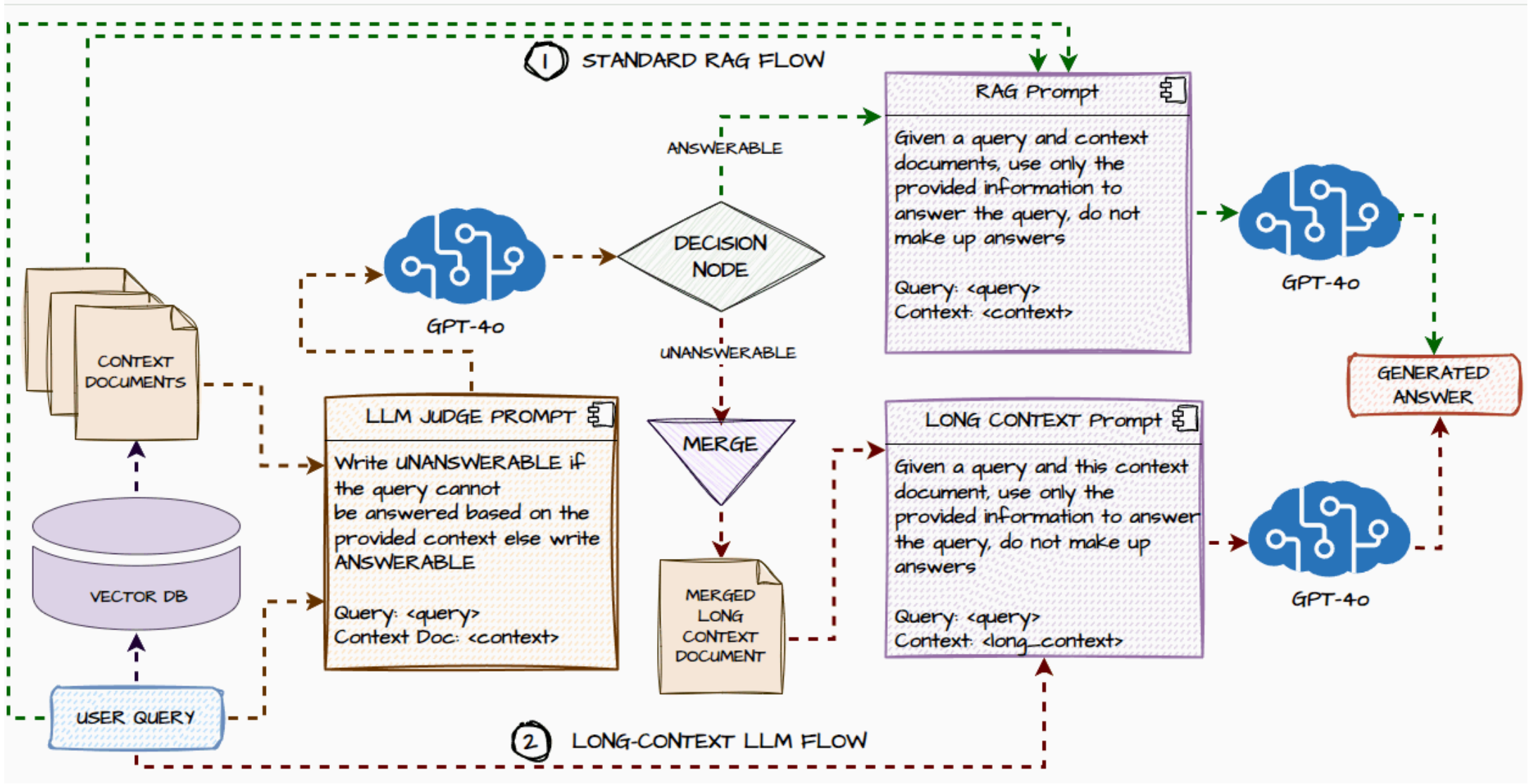
