Agentic RAG

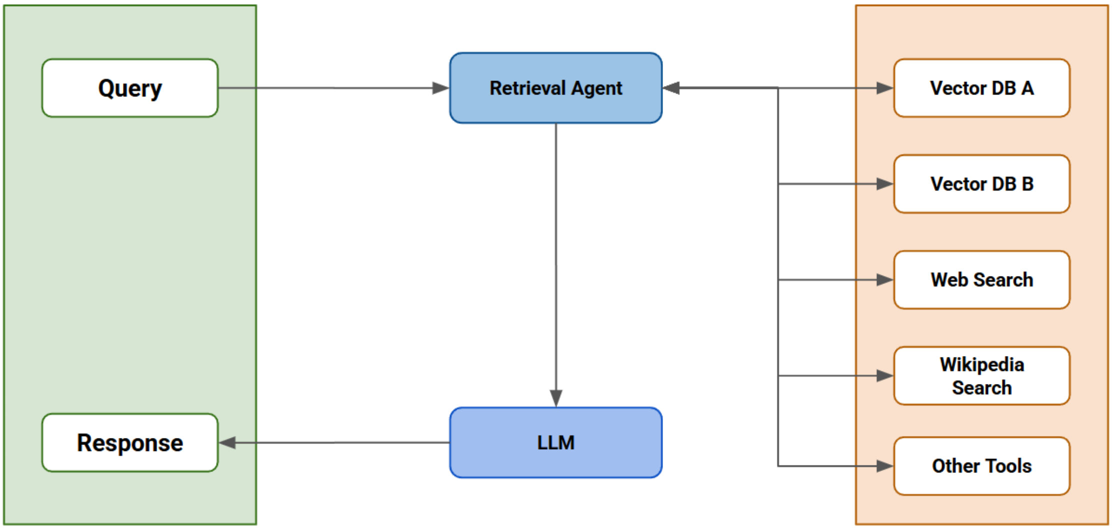
Agentic RAG

Agentic RAG (Retrieval Augmented Generation) is a framework that enhances traditional RAG systems by incorporating intelligent agents to handle complex tasks and make decisions dynamically.

Use an agent to figure out how to retrieve the most relevant information before using the retrieved information to answer the user's question

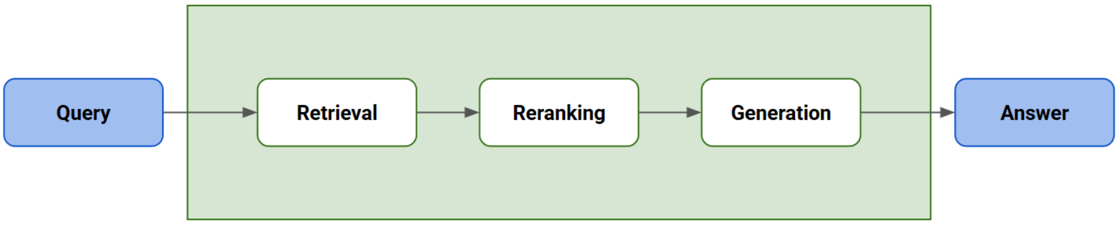
Retrieval Agents are useful when we want to make decisions about whether to retrieve from an index.

To implement a retrieval agent, we simply need to give an LLM access to a retriever tool.



Traditional RAG

Traditional RAG systems follow a relatively straightforward pipeline. A user query is vectorized (embedded) and used to retrieve relevant documents or data from a knowledge base (often via a vector similarity search). The retrieved context is then prepended or appended to the query, and an LLM generates a response using both the query and the retrieved context as input. The entire process is typically a single-pass retrieval followed by a single generation step (i.e., no iterative reasoning or multiple tool calls).


Query
Retrieval
Reranking
Generation
Answer