Analytics Vidhya

Real-World RAG System Architectures

Instructor

Dipanjan Sarkar

Head of Community & Principal AI Scientist at Analytics Vidhya

Google Developer Expert - ML & Cloud Champion Innovator

Published Author

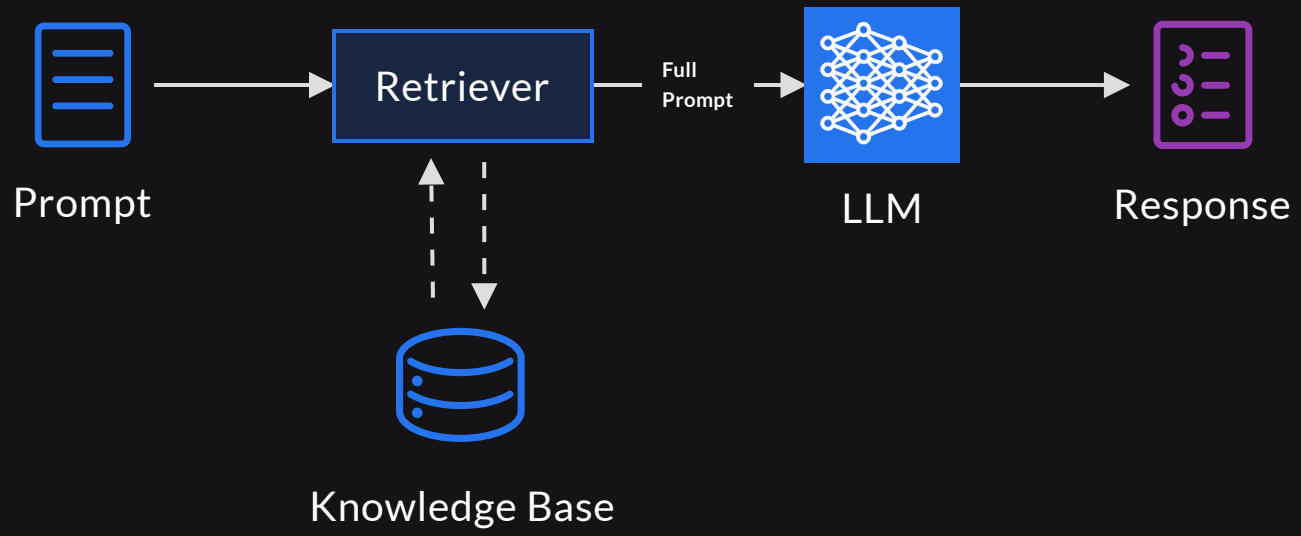
Outline

- Vanilla RAG System
- RAG System with Sources & Citations
- Conversational RAG System
- GraphRAG System
- Multimodal RAG System
- Agentic Corrective RAG System
- Agentic Self-Reflective RAG System

Analytics Vidhya

Vanilla RAG System

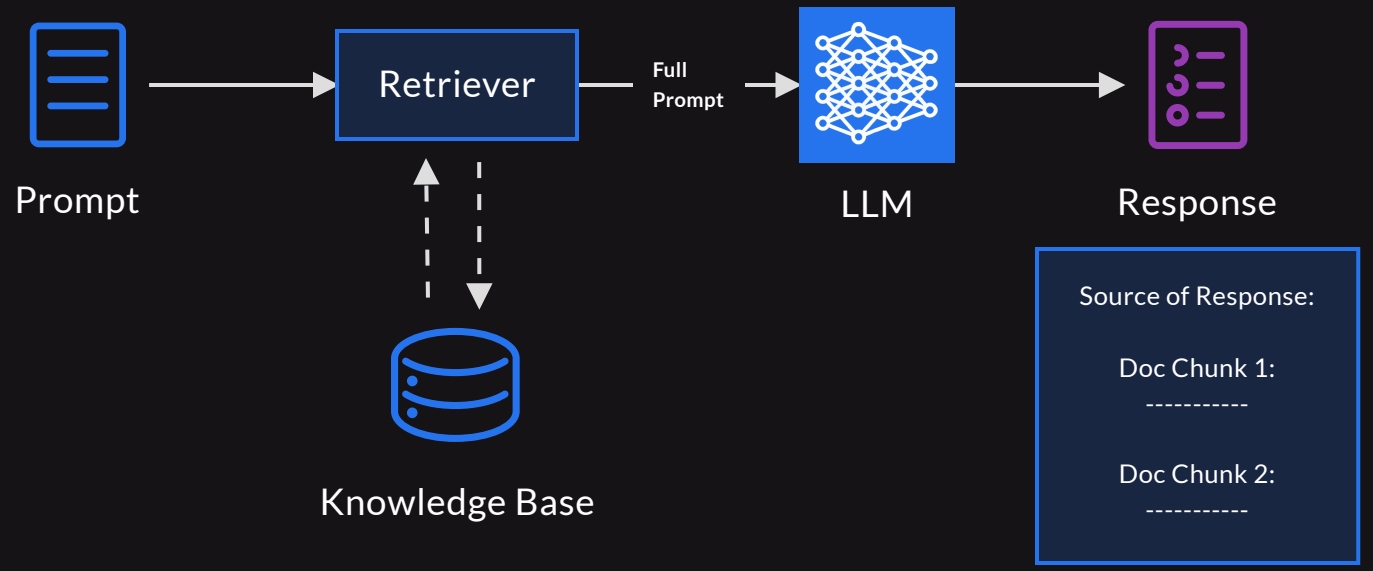
- Standard RAG Workflow connecting a vector database to an LLM
- Retrieve relevant context and generate responses to your query



