

Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is a modern technique used in natural language processing.

It combines information retrieval systems with large language models.

Instead of relying only on the internal knowledge of a language model, RAG first retrieves relevant documents from an external data source. These retrieved documents are then provided as context to the language model.

This approach helps reduce hallucinations and improves factual accuracy, especially in domain-specific applications such as healthcare, law, and finance.

RAG systems are widely used in question-answering systems, chatbots, and enterprise search applications.