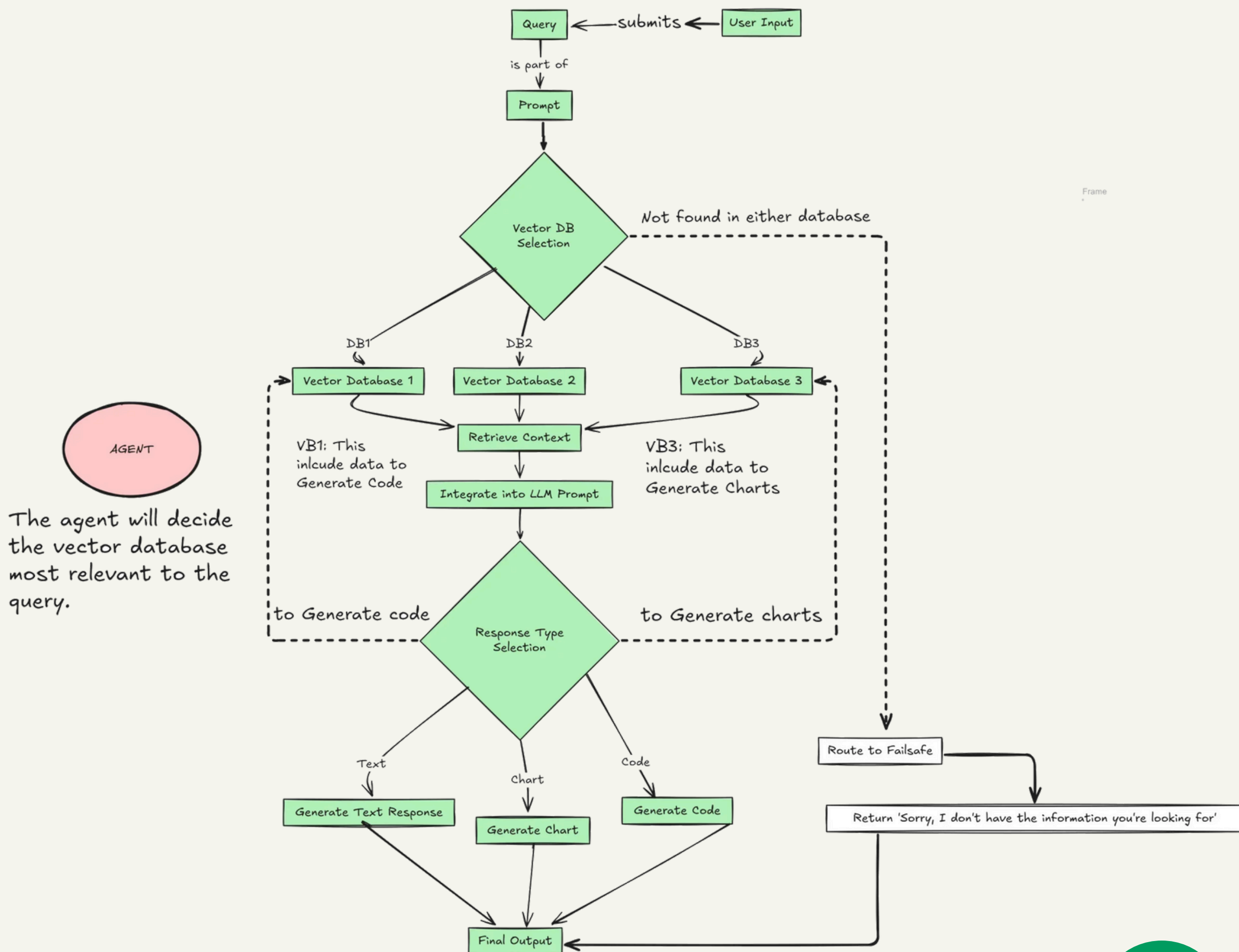
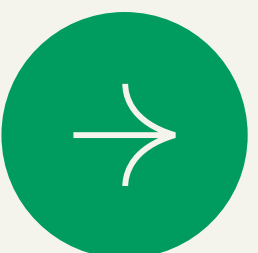