Analytics Vidhya

RAG System with Sources & Citations

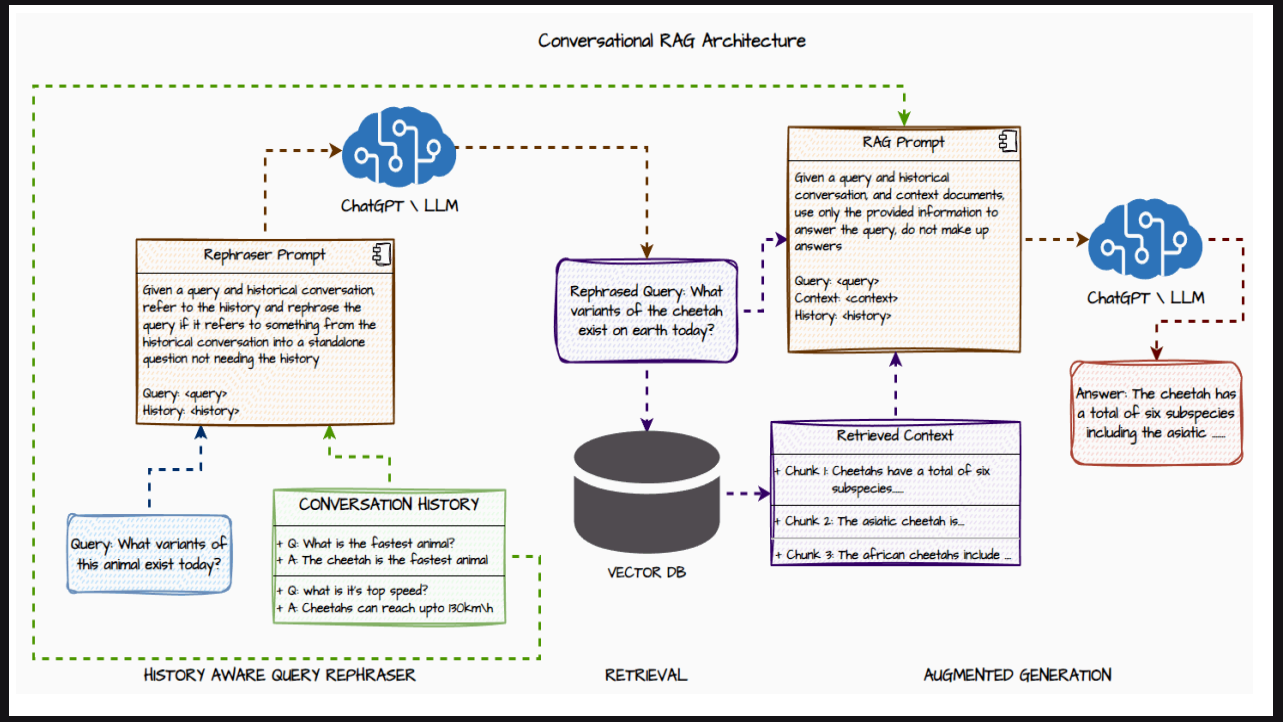
- Standard RAG Workflow connecting a vector database to an LLM
- Retrieve relevant context and generate responses to your query
- Also shows necessary document chunks which were used to generate the final response as sources or citations



