Evaluating the Spatial Reasoning Capabilities of Large Multimodal Models on Chest X-Ray Anomaly Detection

Lindsay Skylar Li
Kent Place School
Summit, NJ 07901
lilskylar26@gmail.com

Abstract

Large language models have shown significant performance on text-based medical evaluations like the USMLE. Since medical data inherently involves more than textual data, large multimodal models (LMMs) have the potential to have broad applications in healthcare—from generating radiology reports to recommending treatment options for patients. Recent studies have shown promising results on a variety of tasks, from answering questions based on images provided to solving complex 3-D protein structures. While current results show potential in LMM-based diagnosis, it is unclear if the output of them are backed by strong spatial reasoning capabilities. To evaluate this, I provided GPT-4o with chest X-rays and asked it to return diagnoses and the coordinates of bounding boxes that surrounded any identified abnormalities on the NIH chest X-ray dataset. I find variable performance across different images in the dataset, suggesting the need for further development of the spatial reasoning capabilities of LMMs. The methodology used in this paper can potentially be used in assessing the progress of GPT and other LMMs.

1 Introduction

Large language models have demonstrated remarkable progress on a variety of text-based medical evaluations like MedQA [10], demonstrating notable capabilities in some cases [1, 5, 6, 7]. In addition to text, medical applications often involves various forms of data, such as X-rays and magnetic resonance imaging (MRI). Recent models have been trained with capabilities to interpret different modalities, including image and audio data. Large multimodal models (LMM), as a result, have the potential of playing a significant role in medicine, with applications ranging from radiology report generation to identifying events in videos. Currently, there are multiple LMMs, including OpenAI's GPT-4o, Google's Gemini, and Meta's ImageBind. Although they are capable of taking in medical images and returning text based on the input, quality of results is inconsistent. Even when they are able to recognize images and diagnose problems correctly, it is not clear whether these conclusions are backed up with spatial reasoning and understanding of the image. In order for these models to be trusted, they must demonstrate strong spatial reasoning.

Since models cannot necessarily output spatial information, it is challenging to find a method for evaluation. I evaluate LMMs on object detection: not only do I ask for a diagnosis, but also prompt it to return the coordinates around where it feels there is an abnormality. I use this idea to delve into how precise GPT-4o is able to be at these specific tasks when given chest X-rays: diagnosing images with some background information, and pointing where exactly this abnormality is seen.

Preprint. Under review.

2 Related work

Previous work has involved datasets with chest X-rays, including CheXpert [3] and the National Health Institute's (NIH) ChestX-ray8 [8]. These were two of the datasets used in an evaluation of Gemini's capabilities when given medical scenarios [9]. They tested with a larger variety of images, including ones from different fields such as dermatology and ophthalmology. The focus was on its ability to diagnose and recognize abnormalities in these X-rays, generating a report. They prompted Gemini by giving X-rays and photographic images, asking it to describe them, provide specific diagnoses if present, and suggest treatment options. They report that anywhere between 43% to 96% of AI-based judgements were evaluated as "equivalent or better" compared to those by radiologists.

For chest X-ray specific studies, they focused on common conditions, namely atelectasis, cardiomegaly, consolidation, pulmonary edema, and pleural effusion [4]. Chest X-rays were uploaded, and the model was prompted with: "Given the following chest X-ray, describe the FINDINGS and IMPRESSION in the image." Gemini occasionally missed diagnoses, but they found that it was proficient in diagnosing cardiomegaly and pleural effusion when present.

3 Methods

3.1 Dataset

In this study, the NIH dataset of chest X-rays, called ChestX-ray8, was used [8]. This publicly-available dataset includes 112,120 images from 30,805 unique patients with corresponding labels including the diagnosis (finding label) and values relating to a bounding box drawn around the area of the diagnosed problem. These labels were used to test the accuracy of GPT-4o's diagnoses; the average intersection over union (IoU) score is reported.

