Promptologist : Large Language Model(LLM) for Medical Query Resolution

Aakriti Kinra
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
akinra@andrew.cmu.edu

Aditi Patil
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
apatil2@andrew.cmu.edu

Aryan Singhal
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
aryans@andrew.cmu.edu

Lakshay Arora
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
lakshaya@andrew.cmu.edu

Abstract

Large Language Models(LLMs) have seen an unprecedented rise in industry and academia, most notably with the advent of ChatGPT. They have found use cases in diverse fields like education, healthcare, software development, design, sentiment analysis, among others. These models, trained on very large data(100s of billions of tokens) and have a very high number of parameters(1.8 trillion for GPT4). Their immense potential comes at the cost of requiring a very large amount of data and compute. Thus their wide development and research has been limited to the biggest tech companies and research labs. Model distillation is a technique that allow us to create smaller, more efficient machine learning models that can achieve high accuracy and performance. Specifically in the field of medicine, we look at MedicalGPT [1] which is a medical QA system. We aim to use the responses generated from MedicalGPT as training data and build a smaller model that is able to perform just as well. Our smaller LLM, Promptologist is trained on Chinese and English data. The data consists of real-world doctor-patient conversations. We use common text generation metrics like BLEU score and BERT score to evaluate the correctness of Promptologist.

1 Introduction

Large Language Models (LLMs) have emerged as a transformative force in artificial intelligence, revolutionizing how machines comprehend and generate human-like text. Deep neural networks power these models, and they show an unparalleled capacity for understanding and generating human language. Some prominent examples include OpenAI's GPT (Generative Pre-trained Transformer) series and BERT (Bidirectional Encoder Representations from Transformers). These models are pre-trained on a huge amount of diverse textual data and possess the ability to grasp linguistic nuances, context, and syntactic structures.

LLMs have various applications across the industries, showcasing their flexibility and versatility. They are extremely useful in tasks that involve natural language understanding, like sentiment analysis, named entity recognition, and language translation. In the finance sector, large language models are essential for analyzing the sentiment of the market news, thus helping investors make informed decisions. They are also used in content creation for generating text that resembles how humans write to cater to diverse needs, from creative writing to content summarization.

In the healthcare sector, large language models have become increasingly important tools, offering significant potential for transformative improvements. LLMs can be utilized for clinical documentation. They can be used for generating a summary of patient records, allowing the doctors and other staff to have a clear and concise narrative of the issue and, hence, speed up the workflow. LLMs also show promising performance in the field of biomedical research. They can help researchers process and understand large volumes of scientific literature.

Their ability to understand complex medical terminology and extract meaningful information from medical texts makes large language models particularly useful in literature review and knowledge synthesis. Since LLMs can analyze a large amount of data, they can be used to examine large amounts of patient records, trial outcomes, and literature. Through this, they can assist in patient care and personalize treatments. They can also identify correlations, predict disease outcomes, and recommend tailored treatment plans. These data-driven insights have the ability to revolutionize the accuracy of diagnostics and the efficacy of treatments.

Despite such promising outcomes, the integration of LLMs into the healthcare sector raises some crucial considerations about data privacy, ethical usage, and model interpretability. Ensuring that patient data is handled with the utmost confidentiality and abiding by ethical guidelines is essential for the responsible deployment of LLMs in healthcare.

Now more than ever, there is a growing need to build language models that can act as question-answering systems in the healthcare sector, which provide accurate and relevant responses. Our project aims to replicate and improve upon existing virtual assistants with a model that uses much fewer parameters by exploiting a healthcare-specific dataset.

Building training inference models is a time-consuming, costly, and resource-intensive task. We want to develop a model that's easy to use, is smaller in terms of size and computational complexity, and specifically caters to a particular domain, which in our case is the medical sector. Our model will be designed to answer questions related to medical domains and will be trained on a large dataset of medical information. By addressing these aspects, we create a model that not only fulfills the above requirements but also drives the application of AI in the medical field.

2 Literature Review

The paper Gu et al. (2023)[2] examines a critical aspect of improving the efficiency of complex language models. The researchers talk about the innovative process of knowledge distillation, which involves the transfer of knowledge from a larger model to a smaller model while maintaining or improving the smaller model's performance. They explore the complexities of training the smaller model using both original data and soft labels generated by the larger model. By filtering the essential knowledge from the larger model, the study demonstrates the potential for deploying efficient language models in scenarios where computational resources are limited. This knowledge transfer technique not only enhances computational efficiency but also contributes to the ongoing conversation on resource optimization in the realm of large language models. The paper provides valuable insights for researchers and practitioners looking to balance model performance with computational constraints.

The project MedicalGPT [1] that influenced our project, inspects the application of knowledge distillation in the medical domain. It focuses on pretraining and finetuning Llama's 7 billion parameter version using actual diaglogues between doctors and patients to create an LLM that answers questions asked by patients. The paper Peng et al. (2023) [3] also follows a similar path and also mentions the potential benefits and challenges of using LLMs in the field of medical research and healthcare.