Analytics Vidhya

Conversational RAG System

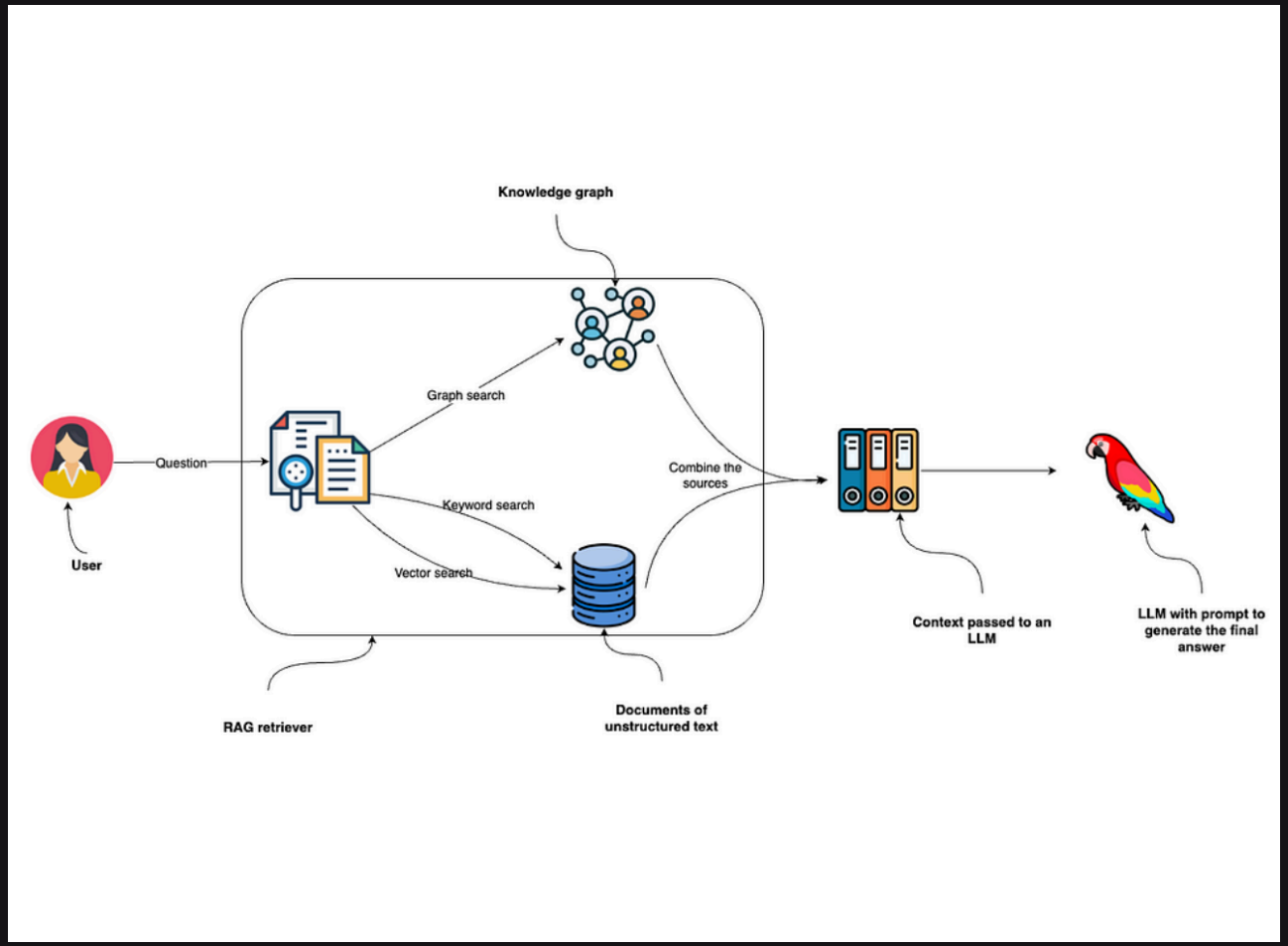
- Use LLMs to rephrase user query for retrieval based on historical conversation context
- Follow standard RAG workflow after that
- Can be extended to multi-user conversation sessions



