Analytics Vidhya

Recap of RAG & Agentic RAG Systems

Dipanjan Sarkar

Head of Community & Principal AI Scientist at Analytics Vidhya

Google Developer Expert - ML & Cloud Champion Innovator

Published Author

What is a RAG System

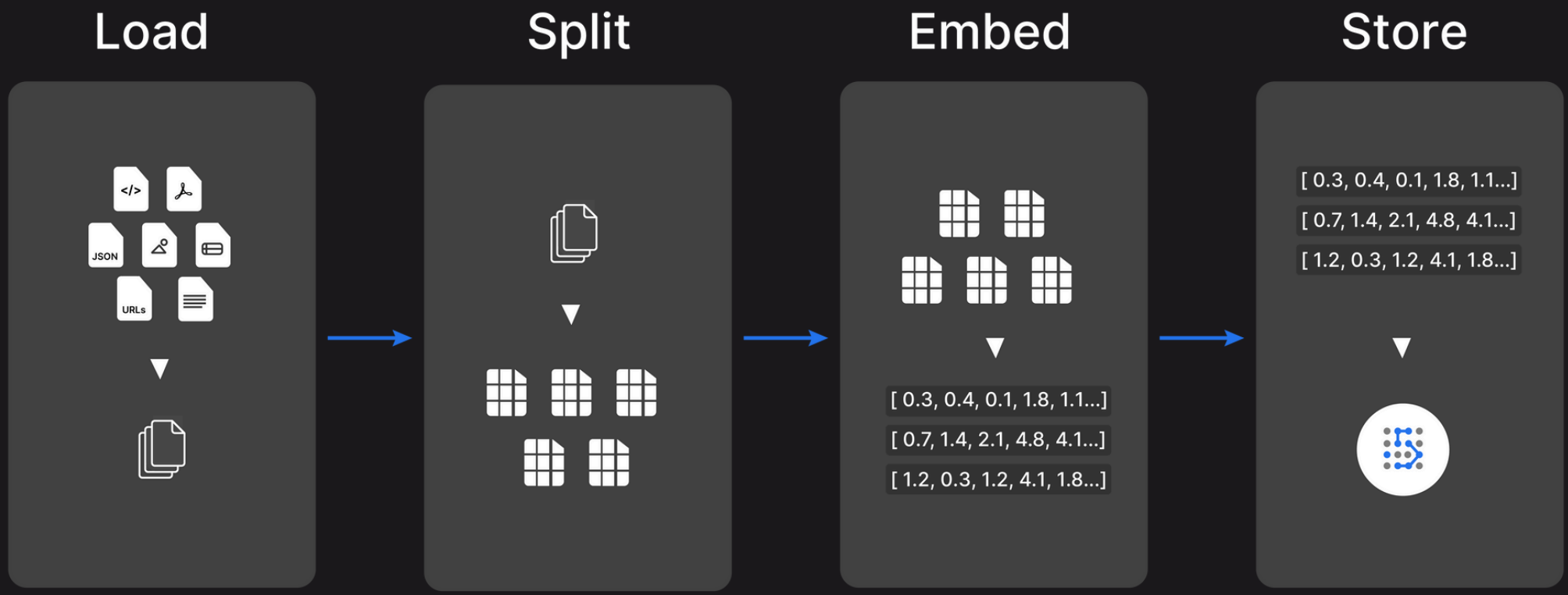
Step 1
Data Processing & Indexing

- Custom documents are processed and chunked
- These chunks are converted to embeddings with a LLM (transformer)
- Chunks and embeddings are stored in a Vector Database index (along with metadata)

