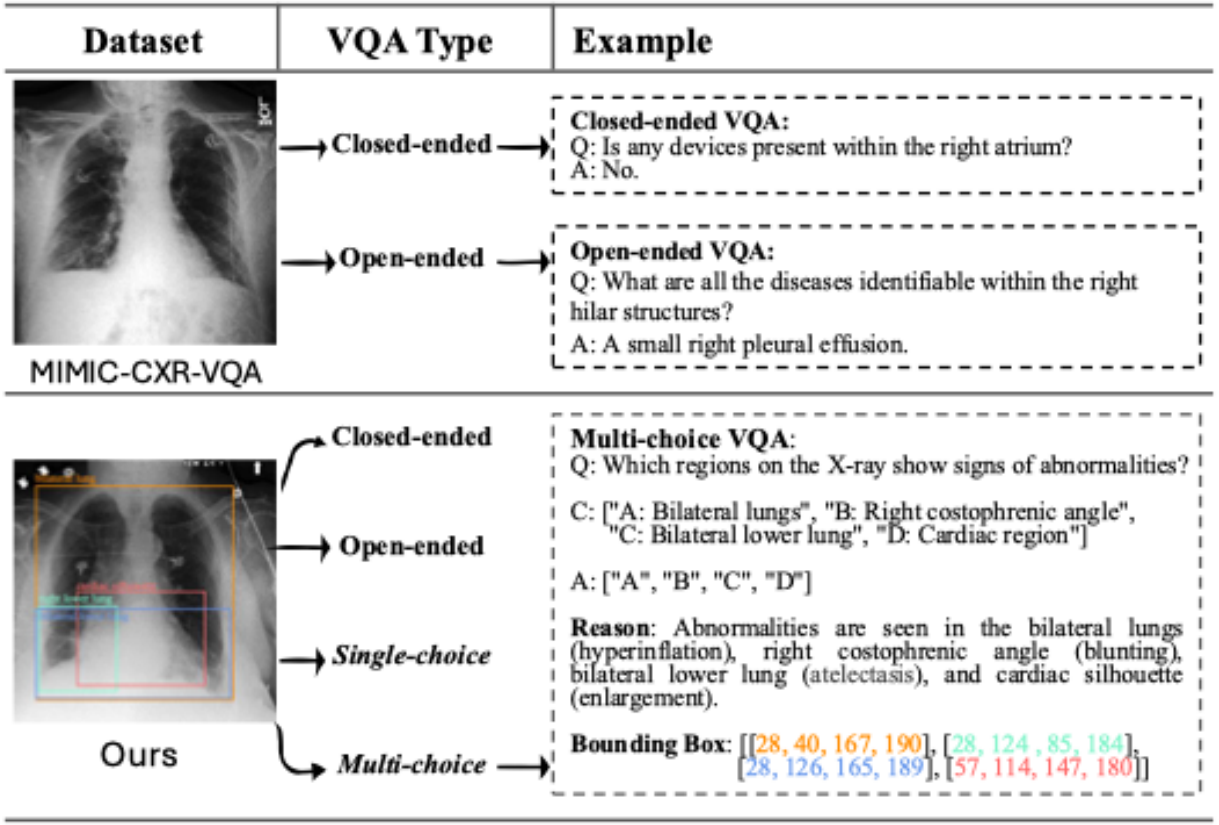
GEMeX: A Large-Scale, Groundable, and Explainable Medical VQA Benchmark for Chest X-ray Diagnosis

Bo Liu[1], Ke Zou[2], Li-Ming Zhan[1], Zexin Lu[1], Xiaoyu Dong[1], Yidi Chen[3], Chengqiang Xie[1], Jiannong Cao[1], Xiao-Ming Wu[1], Huazhu Fu[4]**

[1]The Hong Kong Polytechnic University, Hong Kong S.A.R., China
[2] National University of Singapore, Singapore
[3] West China Hospital of Sichuan University, China
[4] IHPC, Agency for Science, Technology and Research, Singapore

