MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models

Peng Xia[1], Kangyu Zhu[2], Haoran Li[3], Tianze Wang[4], Weijia Shi[5], Sheng Wang[5], Linjun Zhang[4], James Zou[6], Huaxiu Yao[1]

[1]UNC-Chapel Hill, [2]Brown University, [3]Carnegie Mellon University, [4]Rutgers University, [5]University of Washington, [6]Stanford University

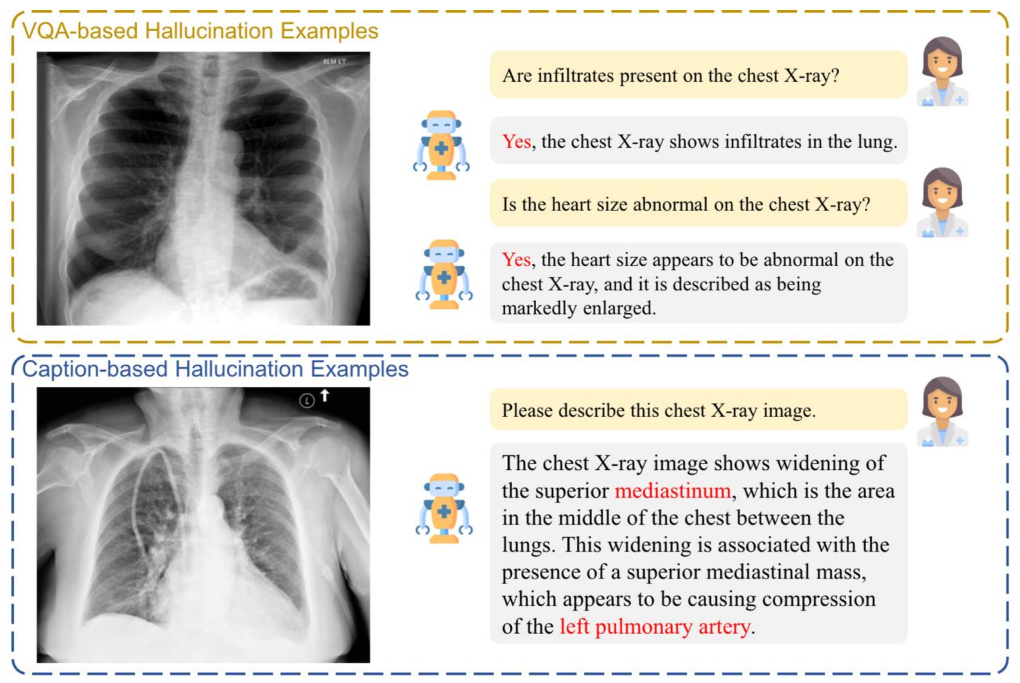
{pxia, huaxiu}@cs.unc.edu

ICLR International Conference On Learning Representations

tl;dr: a multimodal RAG system to improve the factuality for medical large vision language models (Med-LVLMs)

