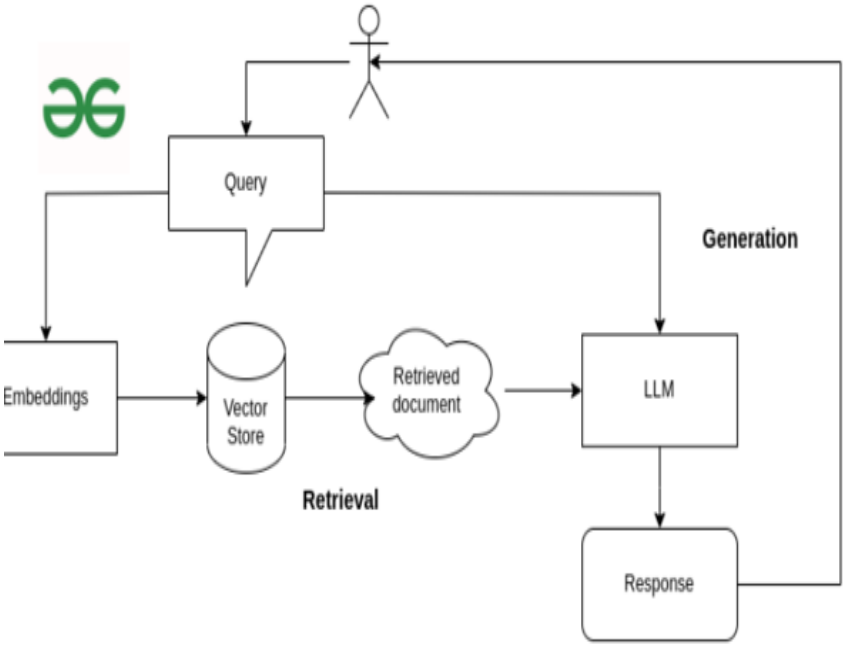
Task 2: RAG Pipeline Diagram



Retrieval augment generation, in short RAG is a mechanism to integrate large language model to a custom data. RAG includes two methods:

- Retriever model
- Generator model

The process includes where a user enters his query. This query is first converted into vector format using embedding model. Based on this input embedding, we look into the existing vector store and retrieve the relevant document/content. This is the process that is involved in retriever model.

In the generator model, we generate a response. During the retriever, we extract the relevant document, this document along with the query is passed to the large language model. Using the LLM intelligence we get a better response to the user query/prompt.