Medical Language Processing – LLM Based Impression Prediction from Radiology Reports

By

Junru Li

Kairan Zhong

Xiran Li

Yue Zhang

Supervisor: Utku Pamuksuz

A Capstone Project

Submitted to the University of Chicago in partial fulfillment
of the requirements for the degree of

Master of Science in Applied Data Science

Division of Physical Sciences

December 2023

Abstract

This research paper introduces the Large Language Model (LLM) specifically fine-tuned for radiology reports. The LLMs are designed to generate accurate impressions by leveraging patients' medical findings and background information. Several models have been explored, and one winning model has been deployed. The primary objective is to assist physicians in optimizing their documentation process, ultimately reducing the time required for generating comprehensive impressions within the radiology report. By employing this technology, physicians can potentially alleviate the risk of burnout associated with extensive and repetitive documentation tasks.

Keywords: Large Language Models, Fine Tuning, Abstractive Summarization

Executive Summary

This research paper presents the fine-tuned LLM for radiology reports. It is designed to produce precise diagnostic impressions by summarizing the textual input of patients' clinical information and findings within the radiology report. The goal of the study is to streamline the documentation workflow for physicians, thereby shortening the time needed to generate impressions. Implementing this model has the potential to mitigate physician burnout resulting from the demands of exhaustive and repetitive documentation duties.

The project starts by identifying the problem and underscoring the importance of efficient documentation in the healthcare industry. It then outlines the data used for this project: the radiology report dataset from UChicago Medicine. Next, the methodology section details the workflow process, beginning with feature engineering, model selection, and model tuning. Then it progresses through the training of LLM base models, integration with the LoRA Adaptor, and evaluation using the ROUGE score. The models are trained to understand the specific nuances and terminology within the radiology domain, enabling them to generate relevant, accurate, and concise impressions. Lastly, key findings and conclusions are presented, showcasing Mistral-7b as the chosen winning model, which has been integrated into a user-friendly interface for demonstration purposes.

Certain limitations of this project's applicability are taken into consideration. Future improvements could focus on exploring additional applications and enhancing the model to further improve its effectiveness in the medical field.

Table of Contents

List of Figures

List of Tables

Introduction

Problem Statement

The radiology report[1] is an essential tool for medical professionals in diagnosing and treating patients. However, generating accurate and comprehensive radiology reports from unstructured narratives in healthcare records can be time-consuming and challenging for radiologists. Current practices rely heavily on manual report generation, which is both costly and prone to human error. According to a study by the American Medical Association, the cost of generating a single medical report manually can range from $6 to $20. Moreover, physicians spend an average of 16 minutes per patient generating medical reports, according to a report by the same organization (American Medical Association, n.d., p. 1). These costs can accumulate, leading to inefficiencies in the healthcare system.

The goal is to develop AI[2] solutions that can efficiently generate radiology impressions[3] from unstructured narratives in healthcare records, thereby overcoming the existing challenge. The solution needs to interpret and analyze the clinical information, techniques used, and findings[4] from the radiology reports to generate accurate and concise impressions.

The client's desired solution involves the implementation of an LLM-based model specifically fine-tuned for radiology reports, accompanied by an AI-powered interface that swiftly and accurately generates impressions from medical inputs. This solution offers multiple benefits. Firstly, it can lead to significant cost savings for healthcare providers by reducing the

[1] Radiology reports are documents generated by radiologists that provide detailed summaries and interpretations of medical imaging studies.

[2] Artificial Intelligence (AI) refers to the field of computer science that focuses on creating intelligent machines capable of performing tasks that typically require human intelligence.

[3] Radiology impressions refer to the diagnostic conclusions and interpretations made by radiologists based on the findings from medical imaging studies.

[4] The findings of a radiology report refer to the specific observations and results obtained from the medical imaging study conducted by the radiologist.

time and resources required for manual report generation. Secondly, it can improve patient outcomes by reducing the risk of human errors and increasing diagnostic accuracy. The utilization of the AI-powered medical report generation solution has the potential to enhance diagnostic accuracy by up to 30%, thereby leading to improved patient outcomes (Frost & Sullivan, 2021, 2). Finally, the solution can help healthcare providers increase revenue by allowing them to see more patients and generate more reports. The use of AI in healthcare is expected to generate $6.6 billion in annual revenue by 2021 (Tractica, 2017, 1).

In summary, the development of the AI-powered solution for radiology report generation is a significant opportunity for healthcare providers to improve patient outcomes, increase efficiency, and reduce costs. By automating the report generation process, healthcare providers can allocate resources more effectively towards other areas of patient care, leading to better utilization of resources and improved patient outcomes.

Analysis Goals

The primary goal of the analysis is to create a medical language processing model that generates clear and concise impressions from complex clinical information and findings within the radiology report, thereby enhancing the accuracy and efficiency of medical diagnoses and treatments. This tool aims to interpret vast amounts of medical data, providing clear and concise impressions to medical professionals.

