BHAVISHYA PANDIT

ADVANCED RAG METHODS

BHAVISHYA PANDIT

WHY RAG - CHALLENGES OF LLMS

Although LLMs are powerful still they have some challenges like:

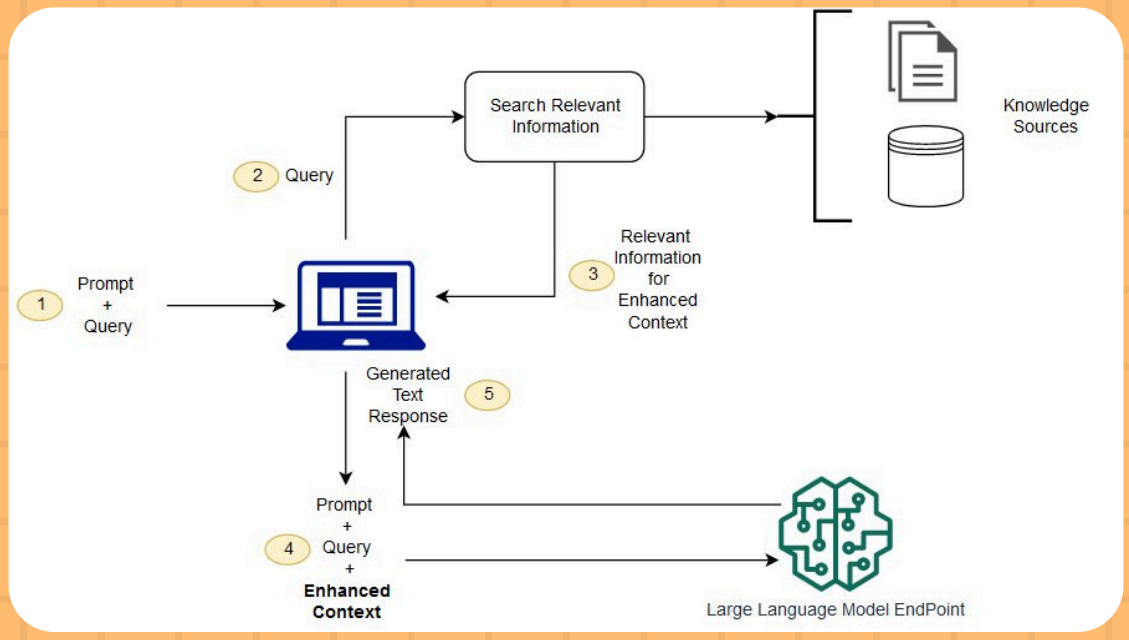
Generic Responses

Hallucinations

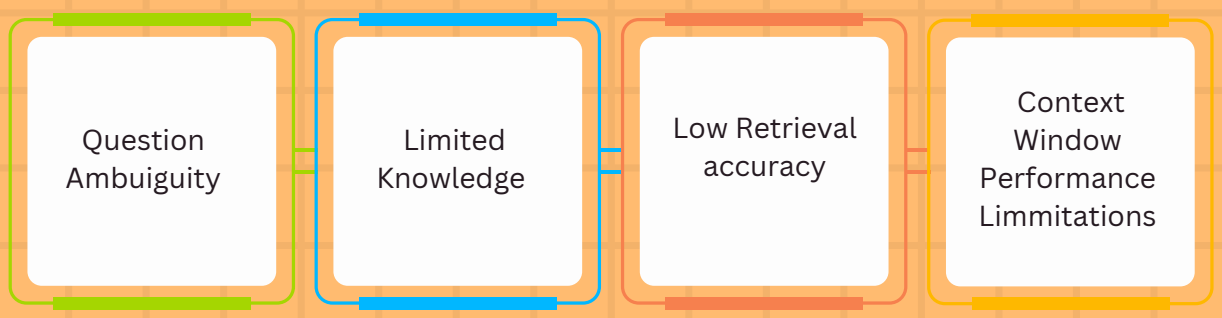
Lack of specific information

RAG is a technique that combines LLMs with external data resources to improve its capabilities.

