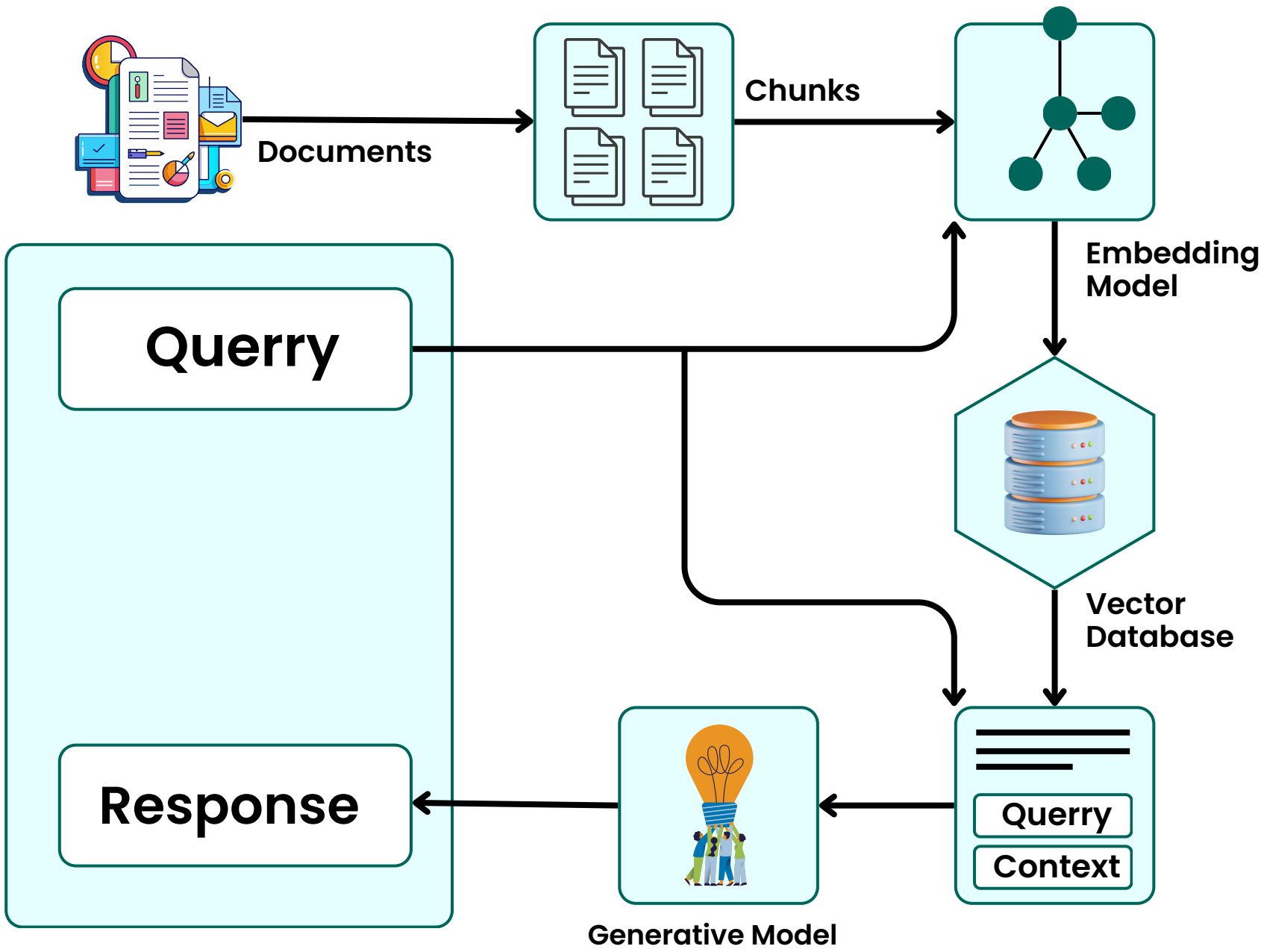
THE RAVIT SHOW

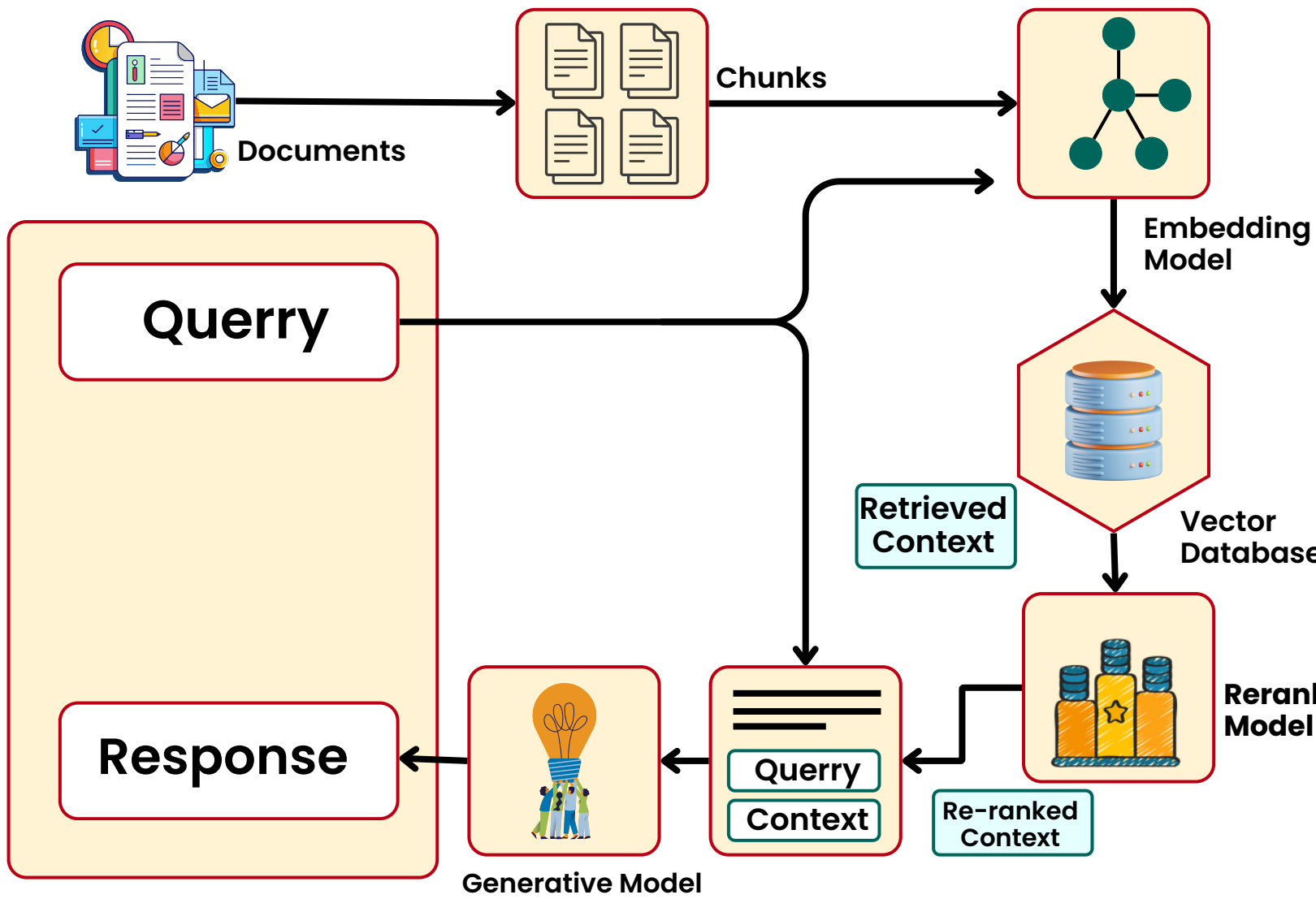
A QUICK OVERVIEW OF RETRIEVAL-AUGMENTED