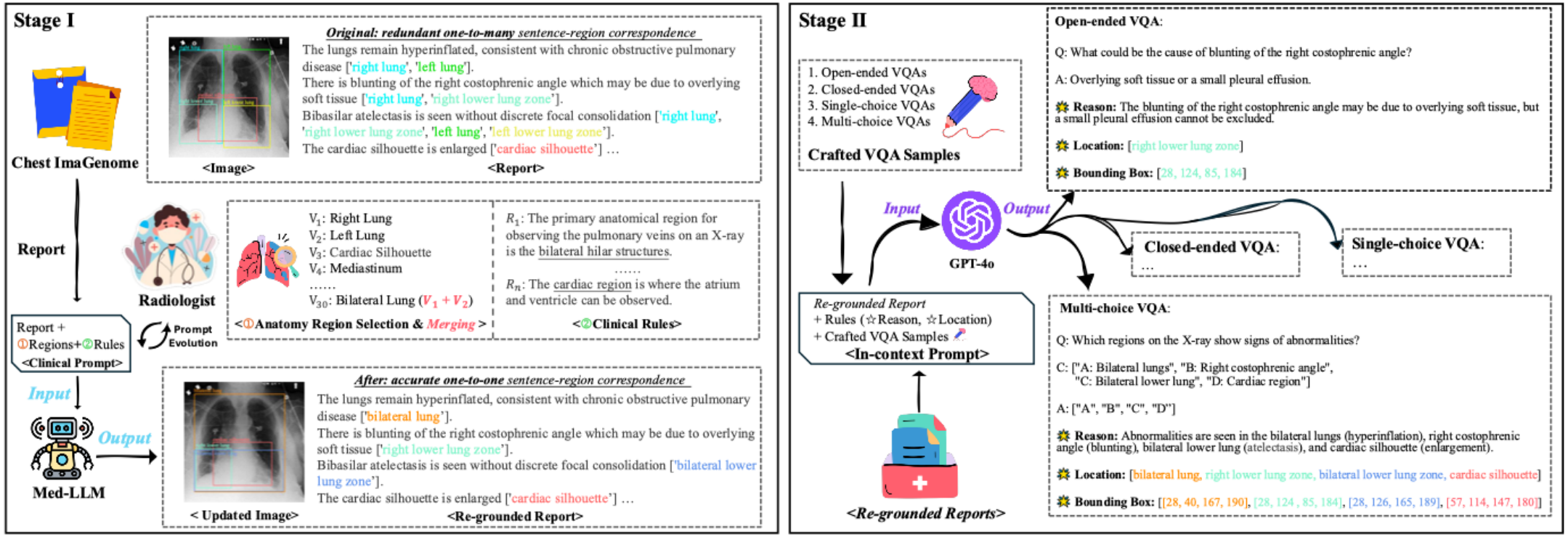
Motivation



a) Popular datasets often lack visual and textual explanations for answers, hindering comprehension for patients and junior doctors

b) They typically offer a narrow range of question formats, like open-ended and closed-ended ones, inadequately reflecting the diverse requirements in practical scenarios

Dataset Construction



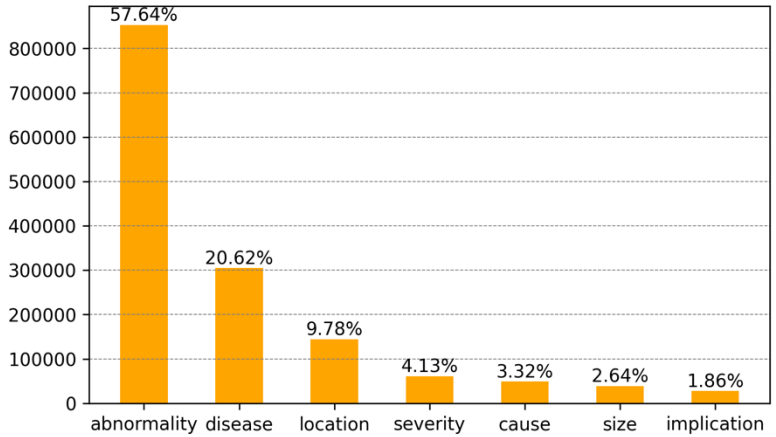
I. In Stage I (left), medical LLM performs re-grounding on the original reports based on the pathological regions and clinical guidance specified by the radiologists, generating more precise sentence-region correspondence.

II. In Stage II (right), the well-crafted prompt enables GPT-4o to generate a high-quality, large-scale Med-VQA dataset with both textual and visual explanations, leveraging the re-grounded reports from Stage I.

