Application of Large Language Models to the Evaluation of an Automated System for Chest Radiograph Interpretation

Sally Yu s8yu@ucsd.edu
Eudora Fong efong@ucsd.edu
Aishani Mohapatra aimohapatra@ucsd.edu

Abstract

In this project, we are building off of an existing image-encoder and LLM framework, LLaVA (Large Language and Vision Assistant), by fine-tuning the model on our dataset of around 23,000 chest X-ray images sourced from patients at UC San Diego Health. Utilizing prompt engineering techniques, our objective is to build a medical LLM that specializes in interpreting X-ray images and generating diagnostic reports that parallel the expertise of professional radiologists. The assessment of model performance will be done through both Natural Language Processing (NLP) methodologies and LLM evaluation to ensure a comprehensive analysis of the model's diagnostic accuracy.

Website:
aimohapatra.github.io/medical-diagnosis-assistant-interpretation/
Code: https://github.com/yushi20/Medical_LLM

1 Introduction

In the radiology world, Radiologists and other physicians must analyze an ever-increasing volume of images on a daily basis due to the continuous rise and use of medical imaging for clinical diagnosis and management. This demand not only poses challenges for accurate diagnosis but also strains the efficiency and effectiveness of healthcare delivery.

Combined neural network architectures with vision encoders and large language models (LLM) have emerged as a potential solution to help automate interpretation. If successful, this technology would be groundbreaking to the field of radiology and clinical medicine. To successfully develop these new technological advancements for use, it is important to develop accompanying technology to assess diagnostic performance of the outputs of these algorithms. However, analyzing the outputs of these algorithms is a complex task, and may itself benefit from LLMs. We explored several approaches to use LLMs to evaluate a prototype algorithm.

2 Methods

In this project, we implemented LLaVA (Large Language and Vision Assistant) to analyze X-ray images and patient history to diagnose a patient (Figure 1). We trained the model with approximately 23,000 X-ray images from patients at UC San Diego Health collected over the years. It includes patient clinical history, radiologist diagnosis, and authorship. The six disease entities anaylzed are: Pneumothorax, Pneumonia, Pleural Effusion, Cardiomegaly, Edema, Rib Fracture. While training, there are two main aspects: the feature alignment stage and the visual instruction tuning stage. During pre-training (feature alignment), a frozen pre-trained vision encoder is connected to a frozen LLM. In the fine tuning stage (visual instruction tuning), the X-ray images are prepared for the model and can directly be used to fit the model.

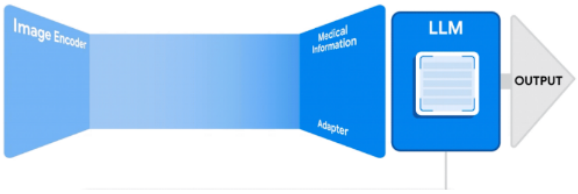

Figure 1: LLaVA framework

After training, we looked into several ways to improve the quality of answers the LLM outputted. One way of doing this is by experimenting with prompting and asking for specific information. We proposed different questions to the model in order to get more specific diagnoses. We also added clinical history, author, and gender to the prompt by training the model on a new set of data that has those three variables corresponding to the X-ray images.

In order to improve the training process of the LLM, one of the ways we reduce noise is through grouping authors less frequently observed in the data into a single miscellaneous author category. In this way, we are able to generalize predictions of a low frequency or never before seen authors.

Another focus of the project is on the evaluation metrics. Due to the semantic complexity of radiology reports, there is a need for a thorough investigation into different methods for assessing model performance. To start, we applied traditional NLP techniques and compared both models with the ground truth (radiologist reports) using similarity scores. Specifically, we vectorized the text reports using Term Frequency - Inverse Document Frequency (TF-IDF) and performed cosine similarity to produce a similarity score that quantifies model performance in terms of sentence structure development and general similarity to real radiology reports. Another method we explored was an LLM-assisted evaluation. For this, we employed the GPT-3.5 Turbo model API developed by OpenAI to assess the accuracy of LLaVA across six disease entities: Pneumothorax, Pneumonia, Pleural Effusion, Cardiomegaly, Edema, Rib Fracture. This involved extracting and comparing disease severity from radiologist and LLaVA generated reports. Leveraging the advanced language capabilities of LLMs in deciphering complex sentence constructions and sentiment, this approach may offer a more nuanced method for assessing LLaVA's performance.

