inference to validate their suitability for deployment on embedded devices.

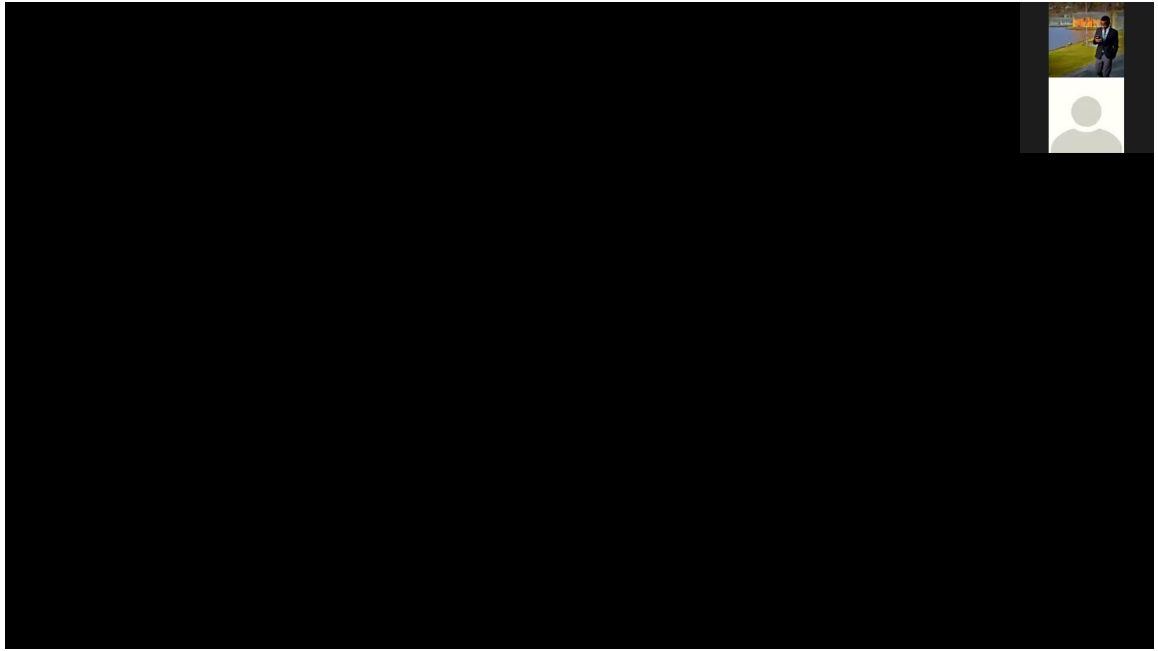


Step 3: Deployment

جامعة نيويورك ابوظبي
NYU ABU DHABI



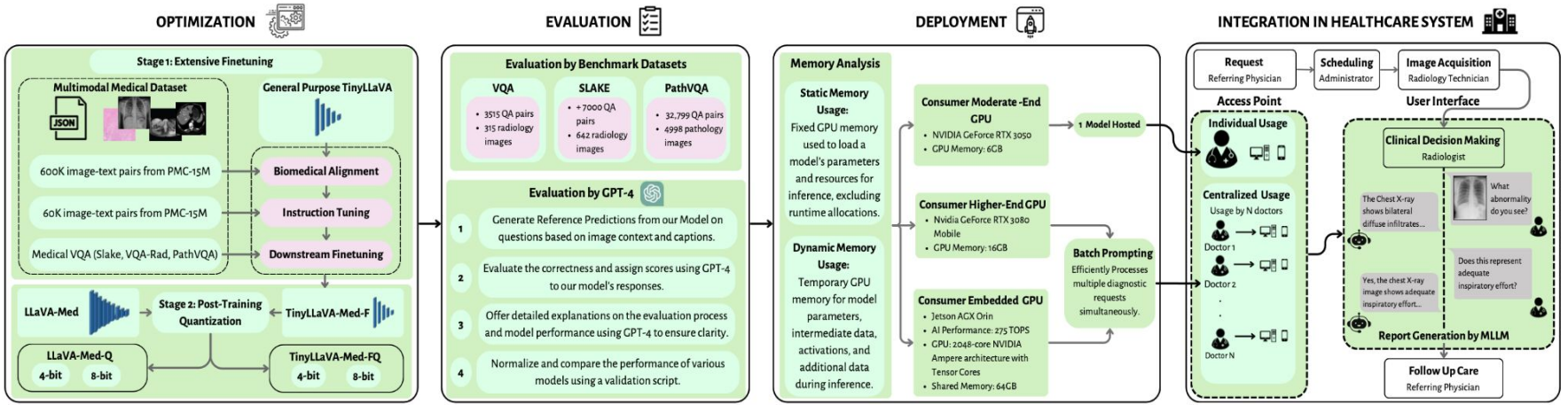
Hardware setup of the TinyLLaVA-Med model on NVIDIA Jetson Xavier, demonstrating the model's deployment and integration into a real-world medical environment.



Video Demo of the user interface for TinyLLaVA-Med model deployment

Step 4: Integration in Healthcare System

جامعة نيويورك ابوظبي
NYU ABU DHABI



- **Centralized Usage in Hospitals:** Proposed deploying the models in a centralized manner, allowing multiple doctors to access a single MLLM, optimizing throughput in high-demand environments.
- **User Interface Design:** Designed access points and interfaces for seamless interaction with the models, enabling healthcare professionals to easily submit diagnostic queries and receive insights.

جامعة نيويورك ابوظبي
NYU ABU DHABI

Lessons Learned and Common Pitfalls

The real timeline

جامعة نيويورك ابوظبي
NYU ABU DHABI

Sep 2023

Initial Project Design

Challenge with getting familiar with foundational MLLM concepts & development environment (Jubail).

Nov 2023

Model and Dataset Selection

Faced challenges in finding suitable open-source models and datasets, particularly due to the complexity of the medical domain.

Dec 2023

Reproducing LLaVA-Med results

The large scale of the PMC-15M dataset made data loading and preprocessing time-consuming.

Dedicated time to thoroughly understand the LLAVA-MED codebase to reproduce its results accurately.

Jan - Mars 2024

Knowledge Distillation of MLLMs

Encountered difficulties with multiple optimization methods due to the complexity of MLLMs.

Flexibility in approach is necessary especially in the case of a tight timeline.

Avril 2024

Switch to Fine-Tuning and Quantization

After challenges with knowledge distillation, decided to focus on fine-tuning and quantization instead.

Deployment in Nvidia Jetson for capstone showcase.

PPTP (July 2024)

Missed Conference Deadlines

Faced difficulties in submitting to conferences due to tight timelines, managing to meet some deadlines but missing others.

Early preparation are crucial for successful conference submissions, especially for year long capstone projects.

جامعة نيويورك ابوظبي
NYU ABU DHABI

Codebase Walkthrough (Project *Git*)

جامعة نيويورك ابوظبي
NYU ABU DHABI

Programming Assignment: Reproducing *TinyLLaVA*


جامعة نيويورك ابوظبي
NYU ABU DHABI
Thank you! Any questions?