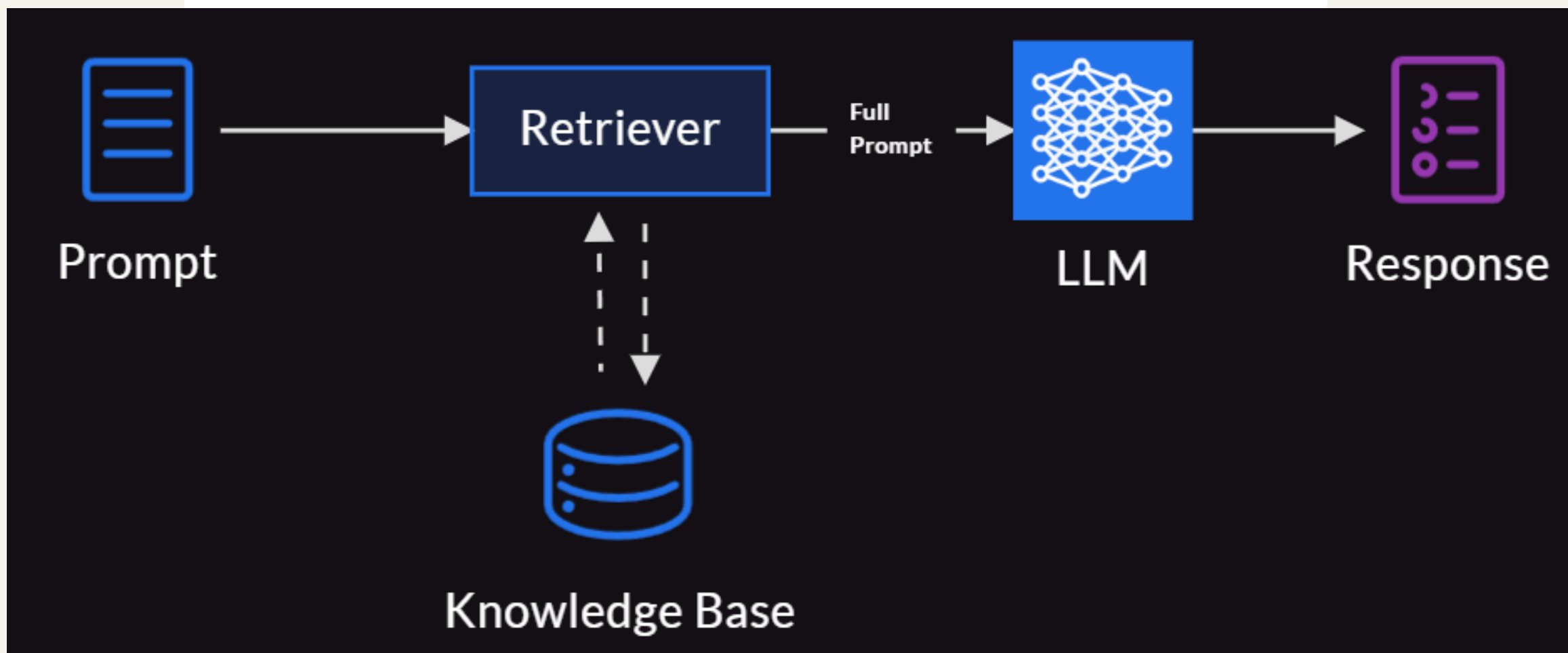