3 Results

When comparing the cosine similarity scores between the model output and radiology reports for both the baseline and final model test sets, we observed a notable improvement in scores in the final model. As seen in Figure 2 and Figure 3, the final observes more unique values and improved cosine similarity scores with the integration of clinical context and an increase in the size of the training data. This stark difference in results parallels the vital importance of patient clinical history in real-world radiologist impressions as well as the impact of training data size on model performance. A comparative analysis of perfor-

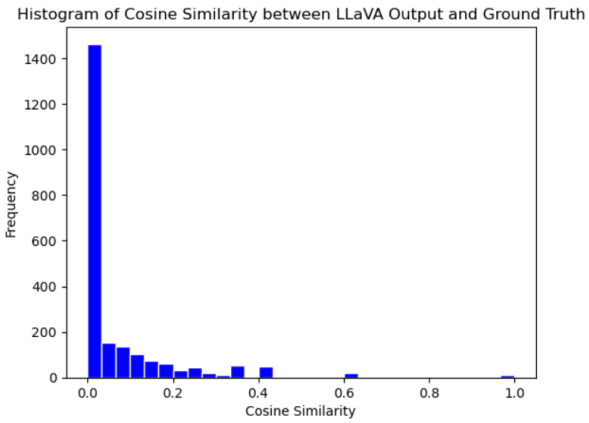

Figure 2: Histogram of LLaVA baseline model reports vs. ground truth radiologist reports cosine similarity

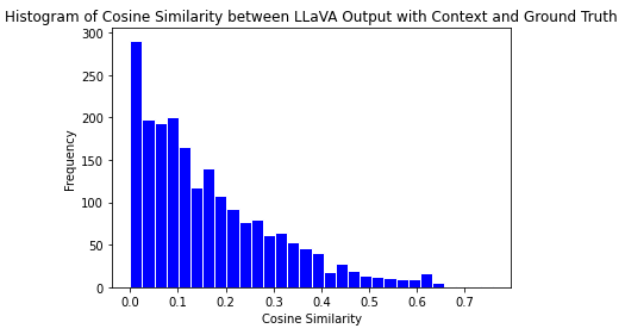

Figure 3: Histogram of LLaVA final model reports with context vs. ground truth radiologist reports cosine similarity

mance between the baseline and final model evaluated by ChatGPT 3.5 Turbo demonstrates an improvement in accuracy across all six disease entities examined. Shown in Figure 4, the highest increase in accuracy is seen in Rib Fracture with a 0.319 improvement in F1 score, while the smallest improvement is observed in Pneunothorax with only 0.11. The final model still demonstrate a bias toward the ”none” class across the severity spectrum for all six disease entities, as illustrated in Figure 5. This inclination may stem from an insufficient address of class imbalance prior to training, given that the dataset is disproportionately skewed toward the ”none” class. Further tuning of the model should therefore integrate techniques to handle the class imbalance issue to improve model precision.

