Retrieval-Augmented Generation (RAG) is an artificial intelligence technique that combines information retrieval with large language models.

RAG retrieves relevant documents from a knowledge base and uses them as context to generate accurate and grounded answers.

RAG helps reduce hallucinations by ensuring responses are based on retrieved factual information rather than purely model memory.

It is widely used in question answering systems, customer support bots, and enterprise search solutions.

A typical RAG pipeline includes document ingestion, chunking, embedding generation, vector storage, retrieval, and answer generation.