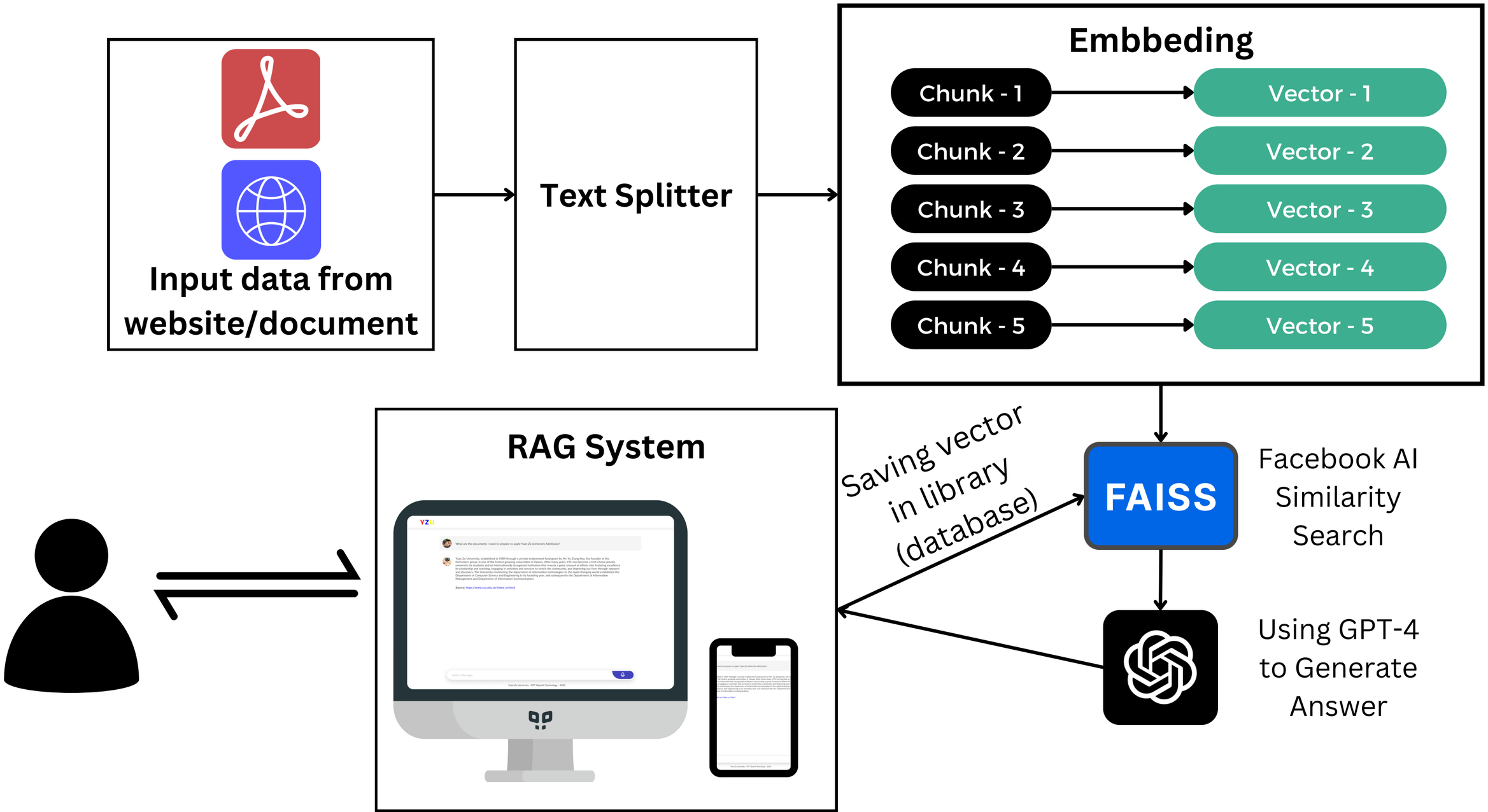
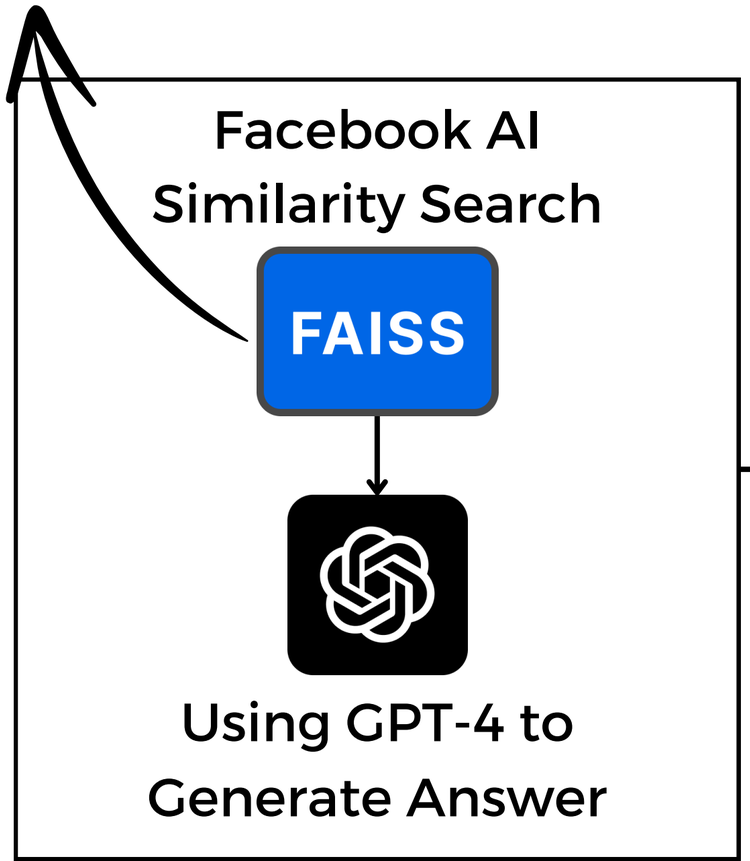
How to Retrieve What Part The Contents Which is Matched Against The Inquiry