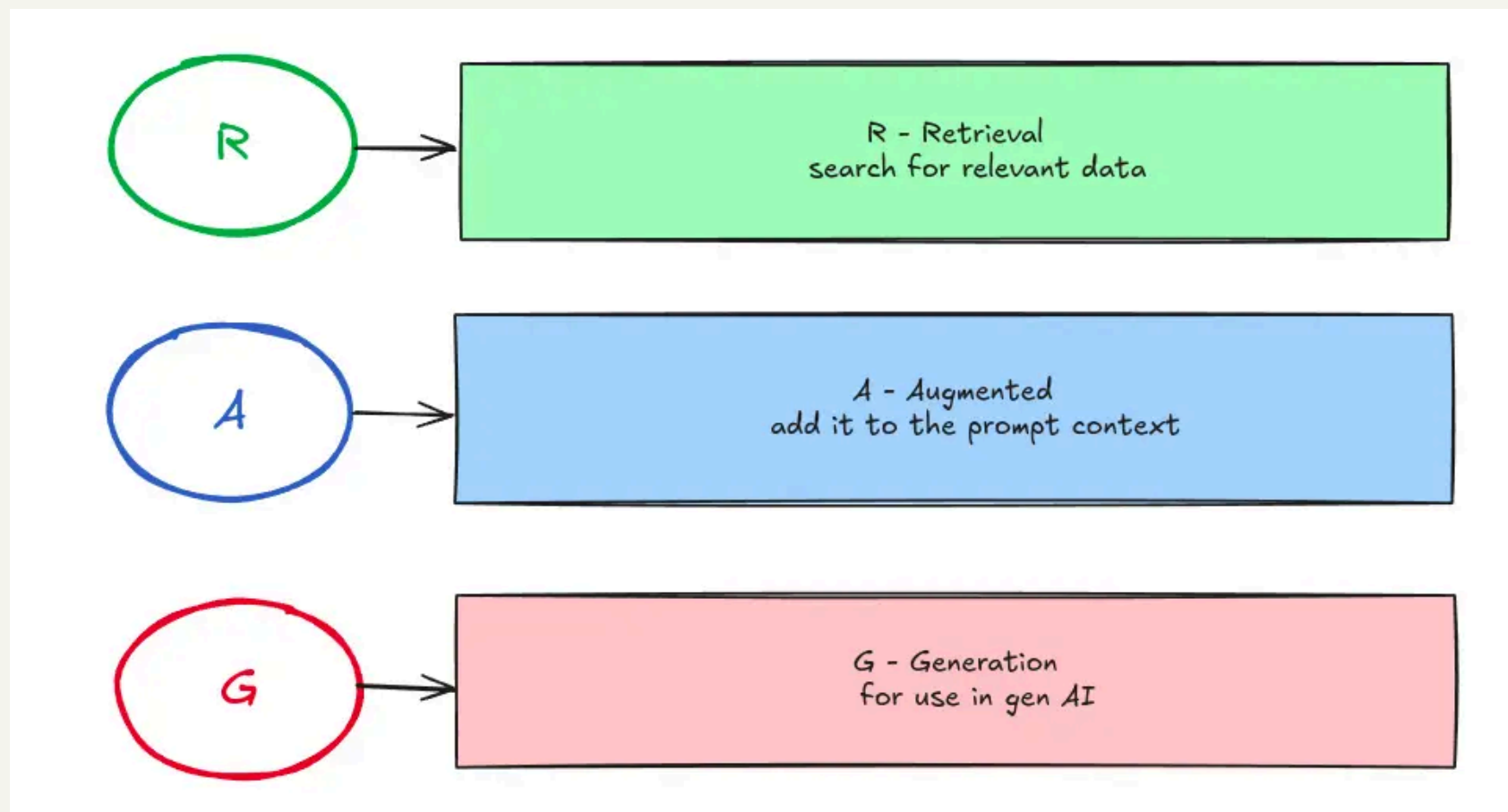
Guide to RAG vs. Agentic RAG Systems



