Medical Language Processing – GPT/LLM Based Impression Prediction from Radiology Reports

By

Junru (Aeris) 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 Analytics

Division of Physical Sciences

May 2023

Abstract

This research paper presents a novel large language model specifically fine-tuned on radiology reports. The model has been designed to generate accurate impressions by leveraging patients' medical findings and background information. The primary objective of this model is to assist physicians in optimizing their documentation process, ultimately reducing the time required for generating comprehensive reports. 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

Our research paper introduces a large language model fine-tuned specifically for radiology reports. The model aims to generate precise impressions by utilizing patients' medical findings and background information. The primary objective of this research is to assist physicians in optimizing their documentation process, leading to a reduction in the time required for generating comprehensive reports. By adopting this technology, physicians can potentially alleviate the risk of burnout associated with extensive and repetitive documentation tasks.

The research project begins by addressing the problem at hand and highlights the significance of efficient documentation in the medical field. The study then presents the key findings and conclusions derived from the implementation of the developed large language model.

The data and methodology employed in this research project involve the fine-tuning of the large language model on a dataset of radiology reports. The model is trained to understand the specific nuances and terminology within the domain of radiology, enabling it to generate highly relevant and contextually accurate impressions.

Further research can be conducted to explore additional applications and enhancements of the model to enhance its effectiveness in the medical field.

Table of Contents

List of Figures

Introduction

Problem Statement

Radiology reports[1] are 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 a time-consuming and challenging task for radiologists. Current practices rely heavily on manual report generation, which is not only costly but also prone to human error.

The challenge is to develop AI[2] solutions that can efficiently generate radiology impressions[3] from unstructured narratives in healthcare records. The solution should be able to interpret and analyze the clinical history, techniques used, and findings[4] from the radiology reports to generate accurate and concise impressions. The significance of the problem is the costs and potential errors associated with manual report generation. 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 client's desired solution is an AI-powered tool that can generate radiology impressions from unstructured narratives in healthcare records quickly and accurately. 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.

benefits of this solution are multiple. Firstly, it can lead to significant cost savings for healthcare providers by reducing the 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 AI-powered medical report generation solutions has the potential to enhance diagnostic accuracy by up to 30%, thereby leading to improved patient outcomes (Frost & Sullivan, 2021). 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).

In summary, the development of AI-powered solutions 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 analysis goals aim to develop a medical language processing model that can generate concise and clear impressions from complex medical findings and reports. The main objective is to improve the accuracy and efficiency of medical diagnosis and treatment by creating a tool that can easily interpret large amounts of medical data and provide clear and concise impressions to medical professionals.

To accomplish the analysis goals, the project will design and develop 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 this language, the model will be able to generate concise and clear impressions. Moreover, the project will design and develop a user-friendly interface that integrates the medical language processing model and make the developed model accessible through an API. This will allow seamless integration of the model into various applications developed 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 modalities like X-Ray, CT, MRI, and Ultrasound. While these modalities are widely used in medical imaging, the scope does not extend to other imaging techniques or specialties.

Single Institution Data: The dataset used for training and development 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 Large Language Models and cutting-edge AI. They have curated a dataset of 1 million de-identified radiology reports, which they use to train their AI-based solution. 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 Large Language Models and Generative Pretrained Models to enable real-time impression generation based on clinical history and findings, covering modalities like X-Ray, CT, MRI, and Ultrasound. The project aims to enhance healthcare workflows and improve patient care.

Literature Review

Radiological 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). 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 (HER)6 systems and streamlining documentation with human review. These applications also reduce computer system interaction time, allowing more focus on patient care, ultimately improving provider workflow and job satisfaction (Korngiebel and Mooney, 2021).

[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.

For Methodology

The utilization of NLP and AI methodologies, such as Large Language Models[7] and Generative Pretrained Models[8], has shown promise in improving the quality and efficiency of radiology reports. Studies have used Large Language Models, such as BERT[9] and LSTMs[10], to analyze radiology reports and predict customized impressions. Gundogdu et al. (2021) 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 interacting with computer systems. Cascella et al. (2023) evaluated the feasibility of using ChatGPT 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[13] medical text de-identification method, was proposed by Liu et al. (2023). 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.