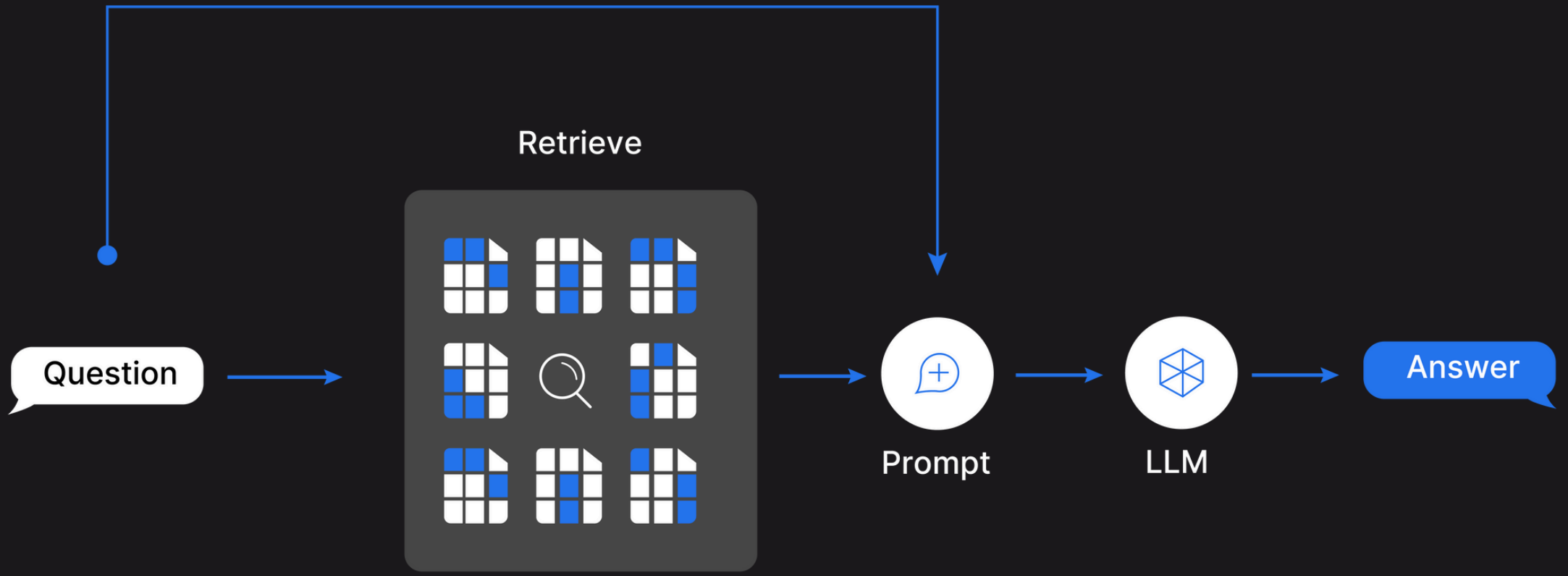
Step 2
Retrieval & Response Generation

- Based on user query the system retrieves relevant document chunks
- Passes these chunks to the LLM along with the query to augment its knowledge
- The LLM generates a human-like response to the user query based on this information