The agent will decide the vector database most relevant to the query.

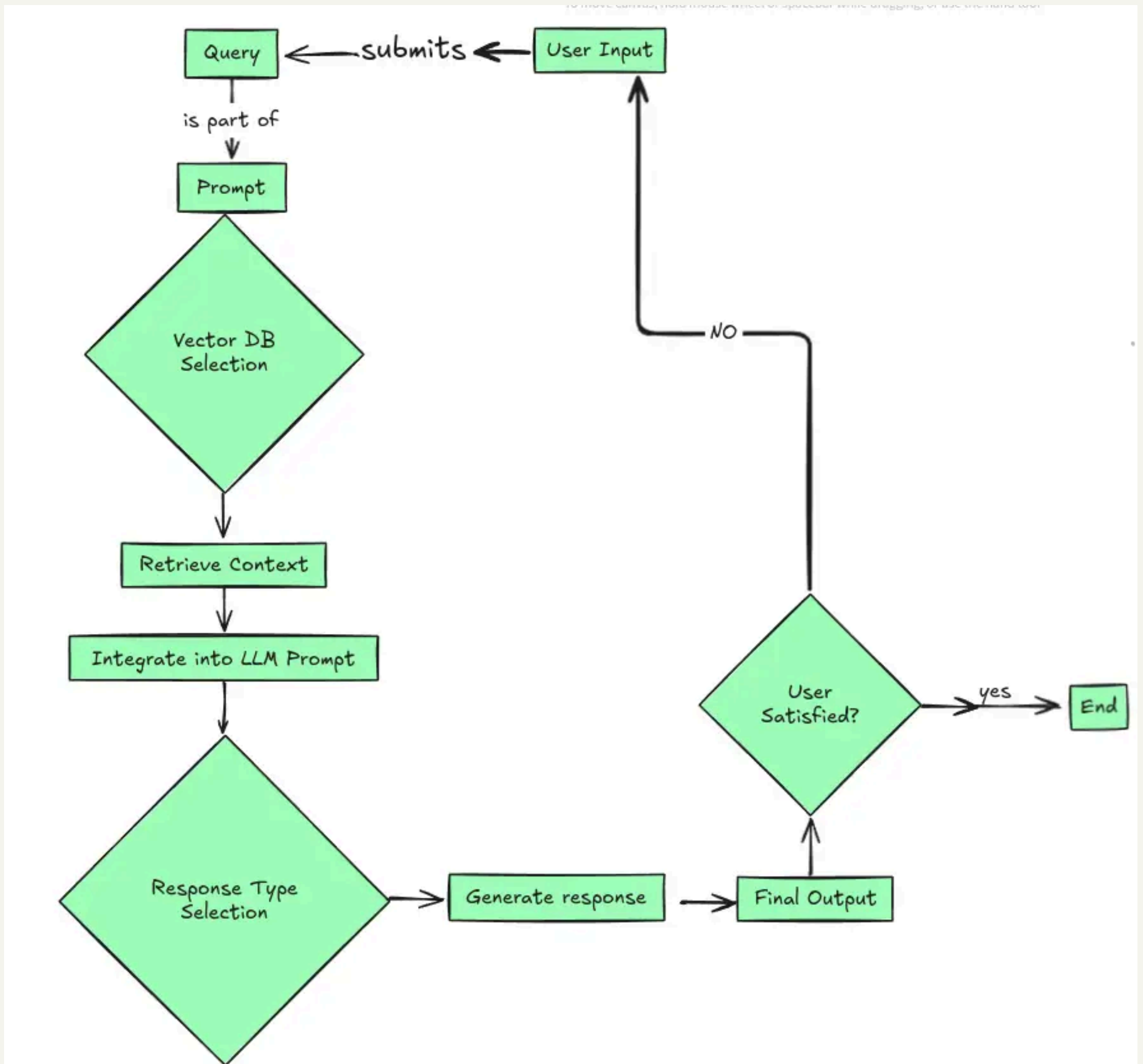


What is a RAG System?



