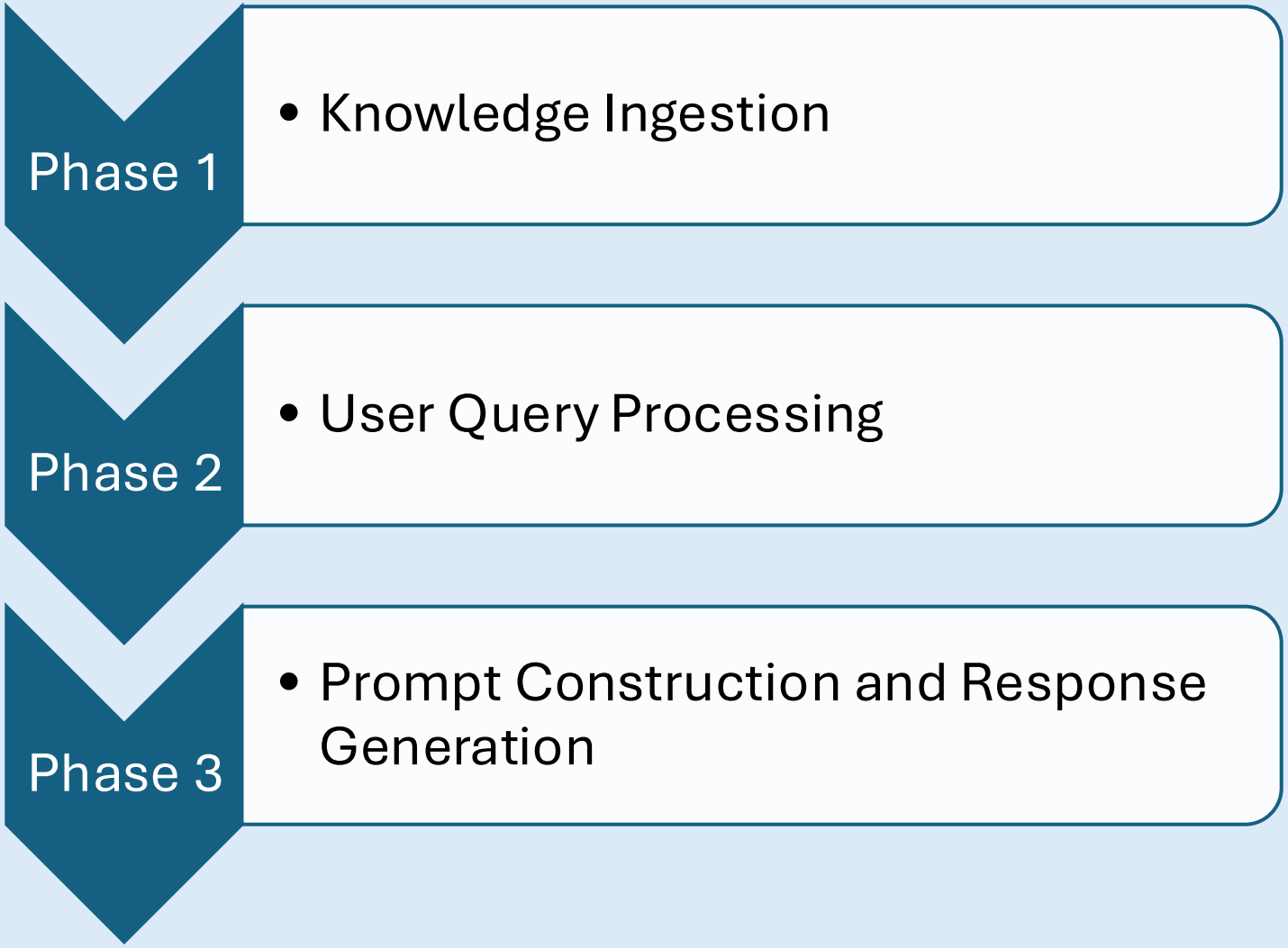
RAG FLOWCHART

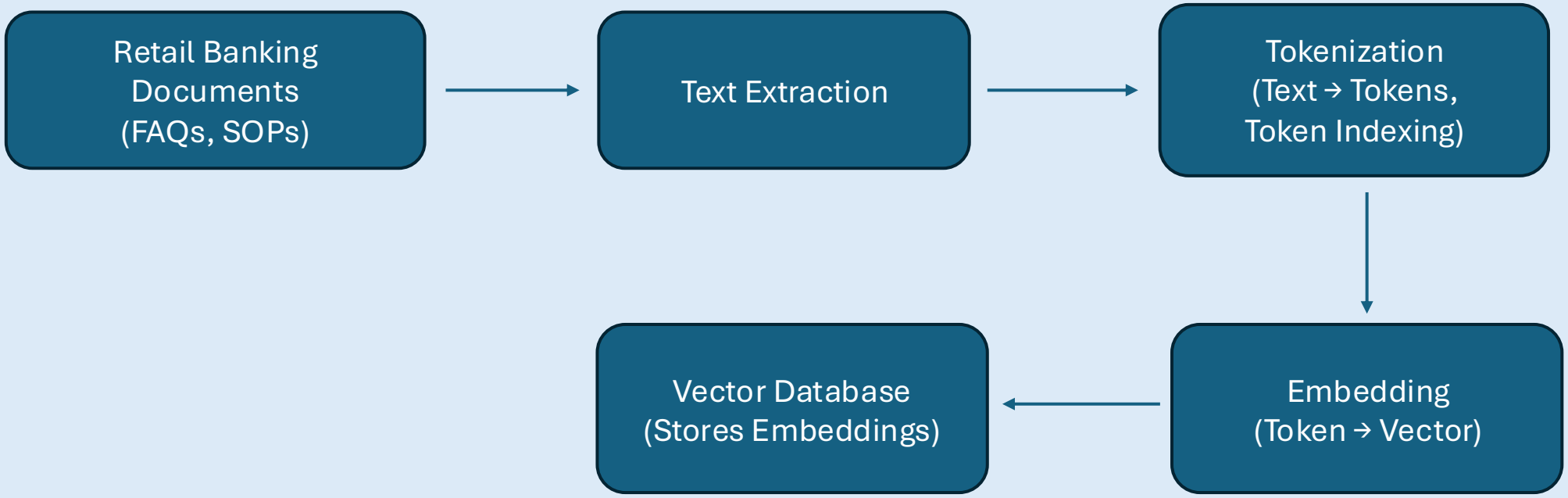
Use Case : Retail Banking

RAG Workflow



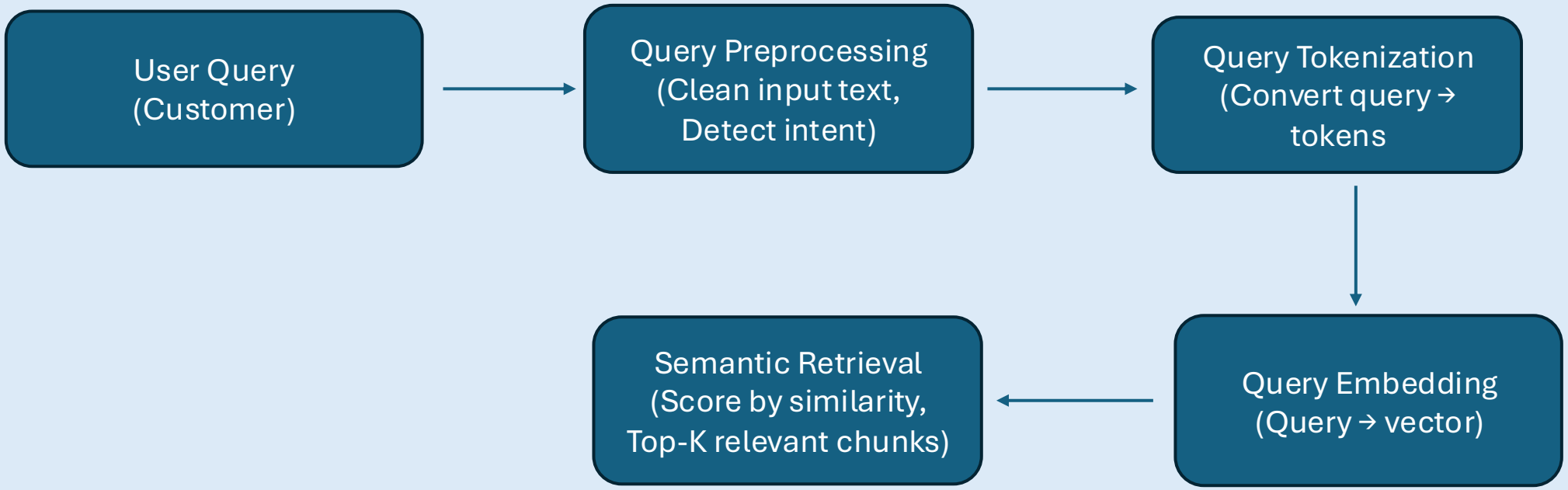
PHASE 1: **Knowledge Ingestion**

- Extract bank-related documents/data
- Split text into tokens and create embeddings/vectors
- Store embeddings in a vector database for fast retrieval



PHASE 2: **User Query Processing**

- Preprocess and tokenize the user query
- Generate query embeddings and perform semantic search using the vector database
- Retrieve top-K relevant document chunks



PHASE 3: **Prompt Construction and Response Generation**

- Combine retrieved context with the user query
- Build the final prompt within token limits
- Generate the response using the LLM