ARCHICTECTURE OF A RAG



The naive implementation of RAG pattern is rarely enough to satisfy production grade requirements due to many factors like :

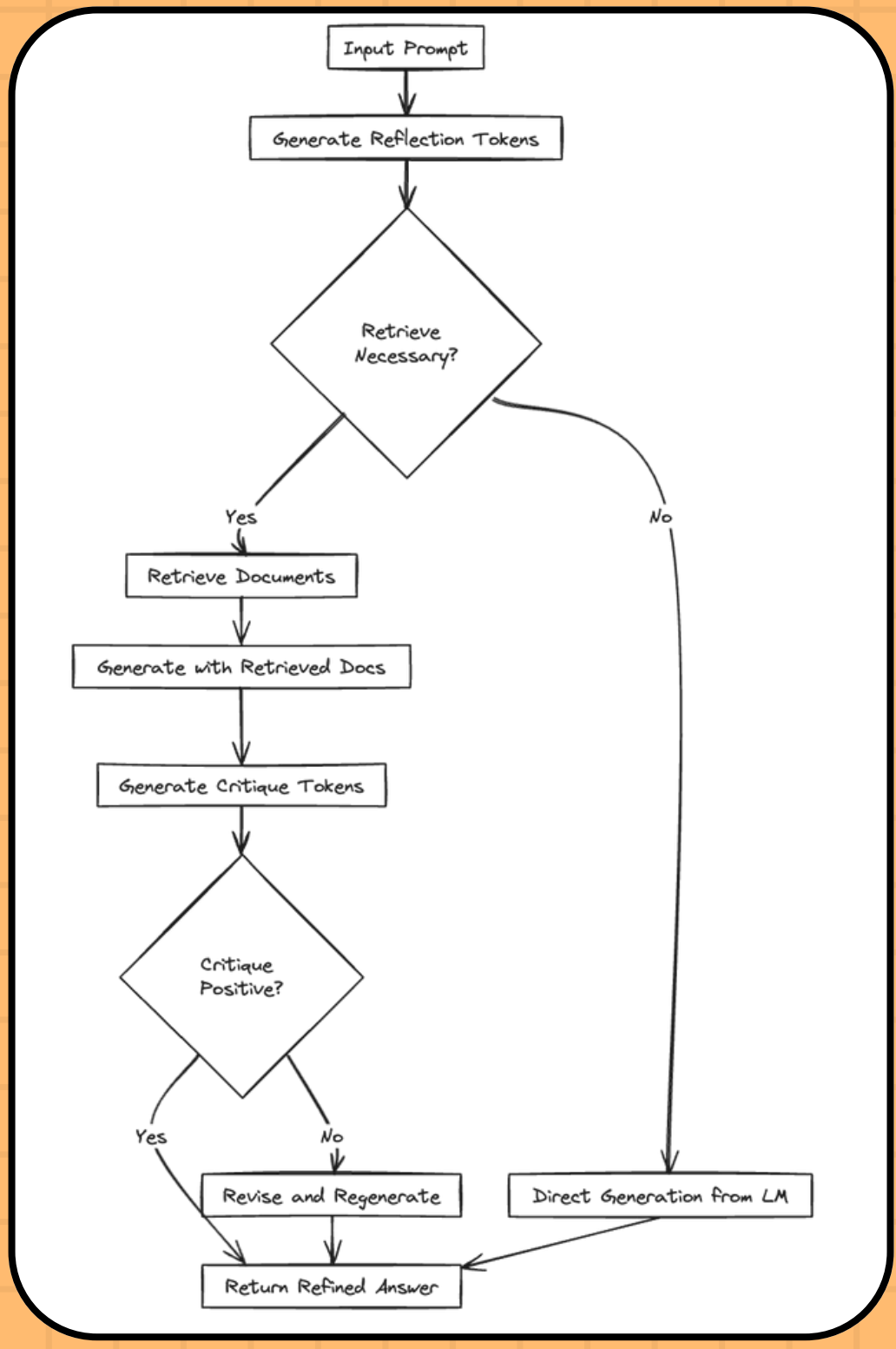


To address these issues some advanced RAG methods are used.