GENERATION (RAG) ARCHITECTURES

Naive RAG



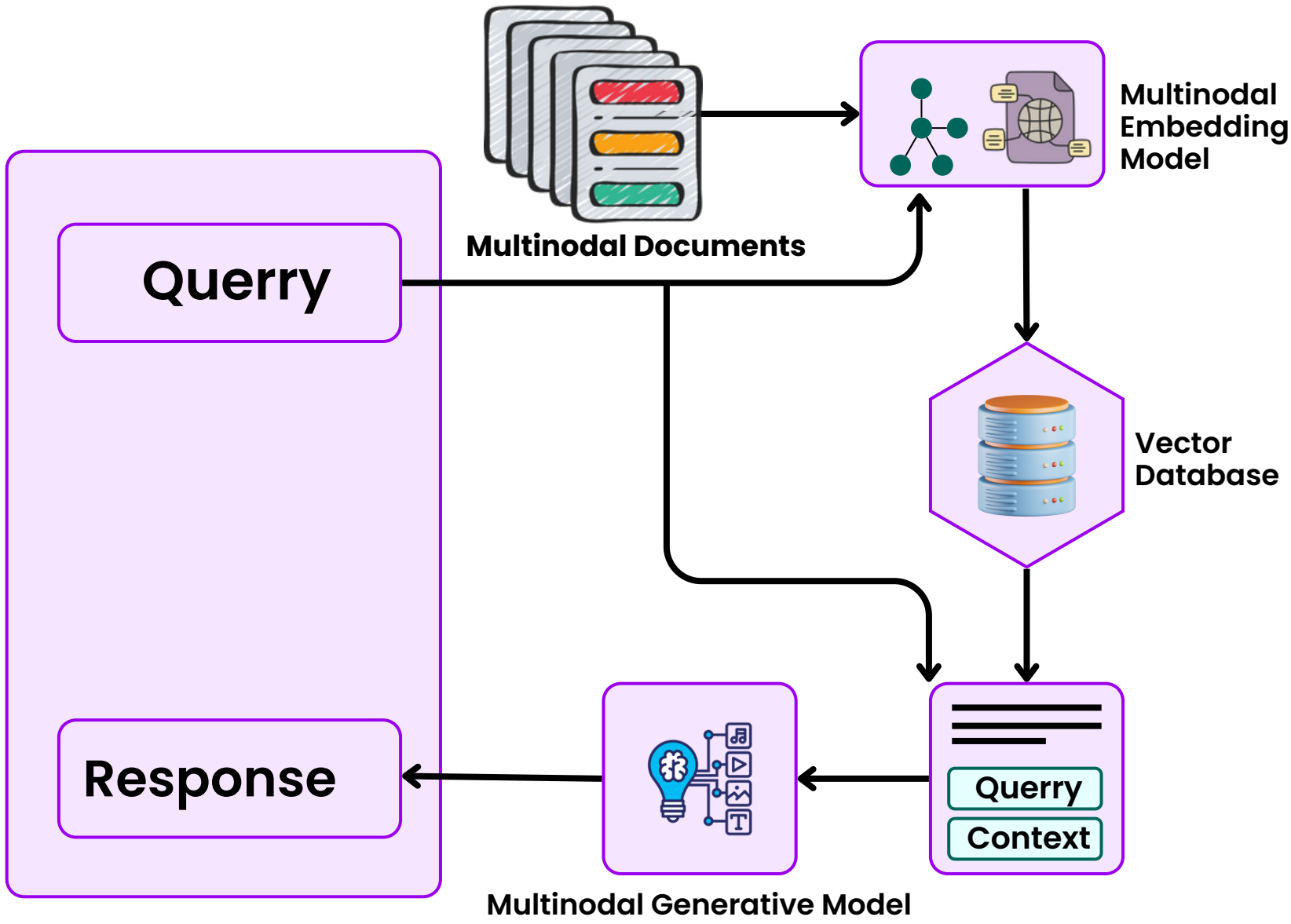
A basic RAG setup where documents are retrieved and split into chunks. These chunks are processed directly to generate a response without additional steps for optimization or ranking.

