Report Data Summarization to Improve Radiology Workflows

Ishan Sinha
Adi Badhwar
Disha Sharma
Mohamed Ibrahim Osman

Abstract

This study explores the application of Large Language Models (LLMs) for summarizing EHR clinical notes, comparing open-source models like Llama2 7B and Mistral 7B against the proprietary GPT-4 Turbo. Employing domain adaptation and one-shot learning techniques, we demonstrate Mistral 7B's superior performance in aligning closer to the GPT-4 Turbo's outputs, which served as a proxy for ground truth. Initial findings, pending clinician feedback, suggest that these LLMs, particularly when domain-adapted, significantly enhance the summarization process, potentially improving radiology workflows. Our research highlights the importance of model customization and the promising role of LLMs in streamlining medical document analysis.

1 Introduction

Electronic Health Records (EHRs) contain crucial information that aids radiologists and other physicians in diagnosis. However, because patient records are extensive and depict complex patient timelines, as illustrated in Figure 1, locating relevant details within the limited time available for each case can pose a challenge. To address this issue, we propose and assess models designed to extract pertinent text snippets from patient records, offering a concise summary to assist radiologists in considering potential diagnoses.

A patient's medical journey is documented through comprehensive free-text narratives stored in Electronic Health Records (EHRs). These narratives encompass input from multiple care teams, specialties, and perspectives, covering various aspects such as symptoms, diagnoses, procedures, interventions, clinical and social histories, and future prognoses. Despite the variability in scope, detail, and structure across these narratives, they collectively provide valuable insights into patient care. However, synthesizing this information into a concise summary, particularly for complex cases, can be daunting and time-consuming.

In the realm of computational linguistics, this challenge can be framed as a multi-document summarization task, where the model must adapt to varying numbers of documents, time intervals between notes, differences in note types, and the diverse aims and focus areas of document authors.

Clinical narratives represent a significant portion of EHR data, yet progress in developing and applying text summarization methods has been relatively slow compared to other areas such as disease prediction and clinical information extraction. Several factors contribute to this, including the complexity of collecting reference summaries, the high stakes nature of AI-driven summarization, and the difficulty in evaluating model performance using automated metrics.

Previous efforts have primarily focused on extractive approaches, leveraging semantic similarity modeling and methods for selecting representative sentences. In radiology, reports are structured into sections, with the impression section serving as the target reference summary for model develop-

ment. Summarizing radiology reports shares similarities with single-document open-domain tasks, emphasizing the modeling of sentence salience and compression.

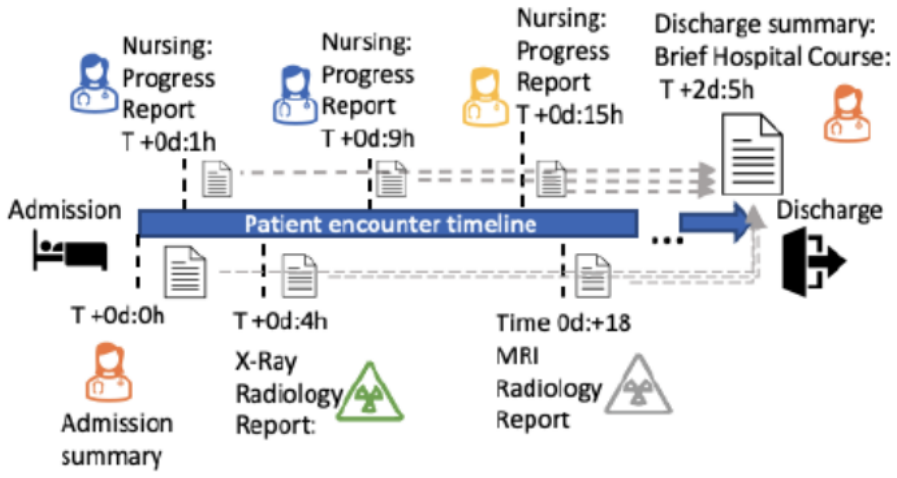

Figure1:Patient clinical notes timeline from Admission to Discharge

2 Related Work

The integration of Large Language Models (LLMs) into radiology is transforming the field by enhancing the interpretation and summarization of radiology reports. Models like Radiology-GPT, specifically trained on medical texts, are proving to be significantly more effective than their general-purpose counterparts.[7] This specialization not only improves accuracy but also efficiency in the radiology workflow. Additionally, innovations such as RadAdapt's domain adaptation strategies further refine these models' abilities to process and summarize clinical texts, making LLMs an indispensable tool for radiologists.[10]

Another breakthrough comes from the application of Direct Preference Optimization (DPO), which optimizes LLM decision-making in clinical settings. This approach, by introducing a new parameterization method, ensures that AI-assisted diagnostics are both precise and tailored to individual patient scenarios.[8] Moreover, the creation of specialized datasets like RadGraph underlines the importance of having tailored training materials. These datasets contain complex language and specific information from radiology texts, crucial for training LLMs to accurately interpret medical reports.[5]

However, the integration of LLMs like GPT4 into radiology workflows is not without challenges. Issues such as data privacy and the need for extensive domain-specific fine-tuning highlight the complexities of deploying these technologies in a medical context[2]. These challenges emphasize the need for ongoing research, development, and adaptation to ensure that LLMs can effectively support radiological practices without compromising patient confidentiality or care quality.

