Study Buddy - Chatbots as Lecture Companions Using LLMs and Retrieval Augmented Generation

Students

Lucien Zimmermann

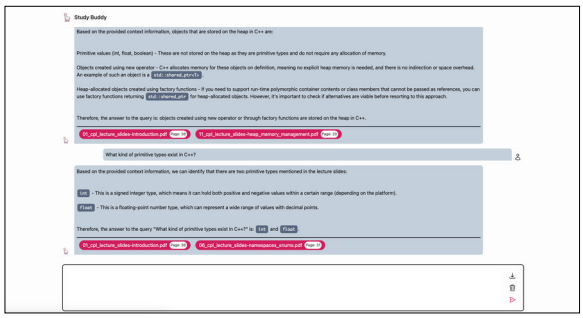
Florian Rohrer

Introduction: The emergence of large language models (LLM) changes the way we search for information. LLMs allow us to ask questions directly and receive answers in natural language. However, the knowledge of LLMs is limited to the information they have been trained on and is therefore often outdated. This limitation can be overcome by using the retrieval augmented generation (RAG) technique. This technique combines the user's prompt with contextual information from a custom knowledge base before asking the LLM to generate an answer. The technique relies on semantic search using embeddings to find relevant content related to the user's prompt in the knowledge base. RAG significantly improves the quality of the answers received from the LLM, especially when specific knowledge beyond what the LLM has been trained on is required.

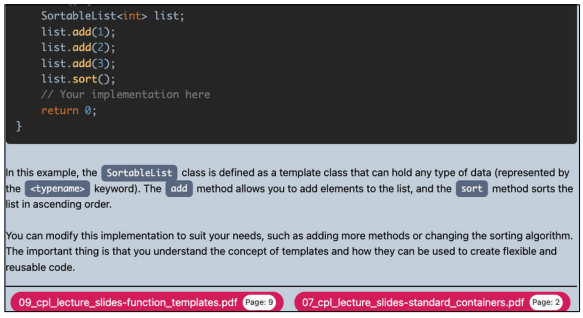
Objective: The goal of this project was to implement a chatbot in Python and React that uses the RAG technique to answer a student's questions about lecture-related content, such as PDF lecture notes. In addition to providing correct answers, the bot should also list the sources used to generate the answers, allowing the student to verify the answer.

Result: A chatbot was implemented using open-source components. The focus was on the LLama2 LLM family and LLamaIndex, a data framework in Python for connecting LLM. The chatbot was tested using slides from the C++ and OOP lectures at OST. We found that the RAG technique works well for answering questions based on text-based notes. However, we encountered difficulties in retrieving relevant context when dealing with bullet points and images in lecture slides, resulting in the LLM generating inaccurate answers. To reduce the impact of these limitations, we conducted tests to evaluate an embedding model that best fits our use case. During our testing we could not find any model, including Llama2, that performed adequate with languages other than English. This problem can only be addressed by fine tuning a model. So we focused our evaluation on English texts. We have also provided a guide for lecturers and students on how to use chatbots like this one efficiently.

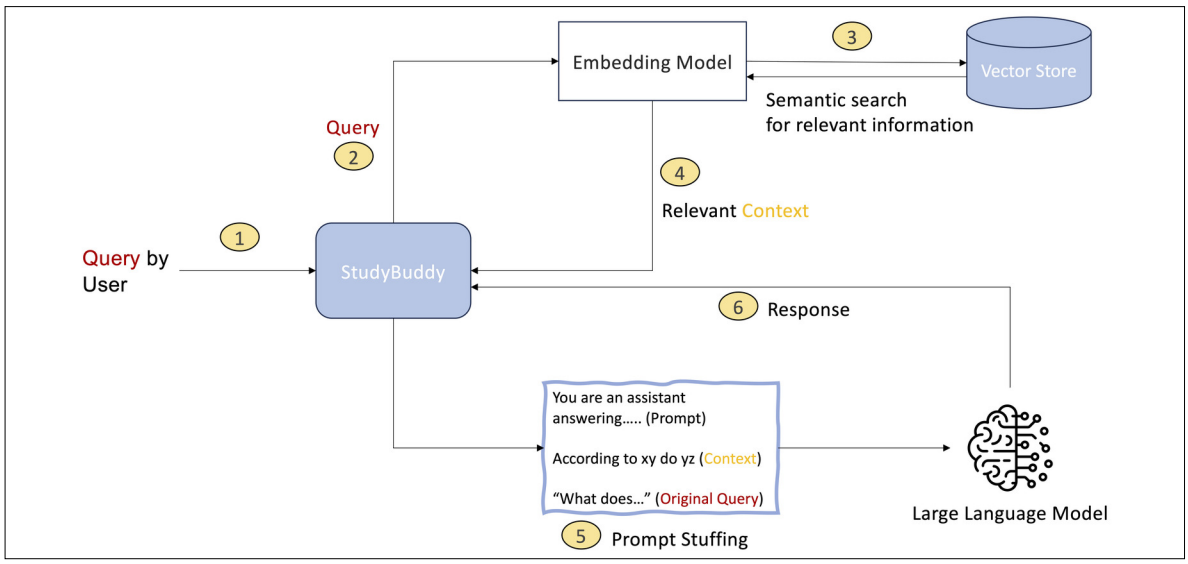
The web app is build using Python and React and is designed with simplicity in mind; it provides a simple chat interface
Own presentment



Answers provided by StudyBuddy are written using natural language and always contain a direct reference to the source
Own presentment



Visual representation of retrieval augmented generation. The query is enriched with relevant context and sent to the LLM
Own presentment



Advisor
Prof. Stefan F. Keller

Subject Area
Software, Application Design