Introduction to RAG Architecture

Retrieval-Augmented Generation (RAG) combines information retrieval with text generation.

Key components:

1. Retriever: Finds relevant documents using vector similarity search
2. Generator: Large Language Model (LLM) that produces final answers
3. Vector Database: Stores document embeddings for efficient retrieval

Advantages:

- Reduces hallucinations by grounding responses in source material
- Allows knowledge updates without retraining the LLM
- Provides source attribution for generated content