Retrieval-Augmented Generation (RAG)

RAG is an architecture that combines information retrieval with large language models. Instead of relying only on model parameters, RAG retrieves relevant documents to answer queries.

Main components of RAG: - Document loader - Text splitter - Embedding model - Vector database - Retriever - Language model

RAG is widely used in enterprise chatbots and document Q&A; systems.