

Evaluating GPT-4o mini for Diagnostic Prediction, Description and Localization in Medical Imaging



Tuna Karacan

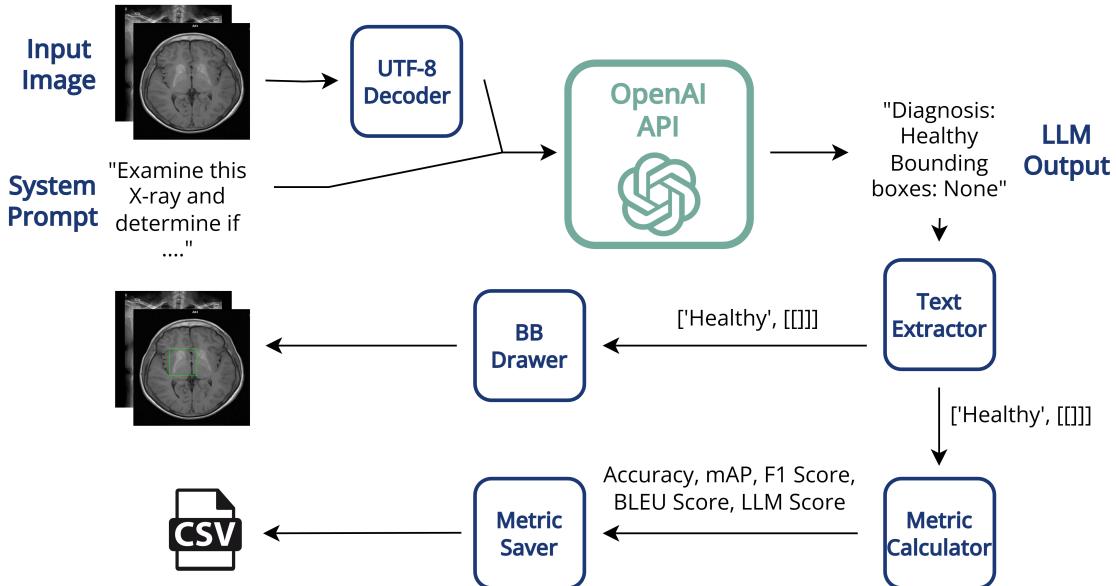
Introduction

🚀 **Motivation:** Large language models (LLMs) have shown surprising capabilities across text, image, and multimodal tasks.

🔧 **Method:** We applied GPT-4o mini to Chest X-rays and MRI brain slices using multimodal prompts and evaluated its outputs across classification, localization, and captioning tasks.

🎯 **Goal:** Assess the viability of zero-shot LLM-based analysis for clinical imaging without domain-specific fine-tuning.

Methodology



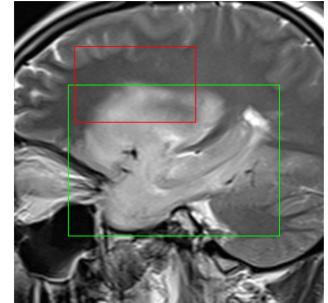
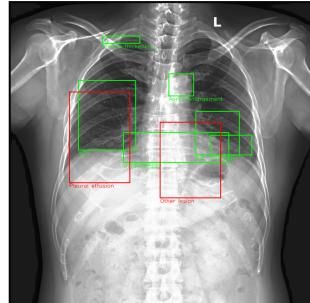
Results

Accuracy	F1- Healthy	F1 Unhealthy	mAP@50:95	mAP@50:95	mAP@50:95
30.00%	14.63%	40.67%	0.00%	0.00%	0.00%

Results for Chest X-Ray Experiments

BLEU-1	BLEU-2	LLM Score	mAP@50:95	mAP@50:95	mAP@50:95
10.80%	1.96%	9.23%	0.80%	0.04%	0.00%

Results for Brain Slice Experiments



Analysis

🧠 **Model Misalignment:** GPT4-o mini is not trained on medical image datasets, limiting its ability to extract clinically relevant visual features.

☒ **Below Random Performance:** Evaluation metrics (e.g., accuracy, mAP, F1) were worse than random guessing, suggesting no meaningful pattern recognition.

🖼 **Poor Visual Grounding:** The model failed to identify disease regions, often producing bounding boxes that were random or unrelated to pathology.

🌐 **Hallucinated Descriptions:** Text outputs included plausible-sounding but incorrect or irrelevant medical terms, highlighting limitations in factual consistency.

☒ **No Fine-Tuning:** Using the model in a zero-shot setting, without domain-specific adaptation, resulted in generic or misaligned responses.

⚙ **Prompting Limitations:** Even carefully engineered prompts could not reliably elicit correct classification or localization behavior from the model.

Conclusion

🚧 **Limitation:** Zero-shot inference, though convenient, cannot replace specialized architectures trained on medical datasets.

💡 **Key Insight:** General-purpose LLMs, even with multimodal capabilities, are currently unsuitable for diagnostic imaging without domain-specific training.

📝 **Future Work:** Incorporate medical vision-language models (e.g., BioViL, Med-PaLM) or fine-tune on labeled datasets like CheXpert or BraTS.