	Baseline Model	Final Model
Pneumothorax	0.953000	0.964000
Pneumonia	0.368000	0.684000
Pleural Effusion	0.661000	0.745000
Cardiomegaly	0.627000	0.766000
Edema	0.593000	0.774000
Rib Fracture	0.651000	0.970000

Figure 4: LLM evaluation of model F1-score by disease entity

4 Discussion

In summary, the enhancement of model accuracy, as evidenced by both similarity scores and individual F1 scores for each disease entity analyzed with the assistance of the ChatGPT 3.5 Turbo model, underscores the value of incorporating clinical contextual data and expanding training data volume. The final model also is shown to generate report with

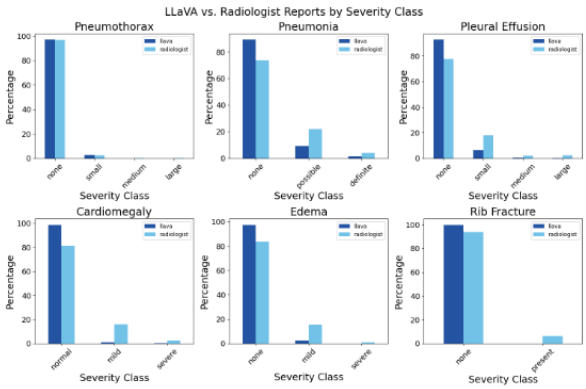

Figure 5: Bar chart of comparison between LLaVA and radiologists reports by disease entity and severity classes.

high accuracies, reaching 95Compared to prior research in this field[3][4], the most notable aspect is LLaVA's ability to generate comprehensive reports, rather than labelers that mainly focus on the presence of diseases (CheXpert). This gives a more detailed analysis of X-ray images that can potentially replace the manual reading that radiologists have to perform. However, there are limitations to this project. One of which is the potential problem with regarding radiologist reports as ground truth. In the preprocessing step, we did not implement any techniques to filter and grade the radiologist reports before training. This could have introduced some compromise to the accuracy of training data, hence accuracy of the model. Due to the lack of ways to grade the accuracy of radiologist reports, steps that may alleviate this problem could be taken to further aid the development of LLaVA, such as the filtering based on only skilled radiologists or giving higher scores to the reports written by skilled radiologists.

5 Conclusion

The traditional NLP method of evaluation provides a way of comparing the overall similarity of the LLaVA generated reports to the ground truth of radiologist reports. However, it may fall short in accurately extracting detailed semantics.Therefore, the incorporation of LLM-assisted evaluation becomes essential. The adeptness of LLMs in parsing complex linguistic structures enables the automated extraction of sentiment from intricate sentence constructions within the reports. This offers a nuanced assessment of the diagnostic accuracy addressing specific diseases, but still exhibits a lack of consistency that needs to be further addressed in subsequent research. For future directions, a more thorough comparative analysis should be done to test various state-of-the-art Large Language Models (LLMs) in evaluating model performance. This can provide valuable insights into the relative strengths and weaknesses of different LLMs on parsing medical reports and assessing diagnostic accuracy of medical LLMs. Additionally it may be valuable to personalize the model output to the user through prompting so that reports are in the voice of a particular radiologist and impressions are true to their x-ray reading style.

6 Contributions

Sally Yu - Maintain codebase on repo; trained the baseline model; ran inference on the validation set; developed and ran evaluation on baseline and final model using OpenAI's API; in charge of writing sections corresponding to the LLM-assisted evaluation on poster, website, and paper; wrote intro and abstract for the paper checkpoint, updated methods, results, conclusion, and references after the checkpoint, prepared the inference examples, and generated plots for LLM-assisted evaluation results.

Eudora Fong - Data preparation; both baseline and final model evaluation with NLP and similarity scoring techniques; in charge of writing sections corresponding to the NLP evaluation on poster, website, and paper; contributed to intro, methods, results, conclusion, built all LaTeX for paper, generated NLP evaluation plots, finalize website.

Aishani Mohapatra - Researched LLaVa documentation and attempted training baseline model; debugged baseline model; developed website, updated introduction and edited sections of the paper for checkpoints and the final product, updated sections of the poster for checkpoints and the final poster, assisted with poster and website aesthetics.

