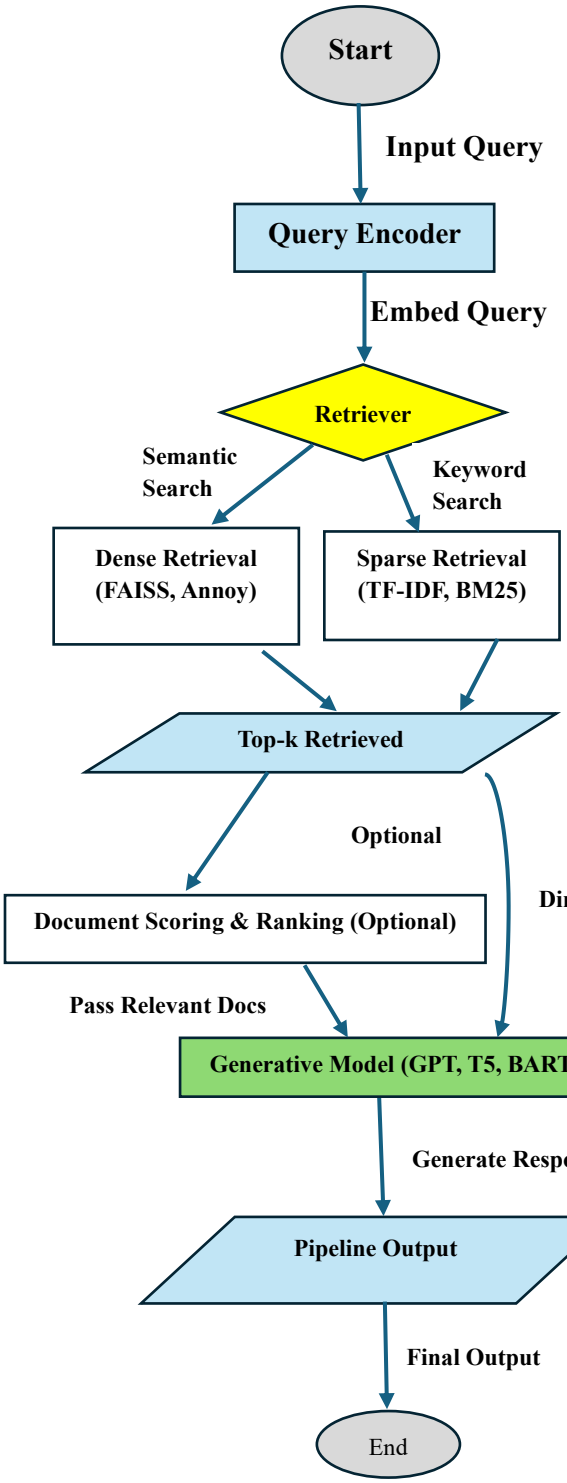
Key Components of the Retrieval-Augmented Generation (RAG) Pipeline

By Atsu Vovor[1]



Query Encoder: The input query is encoded into an embedding using a pre-trained transformer-based model (like **BERT or RoBERTa**). These embeddings help find semantically similar documents in the knowledge base.

Retriever:

- **Dense Retriever:** Uses similarity search **(e.g., FAISS, Annoy)** to match query embeddings against a pre-built vector index of the document corpus.
- **Sparse Retriever:** Uses traditional keyword-based methods **(e.g., TF-IDF or BM25).** Outputs a set of top-k relevant documents.

Generative Model: A pre-trained language model (e.g., GPT, T5, or BART) takes the retrieved documents and the original query as input. It generates the final answer or response, often grounded in the retrieved content.

Document Scoring and Ranking (Optional): Some **RAG** implementations re-rank the retrieved documents to ensure the most relevant ones are prioritized for generation.

Pipeline Output: The final output is a coherent response that combines both the query and the contextual information from the retrieved documents.

[1] Atsu Vovor: Consultant, Data & Analytics Specialist | Machine Learning | Data science | Quantitative Analysis | French & English Bilingual | atsu.vovor@bell.net | https://github.com/atsuvovor/Pub_Data_Analytics_Project | https://public.tableau.com/app/profile/atsu.vovor8645/vizzes