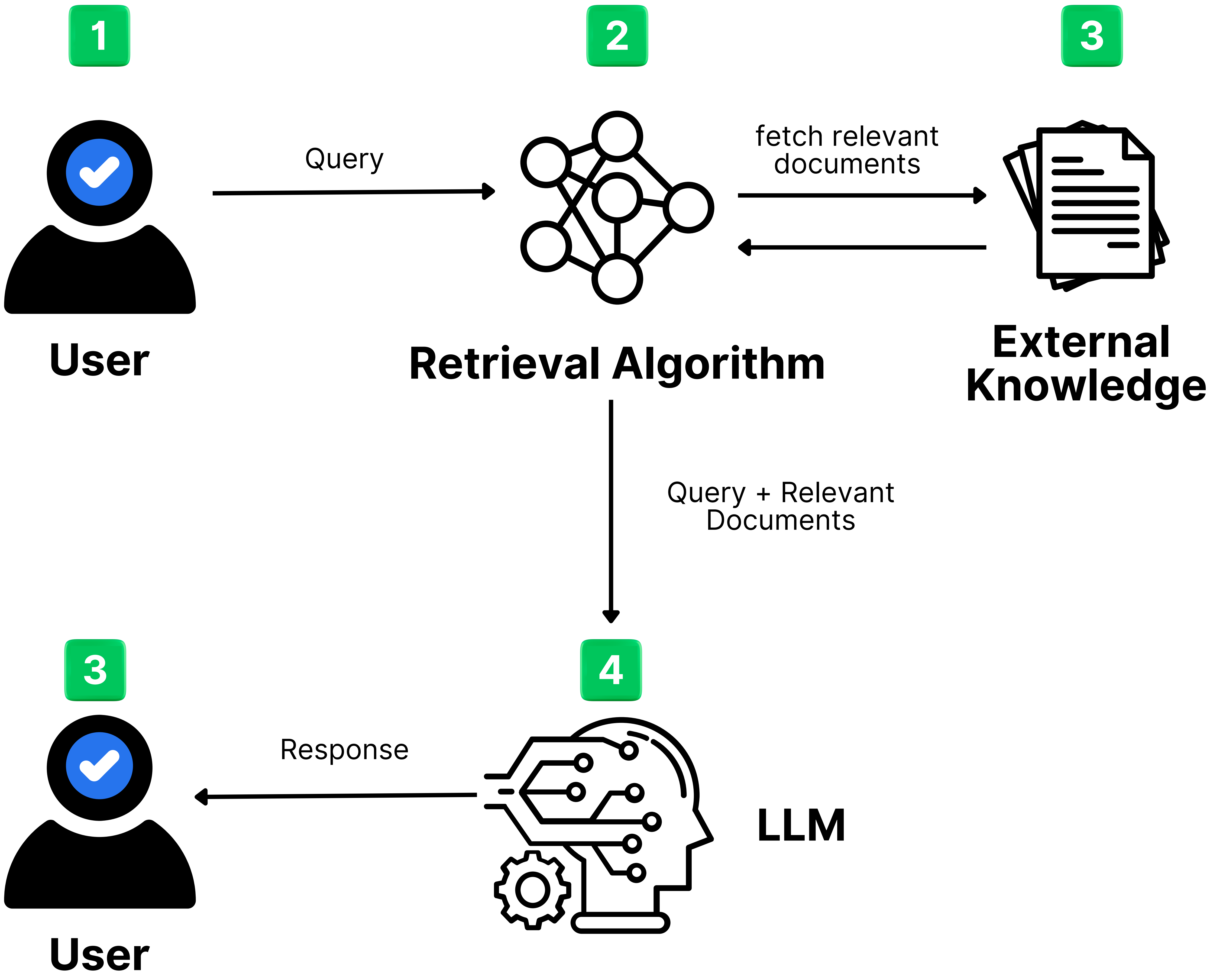
Analytics Vidhya

Mastering RAG

A Perfect Guide to Retrieval Augmented Generation



While building LLMs using Prompt Engineering, there are the following problems that occurs:

- ✓ **The model can be inconsistent.**
- ✓ **The model can hallucinate.**
- ✓ **The model is not up-to date.**

To overcome all these problems, we use one of popular method to build applications powered by LLMs which is

