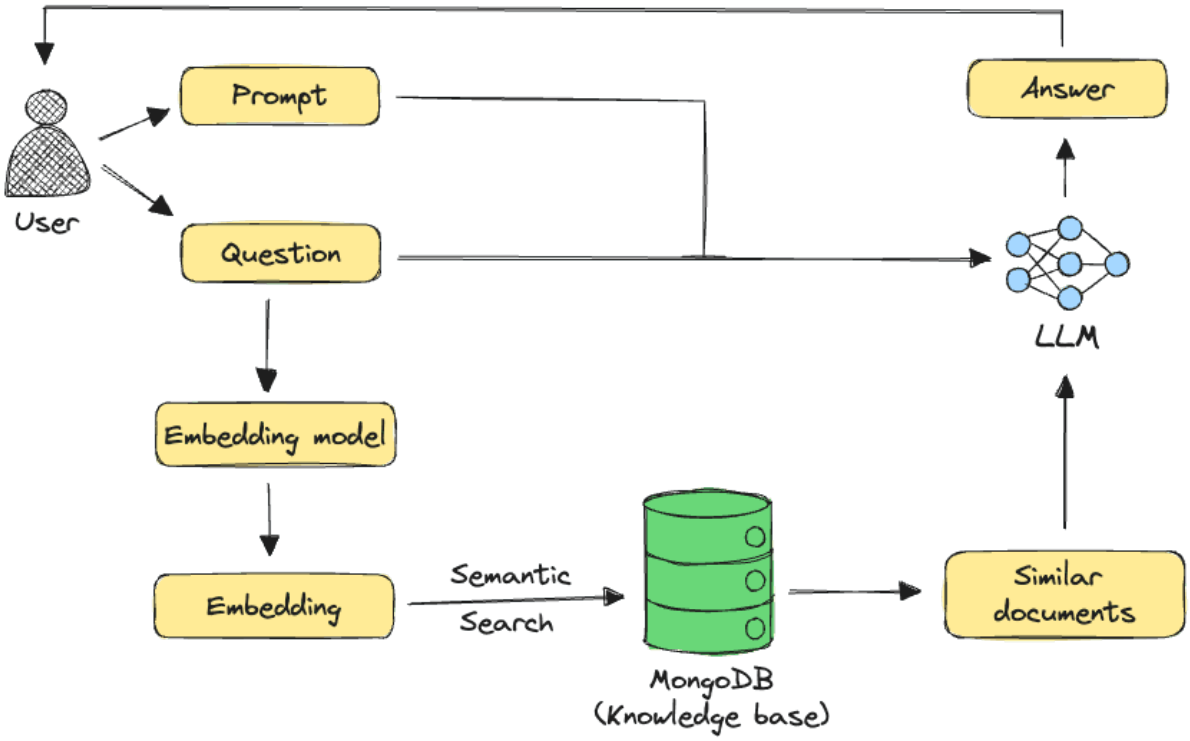
Retrieval-augmented generation, as the name suggests, aims to improve the quality of pre-trained LLM generation using data retrieved from a knowledge base. The success of RAG lies in retrieving the most relevant results from the knowledge base. This is where embeddings come into the picture. A RAG pipeline looks something like this:



In the above pipeline, we see a common approach used for retrieval in genAI applications — i.e., semantic search. In this technique, an embedding model is used to create vector representations of the user query and of information in the knowledge base. This way, given a user query and its embedding, we can retrieve the most relevant source documents from the knowledge base based on how similar their embeddings are to the query embedding. The retrieved documents, user query, and any user prompts are then passed as context to an LLM, to generate an answer to the user's question.