COSE474-2024F: Final Project Proposal
"Prompt Engineering for Medical Question Answering (QA)"

Shin Minseo 2022320338

1. Introduction

The rapid advancements in large language models (LLMs) have opened up potentials for AI-driven assistance across many specialized fields, including healthcare. Developing a reliable medical question answering (QA) system can improve access to healthcare information, as it provides individuals with 24/7 availability of guidance on medical inquiries. However, building a robust medical QA system presents critical challenges: handling complex medical terminology, ensuring safety and ethical responsibility and providing accurate, relevant answers without misleading the user.

In this project, I will leverage prompt engineering techniques to adapt a pre-trained large language model, Llama, to the specialized medical QA domain. By prompt engineering, it means to adjust the structure of input prompts in order to guide Llama's response in a healthcare-related question, without requiring extensive model fine-tuning. The goal is to design prompts that address the challenges of medical QA, and improve Llama's ability to provide accurate, safe, and sound answers.

2. Problem definition & challenges

The objective of this project is to develop a medical QA system capable of interpreting and responding to medical queries in the most appropriate way. Focusing on prompt engineering, I aim to make Llama's pre-trained general knowledge to become specialized in specific demands in medical domain. An initial testing on Llama was done to find out the key challenges when medical QA is used.

The first challenge is that since Llama was trained on general text corpora, it lacks depth of medical knowledge necessary for precise responses. When questions from patients were asked, the responses were found to be incomplete of overly general. For medical QA, accuracy and completeness are the key. Medical inquiries need to be handled cautiously. Initial test with very simple questions like "Is it safe to take ibuprofen every day?" lacked cautionary advise, raising ethical concerns. Lastly, initial test also showed high sensitivity to the way users prompt phrases, in which slight changes in wording led to different responses.

3. Related Works

Pre-trained foundation models for healthcare Many studies are focusing on adapting pre-trained language models to specialized domains. BioBERT is a foundational example of biomedical domain models built on BERT architecture, which significantly outperforms BERT and previous SOTA models in a variety of biomedical text mining tasks: biomedical named entity recognition, biomedical relation extraction and biomedical question answering.

Prompt Engineering in LLM Prompt engineering as a tool for guiding large language models emerged from studies on LLM such as Chat GPT. "Prompt Engineering Guide" introduces several concepts including few-shot, zero-shot, and one-shot learning, which crafts the prompts according to the desired response. It demonstrates that prompt phrasing alone can significantly affect model behavior. According to the guide, few-shot prompting provides a few examples in the prompt to guide the model in completing similar tasks, while zero-shot prompting uses descriptive task-specific instructions without examples. Prompt engineering has been applied to other specialized domains, but there are limited research focusing explicitly on medical QA. By exploring prompt design tailored for healthcare, I will aim to use expand on these techniques for domain-specifc tasks.

Prompt Engineering for medical QA While models like BioBERT and PubMedBERT are fine-tuned on biomedical text, being able to understand medical terminology, they require a high cost of computational resources and data for training. While fine-tuning demonstrated strong results on evaluating LLMs on medical benchmarks, recent studies shows that applying robust prompt engineering to optimize performance on foundation models produce similar level or outperforms when tested on medical benchmarks.

4. Datasets

As primary dataset, I will use "ChatDoctor - HealthCareMagic - 100k" dataset from Huggingface, which consists about 100,000 medical Q-A pairs. It has a broad base of medical interactions that relates to common inquiries, symptoms, conditions and general medical knowledge.

5. State-of-the-art methods and baselines

MedQA is a widely used benchmark dataset for medical QA containing multiple choice questions based on real medical exams. As baseline models for MedQA, I will use BioBERT and MedGPT models for comparison. BioBERT's fine tuning on biomedical data will make it a useful baseline to demonstrate the effectiveness of domain-specific language adaptation.

6. Schedule & Roles

Week 1: - review papers on prompt engineering and SOTA models for MedQA. Explore the chosen dataset to understand patters.

Week 2: - Set up Llama, prepare coding environment and develop initial hard prompt templates.

Week 3-4: Implement prompt variations, and test and evaluate them.

Week 5: Analyze evaluation metrics then refine the prompt structures to improve result.

Week 6: Document findings, analyze limitations, finalize paper.