Retrieve-and-Rerank RAG



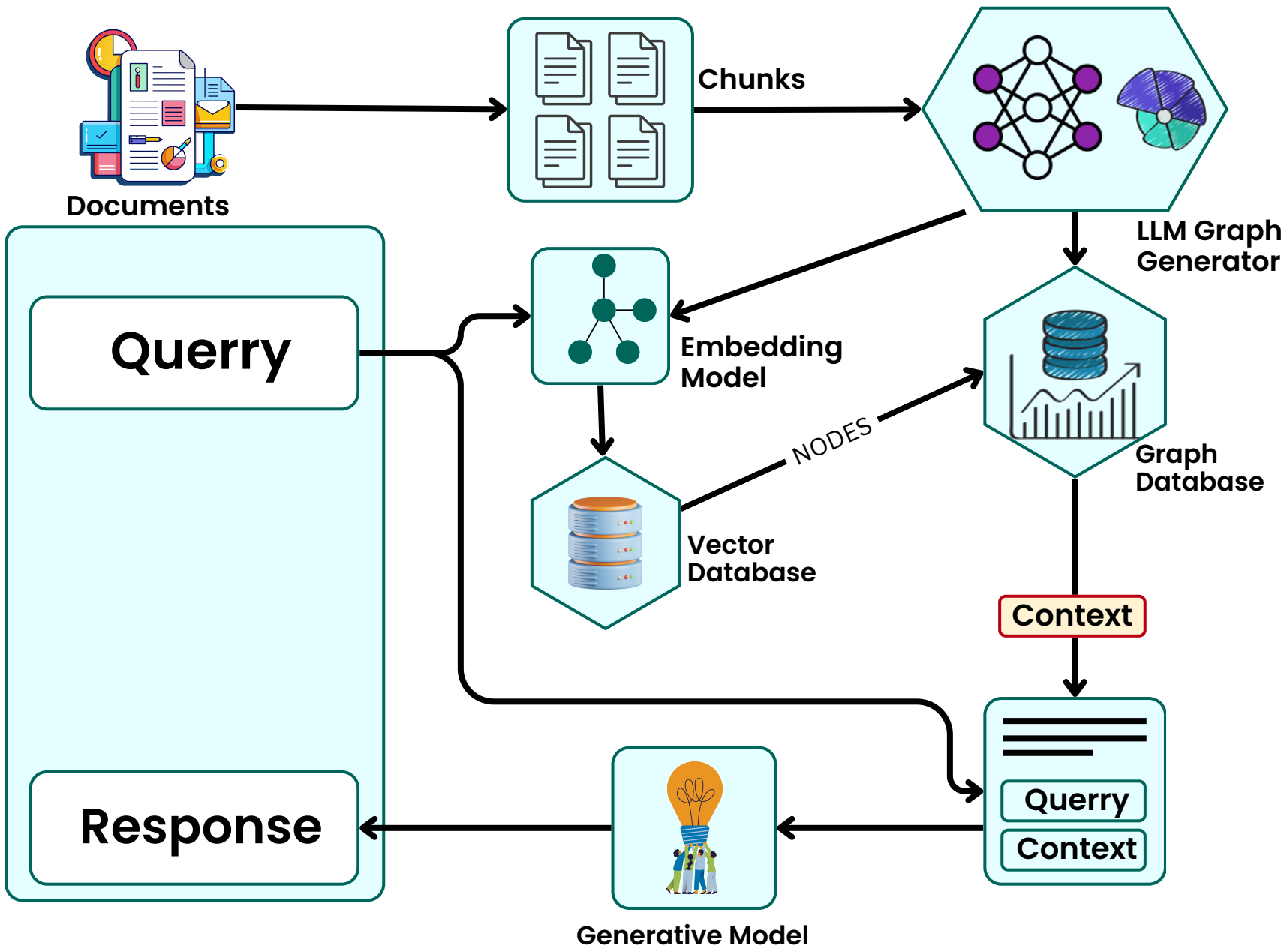
Enhances the retrieval process by re-ranking the retrieved context using a reranker model, ensuring that the most relevant chunks are prioritized for response generation.