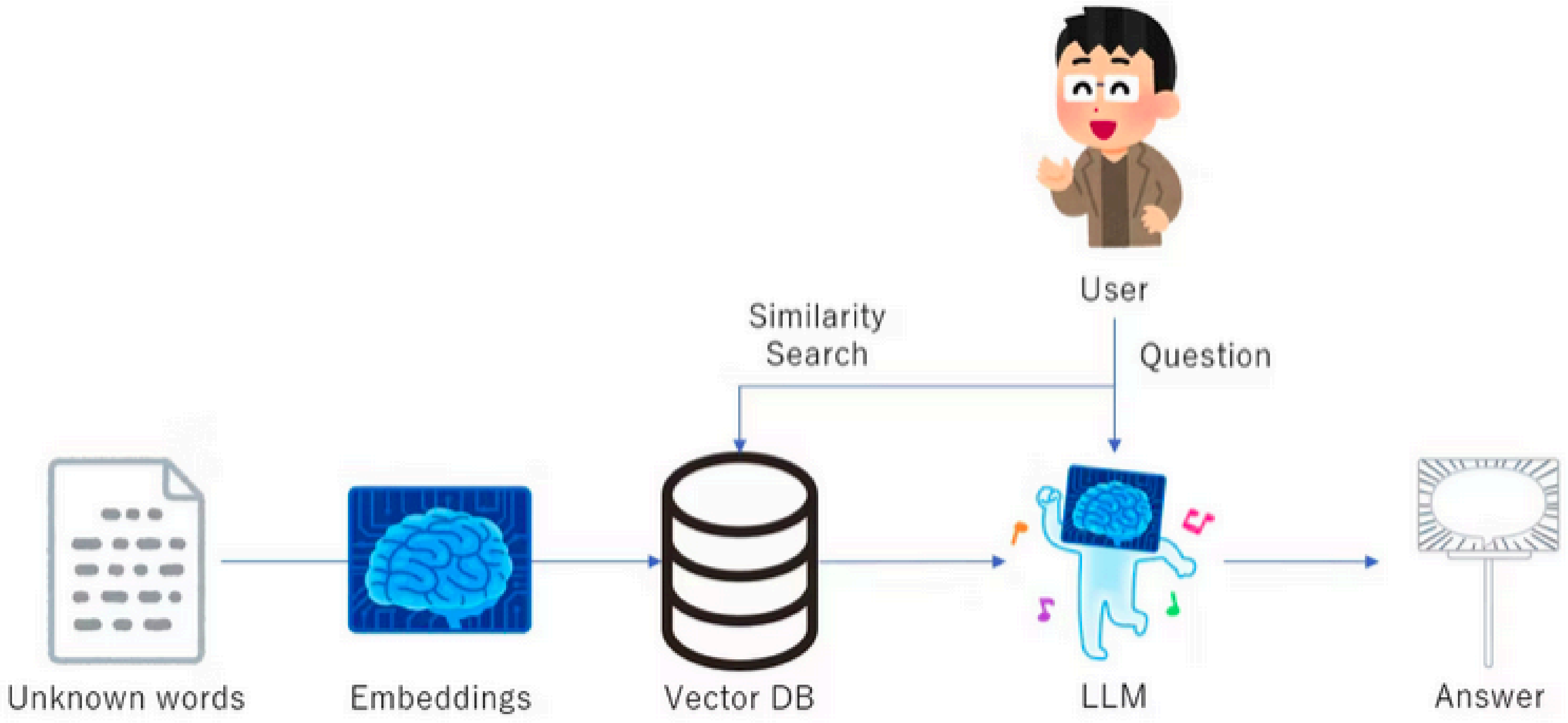
RAG

Ingest custom knowledge to LLMs

What is RAG?

- Retrieval-Augmented Generation (RAG) is an innovative approach in natural language processing that combines two primary components: a retrieval mechanism and a generative model.

- The retrieval component searches a large database of documents to find relevant information, which the generative model then uses to produce a coherent and contextually appropriate response.



How RAG Differs from LLMs?

- Traditional language models (LLMs) generate responses based solely on their training data and the input query.

- While they can be remarkably effective, they often struggle with providing up-to-date or specific information not present in their training data.

- On the other hand, RAG systems augment their generative capabilities with real-time retrieval of information, ensuring responses are fluent, factually grounded, and relevant.

The Importance of RAG

RAG systems are particularly useful in scenarios where up-to-date and specific information is crucial. Some notable applications include:

- **Customer Support**: Providing accurate and timely responses to customer queries by retrieving relevant information from a knowledge base.