[7] Large language models 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 large language models 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.

[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 large language models 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] 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.

For Model Evaluation

In the evaluation of language models, there are two broad categories of techniques: non-interactive and interactive. Non-interactive techniques measure 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 measuring similarity using metrics such as ROUGE and BLEU.[14] Gundogdu et al. (2021) 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) proposed a framework for evaluating HALIE, a special evaluation metric which 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.

For Model Improvement

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

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). In addition,

[14] 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.

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).

For Deployment Process

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 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). Meanwhile, GPT-3 application in the United States would likely require Food and Drug Administration (FDA)[15] approval, and a comprehensive evaluation should encompass a diverse patient population (Korngiebel and Mooney, 2021).

For Abstractive Summarization

Abstractive summarization[16] 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).

[15] 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.

[16] 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.

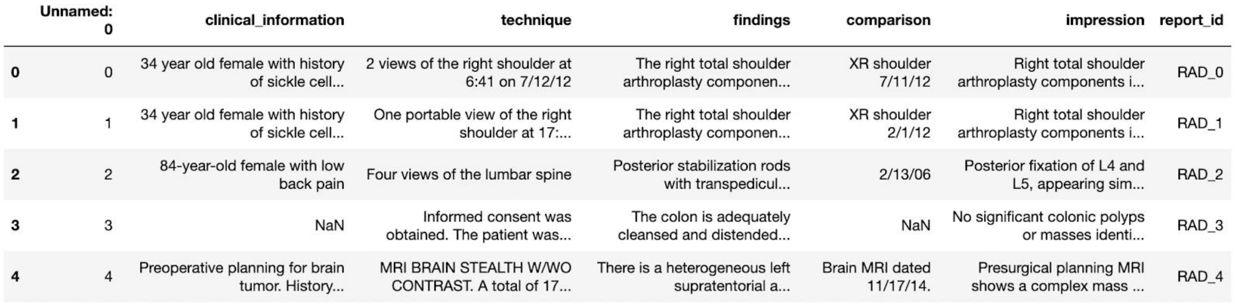
Another approach is the semantic-enhanced generative adversarial network (GAN)-based method for abstractive text summarization (SGAN4AbSum), which uses an adversarial training strategy to train the generator and discriminator to simultaneously handle summary generation and distinguishing the generated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

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. Our objective is to generate accurate wording for the impression section by using the clinical history, techniques used, and findings text input.

Figure 1. *Data Overview*

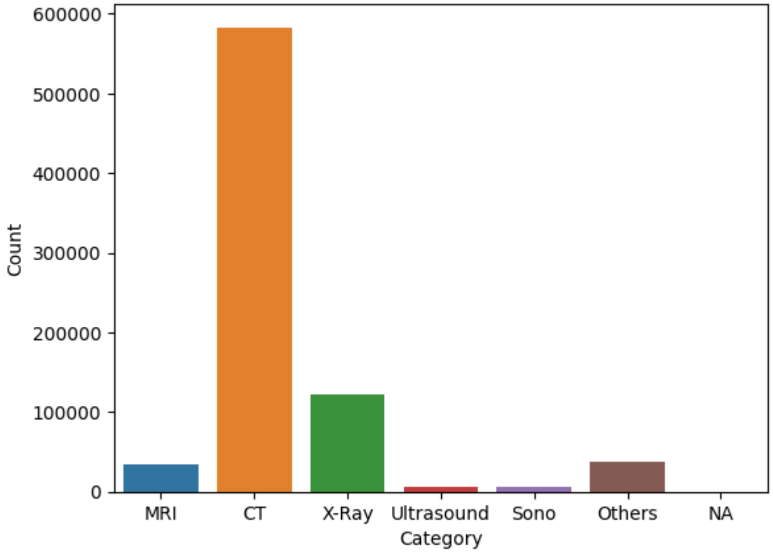


	Unnamed: 0	clinical_information	technique	findings	comparison	impression	report_id
0	0	34 year old female with history of sickle cell...	2 views of the right shoulder at 6:41 on 7/12/12	The right total shoulder arthroplasty componen...	XR shoulder 7/11/12	Right total shoulder arthroplasty components 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

We first performed some EDA, such as simplifying the technique column by classifying the techniques into MRI, X-ray, CT, Ultrasound, and Sono. The distribution of the modalities can be seen in Fig 2. Among all data records, the majority of modalities used were CT (582951), followed by X-ray (122431), Others (36993), MRI (33894), Sono (6577), and Ultrasound (6433).

