Retrieval-Augmented Generation (RAG) is a technique that combines a retriever and a generator.

The retriever finds relevant documents from a knowledge base.

The generator, usually a language model, uses those documents as context to produce an answer.

RAG is useful when the model needs up-to-date or domain-specific information.