To accomplish the analysis goals, the project focuses on developing a robust medical language processing model capable of comprehending complex medical jargon, medical abbreviations, and various medical terminologies found in medical reports. By effectively processing and understanding the medical language, the model will be able to generate concise and clear impressions. Furthermore, the project will involve designing a user-friendly interface that integrates the medical language processing model, making it accessible via an API. This approach will facilitate the model's seamless integration into various applications used by other stakeholders.

In summary, the project's ultimate objective is to deliver a reliable and efficient medical language processing tool that improves medical diagnosis and treatment outcomes. By enabling accurate analysis of medical data and generating concise and clear impressions, medical professionals will be empowered to make well-informed decisions, ultimately enhancing patient care and improving overall medical practices.

Scope

Limited to Radiology Reports: The project focuses specifically on generating impressions from radiology reports. It does not encompass other types of medical reports or documents, such as pathology reports or clinical notes.

Specific Modalities: The project aims to generate impressions for the modality of MRI, considering its sufficient and applicable data size. While additional modalities are utilized in medical imaging, this scope does not encompass other imaging techniques or specialties.

Single Institution Data: The dataset used for training is obtained solely from the University of Chicago Medicine. This limitation restricts the generalizability of the solution to other healthcare institutions and may introduce biases associated with a single source of data.

Background

The client, Inference Analytics Inc., is a healthcare AI startup based in Chicago, specializing in deep learning solutions for medical image and language processing. Their mission is to transform the healthcare industry by leveraging LLMs and cutting-edge AI. Their platform provides real-time clinical decision support to radiologists, improving accuracy, productivity, and clinical quality while reducing physician burnout. For the latest project, Inference Analytics Inc. focuses on developing LLMs and Generative Pretrained Models to enable real-time impression generation based on clinical information and findings, covering modalities like X-ray, CT, MRI, and Ultrasound. This project aims to enhance healthcare workflows and improve patient care.

Literature Review

Radiology reports are critical in medical diagnosis and treatment, and their quality can significantly impact patient outcomes. Clarity, brevity, and clinical correlation are important in radiological reports. Referring physicians depend on radiological reports for timely and accurate diagnosis and treatment of their patients (Lafortune, Breton, and Baudouin, 1988, 4). AI Natural Language Processing (NLP)[5] applications have the potential to enhance the efficiency and satisfaction of radiologists and providers. They can automate tasks such as navigating complex Electronic Health Records (EHR)[6] systems and streamlining documentation with human review. These applications also reduce computer system interaction time, allowing more focus on patient care and ultimately improving provider workflow and job satisfaction (Korngiebel and Mooney, 2021).

Methodology: BERT, LSTMs, and Generative Models

The utilization of NLP and AI methodologies, such as LLMs[7] and Generative Pretrained Models[8], has shown promising potential in improving the quality and efficiency of radiology

---

[5] Natural Language Processing (NLP) is a field of AI that focuses on the interaction between computers and human language, enabling machines to understand, interpret, and generate human language in a meaningful way.

[6] Electronic Health Records (EHR) are digital versions of patients' medical records that provide comprehensive and accessible information about their health history, diagnoses, treatments, and other relevant healthcare data.

[7] Large language models (LLM) refer to advanced artificial intelligence (AI) models that are trained on vast amounts of textual data and utilize deep learning techniques to acquire a deep understanding of language patterns, grammar, and semantics.

[8] Generative Pretrained Models (GPT) are LLMs that have been trained on vast amounts of text data and can generate human-like text based on the patterns and knowledge acquired during training.

reports. Studies have used BERT[9] and LSTMs[10] to analyze radiology reports and predict customized impressions. Gundogdu et al. (2021, 8) developed a model that used BERT and LSTMs to predict customized impressions from radiology reports. The results showed that the proposed model outperformed other models and improved the accuracy of impression prediction. Generative Pretrained Models, such as GPT[11] and T-5[12], have been used to improve work satisfaction for providers and reduce time spent in interacting with computer systems. Cascella et al. (2023, 1) evaluated the feasibility of using ChatGPT[13] in healthcare by analyzing multiple clinical and research scenarios. The results demonstrated that ChatGPT could be used effectively in various healthcare settings, including patient counseling and health education. Moreover, Deid-GPT, a zero-shot[14] medical text de-identification method, was proposed by Liu et al. (2023, 1). The study used GPT-4 to remove sensitive information from medical text without the need for prior training data. The results showed that Deid-GPT could effectively de-identify medical text without sacrificing the quality of the text.

Evaluation: Non-Interactive Metrics vs. Human-AI Interaction

In the evaluation of language models, there are two broad categories of techniques: non-interactive and interactive. Non-interactive techniques measure the similarity between machine-generated and human-written text, while interactive techniques involve human-AI language-based interaction evaluation (HALIE). One typical example of non-interactive evaluation is

[9] BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model that uses transformer architecture and bidirectional training to understand the context of words and generate contextualized word representations.

[10] LSTMs (Long Short-Term Memory) are a type of recurrent neural network (RNN) that can process sequential data, such as text, by retaining long-term dependencies through a memory cell and gating mechanisms.

[11] GPT (Generative Pre-trained Transformer) is a series of LLMs developed by OpenAI that use transformer architecture and pre-training on massive amounts of text data to generate human-like text.

