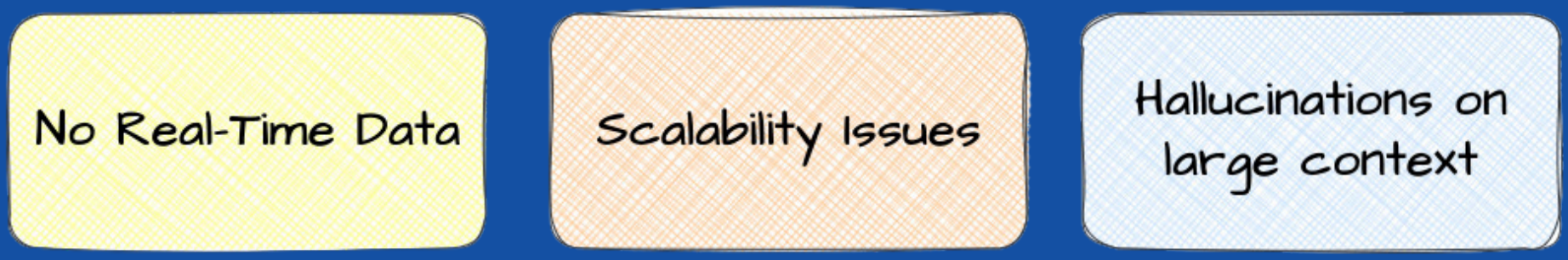
AI AGENTIC RAG

RAG and the problems with it

RAGs let you integrate your data easily and provide more contextual answers to your queries. However there are some limitations with traditional RAG approach:-

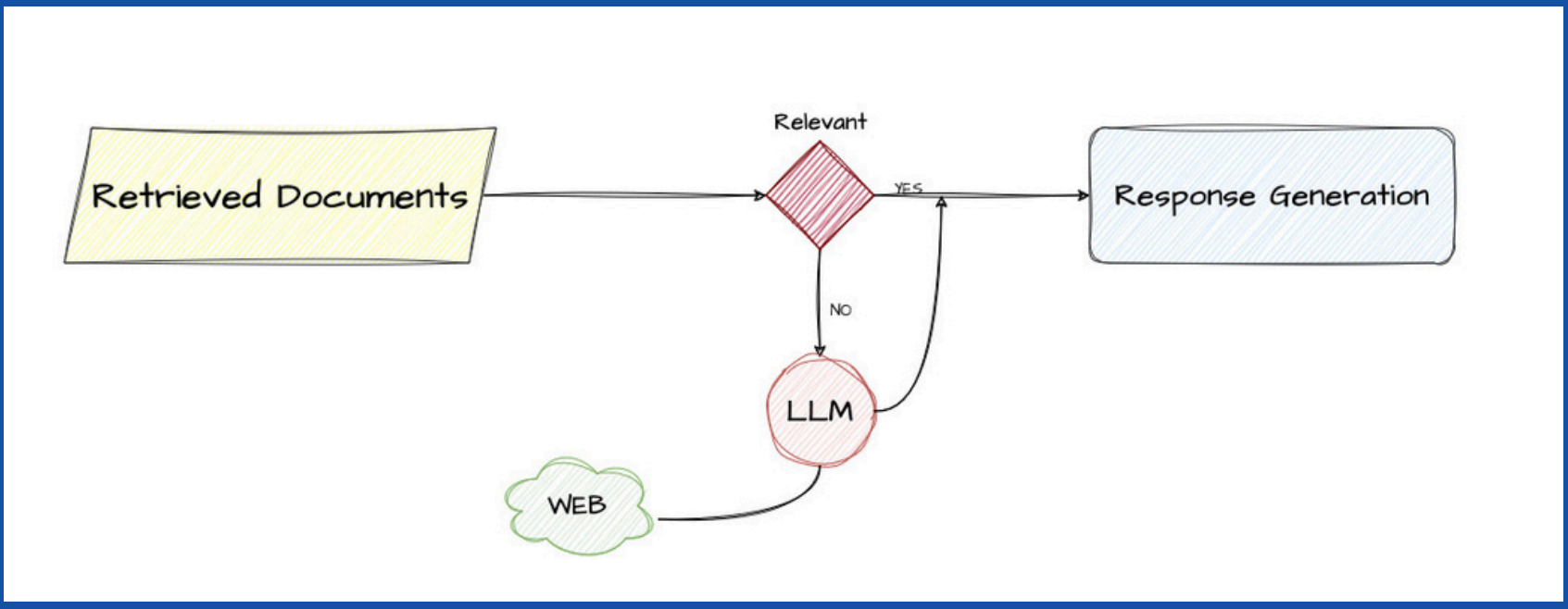


To address all these issues a new Agentic Corrective RAG System is proposed. It incorporates a document grading step to check the relevance of retrieved documents to the query.

