LLM, RAG – All About RAG

What is it?

Retrieval-Augmented Generation(RAG) - These applications are used to integrate large language models (LLMs) into our own applications. RAG applications involve integrating LLMs, prompts, user data which can be our own (.pdf) files , and query history to ask the LLM for specific outputs or answers based on the user's query.

Why RAG applications?

RAG applications are mainly used to provide up-to-date information, providing the relevant information about specific topics and things, used to customize the Large language model to give Answers for the users on topics based on the given resource that we have fed to the RAG application.

RAG Advantages:

Increased Accuracy: RAG leverages reliable external data sources, reducing the risk of inaccuracies and "hallucinations" in generated content12.

Enhanced User Trust: By allowing language models to cite their information sources, RAG adds transparency and builds trust with users3.

RAG Disadvantages:

Dependency on High-Quality Data: RAG relies on high-quality external knowledge sources. Poor-quality data can negatively impact accuracy and diversity of outputs1.

Complex Implementation: Integrating RAG into workflows requires careful consideration and expertise4.

In summary, RAG offers accuracy and transparency benefits, but its success hinges on data quality and implementation considerations