Retrieval-Augmented Generation (RAG) is an approach that combines information retrieval with natural language generation. In RAG, the system first retrieves relevant document snippets from a knowledge base and then uses a language model to generate an answer based on the retrieved content.

RAG is useful in scenarios where the language model does not have access to all required information during training, such as answering queries about private or domain-specific documents.