[12] T-5 (Text-to-Text Transfer Transformer) is a transformer-based model that has been pre-trained on a range of language tasks and can be fine-tuned for various natural language processing tasks.

[13] Chat GPT stands for Chat Generative Pre-Trained Transformer and was developed by an AI research company, Open AI.

[14] Zero-shot learning refers to an approach where a model is able to recognize and classify objects or concepts it has never been explicitly trained on.

measuring the similarity using metrics such as ROUGE and BLEU.[15] Gundogdu et al. (2021, 7) used ROUGE and BLEU metrics to evaluate the performance of their customized impression prediction model based on BERT and LSTMs for radiology reports. In contrast, interactive evaluation methods involve human evaluation of the language generated by the machine. Lee et al. (2022, 1) proposed a framework for evaluating HALIE, a special evaluation metric that involves measuring how well the machine interacts with humans in natural language. The framework takes into account both the quality of the machine-generated response and the user's satisfaction with the interaction.

Model Improvement: Enhancing Clinical Text Analysis with AI and Data Augmentation

Recent research has focused on improving the modeling of clinical text using transformer models, such as BERT, GPT, and T-5. These models have proven effective in handling large datasets and improving prediction accuracy in various healthcare applications, including disease diagnosis and treatment recommendation (Zhang et al., 2022, 7).

Another approach to improving the accuracy of machine learning models in healthcare is through data augmentation. Deep generative models and federated learning have been used to generate synthetic data and improve the robustness of models (Zhang et al., 2022, 4). In addition, text generation using the ChatGPT API has been proposed as a tool for data augmentation in supervised machine learning tasks in healthcare NLP (Edjinedja et al., 2023, 1).

Deployment Process: Compliance, Trust, and Access

In the healthcare industry, analytics has been used for solving various problems such as disease prediction and diagnosis. However, the implementation of AI models such as the

[15] ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy) are commonly used evaluation metrics in NLP to assess the quality of generated text or translations by comparing them to reference or ground truth text.

Generative Pre-trained Transformer 3 (GPT-3) in the US healthcare system requires careful consideration of various operational and implementation factors. Considerations for implementing GPT-3 in the US healthcare system include processing needs, information systems infrastructure, operating costs, model biases, and evaluation metrics. Furthermore, three key operational factors driving GPT-3 adoption are maintaining Health Insurance Portability and Accountability Act compliance, fostering trust with healthcare providers, and expanding access to GPT-3 tools (Sezgin et al., 2022, 1). Meanwhile, GPT-3 application in the United States would likely require Food and Drug Administration (FDA)[16] approval, and a comprehensive evaluation should encompass a diverse patient population (Korngiebel and Mooney, 2021, 1).

Abstractive Summarization: Sequence-to-Sequence and GANs

Abstractive summarization[17] is a challenging task that has been tackled using various methods. One common approach is to use a sequence-to-sequence model, which applies an encoder to map the source text to vectors and then uses a decoder to restore the vector to the summarization. This method has been implemented using models such as T5, BART, and PropheNet (Teng, 2023, 2).

Another approach is the semantic-enhanced generative adversarial network (GAN)-based method for abstractive text summarization (SGAN4AbSum), which uses an adversarial training strategy to train the generator and discriminator to simultaneously handle summary generation and distinguish the generated summary with the ground-truth one (Vo, 2023, 1). These methods have shown promise in improving the quality of abstractive summarization and have the potential for further development in the future.

[16] Food and Drug Administration (FDA) is a regulatory agency responsible for ensuring the safety, efficacy, and quality of food, drugs, medical devices, and other products in the United States.

[17] Abstractive summarization is a technique that generates concise and coherent summaries by understanding the content of the source text and generating new phrases that capture the key information.

Data

Data Sources

The data for this project was obtained from the University of Chicago Medicine, consisting of 789,280 de-identified radiology reports. An overview is shown in Figure 1. The dataset includes columns for clinical information, techniques used in the exam, findings (descriptive observations from the Radiologist), and impressions (the interpretation and conclusion about the exam), which are all text data. The objective is to generate accurate wording for the impression section by using the text input from clinical information, techniques used, and findings.