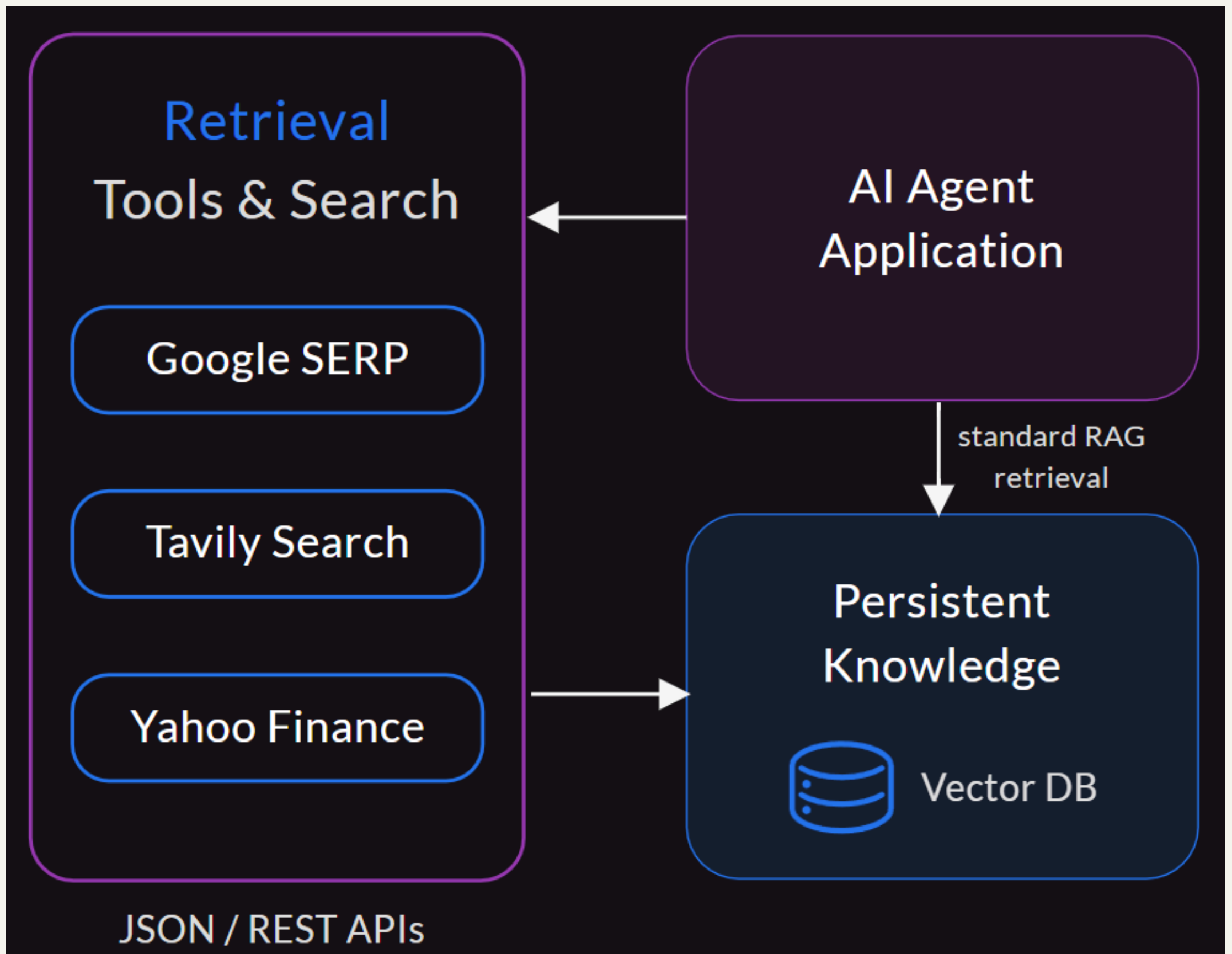
- RAG connects an external knowledge base to augment the existing knowledge of a LLM
- RAG leverages a vector database to first retrieve relevant context for a query and makes the LLM use this context to answer queries
- RAG is beneficial in situations requiring the latest information or answers involving custom enterprise data on which the LLM was never trained