As opposed to that, Gilbert et al. (2023) [4] discusses the potential implications of using artificial intelligence (AI) chatbots as medical devices. The authors argue that since these chatbots are designed to diagnose or treat medical conditions, they should be subject to the regulations and approval processes required of medical devices. The authors point out that these AI chatbots have the potential to revolutionize healthcare by providing accessible and affordable diagnostic and treatment options. However, they also note that using these chatbots raises significant ethical and safety concerns. For instance, if chatbots are not adequately regulated, they may be used to diagnose or treat conditions incorrectly, leading to serious harm to patients. The authors argue that large language model AI chatbots should be subject to the same rigorous regulatory processes as other medical devices to avoid such risks. This would ensure these chatbots are safe and effective before making them available to the public. The authors also suggest that regulatory agencies should work closely with AI developers to create appropriate regulatory frameworks that ensure the safety and efficacy of these chatbots.

3 Dataset

We will be using the medical dataset used by Medical GPT [1][5][6] that has both Chinese and English data. This dataset was created during the Supervised fine tuning of the MedicalGPT model. It consists of medical queries asked by patients and the response MedicalGPT[1][7] gives to these questions. It has 2.08GB of data with 2.4 million rows of data including the pretrain, finetune and test data.

The MedicalGPT dataset containes real conversations between patients and doctors sourced from different datasets. For the purpose of our baseline, we will only be using the English conversations for training, testing and validation. The data is originally in the form of json objects with input and output being the patient's prompt and doctor's response respectively. We pre-process this data by converting the json into patient-doctor texts.

4 Model Selection

We conducted experiments using different models: a simple bigram language model, an n-gram model without attention, and an n-gram model with attention. As a baseline, we chose a simple decoder-only n-gram model with attention. Our encoding approach involves character-level tokenization. As part of experimentation we initially attempted word-level tokenization. However, the resulting vocabulary size proved too extensive for our compute. We dropped the proposal of using common NLP preprocessing techniques such as lemmatization, which would result in a reduced dictionary size. However, considering the output requirements of a generative model, need the words in their proper form, and not their lemmatized representation. Yet, upon training for a few epochs, we observed that the generated outputs lacked grammatical correctness in conversational English. Ultimately, character-level tokenization, where each character is encoded with its ASCII value, proved to be the most effective strategy for our scenario. This method, with a smaller vocabulary size, although a larger timestep dimension, enabled the generation of relevant responses in nearly grammatically correct conversational English. The GitHub repository for this project is referenced here: [8]

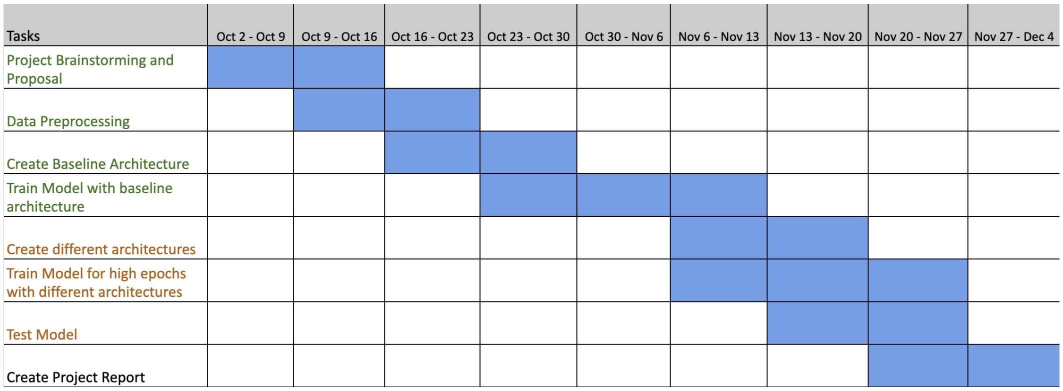
5 Results

We use average cross-entropy loss and BERT score to evaluate the accuracy of our model. We tried BLEU score as an evaluation metric, which we soon realized was not the right metric in our case as it calculates the score by exactly matching the prediction and output, where our aim is to generate outputs that are similar in essence and relevant to the patient's prompt. Thus, we use BERT that learns contextual representations of words or tokens in a large text corpus and calculates the score using cosine-similarity.Our baseline - n-gram model with multi-head attention achieves a precision of 0.7781 and recall of 0.8055 with a F-1 score of 0.7916 (as mentioned in the table below)

Table 1: Average Cross Entropy Loss and BERT score of Testing Set

Models	Avg. Cross Entropy Loss	BERT Precision	BERT Recall	BERT F-1
bi-gram	2.6472	0.4282	0.4567	0.4420
n-gram w/o attention	1.8298	0.6622	0.6582	0.6602
n-gram with attention	**1.3273**	**0.7781**	**0.8055**	**0.7916**

