RAG Architecture Overview

A Retrieval-Augmented Generation system consists of two main stages:

retrieval and generation.

The retrieval stage finds relevant documents from a knowledge base using vector similarity.

The generation stage uses a language model to answer questions based on the retrieved context.

Advanced RAG pipelines improve retrieval using MMR and reranking.

These techniques reduce redundancy and improve answer quality.