RAG System Workflow



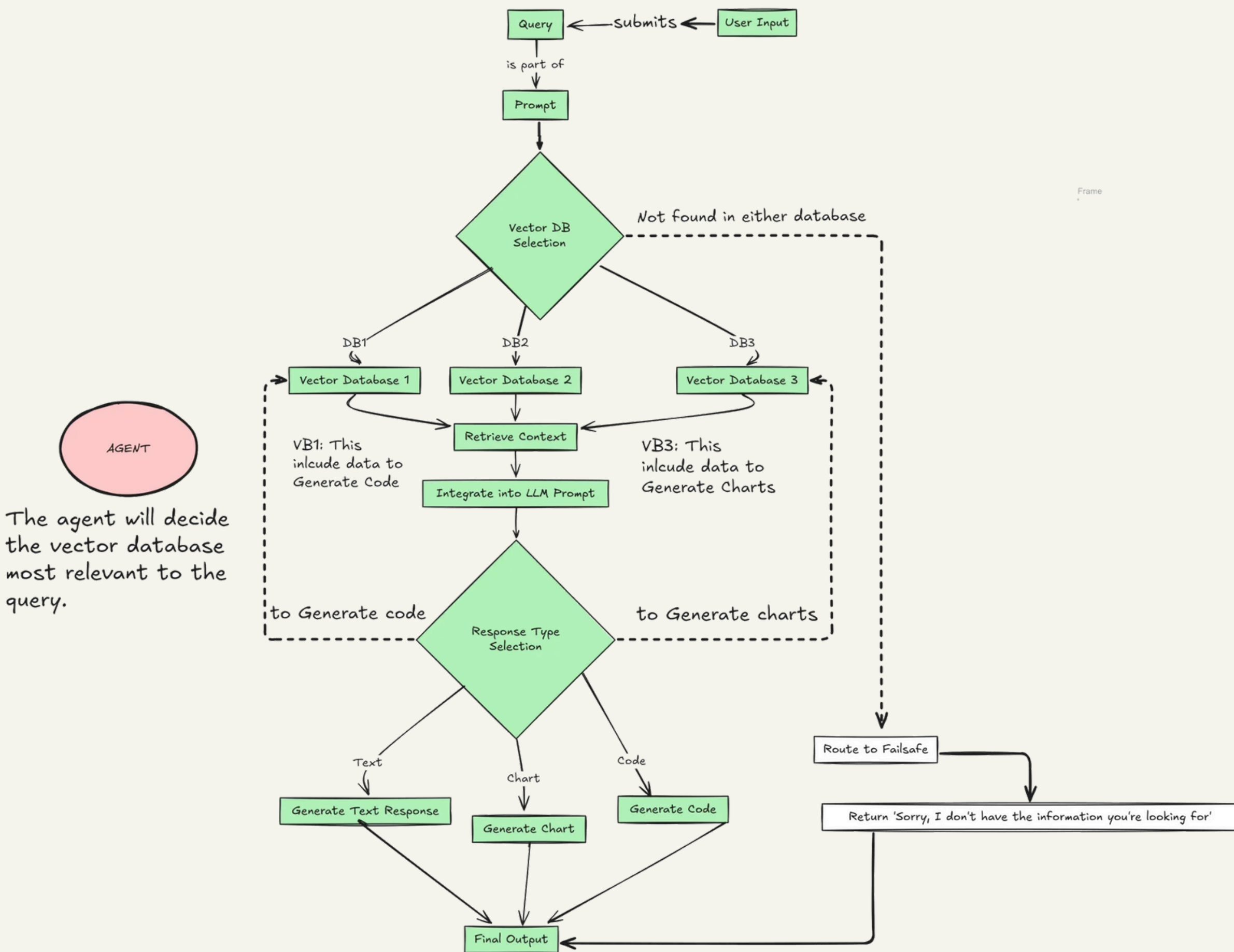
- Documents are processed, chunked, embedded and stored in a Vector DB
- User query is embedded and relevant context documents are retrieved from the DB
- Instruction prompt along with query and context is sent to the LLM to generate a contextual response

What is Agentic RAG?



- Agentic RAG is a combination of AI Agents and RAG Systems
- Leverages retrieval and search tools to access live real-time data besides the vector database
- Can be extended to add in multiple levels of complex flows to validate retrieval, response generation and check for hallucinations
- Examples include Agentic Corrective RAG, Self-Reflective RAG and more