Figure 2. *Distribution of Modalities*



Methodology

Feature Engineering

The data is well-structured and there is no missing value, which means that we can spend less time on data cleaning and feature engineering.

After classifying the reports based on modality (MRI, X-ray, CT, ultrasound, sonograph) and getting a brief overview of the modality distribution, we plan to preprocess the data by tokenizing the text and encoding it in a format that can be used as input to the language models.

We will then build and tune a large language model/generation technique for a certain type of modality and evaluate its performance using either a train/test set or cross-validation. Our final output should be an API consisting of all these models that target different modalities.

Modeling Frameworks

Models and Model Implementation

The models that will be used are Large Language Models and Generative Pretrained Models, such as GPT and T-5, which have been shown to be effective for natural language processing tasks, including text generation and summarization. To be more specific, Large Language Models work by utilizing deep learning architectures, to generate coherent and contextually relevant text based on vast amounts of training data, while Generative Pretrained Models leverage pretraining techniques to generate diverse and high-quality content across various domains.

One main approach we plan to use is Abstractive Summarization, which involves generating a summary of a longer text by paraphrasing and compressing the key information. This approach can be particularly useful in the medical domain, where reports can contain large amounts of complex information that need to be distilled into concise and actionable summaries.

For the training process, we will train the model using mini-batch and epoch techniques. This approach involves dividing the training data into smaller subsets called mini-batches, which are processed iteratively through multiple passes known as epochs.

In general, by training Large Language Models and Generative Pretrained Models on mini-batch and using techniques such as Abstractive Summarization and transfer learning, we aim to build a high-functioning model that can accurately predict radiologist IMPRESSIONS from unstructured clinical text data.

Model Validation and Tuning

We will utilize either train-test split or cross-validation methods to execute our model, which enable us to assess the model's performance for further analysis (evaluation metrics will be discussed in the Finding section). Moreover, it is essential for a model to demonstrate accuracy and generalization capabilities, striking a balance between complexity and accuracy. Therefore, to enhance the model's performance, we will employ techniques such as Fine-Tuning and Hyperparameter Tuning. And to constrain the model's complexity thus preventing potential issues associated with overfitting, regularization will be employed.

Model Optimization

After finally establishing the models and evaluating their performance, we plan to do some Optimization. Firstly, since the clinical text contains a relatively large number of out-of-vocabulary (OOV) terms[17], such as typing errors, or merged words that occurred while combining separate reports into one report, we will leverage BERT embeddings[18], which utilize WordPiece tokens [19]to reduce the OOV rate. Then, for the repetitions of some words or phrases in the output text, also known as the Over Generation Problem[20]. We are going to propose a post-processing step algorithm[21], which will make sure to keep the recall unchanged while increasing the readability of the generated text.

Findings

Model Evaluation and Comparison

In this finding section, we will first compare all the models we built and discuss their performance. Based on the methodology described, the results of implementing the model will

[17] Out-of-vocabulary (OOV) terms refer to words or terms that are not present in the vocabulary or training data of a language model, requiring special handling or treatment during text processing or analysis.

[18] BERT embeddings are contextualized representations of words or sentences generated by the BERT model, capturing the semantic meaning and context of the input text.

[19] WordPiece tokenization is a subword-level tokenization technique that breaks words into smaller units, capturing both known and unknown words and improving the coverage of rare or out-of-vocabulary terms.

[20] The over-generation problem refers to the issue in natural language generation where a model tends to produce excessive or irrelevant output, often resulting in verbose or incoherent text.