3.2 Prompting methods

I provided GPT-4o with chest X-rays as input and a system prompt outlining the process it was to take to analyze the image given. The process was to first identify the problems in the X-ray, if any, determine the most pressing one, and return the coordinates of a bounding box around any abnormal areas. The prompt also included possible diagnoses of the images: atelectasis, effusion, cardiomegaly, infiltrate, mass, nodule, pneumonia, and pneumothorax. Asking the LMM to perform more sophisticated tasks, specifically object detection, tests its capability to make medical diagnoses.

```
You are a professional radiologist and pulmonologist capable of analyzing chest
↪  x-ray images and identifying what the most pressing issue is. You will be given
↪  a 1024 (max_y) x 1024 (max_x) chest x-ray as input.

You are to use the following procedure on each of the chest x-rays:
    1) Search for the following abnormalities in the chest x-ray: atelectasis,
    ↪  effusion, cardiomegaly, infiltrate, mass, nodule, pneumonia, pneumothorax.
    2) Find the coordinates and size of the area that is most affected by that
    ↪  abnormality. If there are multiple problematic areas present, select the one
    ↪  that is most prominent.

Return: coordinates <x_top, y_top> and size: <x_size, y_size>.

Include an explanation of what you see, and then you must include your final
↪  diagnosis in the format:

<diagnosis>
ABNORMALITY: {True/False} (true if abnormal, false if not abnormal)
COORDINATES: (x_top, y_top)
SIZE: (x_size, y_size)
</diagnosis>
```

I only include evaluations for labeled images that GPT-4o returned a positive diagnosis for. I parsed its response to determine whether it thought there was an abnormality, and the coordinate values and dimensions of its bounding box, if any. I evaluated how well the bounding box was to the ground truth label by computing the IoU. If the abnormality value was false, meaning there was no problem identified, no IoU was returned. The code can be found in a GitHub repository here: https://github.com/SkylarL26/Medagent.git.

4 Results

Dataset	Number of images	Mean IoU (min, max)	Standard deviation IoU
ChestX-ray8 [8]	455	0.067 (0, 0.523)	0.115

Table 1: Result values

The resulting mean IoU of all test examples was low, as a perfect bounding box has an IoU of 1. GPT-4o was often able to identify the problems of the images, but was not able to select the most severe one, causing it to diagnose the incorrect abnormality, and draw a box around the wrong area. There were also instances where GPT-4o did not recognize any problems even when there were abnormalities present.

When GPT-4o was able to reach the correct diagnosis, the bounding box drawn was variable in how close it was to the ground truth label. Below is an example of this type of result1(a): it identifies correctly that the abnormality is cardiomegaly, and the IoU is about 0.384. Although it was able to correctly identify the diagnosis, the box it returned was too small but roughly in the right area.

```
Upon examining the chest X-ray, the following observations are made:
1. **Cardiomegaly**: The heart appears enlarged, occupying more than half of the thoracic width.
2. **Effusion**: There is a suggestion of pleural effusion, particularly on the left side, as indicated by the blunting of
↪  the costophrenic angle.
3. **Infiltrate**: There are areas of increased opacity in the left loIr lung field, which could indicate an infiltrate.

The most prominent abnormality appears to be the cardiomegaly.

**Coordinates and Size of the Most Affected Area:**
- **Coordinates**: (400, 500)
- **Size**: (300, 300)

**Explanation**: The heart shadow is significantly enlarged, suggesting cardiomegaly. This is the most pressing issue as it
↪  can indicate underlying heart disease or failure.

<diagnosis>
ABNORMALITY: True
COORDINATES: (400, 500)
SIZE: (300, 300)
</diagnosis>
Image: 00000661_000.png, Abnormality: True, IoU: 0.38372330991199943
```

In other instances, GPT-4o was not able to make a correct diagnosis, and either recognized that the correct diagnosis was a problem or ruled out the possibility altogether. The latter is illustrated in the example below1(b): cardiomegaly is the correct diagnosis, and GPT-4o claims that it is not present. Usually, this mistake causes the returned bounding box to be disparate, resulting in an IoU that is or is close to 0. However, in the case below, although the diagnosis is wrong, the IoU value is high. But, this number is irrelevant to GPT-4o's reasoning capabilities, because the box it drew happened to overlap with the correct area. In this case, the incorrect diagnosis is more significant, as it proves its lack of proficiency to diagnose chest X-rays.

```
Upon examining the chest X-ray, the following observations are made:

1. There is a significant opacity in the right lung field, which appears to be consistent with an infiltrate or
↪  consolidation.
2. The heart size appears to be within normal limits, so cardiomegaly is not evident.
3. There is no clear evidence of a pneumothorax.
4. There is no visible mass or nodule.
5. The left lung field appears relatively clear compared to the right lung field.
6. There is no significant pleural effusion noted.
7. The presence of medical devices (e.g., central venous catheter, pacemaker) is noted, but they do not interfere with the
↪  primary findings.

The most prominent abnormality is the infiltrate in the right lung field.

Coordinates and size of the most affected area:
- Coordinates: (x_top, y_top) = (300, 200)
- Size: (x_size, y_size) = (400, 600)

<diagnosis>
ABNORMALITY: True
```

```
COORDINATES: (300, 200)
SIZE: (400, 600)
</diagnosis>
Image: 00009608_024.png, Abnormality: True, IoU: 0.40114716725806265
```

