

Retrieval-Augmented Generation (RAG) is a technique that combines information retrieval with large language models. Instead of relying only on pre-trained knowledge, RAG retrieves relevant documents from a vector database and provides them as context to the LLM.

The system typically involves:

- Document chunking
- Embedding generation
- Vector storage (e.g., FAISS)
- Similarity search
- LLM-based response generation

RAG reduces hallucinations and improves answer accuracy.