Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is a technique in natural language processing (NLP) that combines the strengths of retrieval-based and generative models to generate high-quality responses to user queries. RAG integrates a retriever component, which retrieves relevant passages or documents from a large knowledge base, with a generative model, which generates responses based on the retrieved information. By leveraging both retrieval and generation, RAG models can produce coherent and contextually relevant responses that are grounded in factual knowledge from external sources. This approach enhances the quality and diversity of generated text compared to traditional generative models alone. RAG has applications in question answering, dialogue systems, content generation, and conversational agents, where generating informative and accurate responses is crucial.