Phu (Jack) Nguyen
CSCE 482-934
8/29/2023

Product Introduction

Our project, sponsored by Kyndryl - the largest IT infrastructure services provider, aims to develop a system to train a large language model (LLM) using private data. The project will be immediately beneficial to business organizations as they can train the LLM model on their own private corporate data to get more tailored and relevant responses for their business needs. Aside from businesses, nonprofits, and academic organizations, as well as normal everyday users, who have an adequate amount of private data available, may utilize and benefit from this project as they see fit.

In terms of data, the user provides the system with their desired data in one of the supported formats. The intention is for the user to input private data - high-quality, specific data that is generated by them or their internal sources - that is free from any irrelevant or garbage values commonly found in public datasets that could negatively impact the LLM. With the provided private data, the user can expect the system to process that data to build an LLM that would provide more relevant and tailored responses that align with their organization's specific needs. The user's expectation is that the system will generate more accurate and applicable results compared to an LLM trained with public data since they can utilize industry or organization-specific insights and context to enhance the system's performance.

From the system's perspective, there are a few major steps required to achieve this task. First, the user will upload their private data (commonly text data) in one of many supported formats to a designated front end (website) to be received and ingested by the system. Then, the system will process the data by extracting and tokenizing the text and creating embeddings from those tokens with Azure OpenAI, an enterprise-grade OpenAI service by Microsoft. Embeddings are meant to capture the semantic meaning of texts and represent them in a numerical vector format and are the core component of training LLMs. Those embeddings are then saved into Pinecone DB, which is a scalable and efficient database for storing and indexing large volumes of vector data that allows for quick and accurate retrieval. This essentially will serve as the knowledge base for our LLM. In order to use the LLM, the user would interact with the system and its knowledge base through a chat user interface where they can input prompts and queries that would trigger the LLM to search the embeddings vector database. Once the system retrieves the relevant embeddings from Pinecone DB, it then uses it to generate tailored responses to be displayed back to the user.

By following these steps with the use of the above external technologies, the project accomplishes the goal of creating a system to train an LLM using private instead of public data, for organizations, whether business, nonprofit, academic, or just everyday users. This privately trained LLM can help these users get better-tailored responses and would perform much better than publicly trained LLMs, as it will not be negatively impacted by garbage or irrelevant values commonly found in public training datasets.