LLM Apps

Basic RAG App: QA of a Document

Problem: the context window limit

- Foundation LLMs are limited by their context window.
- The free version of ChatGPT, for example, cannot handle a text of more than 4097 tokens.
- What if we want to ask questions about a document longer than that limit?

Solution

- We will create a basic RAG application:
  - Split the document into small fragments.
  - Convert those fragments into numbers (called "embeddings").
  - Load the embeddings into a vector database.
  - Create a retrieval system using a predefined LangChain chain.

Process

- Load the text document with a document loader.
- Split the document into fragments with a text splitter.
- Convert the fragments into embeddings with OpenAIEmbeddings.
- Load the embeddings into a FAISS vector database.
- Create a RetrievalQA chain to retrieve the data.