Background

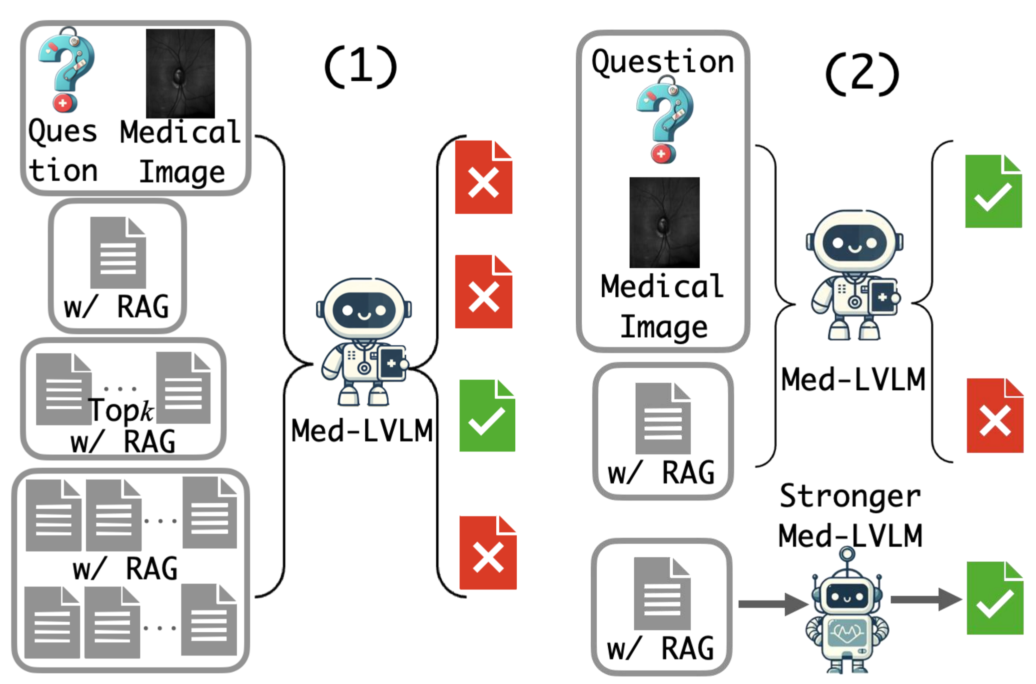
Hallucination in Med-LVLMs



Goal: Build a Reliable Med-LVLM to Generate Factual Responses

Motivation

Recent RAG-based Method



Challenges

1 lack of sufficient high-quality labeled data for fine-tuning ⇒ RAG

2 distribution gap exists between the training data and the real-world data

3 dataset-specific: reducing the generalizability ⇒ MMed-RAG

4 misalignment issues: cross-modality and overall alignment

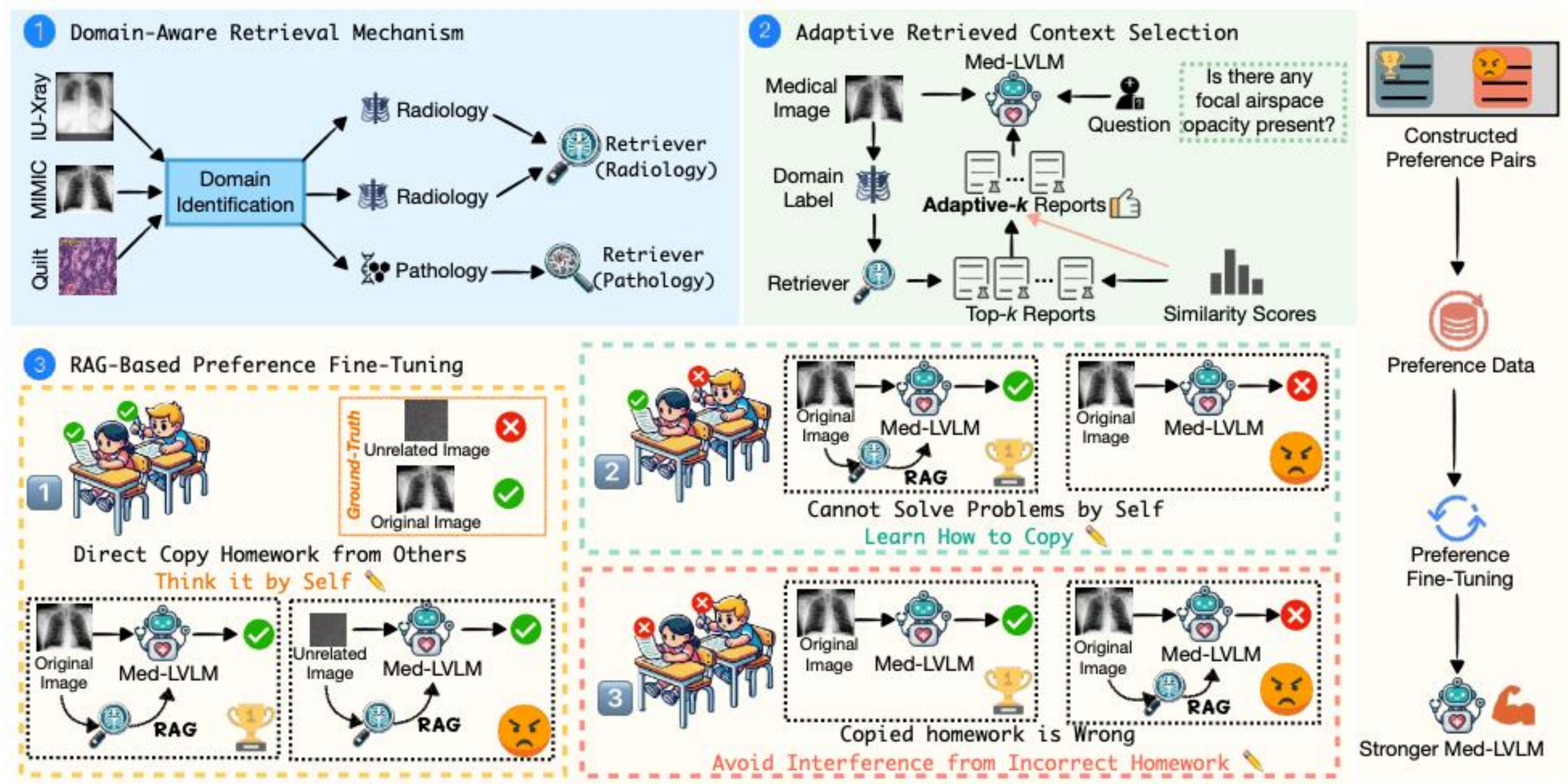
Methodology

(1) Domain Identification

Domain-aware retrieval mechanism: select the best retriever

(2) Adapted Retrieved Context Selection

Dynamically adjusts retrieved info based on similarity scores



(3) RAG-Based Preference Fine-Tuning

1 Direct Copy Homework from Others ❌ Think it by Self ✅

Avoid blindly copying external information by encouraging the model to rely on its own visual reasoning when solving complex problems

2 Cannot Solve Problems by Self ❌ Learn How to Copy ✅

When Med-LVLMs are unsure, MMed-RAG teaches the model to intelligently use retrieved knowledge, pulling in the right information at the right time

3 Copied Homework is Wrong ❌ Avoid Interference from Incorrect Homework ✅

Prevent models from being misled by incorrect retrievals, reducing the risk of generating inaccurate medical diagnoses