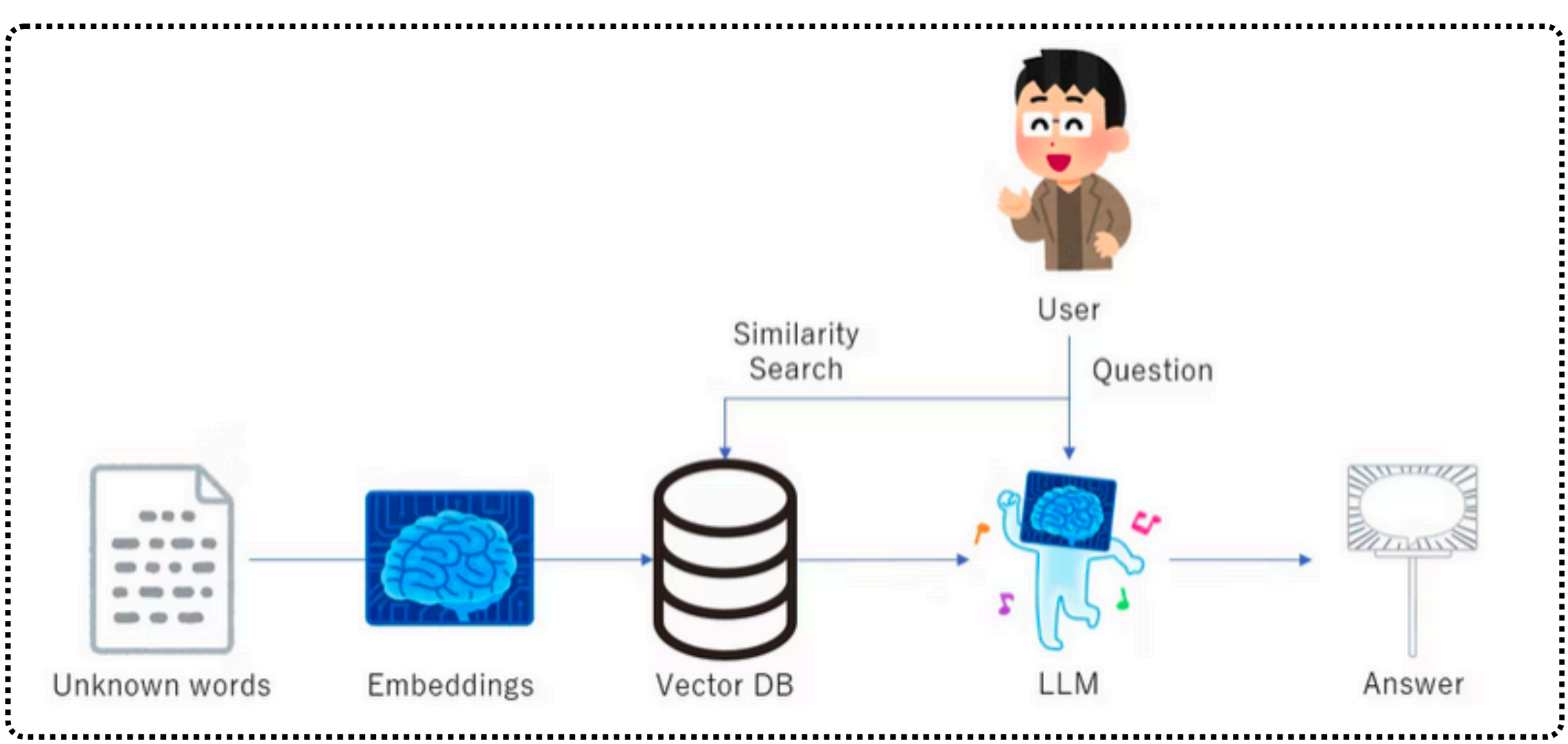
Analytics Vidhya