6 Project Timeline



Tasks	Oct 2 - Oct 9	Oct 9 - Oct 16	Oct 16 - Oct 23	Oct 23 - Oct 30	Oct 30 - Nov 6	Nov 6 - Nov 13	Nov 13 - Nov 20	Nov 20 - Nov 27	Nov 27 - Dec 4
Project Brainstorming and Proposal	🟦	🟦							
Data Preprocessing		🟦	🟦						
Create Baseline Architecture			🟦	🟦					
Train Model with baseline architecture				🟦	🟦	🟦			
Create different architectures						🟦	🟦		
Train Model for high epochs with different architectures						🟦	🟦	🟦	
Test Model							🟦	🟦	
Create Project Report								🟦	🟦

7 Future Work

We have plans to further enhance the performance of our model by increasing the amount and diversity of training data. To achieve this, we intend to translate over 2GB of Chinese (Mandarin) data from MedicalGPT[1] into English and ensure the accuracy of the translation. Additionally, we aim to experiment with various n-gram sizes to find the perfect balance between precision and recall, which will help us improve our n-gram model. We will also explore different smoothing techniques to handle unseen test data more effectively.

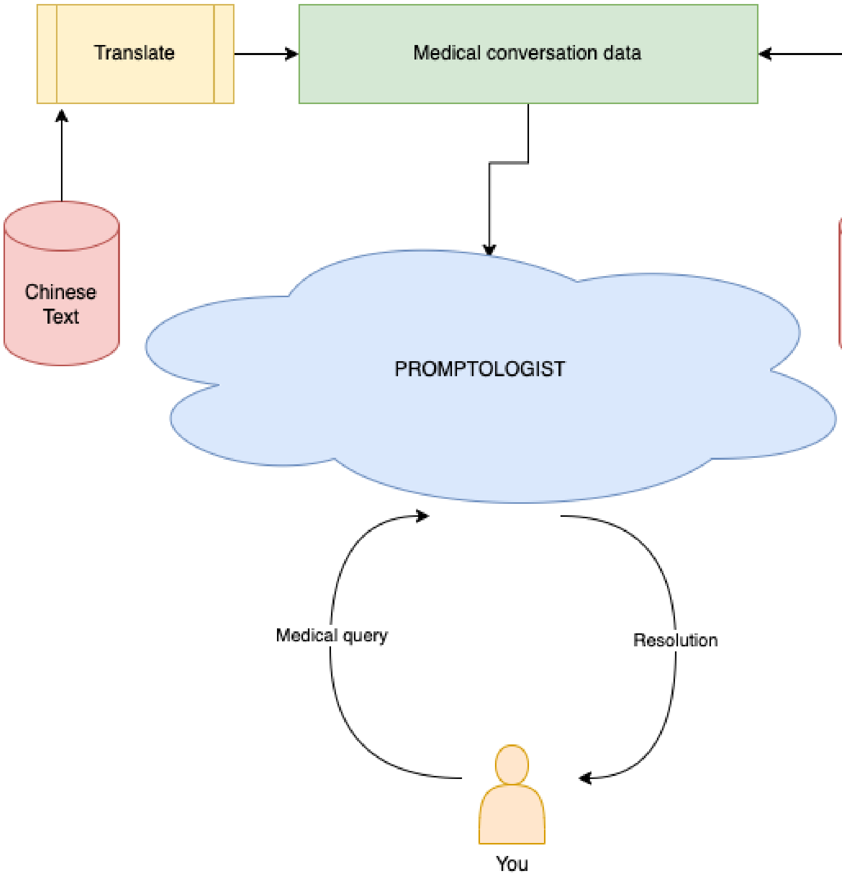


Figure 1: Future plan

References

[1] Ming Xu. Medicalgpt: Training medical gpt model. `https://github.com/shibing624/MedicalGPT`, 2023.

[2] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

[3] Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. A study of generative large language model for medical research and healthcare. *arXiv preprint arXiv:2305.13523*, 2023.

[4] Stephen Gilbert, Hugh Harvey, Tom Melvin, Erik Vollebregt, and Paul Wicks. Large language model ai chatbots require approval as medical devices. *Nature Medicine*, pages 1–3, 2023.

[5] Chinese medical dialogue data. `https://github.com/Toyhom/Chinese-medical-dialogue-data`.

[6] Huatuo-26m. `https://github.com/FreedomIntelligence/Huatuo-26M/blob/main/README_zh-CN.md`.

[7] Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. Huatuo-26m, a large-scale chinese medical qa dataset, 2023.

[8] Aakriti Kinra Aditi Patil Lakshay Arora, Aryan Singhal. Promptologist. `https://github.com/lucky-119/Promptologist`, 2023.