1 SELF REFLECTIVE RAG

The SELF-RAG paper describes fine-tuned model that incorporates mechanisms for adaptive information retrieval and self critique.

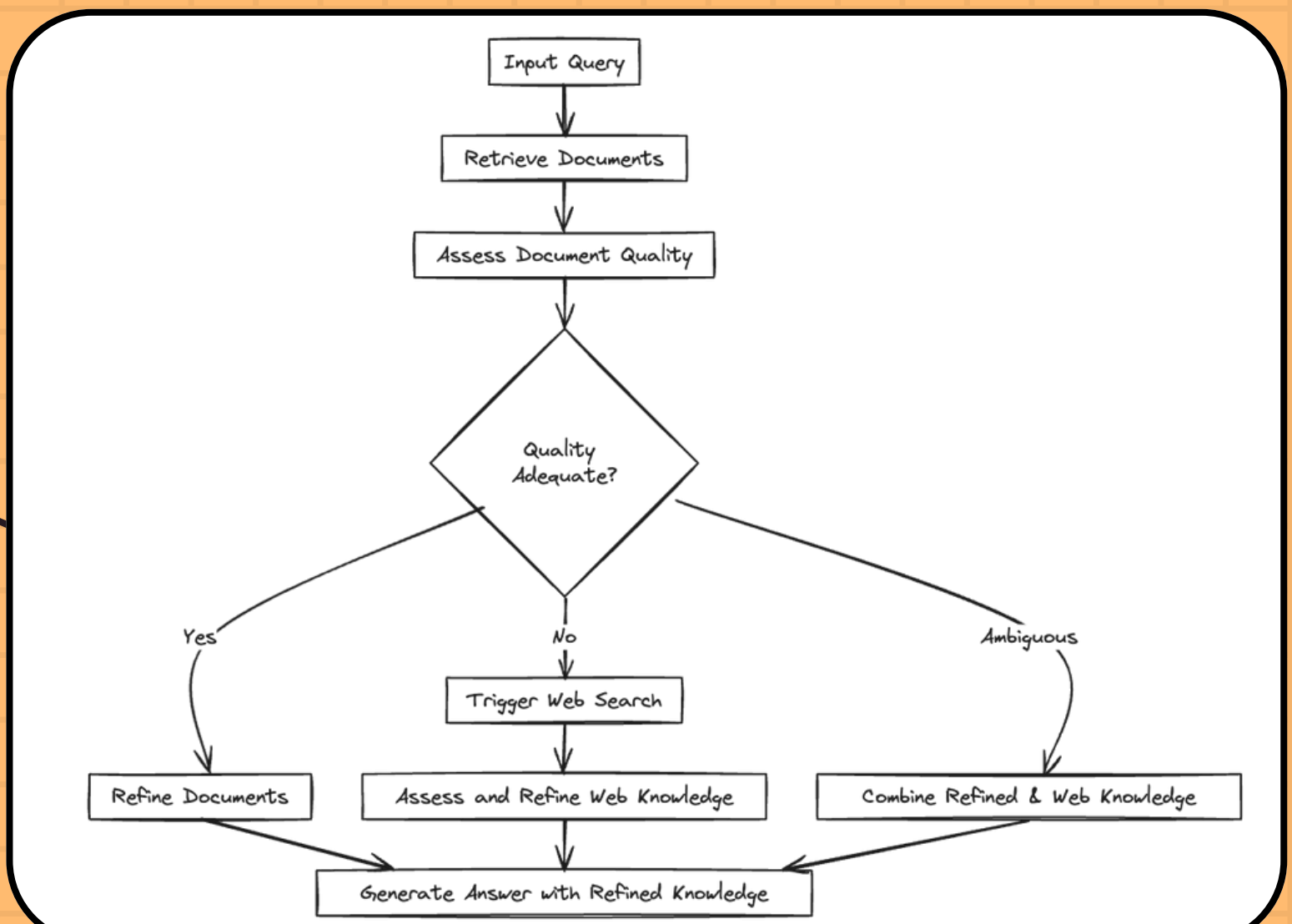
The model can dynamically determine when external information is needed and can critically evaluate its generated responses for relevance and factual accuracy.



