Generalist or Specialist? A Comparative Study of Gemini and Medical VLMs in Medical Image Understanding

Jingyi He, Bivek Panthi

Motivation

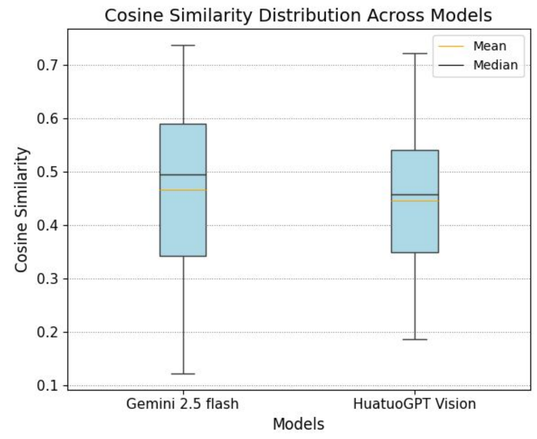
- Evaluate the performance of a general purpose LLM, such as **Google Gemini 2.5 Flash**, on medical VQA tasks, and compare it with domain-specific VLMs: **Maira-2** (Chest X-rays) and **HuatuoGPT-Vision** (Brain MRIs).
- Use **VLM-Seminar25-Dataset** as a benchmark across modalities.
- Explore whether medical specialization leads to **better diagnostic performance**.

