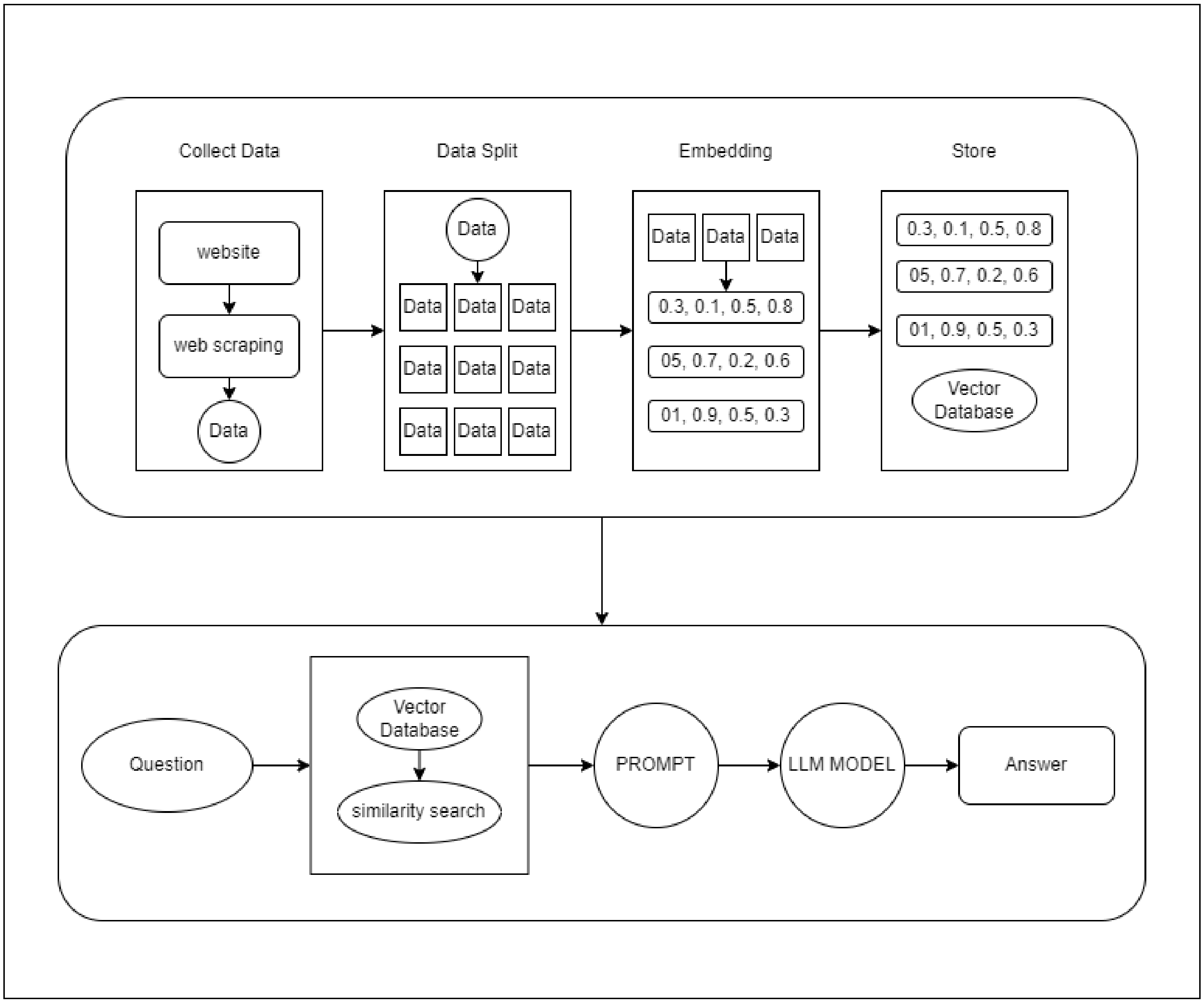
Chatbot Task Report



Chatbot Pipeline

Explanation:

1. Indexing

- Load: The first step involves gathering data through web scraping.

- Split: Large documents or datasets are then broken down into smaller, manageable chunks using text splitters. I am using a 10000-word chunk and 1000 overlap.

- Embedding: Embedding those words for machine understanding.

- Store: The split chunks are stored and indexed, typically in a Vector Store. I am using Faiss vector.

2. Retrieval and Generation

- Retrieve: When a user inputs a query, the system retrieves the most relevant chunks from storage using a Similarity search, which searches over the indexed data.

- Generate: Finally, a Chat Model or Large Language Model (LLM) generates an answer. The model is prompted with the user query and the retrieved data, allowing it to produce a contextually relevant response.

Challenges:

The process of indexing and retrieval in an RAG system involves several challenges. During indexing, data quality can be compromised due to web scraping, and finding the optimal chunk size is critical to balance detail and context. Embedding large datasets is resource-intensive and might miss complex meanings while managing vector stores like Faiss requires optimization for efficient retrieval. LLMs may struggle with multiple chunks and domain-specific queries. Scalability is a concern as data grows, potentially slowing down the system, and delivering a user-friendly experience requires ongoing refinement. Balancing these factors is essential for an effective RAG system.