Lets understand how it works it simple terms:

Corrective RAG

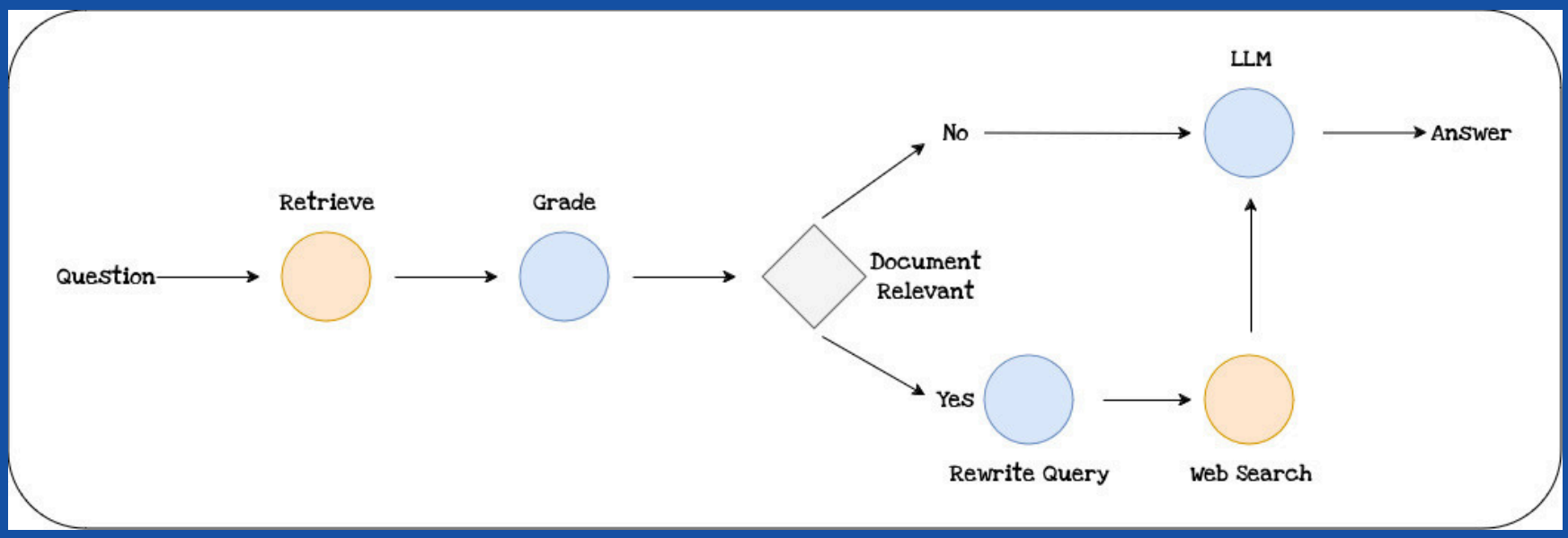
The key behind corrective RAG is to retrieve document chunks from the vector database as usual and then use an LLM to check if each retrieved document chunk is relevant to the input question.



Now if the retrieved document chunks are relevant then normal response generation is carried out. Otherwise , we rephrase the input query, search the web to retrieve new information related to the input question, and then send it to the LLM to generate a response.

Corrective RAG Agent

Below is a high-level workflow of the main components in an Agentic RAG system and the execution flow among these components.



Lets breakdown the diagram:

1. Each node here represents a process that is been carried out.
2. The first node retrieves the relevant documents from the vector database.
3. Grade node grades the document chunks whether the retrieved chunks are relevant or not.
4. The workflow divides in two based on the grading of the document chunks.

If the chunks are not relevant then an LLM is used to rewrite the user query to search on the web then the retrived documents fro the web along with the query are sent to an LLM to genrate final response.

Did you check the other 3 posts on RAG

Link to those in the comments below

Follow for more AI/ML posts