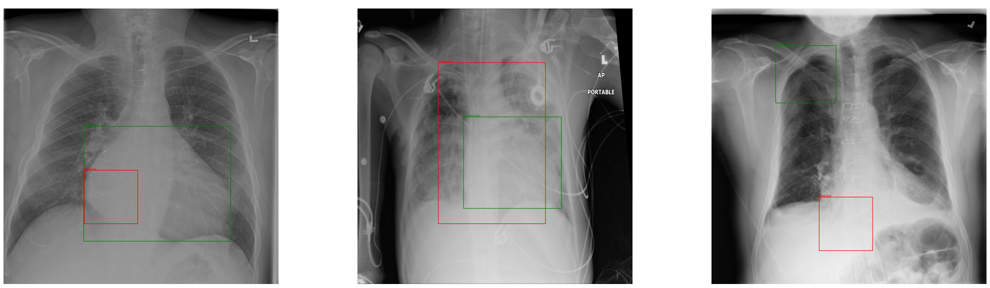
The case illustrated above is an outlier. Usually, when GPT-4o's diagnosis is incorrect, the IoU is, or is close to, 0, since it is unlikely that the diagnosis and the actual abnormality lie in the same region of the X-ray. This is represented in the example below1(c). This represents GPT-4o's reasoning capabilities better than the previous example, as the IoU captures the inaccuracy of its diagnosis.

```
Upon examining the chest X-ray, the following observations are made:

1. There is a noticeable opacity in the right loIr lung field, which could indicate the presence of an infiltrate or
↪  consolidation.
2. The heart size appears to be within normal limits, suggesting no cardiomegaly.
3. There are no visible signs of pneumothorax, as the lung fields appear to be fully expanded.
4. There is no evidence of pleural effusion, as the costophrenic angles are sharp.
5. No distinct masses or nodules are observed in the lung fields.
6. The presence of surgical clips is noted, which may be related to a previous surgical procedure.

The most prominent abnormality is the opacity in the right loIr lung field.

<diagnosis>
ABNORMALITY: True
COORDINATES: (450, 700)
SIZE: (200, 200)
</diagnosis>
Image: 00021896_003.png, Abnormality: True, IoU: 0.0
```



(a) Annotated X-ray Image 00000661_000.png

(b) Annotated X-ray Image 00009608_024.png

(c) Annotated X-ray Image 00021896_003.png

Figure 1: Annotated X-ray Images
(Green box represents ground truth, red box represents predicted.)

The first two examples show instances where the IoU was on the higher side. Most of the time, the IoU was between 0 and 0.1, as represented in the mean value.

5 Conclusion

After analyzing these results, it is apparent that GPT-4o is not yet proficient at accurately analyzing and diagnosing chest X-rays, shown with the low mean IoU value. But, limitations of my method could have also contributed to these results. For example, there is a chance that improvement to my prompting could lead to better execution of the task. An interesting way to address this would be to place some examples in the prompt, as language models tend to do better when provided with samples[2]. An interesting follow up investigation would be to include images with their corresponding labels or radiology reports to see if performance would improve. Furthermore, it should be noted that the evaluation methodology used in this research is relatively stringent, as it requires substantial overlap between the ground truth and predicted bounding boxes. However, it still illustrates the current lack of spatial reasoning by GPT-4o when analyzing images. With the rapid advancement of various LMMs, such as GPT-4o, the application of them in healthcare is hopeful and promising with the possibility of providing accurate diagnoses of medical images and helping physicians better serve patients. The methodology used in this paper can potentially be used in assessing the progress of GPT and other LMMs.

Acknowledgments and Disclosure of Funding

I want to express my gratitude to Linden Li, for helping to review my code and provide consultation to this project.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[3] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[4] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

[5] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[6] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[7] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[8] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[9] Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*, 2024.

[10] Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.