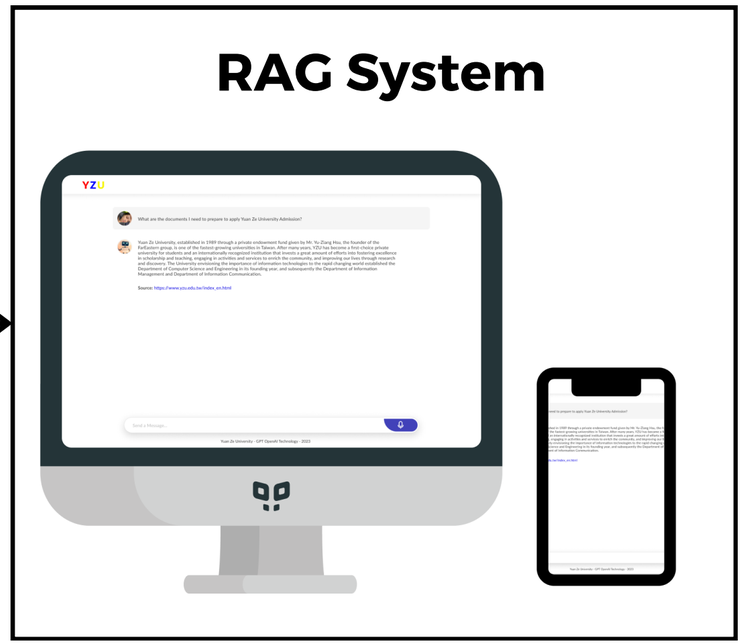
Data Process



FAISS Process

FAISS is a library for efficient similarity search and clustering of vectors. It's commonly used for tasks like finding nearest neighbors or similar items in large datasets of vectors.





How the system can be used to choose the document that matches with user queries?

1. Indexing Documents: We need to index the vector representations of the documents using FAISS.
2. Performing Similarity Search: When a user submits a query or question, we need to convert the query into a vector representation using the same embedding model. Then, we can use FAISS to perform a similarity search to find the most similar vectors (documents) to the query vector in the FAISS index.
3. Choosing the Best Match: Once FAISS returns the indices of the most similar documents, GPT-4 will choose the document that has best matches the user query.
4. Retrieving the Matched Document: Finally, the system will present or generate the answer to the user as the system's response to their query.

We use semantic search before for the process to choose relevant document related with users queries. However, we need to use database to save the vector. By using FAISS, we can save the vector and find the document related with users queries. Lastly, GPT-4 will choose the most similar documents and generate the answer.