Multimodal RAG



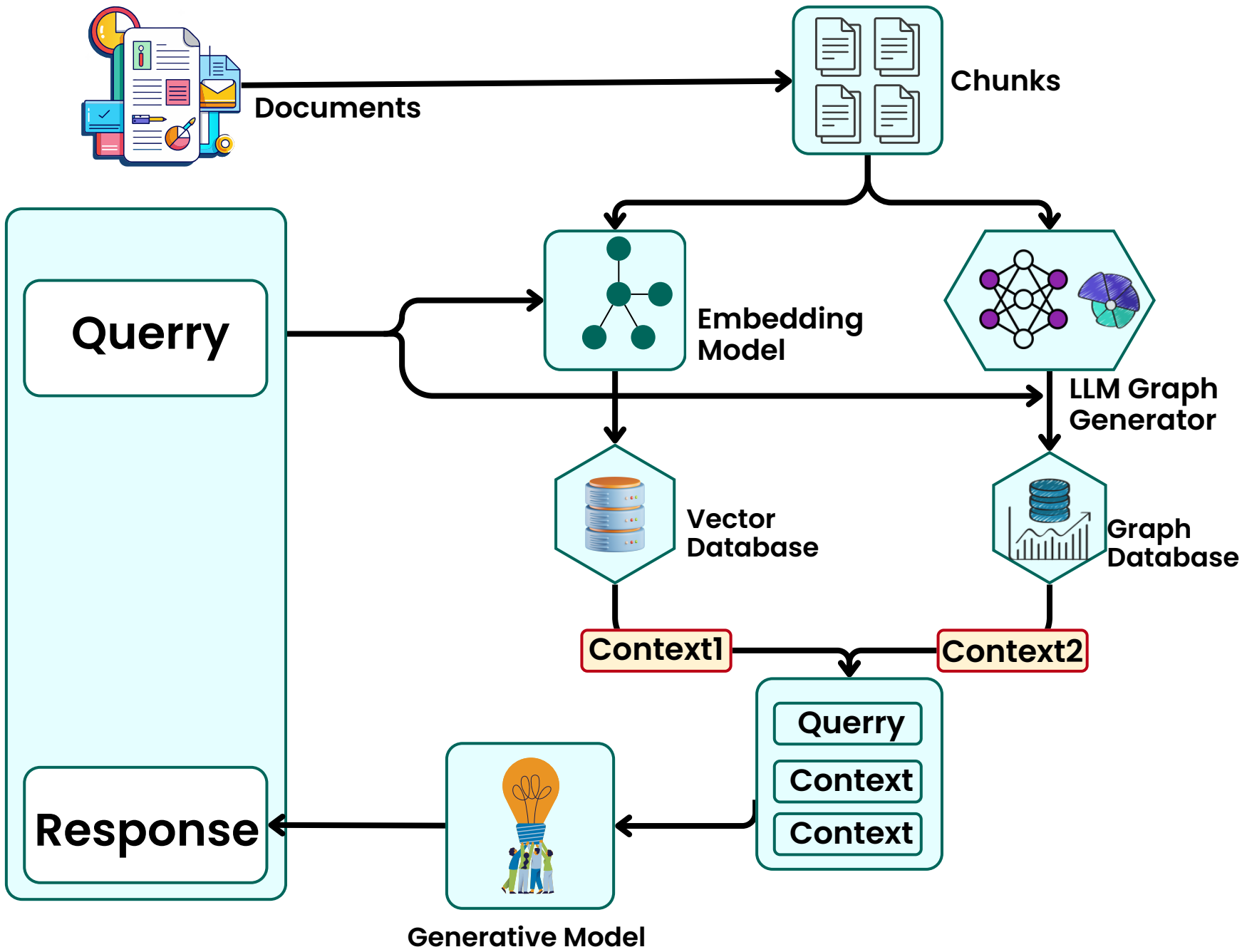
Integrates multiple types of data (e.g., text, images, or video) into the retrieval and generation process, leveraging multimodal embedding and generative models to handle diverse input types.