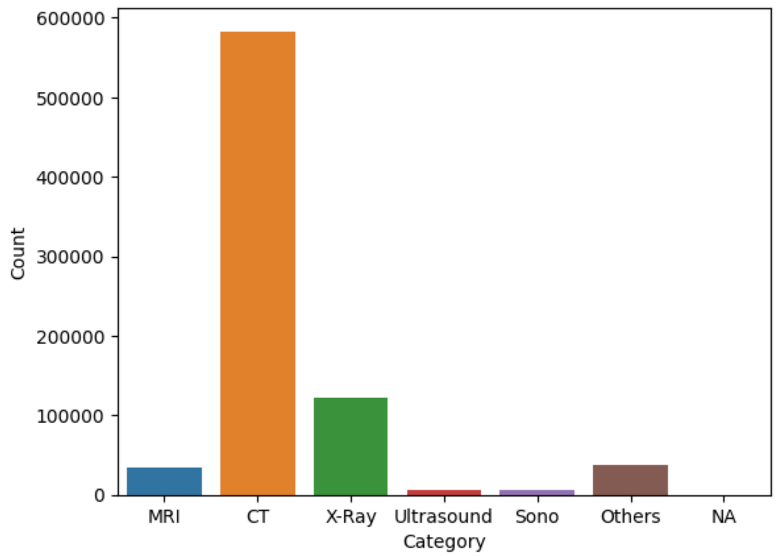
Figure 1. *Data Overview*

	Unnamed: 0	clinical_information	technique	findings	comparison	impression	report_id
0	0	34 year old female with history of sickle cell...	2 views of the right shoulder at 6:41 on 7/12/12	The right total shoulder arthroplasty componen...	XR shoulder 7/11/12	Right total shoulder arthroplasty components i...	RAD_0
1	1	34 year old female with history of sickle cell...	One portable view of the right shoulder at 17:...	The right total shoulder arthroplasty componen...	XR shoulder 2/1/12	Right total shoulder arthroplasty components i...	RAD_1
2	2	84-year-old female with low back pain	Four views of the lumbar spine	Posterior stabilization rods with transpedicul...	2/13/06	Posterior fixation of L4 and L5, appearing sim...	RAD_2
3	3	NaN	Informed consent was obtained. The patient was...	The colon is adequately cleansed and distended...	NaN	No significant colonic polyps or masses identi...	RAD_3
4	4	Preoperative planning for brain tumor. History...	MRI BRAIN STEALTH W/WO CONTRAST. A total of 17...	There is a heterogeneous left supratentorial a...	Brain MRI dated 11/17/14.	Presurgical planning MRI shows a complex mass ...	RAD_4

Descriptive Analysis

Initially, Exploratory Data Analysis (EDA) was performed, including simplifying the technique column by classifying techniques into categories such as MRI, X-ray, CT, Ultrasound, and Sono. The distribution of the modalities can be seen in Figure 2. Among all data records, the majority of modalities used were CT (582,951), followed by X-ray (122,431), Others (36,993), MRI (33,894), Sono (6,577), and Ultrasound (6,433).

Figure 2. *Distribution of Modalities*



In light of the substantial dataset, our research will exclusively focus on constructing specialized LLMs tailored for the MRI modality in subsequent phases.

Methodology

Feature Engineering

In general, the dataset exhibits a well-organized structure and lacks missing values, thereby allowing for a reduced emphasis on data cleaning and feature engineering processes.

After classifying the reports based on modality (MRI, X-ray, CT, Ultrasound, Sono…etc.) and getting a brief overview of the modality distribution, several LLMs were experimented on MRI data. This data has been selected for its ample size, which is sufficient for effective model training while also taking into account the constraints of training duration and computational resources.

Modeling Frameworks

Model Selection

Several LLMs are experimented as our base models including Alpaca-7b, Medalpaca-7b, Llama 2-7b, Llama 2-13b, and Mistral-7b. This choice was influenced by the model's open-source availability, parameter size, use case, and proven performance.

- **Alpaca-7b:** Alpaca-7b is a fine-tuned version of the seven billion-parameter Llama language model from Meta. The team took Llama-7b and finetuned it on 52k instruction-following[18] demonstrations to create a ChatGPT-like chatbot.
- **Medalpaca-7b:** MedAlpaca expands upon both Stanford Alpaca and AlpacaLoRA[19] to offer an advanced suite of LLMs specifically fine-tuned for medical question-answering and dialogue applications.
- **Llama 2 (Llama 2-7b and Llama 2-13b):** Llama 2 is a powerful open-source language model developed by Meta AI, which was released in July 2023. It is based on the Transformer[20] architecture and offers various versions tailored for specific tasks, ranging from text generation to chatbots. Its accessibility and efficiency make it a valuable tool for researchers and developers. Llama 2 is a collection of pretrained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. This paper utilizes Llama 2-7b and Llama 2-13b as pre-trained models for fine-tuning.

---

[18] Instruction-following models are those like ChatGPT that follow instructions from users to generate output.

[19] Alpaca-LoRA is a smaller version of Stanford Alpaca that consumes less power and is able to run on low-end devices like Raspberry Pie. Alpaca-LoRA uses Low-Rank Adaptation(LoRA) to accelerate the training of large models while consuming less memory.

[20] A Transformer is a type of deep learning architecture that uses an attention mechanism to process text sequences. Unlike traditional models based on recurrent neural networks, Transformers do not rely on sequential connections and are able to capture long-term relationships in a text.

- **Mistral-7b:** Mistral-7b-v0.1 is a small (7.3 billion parameters), yet powerful model adaptable to many use-cases. It is a transformer model that uses grouped-query attention and sliding-window attention to achieve faster inference (low latency) and handle longer sequences. Group query attention is an architecture that combines multi-query and multi-head attention to achieve output quality close to multi-head attention and comparable speed to multi-query attention. Sliding-window attention uses the stacked layers of a transformer to attend in the past beyond the window size to increase context length. Mistral-7b has an 8,000-token context length, demonstrates low latency and high throughput, and has strong performance when compared to larger model alternatives, providing low memory requirements at a 7b model size.