2 CORRECTIVE RAG

Corrective RAG (CRAG) is a method that improves the accuracy of language models by intelligently re-incorporating information from retreived documents.

It uses an evaluator to assess the quality of documents obtained for a query. Then, it decides whether to use, ignore, or request more data from these documents.

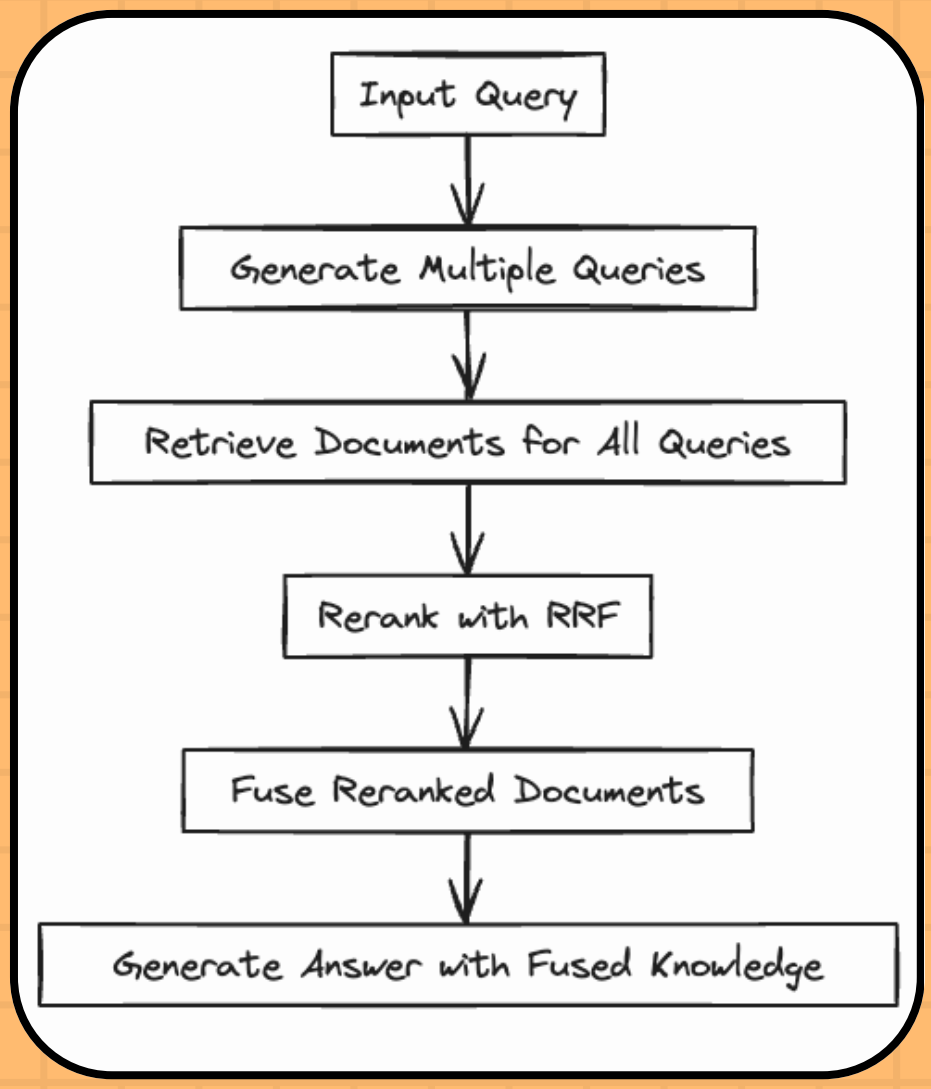


3 RAG FUSION

RAG Fusion method starts with generating multiple queries. Next, a vector search identifies relevant documents for both the original and derivative queries.

After document retrieval the Reciprocal Rank Fusion (RRF) algorithm reranks the documents based on their relevance. These documents are then combined to form a comprehensive and relevant information source.

In the final stage, this combined dataset and all queries are processed by a large language model.



Bhavishya Pandit

FOLLOW FOR MORE AI/ML POSTS!

Bhavishya Pandit