

Retrieval Augmented Generation (RAG) is an architecture that combines the ability of large language models (LLMs) with a retrieval system to enhance the factual accuracy, contextual relevance, and quality of generated response against the query raised by user to a RAG system.

Traditional generative models rely solely on internal parameters for producing responses, which limits their ability to provide up-to-date or domain-specific knowledge. RAG mitigates this by augmenting the generation process with real-time retrieval from external knowledge sources.

Traditional generative models laid the foundation for today's LLMs. They helped us understand how to model processes represent knowledge, user input and generate data. However, they are now mostly replaced or augmented by deep learning-based transformer models, which offer greater accuracy, coherence, and scalability.