Graph RAG



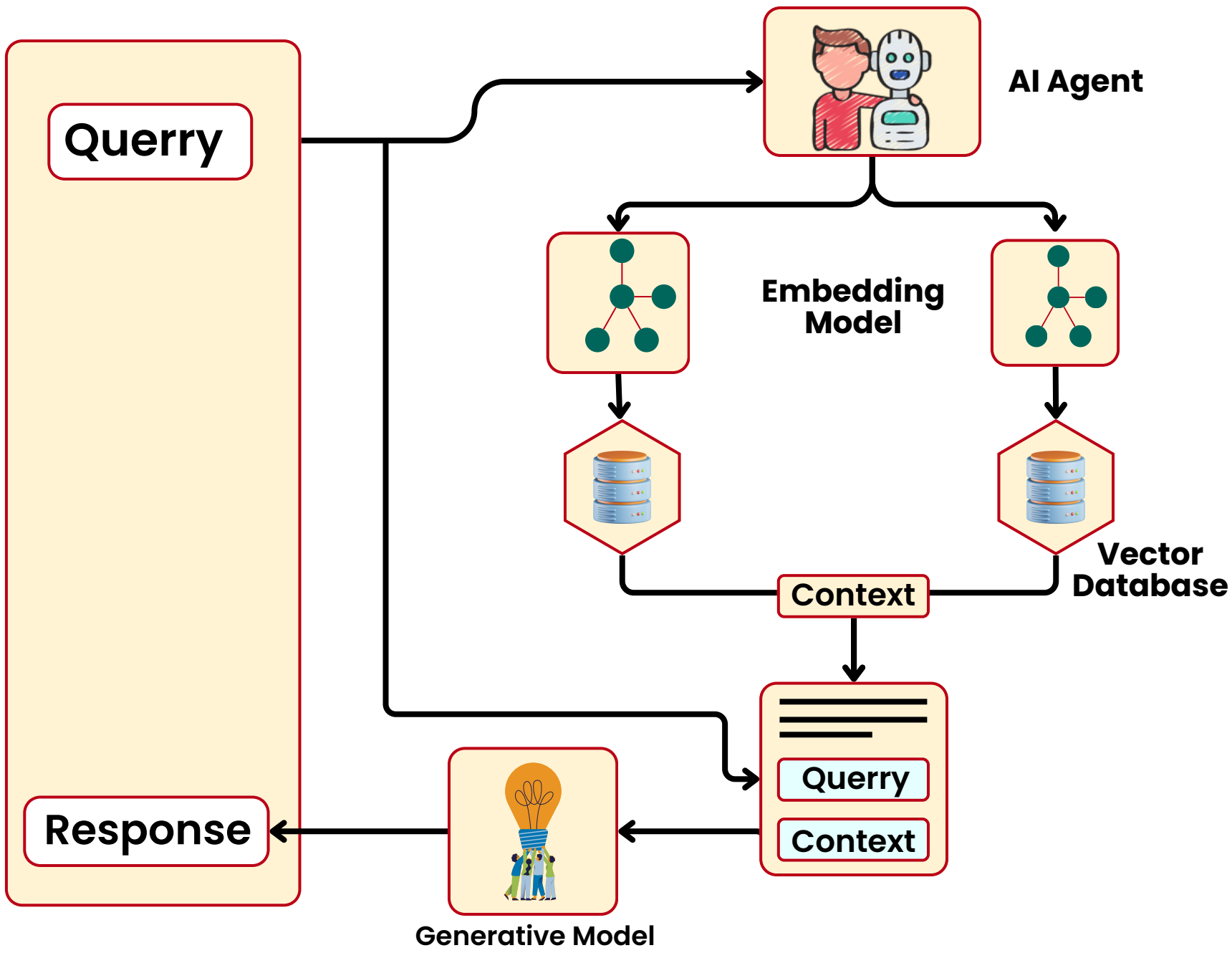
Utilizes graph databases to structure relationships between documents and chunks. This approach enables richer context retrieval by navigating connections and nodes for response generation.