Looking forward, the future of LLMs in radiology hinges on addressing these challenges through continuous innovation and research. Enhancing dataset quality and diversity, along with developing advanced modeling techniques, are pivotal steps in this direction. By doing so, the medical community can leverage LLMs to not only streamline radiology workflows but also to improve patient outcomes significantly. The potential of LLMs to revolutionize radiology practices is immense, promising a future where AI and human expertise collaborate closely for better healthcare delivery.

3 Approach

While closed models like GPT-4 and Claude offer convenience and powerful out-of-the-box solutions, open-source large language models (LLMs) have the potential to address some of the shortcomings of their closed counterparts. Firstly, open-source LLMs provide greater control over data privacy and

security. Healthcare organizations can host and run these models on their own secure infrastructure, ensuring that sensitive patient data remains within their control and complies with strict data protection regulations. Moreover, open-source LLMs can be fine-tuned to the medical domain, which is essential given the significant distribution shift between general domain data and medical data. With access to the source code and model weights, we can optimize open-source models for specific medical domains, potentially outperforming closed general domain models.

With the arrival of many open-source LLMs, the first task is selecting the most suitable model for our downstream task of summarizing the history in radiology reports. The most popular way of evaluating LLM performance is through benchmarks like the Massive Multitask Language Understanding (MMLU) benchmark [4]. However, while MMLU is a comprehensive benchmark useful for evaluating the general capabilities of LLMs, we believe it is not well-suited for evaluating models when the goal is to optimize for a specific downstream task, particularly for tasks like radiology report summarization. The gold standard is to have a radiologist evaluate the model output. However, it is very time consuming and expensive to have a radiologist evaluate and compare the outputs from all the open-source models. In this paper, we propose a more comprehensive automated evaluation framework for choosing open-source LLMs tailored to our downstream task of summarizing the history in radiology reports. We demonstrate this framework using two popular open-source models, namely Llama 7B[9] and Mistral 7B[6].

Since we do not have access to the ground truth (i.e., a summarized history report from a radiologist), we use GPT-4's output as a proxy for the ground truth. As demonstrated by the MMLU benchmark, GPT-4's model performance is substantially better than that of Llama 7B and Mistral 7B. Therefore, we believe GPT-4's output is much closer to the ground truth than the outputs from these open-source models. This assumption is also supported by our experience with the few examples we examined.

Now that we have a proxy for the ground truth, we can use semantic similarity metrics to evaluate the accuracy of the open-source models. For instance, if the output from Llama 7B has a higher semantic similarity to the proxy ground truth than Mistral 7B, we assume Llama 7B's output is more accurate. To calculate the semantic similarity between two outputs, we first need to embed them and then compare the embedding vectors. We used ClinicalBERT[1], a model initialized from BERT and fine-tuned on a large medical dataset, to embed the outputs.

The use of GPT-4 as a proxy for the ground truth and the employment of semantic similarity metrics with a domain-specific embedding model like ClinicalBERT allow us to evaluate the performance of open-source LLMs on our specific downstream task of summarizing radiology report histories. This approach provides a more comprehensive and tailored evaluation framework compared to general benchmarks like MMLU.

4 Experiments with Results

We first created an evaluation dataset of 100 randomly selected data points from the larger dataset. While using the entire dataset would have yielded more reliable results, we were limited by the cost of calling the GPT-4 APIs, which would have been prohibitively expensive for the full dataset. While conducting prompt engineering, evaluations of GPT4-generated summaries focused on identifying significant inaccuracies, unusual terminology, and structural errors. The inclusion of patient diagnosis codes, clinician speciality and parsing input data contributed to enhancements in performance.

We used the following system prompt as default for all models:

> You are an expert medical specialist. Write a short two summary in bullet points for the clinical note the user provides.

For the open-source models, we passed this as a system prompt and the radiology report as a user input. For GPT-4, we combined the system prompt and radiology report into a single user prompt since we cannot modify GPT-4's system prompt. For the open-source models, we set the following prompt parameters: $top\_p = 0.8, temperature = 0.5, max\_new\_tokens = 1024$. We set relatively low temperature and top-p values to generate more controlled and coherent output.

We first obtained the GPT-4 output, which would serve as a proxy for the ground truth. We then obtained results from the out-of-the-box Llama2 7B and Mistral 7B models. Using ClinicalBERT, we

embedded the outputs and calculated their cosine similarity to the proxy ground truth (i.e., the GPT-4 output embeddings), with the results shown in Table 2.

Next, we observed a significant domain shift between the general domain data the open-source models were pretrained on and the clinical data. To bridge this gap, we continued pretraining the LLMs on the clinical data using the masked language modeling objective. The details of the fine-tuning process for each model are provided in Table 1. After fine-tuning the models, we evaluated their performance on the evaluation dataset and reported the results in Table 2.

