جامعـة نيويورك ابـوظـبي

NYU | ABU DHABI

Advancing Healthcare in Low-Resource Environments Through an Optimization and Deployment Framework for Medical Multimodal LLMs

Aya El Mir, Lukelo Thadei Luoga, Boyuan Chen, Muhammad Abdullah Hanif, Muhammad Shafique

{ae2195, ltl2113, bc3194, mh6117, muhammad.shafique}@nyu.edu

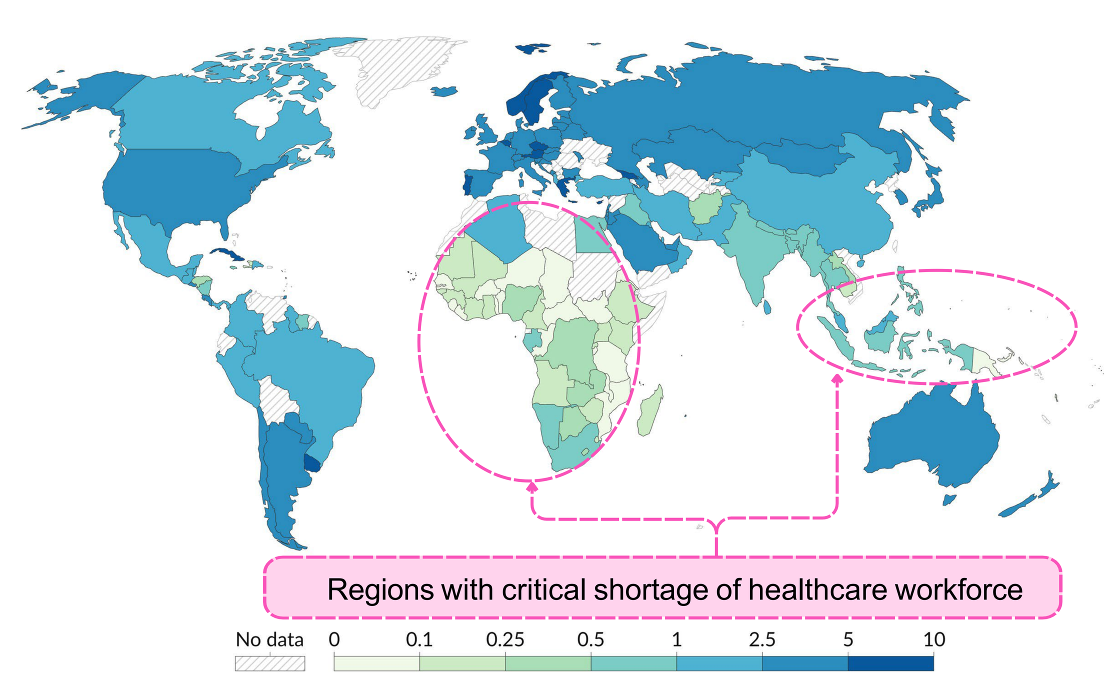
eBRAIN Lab, Division of Engineering, New York University Abu Dhabi (NYUAD), UAE

Introduction and Motivation

- **Severe shortage of healthcare professionals in low-resource countries:**
  - Example: Niger has only 0.03 doctors per 1,000 people, compared to 2.46 in Canada.
  - Growing patient demands far exceed the number of available healthcare providers.
- **Role of AI in addressing healthcare gaps:**
  - Enhances diagnostic accuracy and efficiency by reducing errors from fatigue, thus, supporting overburdened medical staff.
  - Multimodal Large Language Models (MLLMs), a subset of AI, can combine textual information with medical images to help doctors interpret the images more quickly and accurately in real time.

- **Motivation**
  - How can we enable the use of medical MLLMs in resource-constrained regions?

