CMPE-259 Natural Language Processing

Project Proposal

Natural Language Processing (NLP) has experienced a significant boom with the introduction of Large Language Models (LLMs), and developments in this field have rapidly increased. With each passing day, these models are becoming more robust and efficient at interpreting user queries and providing appropriate and relevant answers. Inspired by this, I plan to develop a Large Language Model-based Virtual Assistant (VA). This VA is designed to answer questions related to the Service Agreement and Customer Agreement of Amazon Web Services (AWS).

The business use case for this VA is to assist both customers and employers using AWS services in understanding its policies more clearly. Both agreements are available on the AWS website, with links to the Customer Agreement and Service Agreement. These pages can be saved as PDF files and used to build a Retrieval-Augmented Generation (RAG) pipeline to train the LLM to answer specific questions related to these agreements.

With the rise of large language models, many different models have been developed by various companies. Some of these models are open source, while others require payment to use their services. For this project, I plan to use Ollama-llama3, Ollama-mistral, and OpenAI. Ollama is a platform that manages and runs open-source LLMs locally. OpenAI developed GPT, which is a paid model and is one of the best LLMs available on the market. For this project, I will use their embedding models to create embedding vectors, which will be stored in ChromaDB, an open-source vector database, and integrated into a RAG pipeline. The first step of this pipeline will be retrieval, where the system will fetch information from the database that closely matches the user query. The next step will be augmentation and generation, where the language model uses both the user query and the retrieved information to generate a response that is grounded in accurate data.

I will develop the model using Ollama-llama3, Ollama-mistral, and OpenAI. After developing the bot, I will ask it various questions and use an LLM to evaluate the generated responses against the expected responses. Comparisons of the performance of different models will be provided, and I will explore whether changing parameters in the RAG pipeline can improve the model's performance.