Mastering RAG

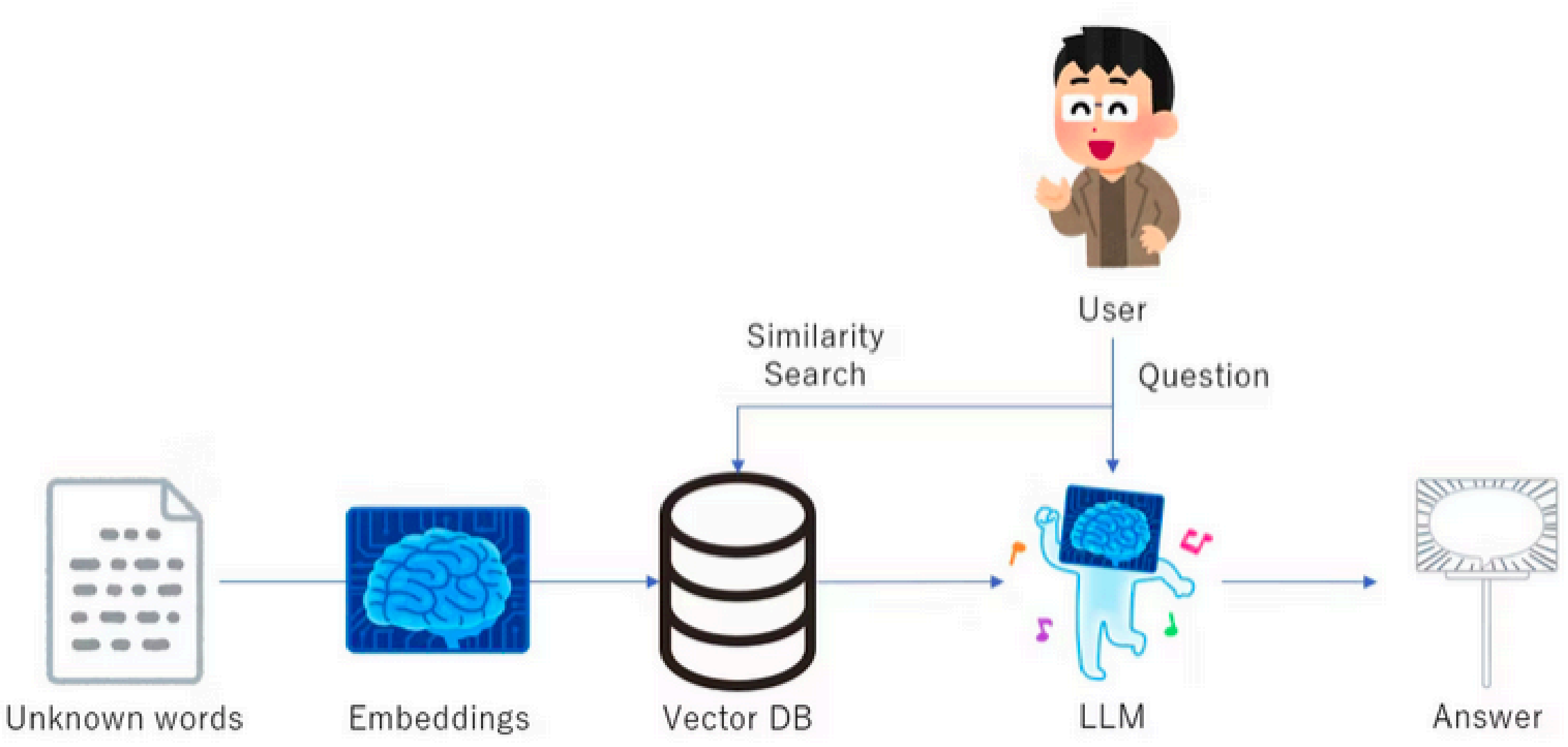
How does RAG work?