Experiments

Medical Visual Question-Answering (VQA)

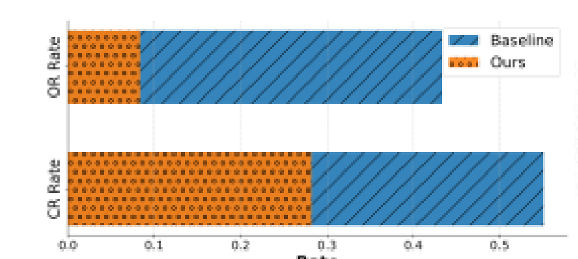
Models	Radiology						Ophthalmology			Pathology					
	IU-Xray			MIMIC-CXR			Harvard-FairVLMed			Quilt-1M			PMC-OA (Pathology)		
	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
LLaVA-Med-1.5	75.47	64.04	67.46	75.79	80.49	68.84	63.03	74.11	63.05	62.80	72.90	60.03	59.28	71.98	54.19
+ Greedy	76.88	65.59	68.74	78.32	86.75	71.13	82.54	85.98	70.09	64.72	70.12	58.75	58.61	70.42	53.10
+ Beam Search	76.91	66.06	68.77	81.56	86.36	73.79	80.93	88.08	68.94	63.52	69.33	57.65	56.29	69.84	52.89
+ DoLa	78.00	66.75	72.19	81.35	85.73	72.73	76.87	85.53	67.10	63.47	69.10	57.58	57.71	70.27	52.95
+ OPERA	70.59	61.54	63.22	69.34	76.66	62.46	71.41	81.37	65.59	60.51	66.32	54.79	55.32	68.30	51.86
+ VCD	68.99	54.35	61.08	70.89	75.57	64.61	65.88	77.20	64.16	61.43	67.39	55.72	55.10	67.94	51.62
+ MedDr	83.33	67.80	77.15	55.16	56.18	58.47	70.17	80.72	64.15	68.15	73.23	67.01	59.97	69.19	57.01
+ FactMM-RAG	84.51	68.51	77.07	77.58	81.86	70.09	83.67	87.21	72.20	69.25	73.62	68.15	60.49	69.38	57.31
+ RULE	87.84	78.00	85.78	83.92	87.49	83.44	87.12	92.89	77.08	68.97	73.80	68.13	61.41	70.36	58.91
MMed-RAG	89.54	80.72	87.13	83.57	88.49	85.08	87.94	92.78	80.81	72.95	76.35	72.25	64.54	73.09	61.42