[21] A post-processing step algorithm is a computational procedure applied after the main processing task to refine, filter, or enhance the output data or results.

be evaluated by two methods. The first one is a statistical metric, here we will use ROUGE metrics[22], and report the ROUGE-1, ROUGE-2 and ROUGE-L performances of our models for evaluation and comparison. The findings are expected to indicate that the model is able to generate accurate and concise impressions that are aligned with the practice style for various modalities. In addition to the statistical ROUGE metric, we will also employ radiologists' evaluations, using human assessment to validate the predicted impressions. Since the ROUGE metric is only measuring word and sequence level similarity of the predicted and actual impressions, it comprises the visible deficiency of missing the factual correctness and the utility of the inference.

Expected Findings and Output

The proposed model is expected to have several important findings that could lead to significant benefits for medical professionals and patients alike. By exploiting domain-specific knowledge related to various imaging modalities, the model is designed to generate accurate impressions. This means that the model will be able to learn and understand the unique features of different imaging techniques, and will generate impressions that are tailored to each modality. Our models are expected to generate concise and accurate impressions that summarize the crucial conclusion(s) of each report. The length of the impression will be proportional to the amount of information in the findings and clinical information sections, ensuring that the generated impressions are comprehensive and informative.

Our expected output is an API to generate tailored impressions in real-time, reducing the time spent on dictations for each report. Furthermore, the model's batch processing capabilities

[22] ROUGE measures the similarity between the generated text and the reference text based on the recall of n-grams (contiguous sequences of n words) in the generated text that are also present in the reference text.

could be used by practitioners to audit previous impressions, correct errors, and even use it as a support system to train inexperienced radiologists in the impression generation process. This could improve the overall quality of radiology reports and reduce the risk of errors and omissions.

Overall, the proposed model is expected to provide significant benefits for medical professionals and patients, by generating accurate, tailored, and comprehensive impressions in real-time, reducing the risk of errors, and improving the efficiency of radiology reporting.

Discussion

TBD

Conclusion

TBD

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

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, from https://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: Title of Appendix A

TBD

Interim Report

Models

We performed fine-tuning on our base model using sonograph data (6577 records in total). The training process consisted of 300 steps and 6 epochs. Additional training steps can be implemented to improve the model's fitting.

Alpaca-7b

Size of the test data: 1000 (100 samples drawn from test set for evaluation)

Table 1. *Rouge Metrics of Alpaca-7b*

Type	Low			Mid			High		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
rouge1	0.1148	0.4644	0.1543	0.1358	0.5262	0.1744	0.1621	0.5885	0.1961
rouge2	0.0478	0.2843	0.0705	0.0603	0.3559	0.0877	0.0779	0.4229	0.1087
rougeL	0.0925	0.4083	0.1278	0.1076	0.4708	0.1441	0.1276	0.5371	0.1673
rougeLsum	0.0916	0.4057	0.1281	0.1076	0.4737	0.1447	0.1284	0.5378	0.1675

Medalpaca7b

Size of the test data: 400 (100 samples drawn from test set for evaluation)

