Retrieval Augmented Generation (RAG) is a technique that combines information retrieval with large language models.

In a RAG pipeline, documents are first loaded and converted into text. The text is then split into smaller chunks so that embedding models can process them efficiently.

Each chunk is converted into a numerical vector using an embedding model such as Sentence Transformers. These vectors represent the semantic meaning of the text.

The vectors are stored in a vector index like FAISS. FAISS allows fast similarity search between the user query embedding and document embeddings.

When a user asks a question, the query is embedded and compared with stored vectors. The most similar chunks are retrieved.

These retrieved chunks are provided as context to a language model. The language model then generates an answer based on both the question and the retrieved context.

RAG improves factual accuracy because the model uses external documents instead of relying only on its internal knowledge.

Common applications of RAG include chatbots, document question answering, knowledge assistants, and enterprise search systems.

RAG systems typically consist of these components: document loader, chunker, embedder, vector store, retriever, prompt builder, and generator.

FAISS is a popular local vector store library that does not require authentication or cloud services. It runs entirely on your machine.

Sentence Transformers are widely used for creating high-quality embeddings for semantic search.

Building RAG from scratch helps engineers understand each component deeply and provides full control over the pipeline.