Brain MRI Description Analysis

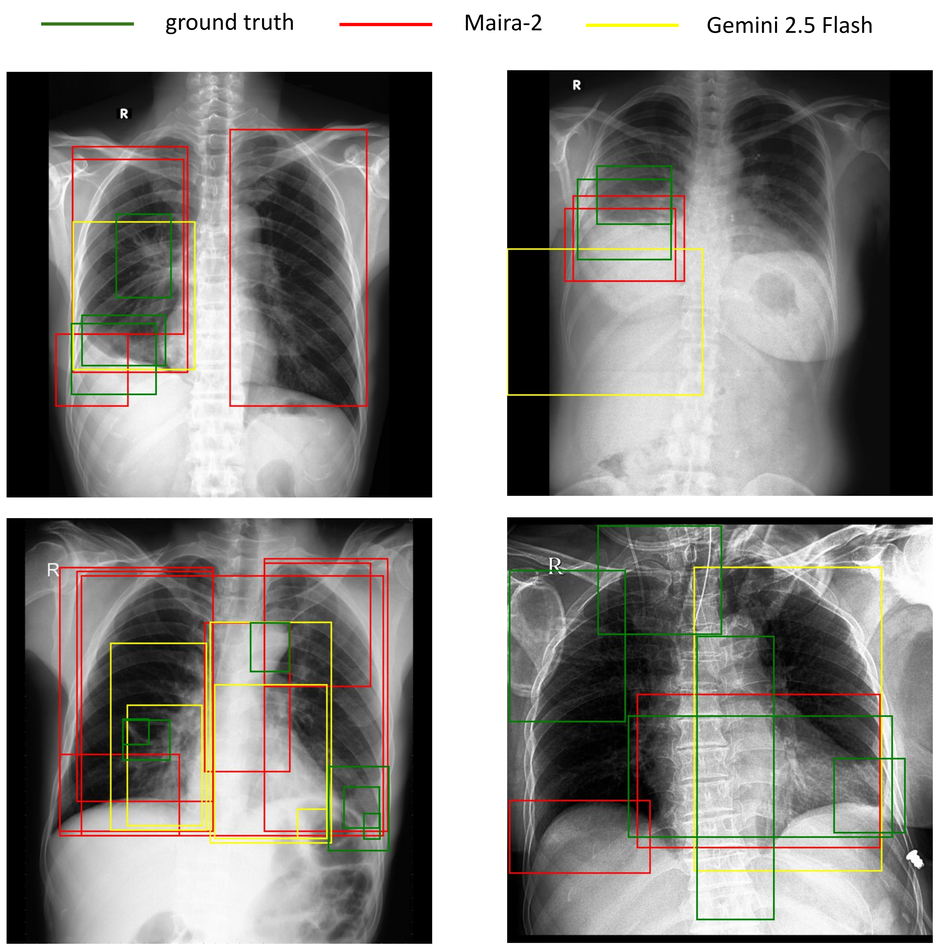
Llama-3.1-8B

all-MiniLM-L6-v2

	Gemini	HuatuoGPT-Vision
Accuracy	59.6%	58.7%

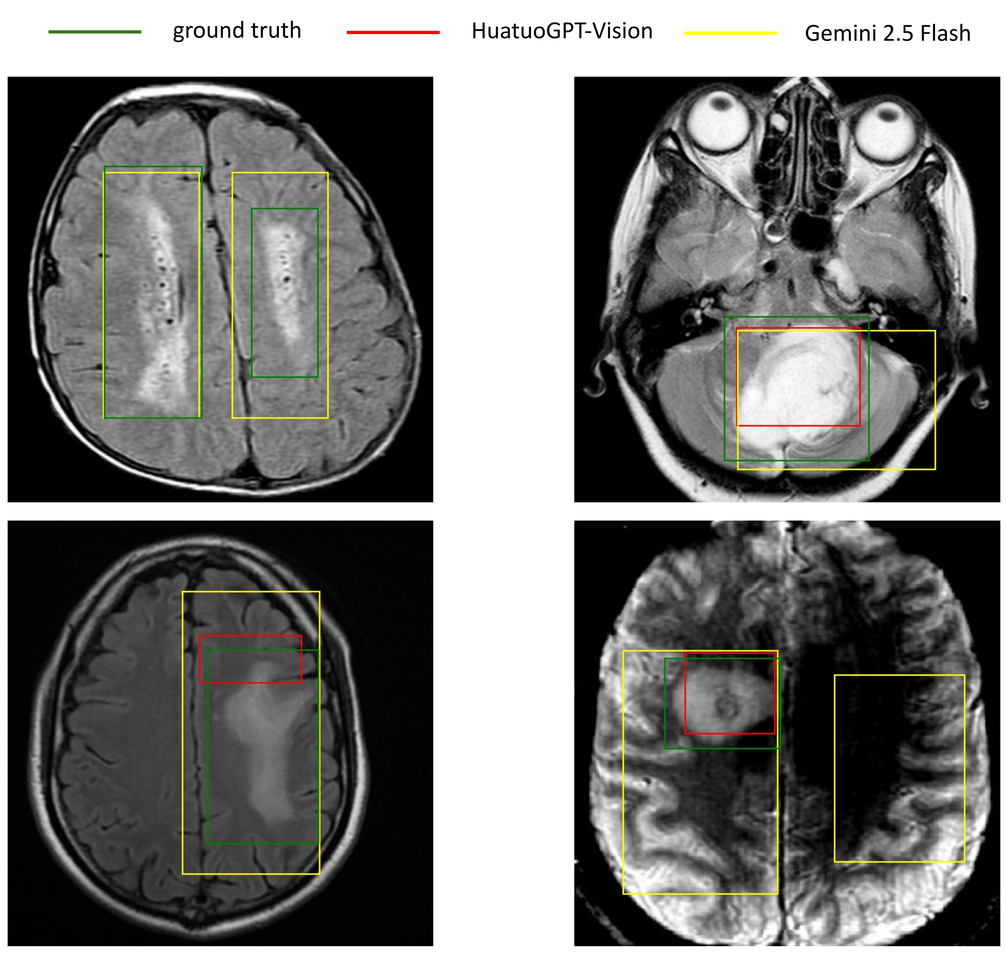


Chest X-rays Abnormality Grounding



	Gemini	Maira-2
Average mAP@50:95	0.00%	0.43%
Average mAP@50	0.00%	1.66%
Average mAP@75	0.00%	0.07%
RoDeO/localization	9.35%	21.99%
RoDeO/shape_matching	7.20%	17.85%
RoDeO/classification	3.26%	23.63%
RoDeO/total	5.43%	20.08%

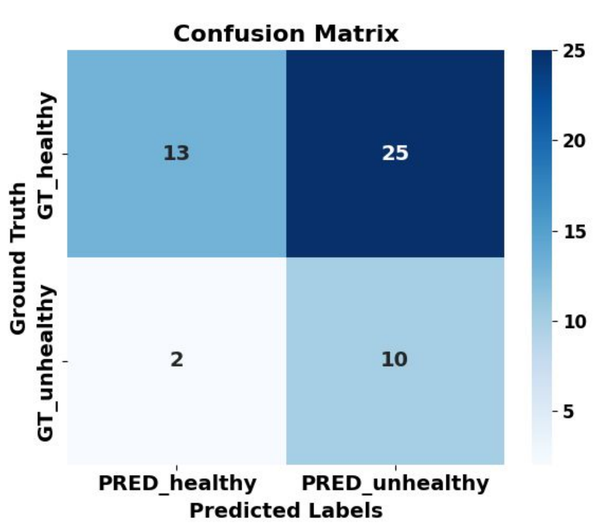
Brain MRI Abnormality Detection



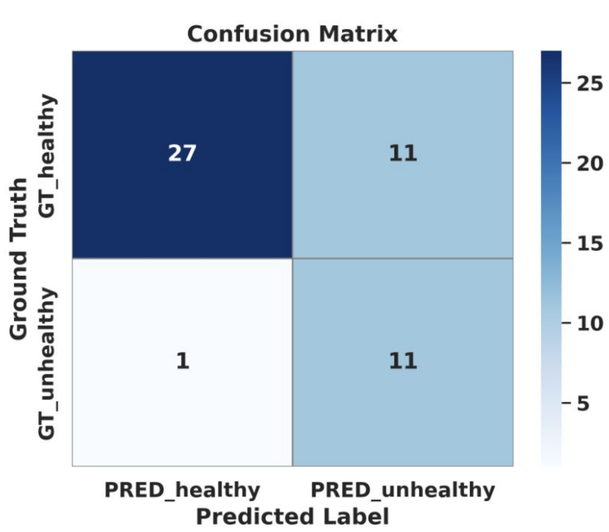
	Gemini	HuatuoGPT-Vision
Average mAP@50:95	3.38%	1.94%
Average mAP@50	11.90%	8.33%
Average mAP@75	0.55%	0.00%
RoDeO/localization	37.07%	49.25%
RoDeO/shape_matching	28.36%	32.90%
RoDeO/classification	41.47%	52.93%
RoDeO/total	34.75%	41.36%

Chest X-rays Classification Task

Gemini 2.5 Flash



Maira-2



Brain MRI Disease Diagnosis

Llama-3.1-8B

	HuatuoGPT-Vision
Accuracy	72%

Gemini 2.5 Flash

	Gemini 2.5 Flash
Accuracy	48%

Jingyi He, Bivek Panthi