References

1. H. Liu, C. Li, Q. Wu, & Y. J. Lee. (2023, December 11). Visual instruction tuning. *arXiv.org*. Retrieved from https://arxiv.org/abs/2304.08485
2. Haotian-Liu. (n.d.). Haotian-liu/llava: [NeurIPS'23 oral] visual instruction tuning (llava) built towards GPT-4V level capabilities and beyond. *GitHub*. Retrieved from https://github.com/haotian-liu/LLaVA
3. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019, January 21). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv.org*. Retrieved from https://arxiv.org/abs/1901.07031
4. Xu, S., Yang, L., Kelly, C., Sieniek, M., Kohlberger, T., Ma, M., Weng, W.-H., Kiraly, A., Kazemzadeh, S., Melamed, Z., Park, J., Strachan, P., Liu, Y., Lau, C., Singh, P., Chen, C., Etemadi, M., Kalidindi, S. R., Matias, Y., … Sellergren, A. (2023, September 7). ELIXR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and Radiology Vision Encoders. *arXiv.org*. Retrieved from https://arxiv.org/abs/2308.01317

Appendices

A.1 Project Proposal

In the radiology world, radiologists must analyze an ever-increasing volume of X-ray images on a daily basis due to the exponential rise of medical imaging. This demanding pace not only poses challenges for accurate diagnosis but also strains the efficiency and effectiveness of healthcare delivery. Recognizing the pressing need for transformative solutions in this domain, we aim to automate and streamline this process by creating a way to use a Large Language Model (LLM) to analyze X-ray image input and generate a text radiology report, similar to the way that current LLM applications like ChatGPT can generate text from input. By applying this technology to the field of radiology, we seek to augment diagnostic accuracy while optimizing the operational workflow of this area of healthcare for the benefit of both radiologists and patients.

More specifically, we want to build a labeler that consists of an image encoder and a LLM, which can take in X-ray images and output radiology reports. Comparable to the idea of ELIXR, a system made of a language-aligned image encoder grafted onto a LLM that can perform CXR classifications, our model aims to achieve a similar objective where instead of the output being classification of a list of diseases, it generates more comprehensive reports.

This project relates to our quarter 1 project as neural network models can also be used as autoencoders that map the data present in X-ray images into a compressed vector embedding space that LLM can then map to human-readable text as medical labels. Our experience with tuning hyperparameters on different image augmentation techniques and dynamic ranges of input images should be useful as well in helping us fine-tune the system after we get a baseline model. Past work (ELIXR) in this space have achieved relatively good results by combining a visual model with a neural LLM, achieving a mean AUC of 0.850 across 13 findings on zero-shot chest X-ray classification task. It utilizes a medical information adapter that takes the output from the image encoder and maps them to a series of tokens in the LLM's input embedding space as a connector between the encoder and LLM. This is something that we also want to try implementing to our model. Since previous research did not fine-tune the image encoder nor the LLM, we would like to spend time fine-tuning our model to see if performance can be improved.

Our primary output will be a website, as it will portray all the necessary information to readers in a clear and concise manner. This is mainly because we want it to be accessible to as many people as possible, and have a layout that allows a wide range of people to access and understand our project. We are also planning to write a paper in order to reach a more scientific community and explain the intricacies of our model in more detail.

Our project primarily generates data, and the best way to communicate this is through images, graphs, and simplified explanations of our methodology and results. Since the main

component of our quarter two project is a LLMs for radiographic images, we can communicate that by creating a diagram: Image → Text, and showing an example of example text the LLM might generate based on a sample image. The caption that our LLM generates can then be compared with the text from the original data to see how accurate the LLM's interpretation of the image is. Radiographic images are not easy to interpret, especially since radiologists take years and thousands of images to really understand radiographic images properly. The small intricacies in radiographic images can be difficult for someone without professional training to interpret, so showing a comparison of images with and without pulmonary edema with their corresponding descriptions is important. Since we are trying to teach an audience about our approach to caption radiographic images, it would be helpful to show a comparison of the input, output, and an explanation of how they compare. Radiologists are the primary audience our project targets, so to help them detect pulmonary edema and other conditions we could have an option for the user of the website to input their own radiographic image and see what caption our LLM generates for it.