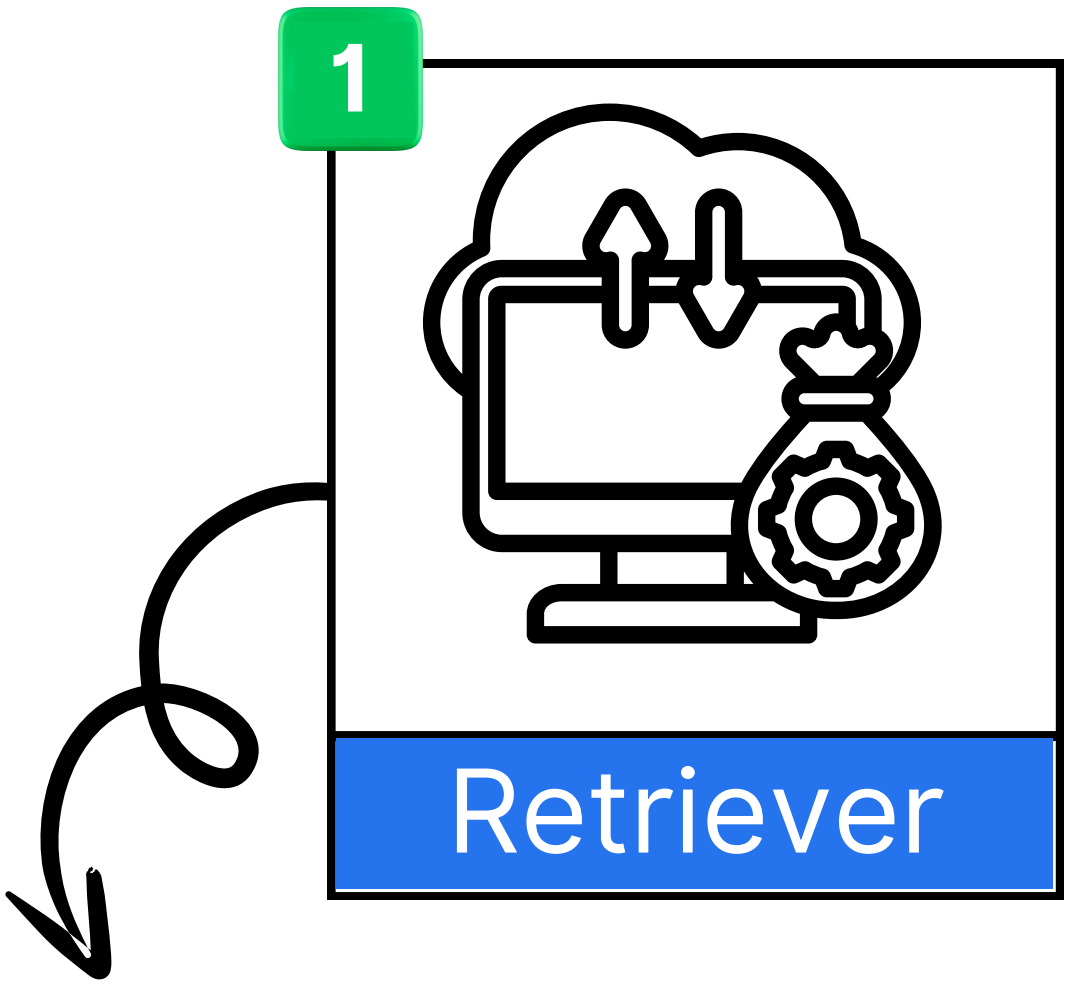
Introduction

- Retrieval-Augmented Generation (RAG) is a hybrid AI framework that enhances the capabilities of large language models (LLMs) by integrating information retrieval with generative text generation. This approach enables models to generate more accurate, informed, and up-to-date responses by fetching relevant external data during inference.