Model Tuning

- **Instruction Tuning:** instruction tuning is a fine-tuning technique applied to our base LLM to enhance its ability to follow and execute specific instructions. This study concentrates on the medical domain, specifically training a machine learning model to autonomously generate diagnostic impressions based on a comprehensive analysis of clinical data and diagnostic findings Therefore, the instruction given to the model is: “Generate impression based on medical inputs”. This directive precedes the input, which comprises two critical components: Clinical Information and Findings. These inputs encapsulate the patient’s background, symptoms, and details from medical imaging reports.
- **LoRA Adaptor Integration:** the LoRA (Low-Rank Adaptation) adaptor is an additional module implemented on top of our base LLM. LoRA enables efficient parameterization by introducing trainable low-rank matrices that capture important adaptations without

altering the pre-trained weights directly. The integration of the LoRA adaptor with our chosen base model involved a strategic insertion of low-rank matrices into specific layers of the model. The selection of these layers was guided by the objective of enhancing the model's ability to process and interpret complex medical data. This approach allows the model to retain its general language understanding abilities while adapting to the specific task of medical impression generation with minimal increase in the number of trainable parameters.

Model Training

Prior to training, the dataset is divided into two subsets: 80% (27,115 records) for training and 20% (6,779 records) for testing.

The training procedure for the model was designed to enable it to precisely associate the clinical information and observations found in MRI data with their relevant medical interpretations. This was achieved through a supervised learning approach, where the model was presented with paired examples of clinical information and the correct medical impression.

The base LLM, augmented with the LoRA adaptor, undergoes a training process where it learns to map the clinical information and findings to the correct impression. Throughout this process, the model's parameters, including those in the LoRA adaptor, are optimized to minimize the discrepancy between the generated impressions and the ground truth provided in the training data.

Model Optimization

- **Hyperparameter Optimization**: considering our computational bandwidth and the unique characteristics of our dataset, the hyperparameters for the models are meticulously tuned hyperparameters. This included calibrating the learning rate, batch size, and the number of epochs. The objective was to achieve an equilibrium between computational expediency and optimal model performance.

- **Regularization**: to mitigate the risk of overfitting, particularly given the limited parameter updates inherent in the LoRA approach, the model's architecture was augmented with dropout layers. Furthermore, the implementation of weight decay was explored as an additional strategy to enhance generalization capabilities.

Model Evaluation

To assess the performance of the models in generating precise impressions from patients' medical findings and background information, ROUGE scores have been selected as the primary evaluation metrics. This choice is based on their widely recognized utility in evaluating the quality of generated text summaries. Four different types of ROUGE scores, specifically ROUGE-1[21], ROUGE-2[22], ROUGE-L[23], and ROUGE-LSum[24], have been employed for model evaluation to provide a comprehensive assessment.

The combination of ROUGE scores is utilized to ensure a multifaceted evaluation that not only considers the fluency and coherence of the generated text but also the model's ability to capture the specific nuances and terminology within the radiology domain. This comprehensive evaluation is pivotal in determining how effectively the model can assist physicians in optimizing their documentation processes and reducing the time required to produce

[21] ROUGE-1 evaluates the overlap of unigrams (single words) between the generated impressions and reference documents.

[22] ROUGE-2 extends this assessment to bigrams (pairs of adjacent words), shedding light on the fluency and coherence of the generated text.

[23] ROUGE-L quantifies the longest common subsequence, which is crucial in capturing the essence and contextual accuracy of the reports.

[24] ROUGE-Lsum is related to the ROUGE-L metric but applies a slightly different calculation method. It applies the ROUGE-L calculation method at the sentence level and then aggregates all the results for the final score.

summarization for reports, ultimately mitigating the risk of burnout associated with repetitive documentation tasks.

Model Implementation

The implementation involves deploying a fine-tuned Mistral-7b model via Gradio to create a user-friendly web interface. Gradio is recognized for its straightforward approach in turning complex machine learning models into accessible web applications. This approach is particularly beneficial for users who may not have extensive technical expertise but require the advanced capabilities of the machine learning model.

Through this interface, users can input clinical background and findings into designated text boxes. Upon submission, the fine-tuned Mistral-7b model processes the input data and generates a medical impression based on the given information.

The generated medical impression is promptly displayed on the same Gradio interface, ensuring a quick and precise response. Overall, this interface presents an efficient, interactive method for generating medical impressions from clinical inputs, enhancing the user-friendliness of LLM technology.

Findings

ROUGE Score

The research aimed to enhance the prediction of customized impressions from radiology reports. All models in this paper are trained on data consisting of MRI modality, with a dataset size of 33,894, and a training-to-test split ratio of 80% to 20%. The performance of various models was evaluated using ROUGE metrics (ROUGE-1, ROUGE-2, and ROUGE-L), and these were compared against a baseline. This baseline, comprising a combination of BERT and

LSTMs, was described in our client's 2021 paper titled 'Customized Impression Prediction from Radiology Reports Using BERT and LSTMs'.