Report Generation

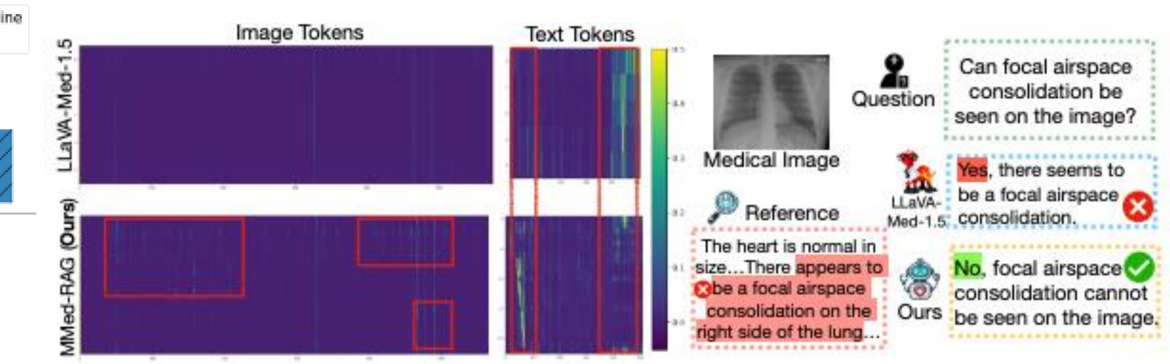
Models	Radiology						Ophthalmology		
	IU-Xray			MIMIC-CXR			Harvard-FairVLMed		
	BLEU	ROUGE-L	METEOR	BLEU	ROUGE-L	METEOR	BLEU	ROUGE-L	METEOR
LLaVA-Med-1.5	9.64	12.26	8.21	12.11	13.05	11.16	18.11	11.36	10.75
+ Greedy	11.47	15.38	12.69	16.63	14.26	14.19	17.98	11.49	13.77
+ Beam Search	12.10	16.21	13.17	16.97	14.74	14.43	18.37	12.62	14.50
+ DoLa	11.79	15.82	12.72	17.11	14.89	14.81	18.26	12.51	14.51
+ OPERA	10.66	14.70	12.01	15.40	12.52	13.72	16.59	11.47	13.63
+ VCD	10.42	14.14	11.59	15.18	12.30	13.38	16.73	11.38	13.89
+ MedDr	12.37	16.45	13.50	18.59	15.72	16.77	19.82	13.72	15.40
+ FactMM-RAG	14.70	18.05	15.92	18.71	15.84	16.82	20.82	14.17	15.31
+ RULE	27.53	23.16	27.99	18.61	15.96	17.42	22.35	14.93	17.74
MMed-RAG	31.38	25.59	32.43	23.25	12.34	20.47	24.82	16.59	19.85

Analysis

Effectiveness in Mitigating Misalignment Issues



OR: Over-Reliance;
CR: Copy-Reference



Ablation Study

Model	IU-Xray		FairVLMed	
	VQA	RG	VQA	RG
LLaVA-Med-1.5	68.99	10.04	66.63	13.41
+DR	77.12	13.23	72.69	15.89
+RCS	79.56	17.92	75.74	17.22
+RAG-PT (Ours)	**85.80**	**29.80**	**87.18**	**20.42**

DR: domain-aware retrieval mechanism;
RCS: adaptive retrieval context selection;
RAG-PT: RAG-based preference fine-tuning

Impact of Preference Data

1: Direct Copy Homework from Others;
2: Cannot Solve Problems by Self;
3: Copied Homework is Wrong

Model	IU-Xray		FairVLMed	
	VQA	RG	VQA	RG
LLaVA-Med-1.5	68.99	10.04	66.63	13.41
+RAG-PT 1	80.19	19.38	79.42	18.37
+RAG-PT 2	80.27	**20.16**	79.35	18.66
+RAG-PT 3	**81.30**	19.43	**80.07**	**18.92**

References

Li C, Wong C, Zhang S, et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. NeurIPS 2023.
Xia P, Chen Z, Tian J, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models.. NeurIPS 2024.
Xia P, Zhu K, Li H, et al. Rule: Reliable multimodal rag for factuality in medical vision language models. EMNLP 2024.
Hu Y, Li T, Lu Q, et al. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. CVPR 2024.