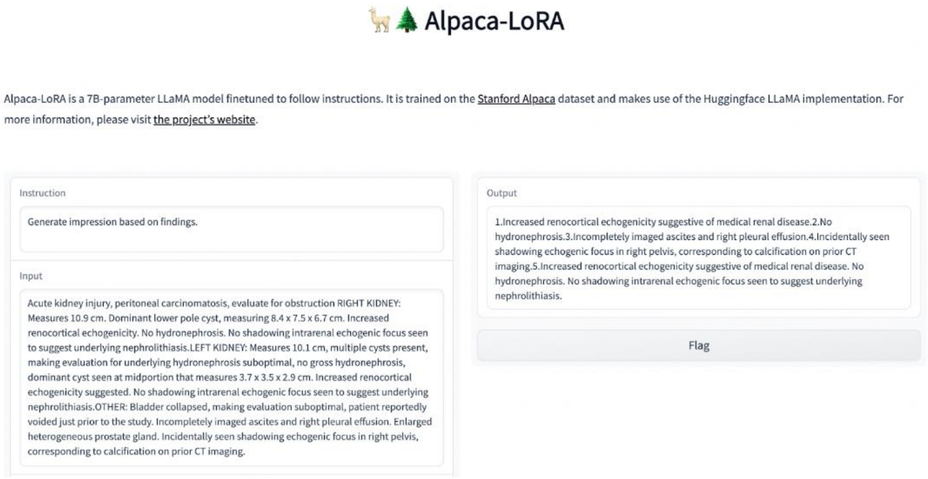
Table 2. *Rouge Metrics of Medalpaca7b*

Type	Low			Mid			High		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
rouge1	0.5040	0.4384	0.4300	0.5721	0.5132	0.5055	0.6405	0.5883	0.5780
rouge2	0.3583	0.3236	0.3292	0.4393	0.4030	0.4097	0.5252	0.4897	0.4952
rougeL	0.4759	0.4178	0.4138	0.5492	0.4971	0.4907	0.6161	0.5728	0.5628
rougeLsum	0.4800	0.4187	0.4180	0.5500	0.4979	0.4914	0.6230	0.5731	0.5649

In general, medalpaca 7b performs much better than alpaca 7b according to ROUGE scores. Further investigation could be conducted to explore the reasons behind.

Demo

Here's an example of output of Alpaca-7b model provided by Hugging Face:



🦙🌲 Alpaca-LoRA

Alpaca-LoRA is a 7B-parameter LLaMA model finetuned to follow instructions. It is trained on the Stanford Alpaca dataset and makes use of the Huggingface LLaMA implementation. For more information, please visit the project's website.

Instruction

Generate impression based on findings.

Input

Acute kidney injury, peritoneal carcinomatosis, evaluate for obstruction RIGHT KIDNEY: Measures 10.9 cm. Dominant lower pole cyst, measuring 8.4 x 7.5 x 6.7 cm. Increased renocortical echogenicity. No hydronephrosis. No shadowing intrarenal echogenic focus seen to suggest underlying nephrolithiasis.LEFT KIDNEY: Measures 10.1 cm, multiple cysts present, making evaluation for underlying hydronephrosis suboptimal, no gross hydronephrosis, dominant cyst seen at midportion that measures 3.7 x 3.5 x 2.9 cm. Increased renocortical echogenicity suggested. No shadowing intrarenal echogenic focus seen to suggest underlying nephrolithiasis.OTHER: Bladder collapsed, making evaluation suboptimal, patient reportedly voided just prior to the study. Incompletely imaged ascites and right pleural effusion. Enlarged heterogeneous prostate gland. Incidentally seen shadowing echogenic focus in right pelvis, corresponding to calcification on prior CT imaging.

Output

1.Increased renocortical echogenicity suggestive of medical renal disease.2.No hydronephrosis.3.Incompletely imaged ascites and right pleural effusion.4.Incidentally seen shadowing echogenic focus in right pelvis, corresponding to calcification on prior CT imaging.5.Increased renocortical echogenicity suggestive of medical renal disease. No hydronephrosis. No shadowing intrarenal echogenic focus seen to suggest underlying nephrolithiasis.

The original output of this example is: “Increased renal cortical echogenicity, compatible with underlying medical renal disease. Bilateral renal cysts, no gross hydronephrosis. Additional findings as above.”

In general, the model is capable of generating appropriate impression text. Further analysis and observations regarding the generated output will be provided in the 'Observations' section, allowing for a more comprehensive assessment.

Observations

Upon examination, it is evident that the Medalpaca model has successfully generated output that captures the essence of the actual impression records from the University of Chicago Medicine. Both sets of text (generated output & actual impression) share common themes and convey similar information, indicating that the model has effectively learned to extract key points from the source material. However, there are some discrepancies in the level of detail and phrasing between the two sets of text. The generated output may, though occasionally, include extraneous information or lack the precision found in the actual impression. In summary, the Medalpaca model demonstrates a strong ability to generate output that closely aligns with the reference material, but there is still room for improvement in terms of accuracy and conciseness to ensure a more faithful representation of the original medical content.

Next Steps

- Also include the clinical historical information as one part of the input
- Consider the order of the impression, which means that in addition to generating accurate impressions, we need to ensure that the order or sequence of the impressions is meaningful and coherent

- Try Other Models: 4 models will be deployed in total and the best one among the one will be selected

- Run the model on other modalities (CT, X-ray, Ultrasound … etc)

- Build and Refine Output API