Retrieval-Augmented Generation (RAG) is a technique that combines retrieval-based approaches with generative models.

It enables language models to ground their responses in external knowledge by retrieving relevant documents before generating an answer.

RAG is useful in scenarios where the model needs to access up-to-date or domain-specific information without relying solely on pretraining.