Analytics Vidhya

GraphRAG System

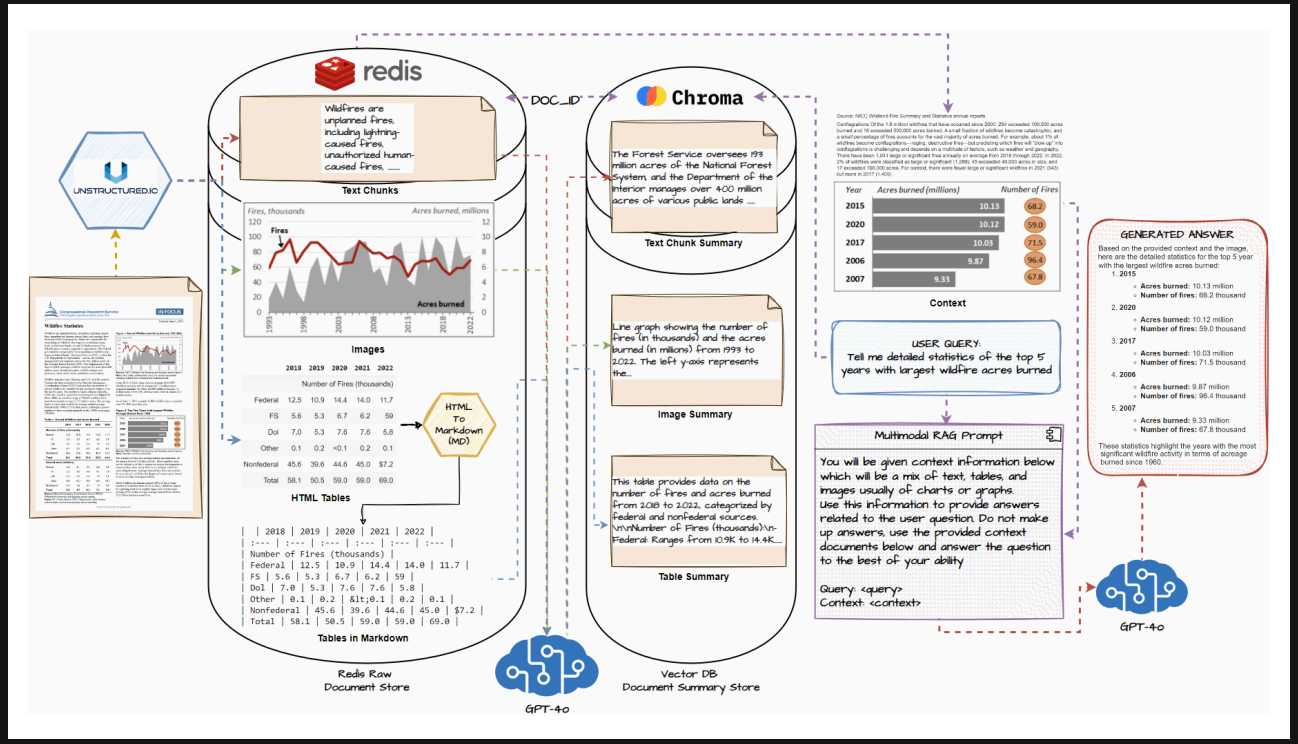
- Store data into a vector database along with relevant entity - relationships into a knowledge graph
- Retrieve relevant information as context from both vector and graph databases
- Combine the context information as use the LLM for response generation