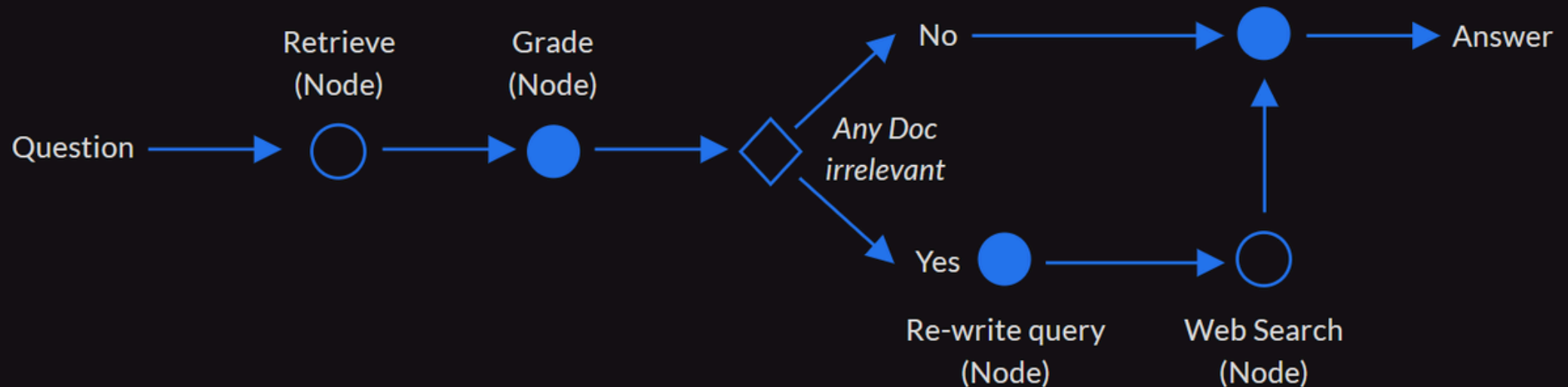
Agentic RAG Workflow



- Agentic RAG Systems have various workflows depending on the use-case
- In this workflow we first create various vector databases based on specific document types and domains
- Based on the user query the LLM will reason and route to the relevant Vector DB
- Context documents are retrieved and the standard RAG flow is executed after that as usual
- Very useful when you have documents related to different domains, departments

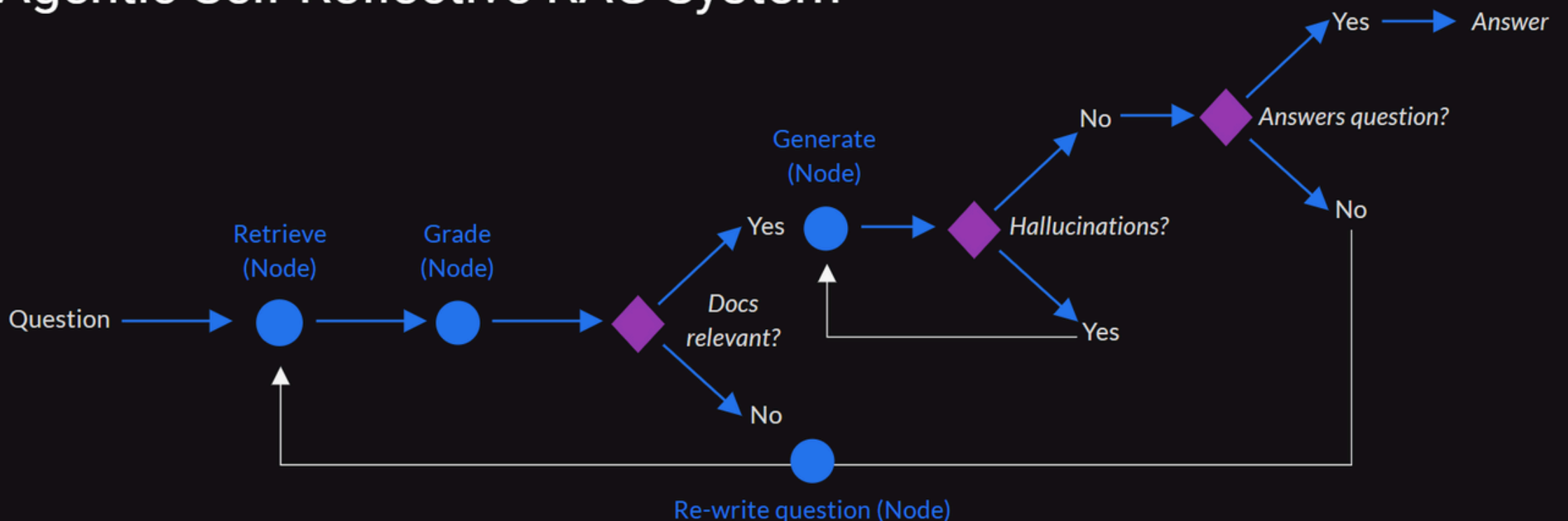
Popular Agentic RAGs

Agentic Corrective RAG System







- Use Agentic flows and build a graph-based network
- Utilize a powerful language model to assess if the retrieved context from the vector database is sufficient to answer the query.
- If it's applicable, follow the standard RAG flow; otherwise, use web search tools to obtain live contextual information to answer the query.