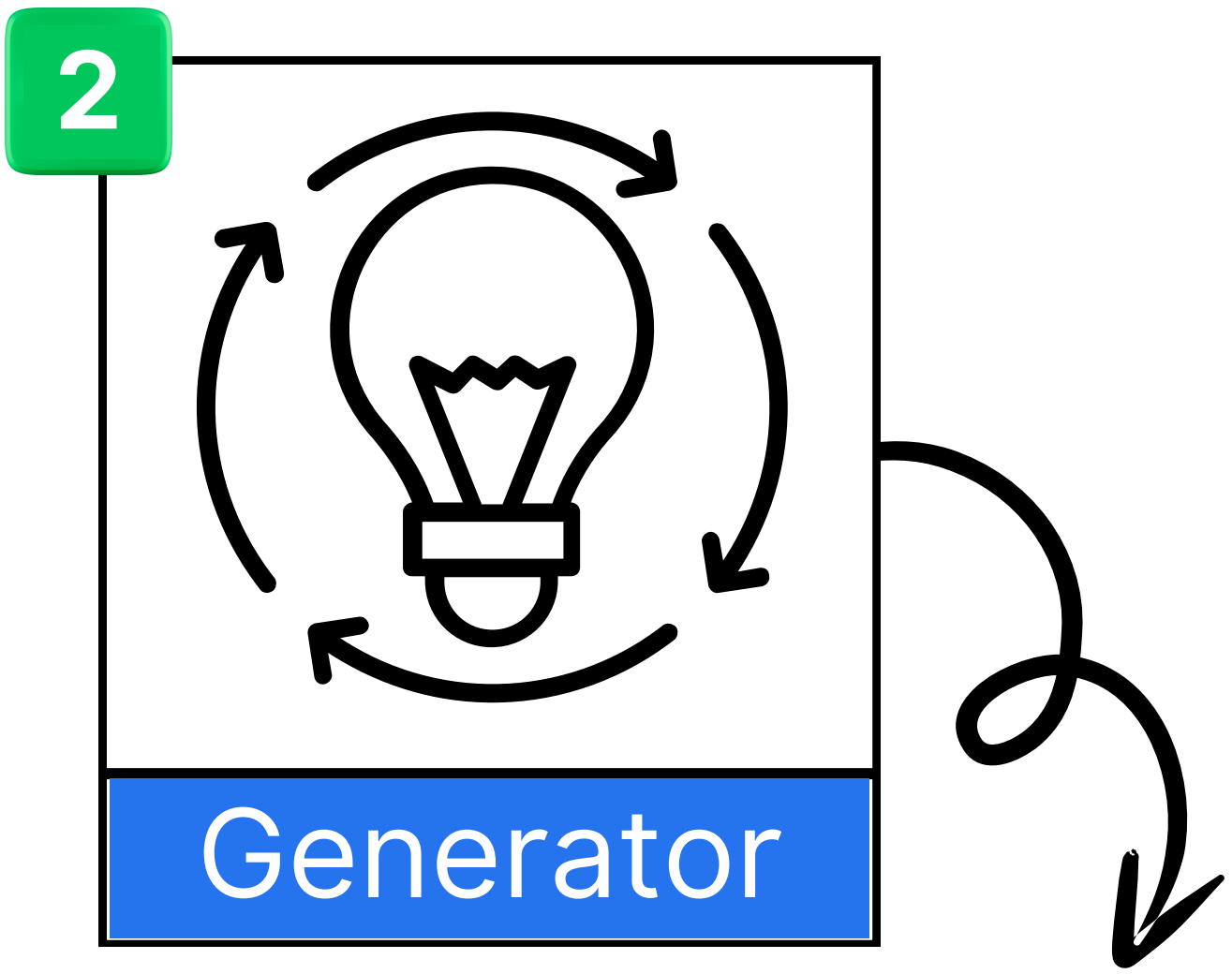


Component of RAGs

There are basically two component in RAG:



- Responsible for fetching relevant documents or data from an external knowledge base.
- Typically uses vector search, BM25, or dense passage retrieval (DPR) to find the most relevant information.
- The retrieval system can leverage structured (databases, APIs) and unstructured (documents, PDFs, web pages) sources.



- A language model (LLM) that takes both the query and retrieved context as input.
- Generates responses by synthesizing retrieved information and leveraging its pre-trained knowledge.
- Uses transformer-based architectures like GPT, BERT, or T5 for coherent text generation.

Working of a RAG

User Query Input

- The user inputs a question or query.

Retrieval Process

- The retriever searches for relevant documents or snippets from an external knowledge base (e.g., vector database, indexed documents, or online sources).

Ranking and Filtering

- Retrieved documents are ranked based on relevance scores.

- Irrelevant or redundant information is filtered out.

Fusion with LLM Context

- The retrieved data is fed into the language model along with the original user query.

Text Generation

- The model generates a response using both retrieved knowledge and its internal pre-trained information.

Response Delivery

- The final output is presented to the user with an informed and contextually rich response.

Advantages of RAG

- **Improves Accuracy**: Enhances responses by incorporating external, updated knowledge.
- **Reduces Hallucination**: Mitigates misinformation by grounding answers in real-world data.
- **Domain-Specific Adaptability**: Works well for specialized knowledge areas like healthcare, legal, and finance.
- **Efficient Information Retrieval**: Speeds up knowledge access without requiring extensive fine-tuning of the LLM.