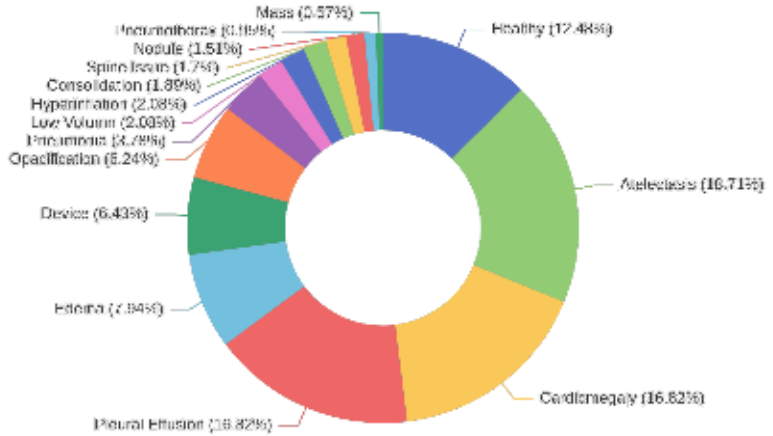
Dataset Statistics

Dataset	# Images	# QA Pairs	# Modalities	# Question Types‡	# Groundable	# Explainable
VQA-RAD [24]	0.315K	3.5K	Diverse†	O. & C.	✗	✗
SLAKE [30]	0.642K	14K	Diverse†	O. & C.	✗	✗
OmniMedVQA [18]	118.010K	128K	Diverse†	O. & C. & S.	✗	✗
PMC-VQA [51]	149.075K	227K	Diverse†	O. & C. & S.	✗	✗
VQA-Med [7]	4.5K	4.5K	Diverse†	O. & C.	✗	✗
PathVQA [16]	149K	33K	Pathology	O. & C.	✗	✗
RadGenome-Chest CT [52]	50.188K	1.3M	Chest CT	O. & C.	✓	✗
MIMIC-Diff-VQA [17]	164.324K	700K	Chest X-ray	O. & C.	✗	✗
MIMIC-CXR-VQA [4]	142.797K	377K	Chest X-ray	O. & C.	✗	✗
GEMeX (Ours)	151.025K	**1.6M**	Chest X-ray	O. & C. & S. & **M.**	✓	✓(Vision & Language)

Comparison of medical VQA datasets



Distribution of question content



Distribution of normality and common abnormality

Evaluation

Models	Open-ended		Closed-ended			Single-choice			Multi-choice			
	AR-score†	V-score	A-score	AR-score†	V-score	A-score	AR-score†	V-score	A-score	AR-score†	V-score	Avg.†
Random	-	-	48.80	-	-	25.85	-	-	7.50	-	-	-
GPT-4o-mini [1]	**97.68**	18.05	59.30	71.14	28.64	59.00	77.47	23.62	49.13	82.91	19.19	82.30
LLaVA-v1 [34]	76.14	-	30.76	38.02	-	-	50.47	-	-	66.52	-	57.79
LLaVA-v1.5 [33]	77.62	-	58.93	57.00	-	47.00	57.05	-	-	65.17	-	64.21
Mini-GPT4-v1 [56]	55.32	-	26.33	31.09	-	-	37.63	-	-	46.65	-	42.67
mPLUG-Owl [48]	76.73	-	27.26	36.70	-	32.00	46.89	-	-	67.92	-	57.06
DeepSeek-VL [35]	79.30	11.00	57.10	59.86	8.28	51.69	62.03	8.57	17.99	70.35	12.98	67.89
Qwen-VL-Chat [6]	78.36	3.17	23.02	45.79	12.25	44.69	59.15	16.69	7.30	67.21	2.26	62.63
LLaVA-Med-v1 [26]	90.34	-	62.62	69.91	-	-	61.74	-	-	68.14	-	72.53
LLaVA-Med-v1.5 [26]	94.43	-	71.82	76.54	-	-	66.04	-	-	67.28	-	76.07
MiniGPT-Med [3]	86.12	-	55.24	65.25	-	-	55.61	-	-	64.33	-	67.83
XrayGPT [40]	81.17	-	-	68.17	-	-	48.33	-	-	55.10	-	63.19
RadFM [43]	88.57	-	58.01	67.91	-	-	57.82	-	-	62.41	-	69.18
LLaVA-Med-GEMeX	97.05	**51.47**	**77.35**	**80.72**	**53.20**	**73.08**	**81.42**	**54.57**	**67.42**	**84.98**	**47.99**	**86.04**

- We evaluate 12 representative large vision language models (LVLMs) on this dataset. The results show suboptimal performance, underscoring the dataset's complexity.
- Most existing LVLMs exhibit weak performance in providing answer explainability, particularly for the visual component. Moreover, when faced with choice-based questions, many models — especially in the medical domain — struggle to deliver definitive answers.
- We propose a strong model by fine-tuning LLaVA-Med-7B on the training set. The substantial performance improvement showcases the dataset's effectiveness.

Thank you!