- **Healthcare**: Assisting medical professionals with quick access to the latest research and clinical guidelines.

- **Education**: Offering detailed explanations and additional resources to students based on their queries.

Advantages

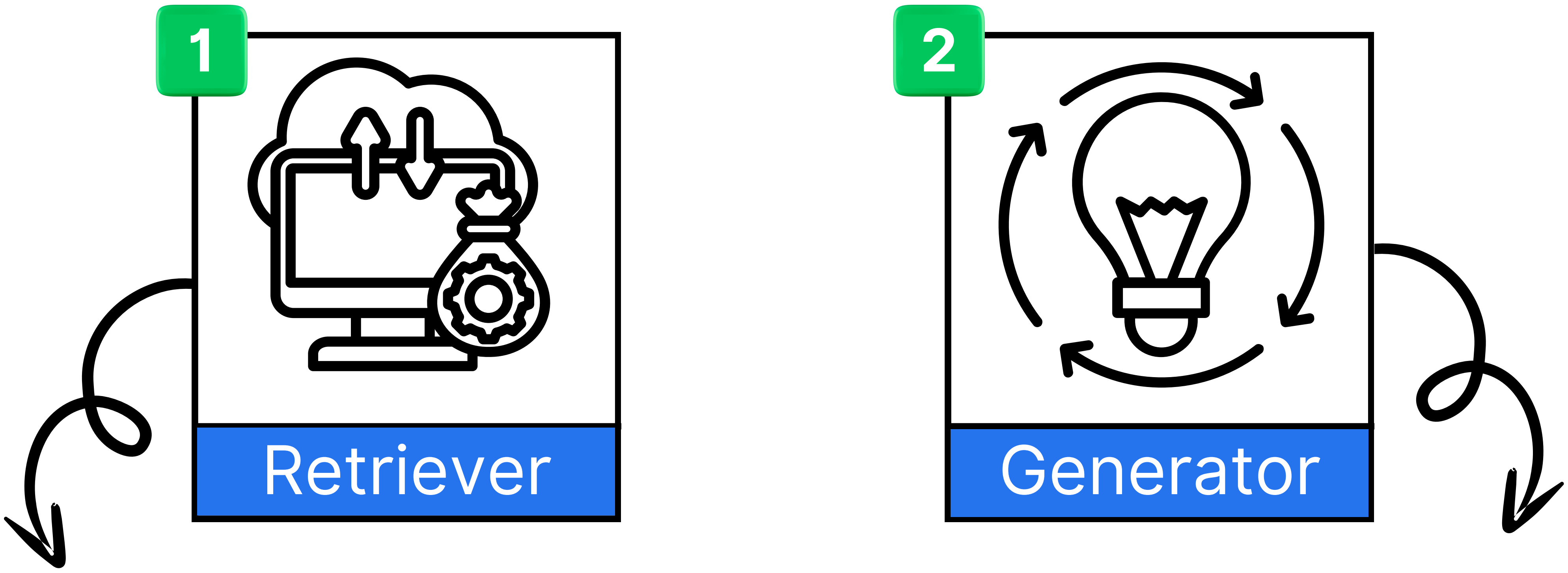
- RAG systems can generate more accurate and contextually appropriate responses by retrieving relevant documents.

- These systems can access the latest information, making them ideal for dynamic fields where knowledge constantly evolves.

- The retrieval mechanism ensures the generated responses are closely aligned with the user’s query, enhancing the overall user experience.

Component of RAGs

✓ There are basically two component in RAG:



1 Retriever

2 Generator

The retriever searches for and finds the best information from databases or websites, like a search engine that quickly pulls up relevant details.

The generator takes that information and combines it with what it already knows to create a clear, human-like response.

Working of a RAG

1

User

In the first step, the user ask the query to the LLM.

2

Retrieval Algorithm

The query is then sent to the **Retrieval algorithm**, that is responsible to fetch the relevant documents from the knowledge base.

3

External Knowledge

This **External knowledge base** is the source from where the Retrieval algorithm is fetching the relevant documents.

4

LLM

The retrieved documents, along with the original query, are sent to the language model (LLM).

5

User

The generator processes both the query and the relevant documents to generate a response, which is then sent back to the user.

Analytics Vidhya

FREE COURSE

RAG

We are offering a FREE Course on “Building Your first RAG System using LlamaIndex”