Retrieval-Augmented Generation (RAG) is a hybrid approach combining information retrieval and natural language generation. It enhances large language models (LLMs) by retrieving relevant document chunks from a knowledge base before generating responses. This method is particularly effective for tasks requiring factual accuracy, such as document summarization or question answering.

RAG systems typically involve two components: a retriever and a generator. The retriever uses vector embeddings to identify relevant text chunks, while the generator, often a pre-trained LLM like GPT or BART, produces coherent outputs based on the retrieved context. This combination reduces hallucinations and improves response quality.