Agentic Self-Reflective RAG System



- Use standard vector database retrieval for context based on query
- Leverages agentic reflection pattern to use an LLM to reflect on the context and check for relevancy
- Also checks for hallucinations and if the question is answered using the same pattern to make the system more accurate

RAG vs. Agents vs. Agentic RAG

Feature		RAG	Agents	Agentic RAG
	Key Role	Combines LLMs with external data retrieval to generate responses	Combines LLMs, tools, and instructions for autonomous task management	Enhances RAG by using agents for intelligent retrieval, response generation, grading, critiquing, and more
	Real-Time Data Retrieval	Not possible in native RAG	Not a core feature but possible with tools	Designed for real-time data retrieval and integration
	Integration with Retrieval Systems	Tied to static retrieval from pre-defined vector databases	Not specifically tied to retrieval, can work with search tools	Deeply integrated with diverse retrieval systems, agents control the process
	Context-Awareness	Limited by the static vector database, no advanced or real-time context-awareness	Moderate, based on the agent's logic and tools	High, agents adapt to user query and retrieve context, including real-time data

Feature		RAG	Agents	Agentic RAG
	Task Complexity	Handles simple query-based tasks but lacks advanced decision-making	Handles complex, multi-step tasks with multiple agents and tools working in coordination if needed	Handles complex, multi-step tasks with multiple tools and agents as needed for retrieval, reasoning, answering, grading, and more
	Decision-Making	Limited, no autonomous decision-making involved	Autonomous decisions based on environment and task, not tied to data retrieval	Agents autonomously decide what data to retrieve, how to retrieve, grade, reason, reflect, and generate responses
	Multi-Step Reasoning	Limited to single-step queries and responses	Capable of multi-step reasoning if designed for complex tasks	Excels at multi-step reasoning, especially after retrieval with grading, hallucination, and response evaluation

- **Agentic RAG systems handle complex, multi-step tasks with multiple tools and agents as needed for retrieval, reasoning, answering, grading, and more**
- **Additionally, Agentic RAG systems excel at multi-step reasoning, especially after retrieval with grading, hallucination, and response evaluation**

Detailed Article

[Free Courses](#)[Learning Paths](#)[GenAI Pinnacle Program](#)[Agentic AI Pioneer Program](#)New[Interview Prep](#) [Career](#) [GenAI](#) [Prompt Engg](#) [ChatGPT](#) [LLM](#) [Langchain](#) [RAG](#) [AI Agents](#) [Machine Learning](#) [Deep Learning](#) [GenAI Tools](#) [LLMOps](#) [Python](#) [NLP](#)[Home](#) > [RAG](#) > [RAG vs Agentic RAG: A Comprehensive Guide](#)

RAG vs Agentic RAG: A Comprehensive Guide

[Pankaj Singh](#)

Last Updated : 27 Nov, 2024

21 min read



Today, I am discussing RAG vs Agentic RAG. In this guide, I will provide you with the comparison and then proceed to the hands-on part.

Firstly, let's understand what RAG is. It is not a piece of old cloth but the framework LLM uses to get **relevant, up-to-date, and context-specific information by combining retrieval and generation capabilities.**

But can we see the limitations of LLMs without RAG? Absolutely! Here, I have asked ChatGpt to give me output on its knowledge without any external searches for Swarm by OpenAI; it cannot provide the right output. This is due to its knowledge cutoff date, which is 2023, and to get the correct output, it has to be updated with new information or access to an external source. Intriguing, right? So, can we augment the LLMs with our own custom data to get the right response? Of course, we can do it with long-context LLMs and **RAG**. Today, we will be talking about RAG.

based on your knowledge without any external searches can you tell me what is SWARM by OPENAI

< 2/3 >



As of my last update, **SWARM** by OpenAI does not appear to be a known product, project, or framework released by OpenAI. It's possible that you're referring to:

1. **Swarm Intelligence:** A concept in artificial intelligence inspired by the collective behavior of decentralized systems, such as ant colonies or bird flocks, which may relate to advanced AI systems like those OpenAI develops.
2. A **new initiative or framework** by OpenAI that was introduced after my last knowledge update in late 2023.

If you have specific details or context about SWARM, feel free to share, and I can provide insights or relate it to existing AI or machine learning concepts!

Instead of relying solely on the [large language model's \(LLM\)](#) pre-trained knowledge, which may be outdated or incomplete, RAG dynamically retrieves the most relevant documents or information from an external knowledge base or database.

CHECK OUT THE
DETAILED ARTICLE
HERE