RAG Workflow - Step 1 - Data processing and Indexing

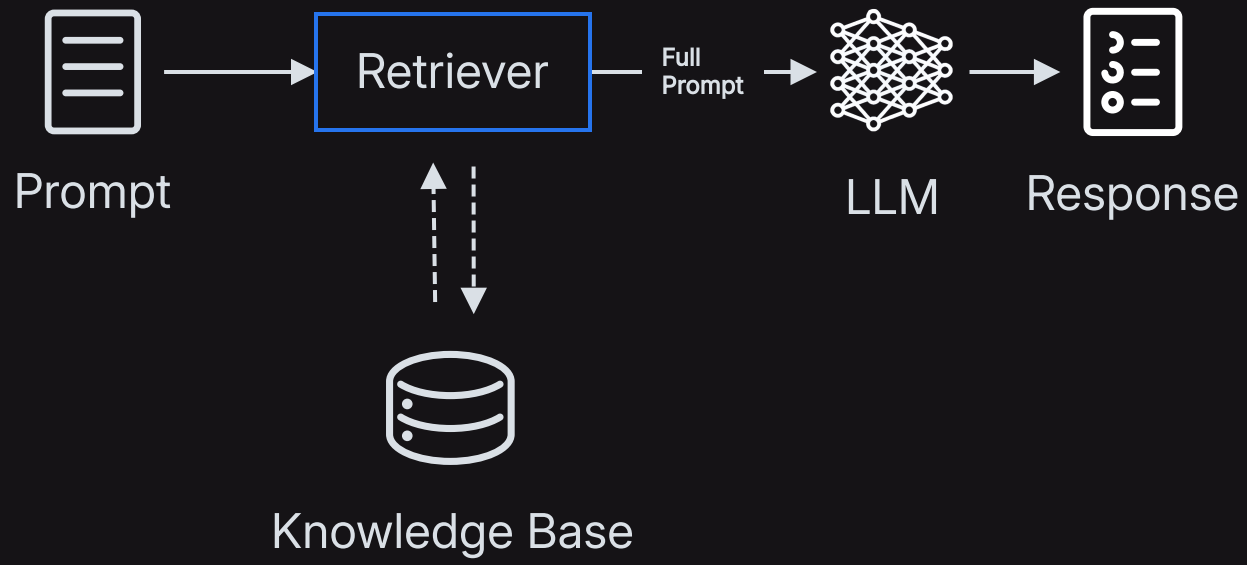


RAG Workflow - Step 2 - Retrieval and Response Generation

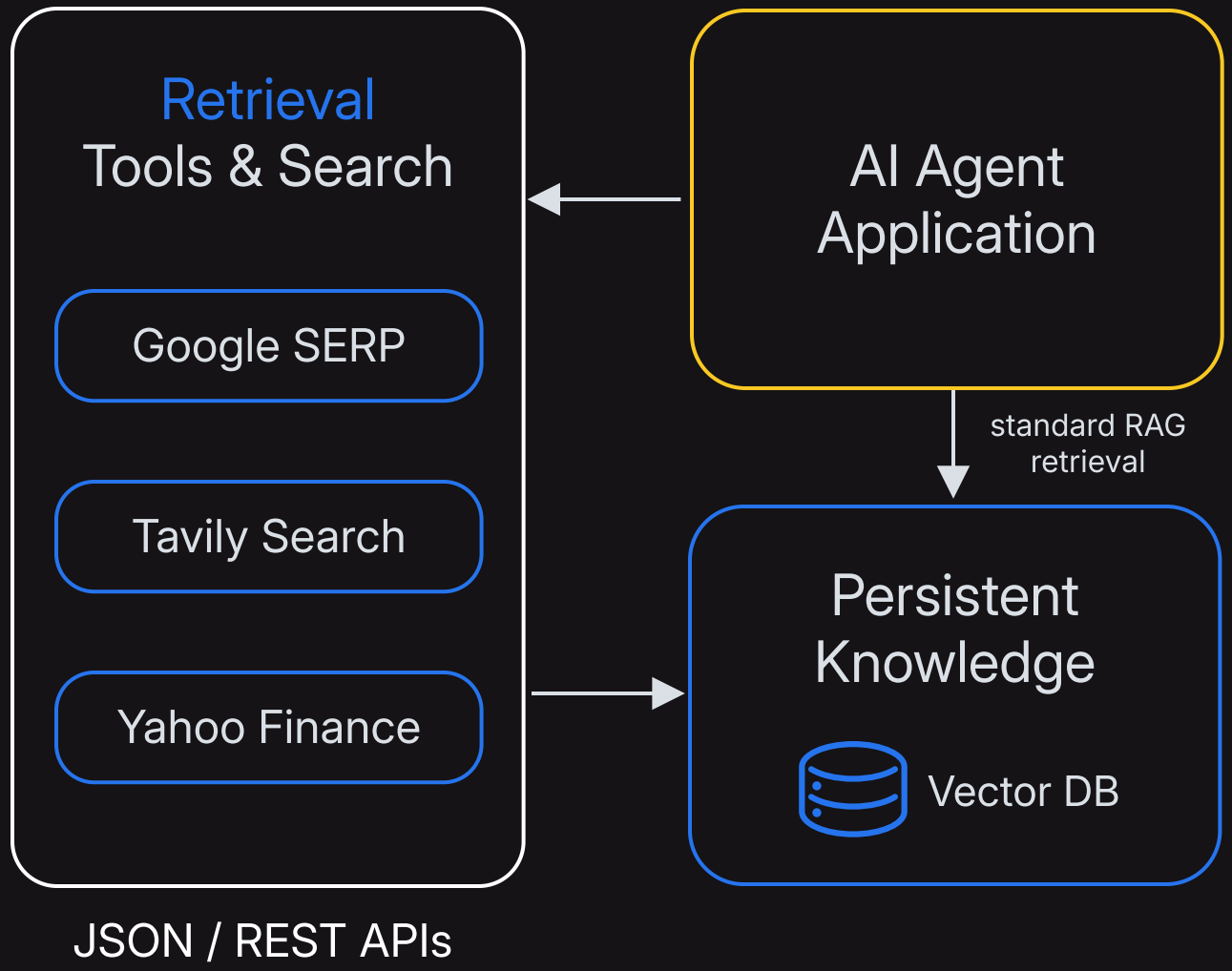


Retrieval Augmented Generation (RAG)

- 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.



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

Thanks