Table 1. *ROUGE Metrics Comparison*

System	ROUGE-1	ROUGE-2	ROUGE-L
PG ⊕ BERT (Baseline Result)	47.17	32.89	45.91
Alpaca-7b	45.63	27.34	38.08
Medalpaca-7b	50.47	31.02	42.10
Llama 2-7b	46.75	28.02	39.07
Llama 2-13b	44.98	26.54	38.53
Mistral-7b	56.42	40.11	48.66

The ROUGE metrics comparison is summarized in Table 1. The most significant improvement was observed with Mistral-7b, which outperformed all other models and baseline results in all ROUGE metrics, indicating its superior capability in generating text closely aligned with reference texts.

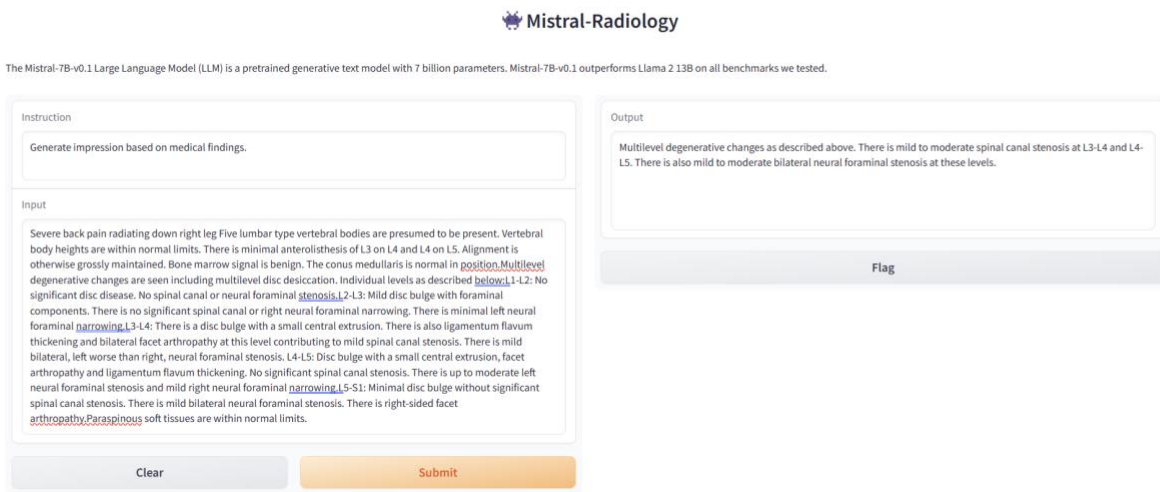
Other models like Medalpaca-7b showed improved performance compared to the baseline but did not reach the levels achieved by Mistral-7b. Meanwhile, models such as Alpaca-7b, Llama 2-7b, and Llama 2-13b scored lower than both the baseline and Mistral-7b across all ROUGE metrics.

Overall, these findings indicate that the Mistral-7b model has a superior capability in text generation tasks within the context of MRI modality data, as measured by ROUGE metrics. This underscores its potential usefulness in applications like generating impressions from radiology reports.

Demo

Below is an example demonstrating our user-friendly interface, developed with Gradio, utilizing the fine-tuned Mistral-7b model.

Figure 3. *Mistral-7b Demo*



This demo highlights the functionality and effectiveness of our interface, emphasizing its capacity to generate medical impressions that align closely with original outputs. In the demo, the original output presented was: 'Multilevel degenerative changes in the lumbar spine, relatively worse at the L3-L4 and L4-L5 levels. There is mild spinal canal stenosis at L3-L4. There is moderate left L4-L5 neural foraminal stenosis. Additional details as above.' The close alignment between the generated and original outputs underscores the model's accuracy and its applicability in clinical settings, as well as the practicality of the interface.

Moreover, the demonstration reveals how the interface's design prioritizes clarity and user-friendliness. It facilitates the straightforward input of clinical data, leading to the instant generation of medical impressions. This prompt and intuitive nature of the interface will streamline the process for physicians, reducing the time required for summarizing reports.

Discussion

The research findings from the effort to advance medical language processing in radiology through LLMs provide valuable insights, but it's important to consider the associated limitations.

Focusing on the objective of generalizability, the project has encountered challenges. While there has been progress in deploying various fine-tuned models, further efforts are needed to ensure these models can be effectively generalized across different clinical settings. A key future goal is to expand the model to include test data from different medical centers, such as Indiana University, enhancing its applicability in varied medical environments.

The evaluation methodology in this research was based on quantitative assessments using ROUGE scores. These assessments have offered substantial insights, but limitations exist, particularly in ROUGE's precise word matching, which may not always capture the nuances of medical language. Additionally, biases in measures based on labeled human judgments have been a concern. To address these issues, the integration of qualitative metrics is planned to provide a more well-rounded evaluation. Expert input from Subject Matter Experts (SMEs) in radiology is also being considered to offer deeper insights and balance the limitations of quantitative analysis.