Hybrid RAG



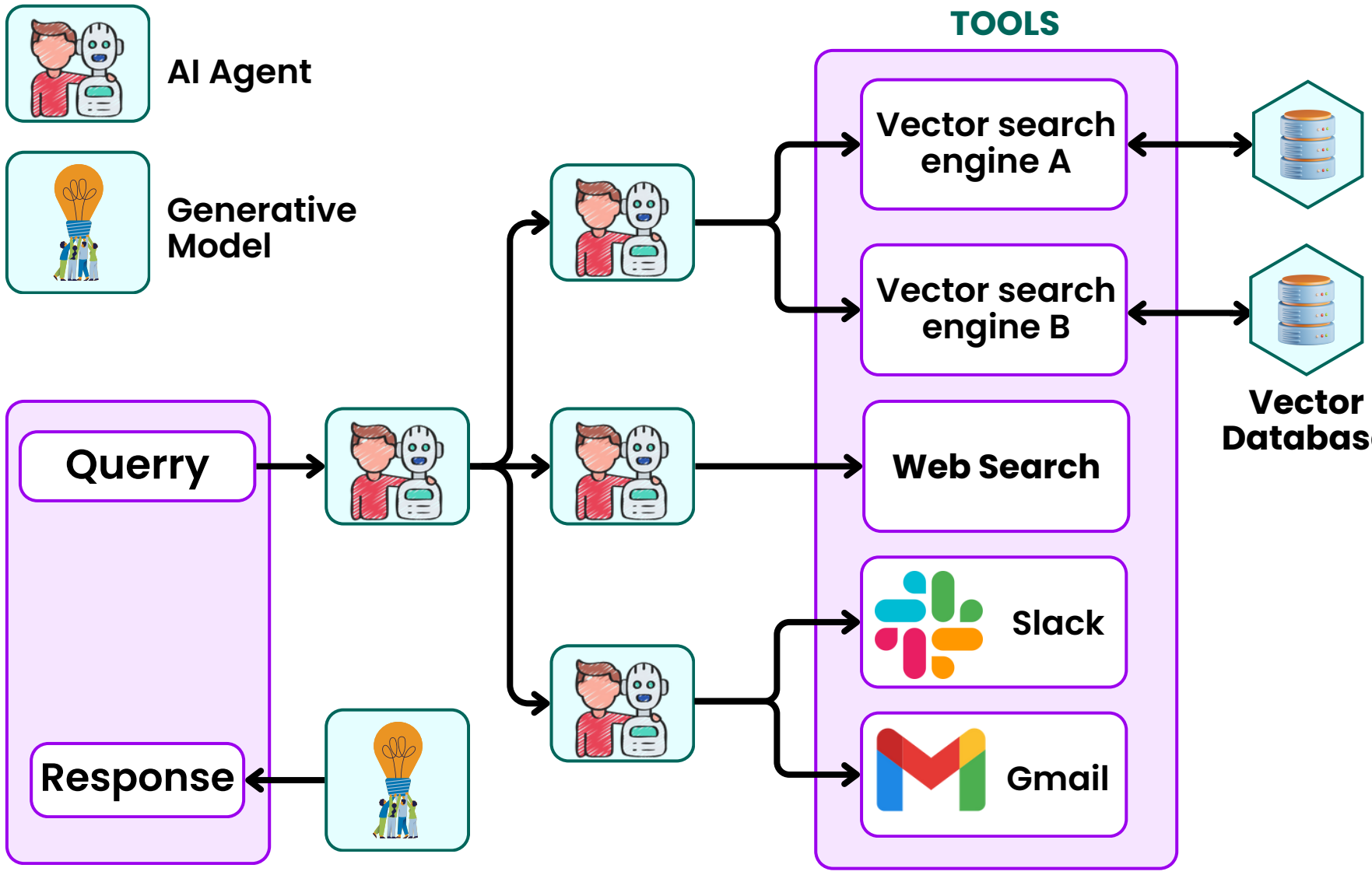
Combines multiple retrieval methods, such as vector search and graph traversal, to enrich the context used in generation. This hybrid approach ensures comprehensive coverage of the data.

Agentic RAG (Router)



Employs a router-based setup to direct queries to the most appropriate models or datasets. This architecture enhances efficiency by optimizing resource utilization for specific tasks.

Agentic RAG (Multi-Agent)



Uses multiple agents or tools (e.g., vector search engines, web search, Slack, Gmail) to aggregate context from various sources. This architecture excels in complex, multi-source query handling.

Don't Forget to
Save For Later

Was it useful?

Let me know in the comments

THE RAVIT SHOW

@theravitshow