

LLM Apps

Basic App: Evaluation of a QA App

Problem

- We want to evaluate the quality of a question and answer application on a document (QA App).
- Evaluating an LLM application is not easy, as the answers can be slightly different for the same questions. A conventional software application evaluation system cannot be applied.

Solution

- We will prepare a list of questions for which we know the answers and ask them to the QA App using a RetrievalQA chain for manual evaluation.
- Then we will use a QAEvalChain so that the App itself checks if its answers match the evaluation answers.

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 including an input_key to identify the user's prompt (the question).
- Create a dictionary with the evaluation questions and answers.
- Use the RetrievalQA chain to manually evaluate the App.
- Use a QAEvalChain chain for the App to evaluate itself.