To sum up, the research has contributed significantly to the field of medical language processing in radiology, yet the acknowledgment of its limitations is imperative. There is a strong commitment to further refine and develop these research findings, with the aim of making substantial advancements in the field.

Conclusion

In conclusion, a robust AI-based solution has been delivered, tailored to generate impressions from radiology reports. This solution is supported by a well-documented and transparent architecture, methodology, and implementation, ensuring its effective employment within healthcare environments.

Five LLMs were examined as base models, with each being fine-tuned using the instruction tuning method. The effectiveness of these fine-tuned models was assessed using ROUGE scores, leading to the discovery that the Mistral-7b model surpassed its counterparts in performance.

The final deliverable is a user-friendly web application, deploying the high-performing Mistral-7b model, created using Gradio. Users can input clinical information and findings into the application and swiftly receive accurate medical impressions. This intuitive and accessible interface significantly enhances the usability and effectiveness of the AI solution, making it a practical tool in medical language processing for radiology.

Quantifying the impact of this work, a significant improvement in operational efficiency within the radiology domain is anticipated. This solution has the potential to reduce the time and effort required for impression generation, enabling healthcare providers to allocate resources more effectively.

The next steps involve the integration of the solution into existing healthcare infrastructures. Additionally, exploring potential extensions of this work, such as adapting the model for other medical specialties or further enhancing its performance, is recommended.

References

American Medical Association. (n.d.). *Improving Care Delivery: Streamlining Clinical Prior Authorization [Fact sheet]*. Retrieved May 10, 2023, from https://www.ama-assn.org/system/files/2019-06/improving-care-delivery-streamlining-clinical-prior-authorization.pdf

Cascella, M., Montomoli, J., Bellini, V., & Bignami, E. (2023). *Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios.* Journal of Medical Systems, 47(1), 1-5.

Edjinedja, L., Omar, E., Barakat, O., Desmettre, T., & Marx, T. (2023). LEVERAGE CHATGPT AS TOOL FOR DATA AUGMENTATION FOR SUPERVISED MACHINE LEARNING TASKS IN HEALTHCARE NATURAL LANGUAGE PROCESSING.

Frost & Sullivan. (2021). *AI-powered medical report generation solutions: Transforming the healthcare industry.* Retrieved May 10, 2023, from https://ww2.frost.com/files/5616/6781/6437/FS_AI-powered_Medical_Report_Generation_Solutions.pdf

Gundogdu, B., Pamuksuz, U., Chung, J. H., Telleria, J. M., Liu, P., Khan, F., & Chang, P. J. (2021). *Customized impression prediction from radiology reports using BERT and LSTMs.* IEEE Transactions on Artificial Intelligence.

Korngiebel, D.M., Mooney, S.D. (2021). *Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery.* npj Digit. Med. 4(1), 93. https://doi.org/10.1038/s41746-021-00464-x

Lafortune, M., Breton, G., & Baudouin, J. L. (1988). *The radiological report: what is useful for the referring physician?* Canadian Association of Radiologists journal. Journal l'Association canadienne des radiologistes, 39(2), 140.

Lee, M., Srivastava, M., Hardy, A., Thickstun, J., Durmus, E., Paranjape, A., ... & Liang, P. (2022). *Evaluating human-language model interaction.* arXiv preprint arXiv:2212.09746.

Liu, Z., Yu, X., Zhang, L., Wu, Z., Cao, C., Dai, H., ... & Li, X. (2023). *Deid-gpt: Zero-shot medical text de-identification by gpt-4.* arXiv preprint arXiv:2303.11032.

Sezgin E, Sirrianni J, Linwood S. (2022). *Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model*. JMIR Med Inform 10(2), e32875. https://medinform.jmir.org/2022/2/e32875

Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. *"Alpaca: A strong, replicable instruction-following model."* Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html 3, no. 6 (2023): 7.

Teng, Z. (2023). *Abstractive summarization of COVID-19 with transfer text-to-text transformer*. ACE Vol. 2: 232-238. DOI: 10.54254/2755-2721/2/20220520.

Tractica. (2017). *Artificial intelligence for healthcare applications.* Retrieved May 10, 2023, fromhttps://www.tractica.com/wp-content/uploads/2017/02/Artificial-Intelligence-for-Healthcare-Applications.pdf

Vo, T. (2023). *A novel semantic-enhanced generative adversarial network for abstractive text summarization.* Soft Comput. https://doi.org/10.1007/s00500-023-07890-x

Zhang, A., Xing, L., Zou, J. et al. (2022). *Shifting machine learning for healthcare from development to deployment and from models to data.* Nat. Biomed. Eng 6, 1330–1345. https://doi.org/10.1038/s41551-022-00898-y

Appendix A: ROUGE Scores for Different Models

Appendix A presents a series of tables delineating the performance metrics of various models as evaluated by ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores. These metrics, including precision, recall, and F-measure, are reported across three distinct thresholds: low, mid, and high. Each table corresponds to a different model, providing a granular view of their performance in summarization tasks.