Finally, we obtained a clinical note with a radiologist summary and used it as a one-shot example in our prompts. We appended this one-shot example to the prompts of three models: GPT-4, the domain-adapted Llama2 7B, and the domain-adapted Mistral 7B. We share the results in Table 2.

Table 1: Hyperparameters for fine-tuning open-source LLMs

Hyperparameter	**Value**
epochs	7
learning_rate	$2e-4$
weight_decay	0.001
LoRa Rank	8
lora_alpha	32
lora_droupout	0.05

Table 2: The cosine similarity between the output embeddings of the open-source models and GPT-4

	LlaMa2 7B	**Mistral 7B**
Out-of-the-box	0.81	0.84
Domain-adapted	0.85	0.86
Domain-adapted + one-shot attempt	0.90	0.92

5 Discussion/Analysis of Results

According to the cosine similarity metric, Mistral 7B outperforms Llama2 7B in all experiments: out-of-the-box models, domain-adapted models, and domain-adapted one-shot attempt models. This is in agreement with their respective performances on the MMLU benchmarks, where Mistral 7B outperforms both Llama2 7B and Llama 13B across all tested benchmarks [6]. We believe Mistral 7B's superior performance is due to its architecture, which employs sliding window attention [[3], [2]] designed to handle longer sequences more effectively. This is critical, as some clinical notes can span many years, resulting in very long input sequences for the model.

Llama2 7B exhibited a 4% increase in cosine similarity after domain adaptation, while Mistral 7B showed a 2% increase. We hypothesize that the domain shift between Llama2 7B's original pretraining dataset and the clinical dataset is larger than the domain shift between Mistral 7B's original pretraining dataset and the clinical dataset. However, we cannot confirm this hypothesis since the pretraining dataset for Mistral is not publicly disclosed.

The one-shot attempt led to the most significant increase in cosine similarity: a 5% increase for Llama2 7B and a 6% increase for Mistral 7B. For the zero-shot attempt, the outputs from the models may vary substantially. However, the one-shot example standardizes the outputs, boosting the cosine similarity. The extent to which this boost translates to an actual increase in the quality of clinical summaries requires evaluation by radiologists.

According to the proposed cosine similarity metric, model performance monotonically increases as we employ techniques like domain adaptation and few-shot prompting, which are known to improve model performance. This supports the effectiveness of the proposed evaluation framework. However, fully proving the framework's effectiveness requires radiologist evaluation of the different outputs. Additionally, using ground truth summaries from radiologists would eliminate the need to rely on GPT-4's output as a proxy.

6 Conclusion

Our investigation into the utilization of open-source and proprietary Large Language Models (LLMs) for summarizing radiology reports has yielded promising initial findings. Mistral 7B, in particular, has demonstrated superior performance across several metrics, outperforming Llama2 7B in both out-of-the-box and domain-adapted scenarios. This superiority is attributed to Mistral 7B's architectural advantages, notably its ability to handle longer sequences more effectively, which is crucial for processing the extensive and complex clinical notes characteristic of radiology reports. The improvements observed through domain adaptation and the one-shot learning attempt further underscore the potential of fine-tuning and model customization in bridging the gap between general domain pretraining and specialized medical applications.

While we are still in the process of gathering feedback from clinicians to fully assess the practical impact of these findings, our results indicate that models like GPT-4 Turbo and Mistral 7B, especially when further optimized through domain-specific fine-tuning, offer significant advancements over traditional methods. The increased cosine similarity scores post-domain adaptation suggest that such tailored approaches can significantly enhance the relevance and accuracy of generated summaries, potentially streamlining radiology workflows and aiding in faster, more informed decision-making.

It's crucial to note, however, that the definitive validation of these models' utility in clinical settings hinges on comprehensive evaluations by radiology professionals. Future work will aim to corroborate these promising automated evaluation results with expert human assessments to ensure that the summaries generated by these LLMs are not only technically accurate but also practically useful in enhancing patient care.

In summary, while awaiting detailed feedback from clinicians, our investigation provides strong evidence of the enhanced capabilities of models like GPT-4 Turbo and Mistral 7B in the radiology domain. These LLMs present a significant step forward in our quest to leverage AI for improving the efficiency and effectiveness of radiology workflows, marking an exciting direction for future research and application in medical informatics.

7 Contributions of Team Members

Adi led the charge on securing data, pulling together the strategy, and running code on OpenAI GPT-4 and Anthropic Claude. Mohamed worked on the open source part of the project. Ishan worked on literature review. Disha worked on providing support from a medical perspective given she is in the medical school. We all contributed to the presentation and report.

References

[1] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[2] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[3] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[5] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

[6] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[7] Z Liu, A Zhong, Y Li, L Yang, C Ju, Z Wu, et al. Radiology-gpt: a large language model for radiology. arxiv [preprint]. 2023 [cited august 21, 2023].

[8] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[10] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. Clinical text summarization: Adapting large language models can outperform human experts. *Research Square*, 2023.