Global distribution of doctors per 1,000 people reveals significant shortages in many African countries



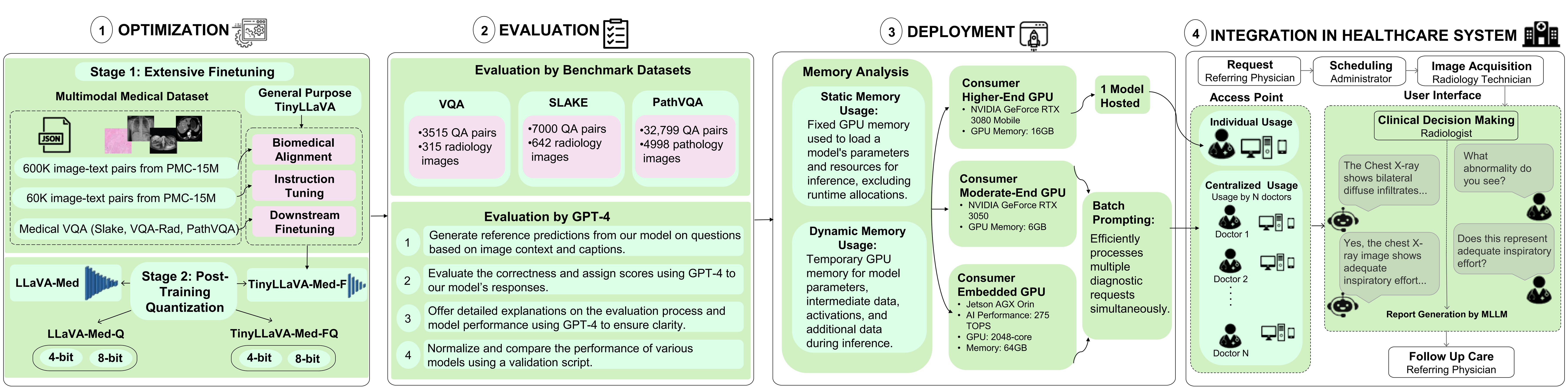
- **Challenges**
  - State-of-Art Medical MLLMs (e.g. LLaVA-Med) have high computational demands usually requiring HPC infrastructures.
  - Low resource areas with only access to consumer-grade GPUs, cannot benefit from MLLMs for the healthcare domain and beyond.

Novel Contributions:

- **Optimized Medical MLLM Framework:** TinyLLaVA-Med-F and Quantized models (4-bit, 8-bit), fine-tuned for efficient deployment on consumer-grade GPUs for the healthcare domain.
- **Performance-Memory Trade-off**: Models on the Pareto front, balancing accuracy and memory.
- **Foundation for Future Research**: Accessible MLLMs for healthcare on consumer GPUs.

Proposed Methodology

- **Optimization:** fine-tuning and quantization to create efficient MLLMs for consumer-grade GPUs.
- **Evaluation** by medical VQA datasets and GPT-4 alongside Memory usage analysis.
- Proposed **deployment** to integrate the MLLMs for medical decision-making support (e.g. radiology).



Experimental Results

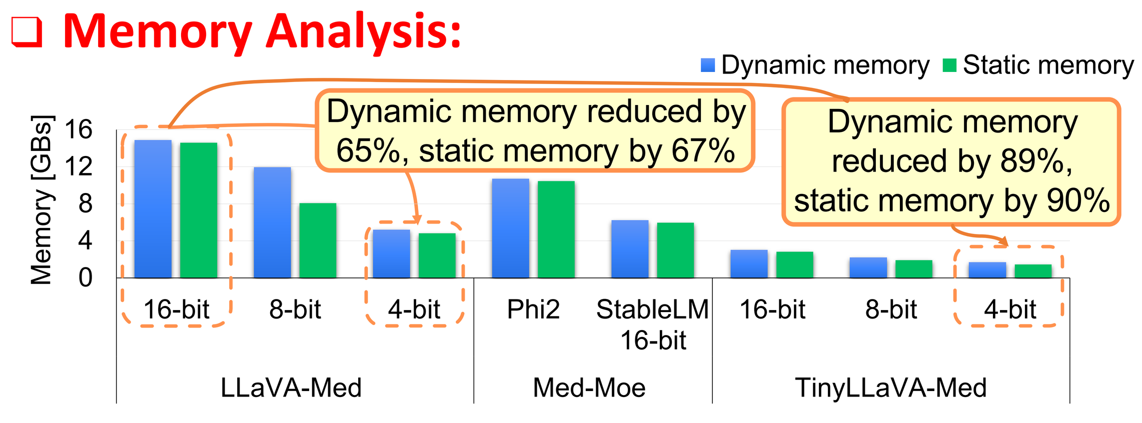
- **Medical VQA Performance:**
  - TinyLLaVA-Med-F and quantized variants (FQ4, FQ8) achieved competitive accuracy with minimal drop.