Analytics Vidhya

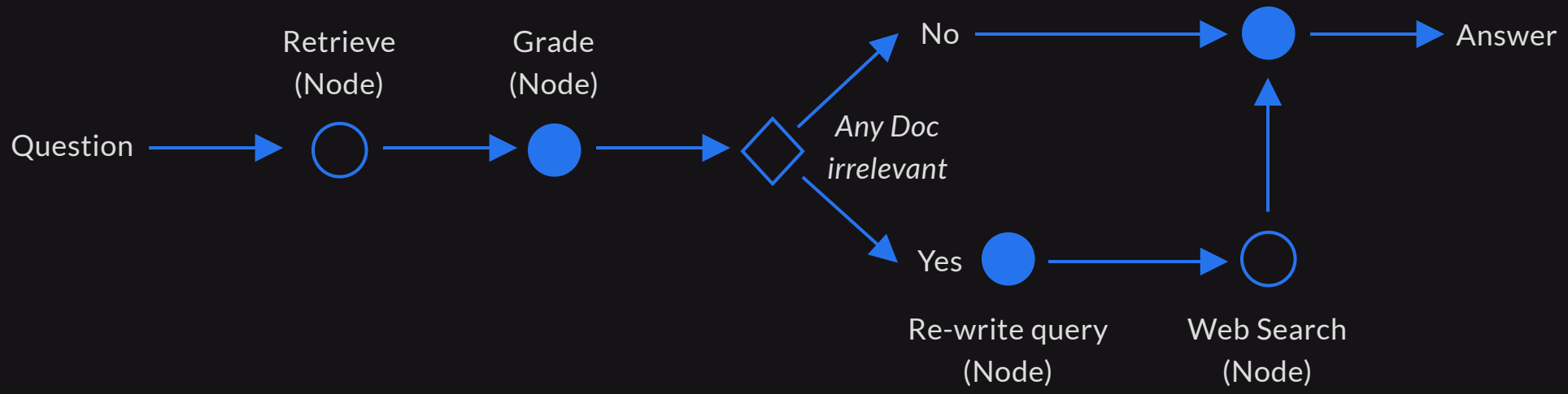
Multimodal RAG System

- Process multimodal data (mixture of text, images, and tables)
- Retrieve relevant context which can be text, images and tables
- Use Multimodal LLMs to answer questions based on multimodal retrieved context information



Analytics Vidhya

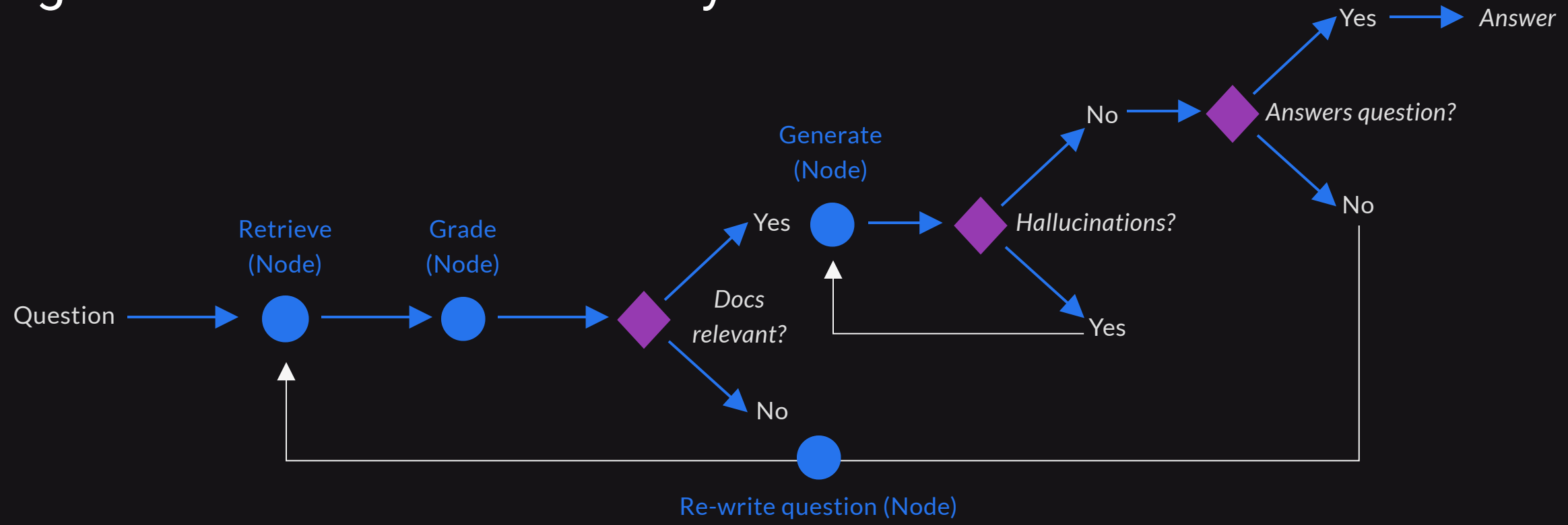
Agentic Corrective RAG System



- 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.

Analytics Vidhya

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

Analytics Vidhya

Thank You

Analytics Vidhya