Table A1. *Baseline ROUGE Metrics*

System	ROUGE-1	ROUGE-2	ROUGE-L
PG ⊕ BERT (Baseline Result)	47.17	32.89	45.91

Table A2. *ROUGE Metrics of Alpaca-7b*

Type	Low			Mid			High		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
ROUGE-1	0.1148	0.3644	0.1543	0.1358	0.4563	0.1744	0.1621	0.4885	0.1961
ROUGE-2	0.0478	0.3843	0.0705	0.0603	0.2734	0.0877	0.0779	0.3229	0.1087
ROUGE-L	0.0925	0.3083	0.1278	0.1076	0.3808	0.1441	0.1276	0.4371	0.1673
ROUGE-L SUM	0.0916	0.3057	0.1281	0.1076	0.3837	0.1447	0.1284	0.4378	0.1675

Table A3. *ROUGE Metrics of Medalpaca-7b*

Type	Low			Mid			High		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
ROUGE-1	0.3543	0.4426	0.3628	0.4198	0.5047	0.4154	0.4829	0.5648	0.4611
ROUGE-2	0.1995	0.2450	0.2008	0.2491	0.3102	0.2530	0.3000	0.3781	0.3046
ROUGE-L	0.2814	0.3595	0.2929	0.3343	0.4210	0.3391	0.3845	0.4850	0.3846
ROUGE-L SUM	0.2868	0.3602	0.2942	0.3332	0.4217	0.3377	0.3829	0.4842	0.3859

Table A4. *ROUGE Metrics of Llama 2-7b*

Type	Low			Mid			High		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
ROUGE-1	0.4716	0.4319	0.4239	0.5084	0.4675	0.4537	0.5455	0.5071	0.4858
ROUGE-2	0.2615	0.2442	0.2358	0.2948	0.2802	0.2665	0.3313	0.3197	0.3022
ROUGE-L	0.3823	0.3585	0.3455	0.4132	0.3907	0.3734	0.4448	0.4270	0.4026
ROUGE-L SUM	0.3803	0.3563	0.3437	0.4125	0.3916	0.3739	0.4489	0.4275	0.4051

Table A5. *ROUGE Metrics of Llama 2-13b*

Type	Low			Mid			High		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
ROUGE-1	0.4239	0.3890	0.3816	0.4913	0.4498	0.4345	0.5589	0.5210	0.4961
ROUGE-2	0.2252	0.2095	0.2031	0.2832	0.2654	0.2558	0.3576	0.3372	0.3228
ROUGE-L	0.3460	0.3238	0.3098	0.4064	0.3853	0.3659	0.4697	0.4495	0.4259
ROUGE-L SUM	0.3412	0.3224	0.3098	0.4068	0.3847	0.3656	0.4711	0.4556	0.4286

Table A6. *ROUGE Metrics of Mistral 7b*

Type	Low			Mid			High		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
ROUGE-1	0.6239	0.4949	0.5238	0.6616	0.5642	0.5593	0.6980	0.5746	0.5950
ROUGE-2	0.4192	0.3404	0.3569	0.4626	0.4011	0.3986	0.5098	0.4308	0.4467
ROUGE-L	0.5301	0.4227	0.4453	0.5695	0.4866	0.4841	0.6101	0.5109	0.5315
ROUGE-L SUM	0.5303	0.4245	0.4461	0.5700	0.4622	0.4837	0.6092	0.5031	0.5249

Appendix B: Additional Evaluative Metrics: Factual Correctness

Appendix B delves into an additional evaluative dimension beyond the ROUGE scores: factual correctness. This criterion assesses the models' capacity to generate outputs that are not only linguistically coherent but also factually accurate. The significance of factual correctness is exemplified in the analysis provided in Figure A1., where two model-generated summaries are contrasted with a human-generated summary. As depicted, Summary A, despite its higher ROUGE-L score, contains a factual inconsistency, whereas Summary B, with a lower ROUGE-L score, maintains factual accuracy.

Figure A1. *A (truncated) radiology report and summaries with their ROUGE-L scores (Zhang et al., 2022, 1)*

Background: radiographic examination of the chest. clinical history: 80 years of age, male … **Findings:** frontal radiograph of the chest demonstrates repositioning of the right atrial lead possibly into the ivc. ... a right apical pneumothorax can be seen from the image. moderate right and small left pleural effusions continue. no pulmonary edema is observed. heart size is upper limits of normal.
Human Summary: pneumothorax is seen. bilateral pleural effusions continue.
Summary A (ROUGE-L = 0.77): no pneumothorax is observed. bilateral pleural effusions continue.
Summary B (ROUGE-L = 0.44): pneumothorax is observed on radiograph. bilateral pleural effusions continue to be seen.

Considering this situation, this appendix reports on the factual correctness of the final Mistral-7b model's outputs, categorizing them into negative, neutral, and positive factual correspondences with the ground truth. The following tables quantify these aspects, providing a nuanced view of the model's performance in terms of factual integrity.

Table A7. *Factual Correctness of Mistral-7b*

Percentage of Negative Items	Percentage of Neural Items	Percentage of Positive Items
13.4%	57.0%	27.9%