Project member: Juan Yu

Project topic: Context Retrieval for Retrieval Augmented Generation

Project description:
Large Language Models(LLMs) are trained on data that was collected as of some prior time, and the training data of most LLMs is not domain specific. For example, the training data of the current ChatGPT3.5 covers a diversified range of domains and is up until January 2022. That means, LLMs may fail to provide up-to-date information or generate a good answer to a domain specific question. Retrieval Augmented Generation (RAG) provides a solution by retrieving contextual data from additional datasets that can be real-time news, domain-specific knowledge, or any other data of interest given a user query and grounding LLMs on the retrieved contextual data for answering that specific query. It can reduce hallucination of LLMs and thus improve its performance as it provides supporting evidence for an answer, and more importantly, it is much more efficient than fine-tuning LLMs on additional datasets. RAG would facilitate the application of LLMs by various real-world industries. For example, financial analysts and traders rely heavily on up-to-date listed company-specific information and real-time market news, while frequent fine-tuning LLMs is not practical. This project focuses on a critical component of RAG, retrieving context from an unstructured text document given a natural language query.