Model	VQA-RAD		SLAKE		PathVQA	
	Open	Closed	Open	Closed	Open	Closed
TinyLLaVA-1.5B (Baseline)	19.15	59.93	35.22	60.1	11.16	63.7
Our Supervised finetuning results (MLLM Based Methods)						
LLAVA	50	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	91.3
TinyLLaVA-Med-F (1.5B)	50.6	81.25	**85.34**	85.43	**39.25**	90.56

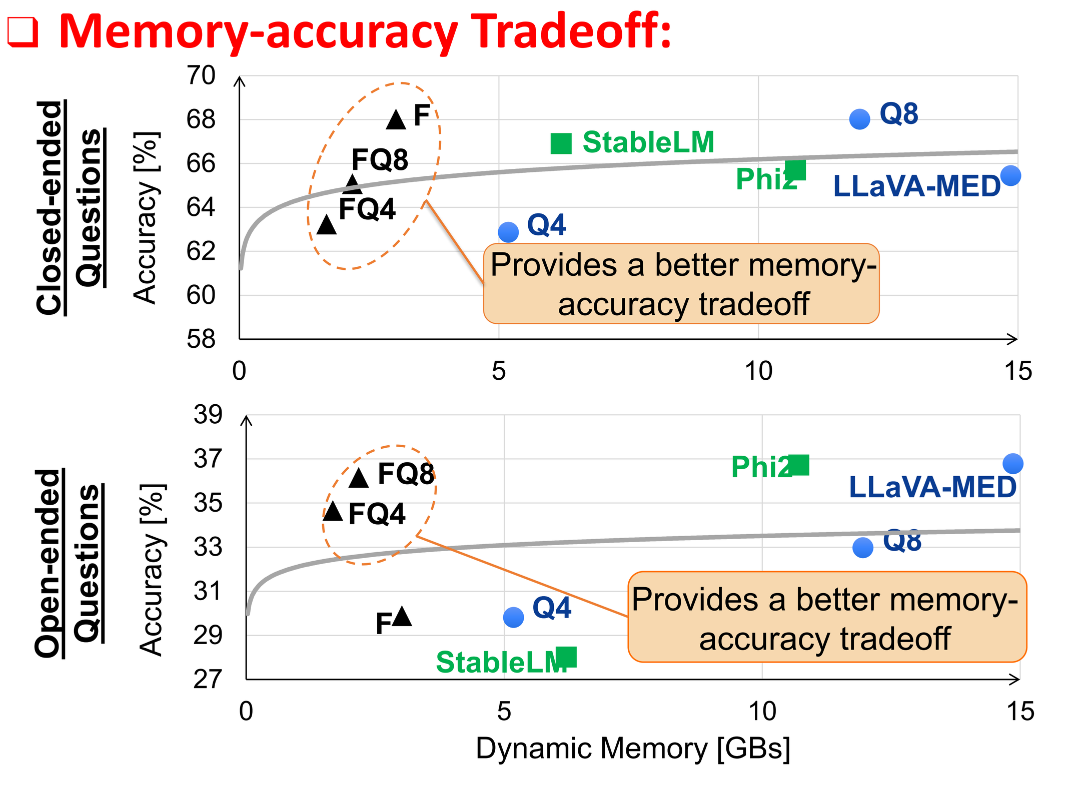
- **GPT-4 Evaluation:**
  - The TinyLLaVA-Med family of models demonstrates overall robust accuracy in medical conversations.

Model	Conv.	Desc.	X-Ray	MRI	Histology	Gross	CT Scan	Overall
TinyLLaVA (1.5B)-Baseline	40.87	35.11	45.08	39.65	39.86	35.03	37	39.38
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	**58.2**	55.22	57.49
LLaVA-Med-Q4 (Mistral7b)	58.65	48.94	61	47.96	53.33	53.33	53.88	56.14
Med-Moe (Phi2:3.6B)	55.49	43.79	60.37	46.68	55.91	47.11	51.4	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 (1.5B)	52.92	41.04	63.85	40.7	51.43	52.02	41.97	49.84
TinyLLaVA-Med-FQ8 (1.5B)	53.8	39.89	63.13	42.09	54.96	46.55	40.83	50.2
TinyLLaVA-Med-FQ4 (1.5B)	51.6	38.07	59.42	41.94	49.43	49.93	40.42	48.09

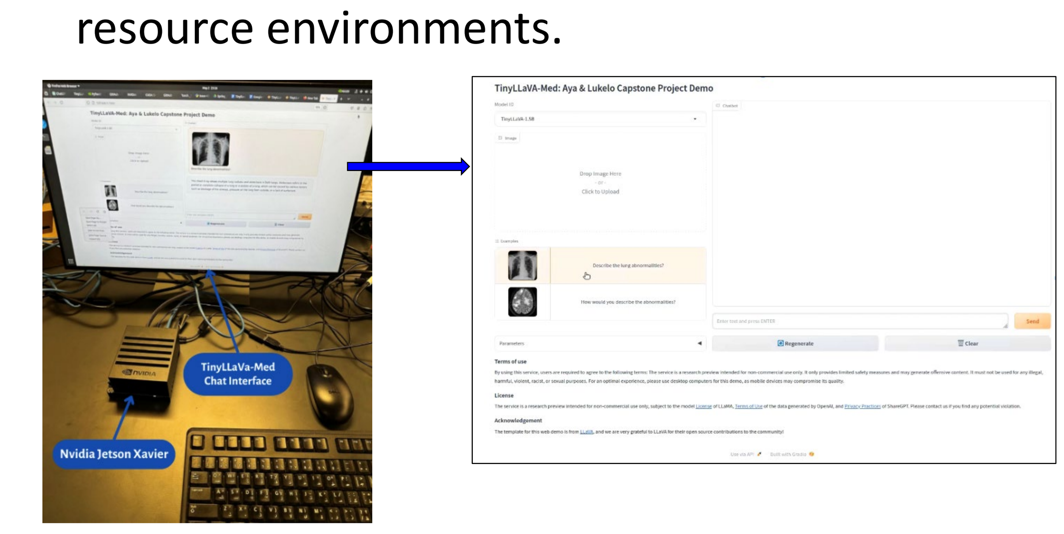
- **Memory Analysis:**



- **Memory-accuracy Tradeoff:**



- **EdgeAI Prototype:**
  - TinyLLaVA-Med deployed on a consumer-grade GPU, enhancing medical AI accessibility in low-resource environments.



Key References:

- A. El Mir and L. T. Luoga et al., "Advancing healthcare in low-resource environments through an optimization and deployment framework for medical multimodal large language models," in IEEE-EMBS BHI, 2024.
- B. Zhou et al., "Tinyllava: A framework of small-scale large multimodalmodels," arXiv preprint arXiv:2402.14289, 2024.
- C. Li et al., "Llava-med: Training a large language-and-vision assistantfor biomedicine in one day," Advances in Neural Information ProcessingSystems, vol. 36, 2024
- S. Jiang et al., "Moe-tinymed: Mixture of experts for tiny medical largevision-language models," arXiv preprint arXiv:2404.10237, 2024
- W. Bank, "Medical doctors per 1,000 people," Jun. 2024, multiplesources compiled by World Bank – processed by Our World in Data

Paper Link

eBRAIN Lab

AI Revolution in Healthcare