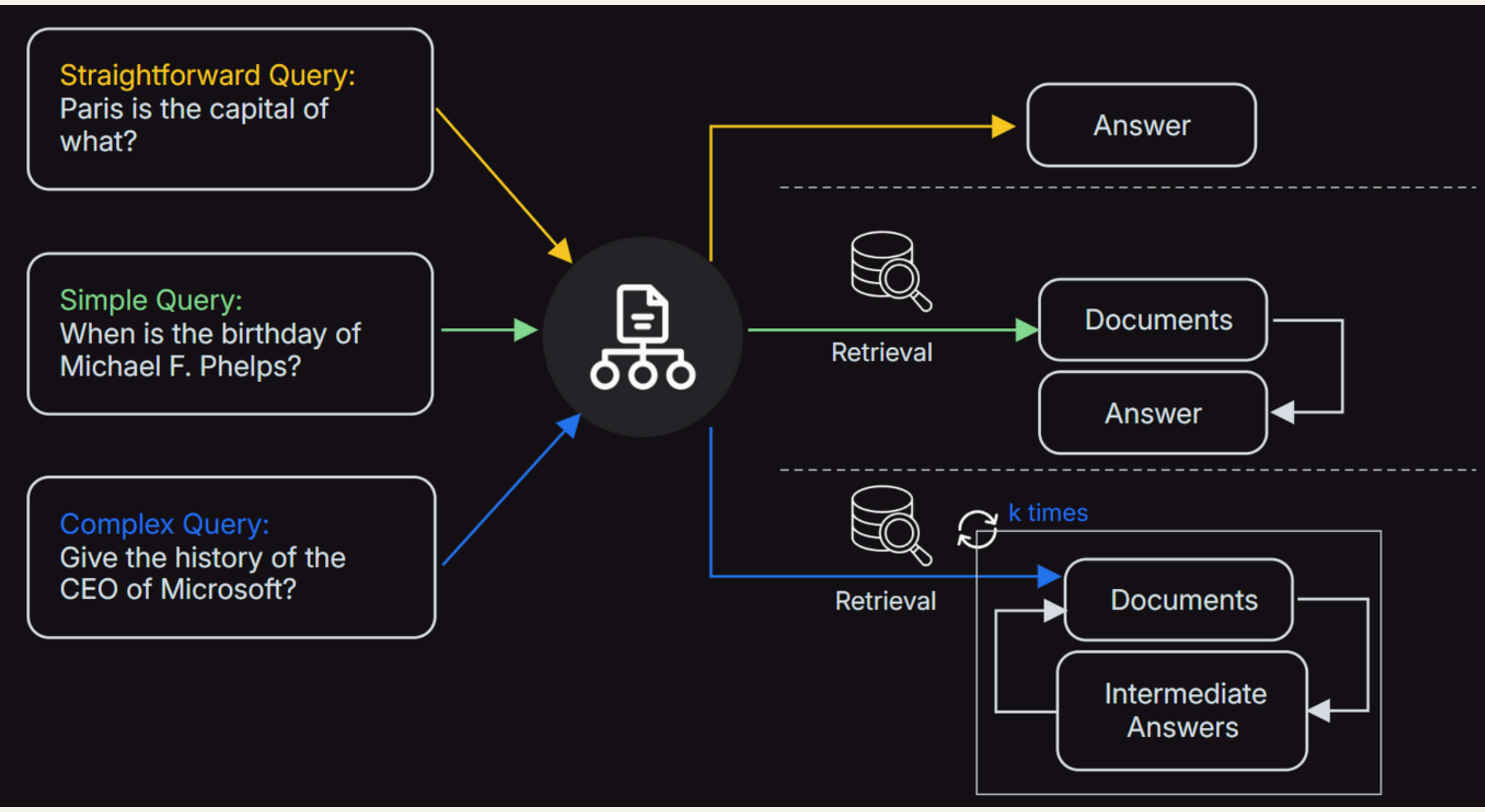
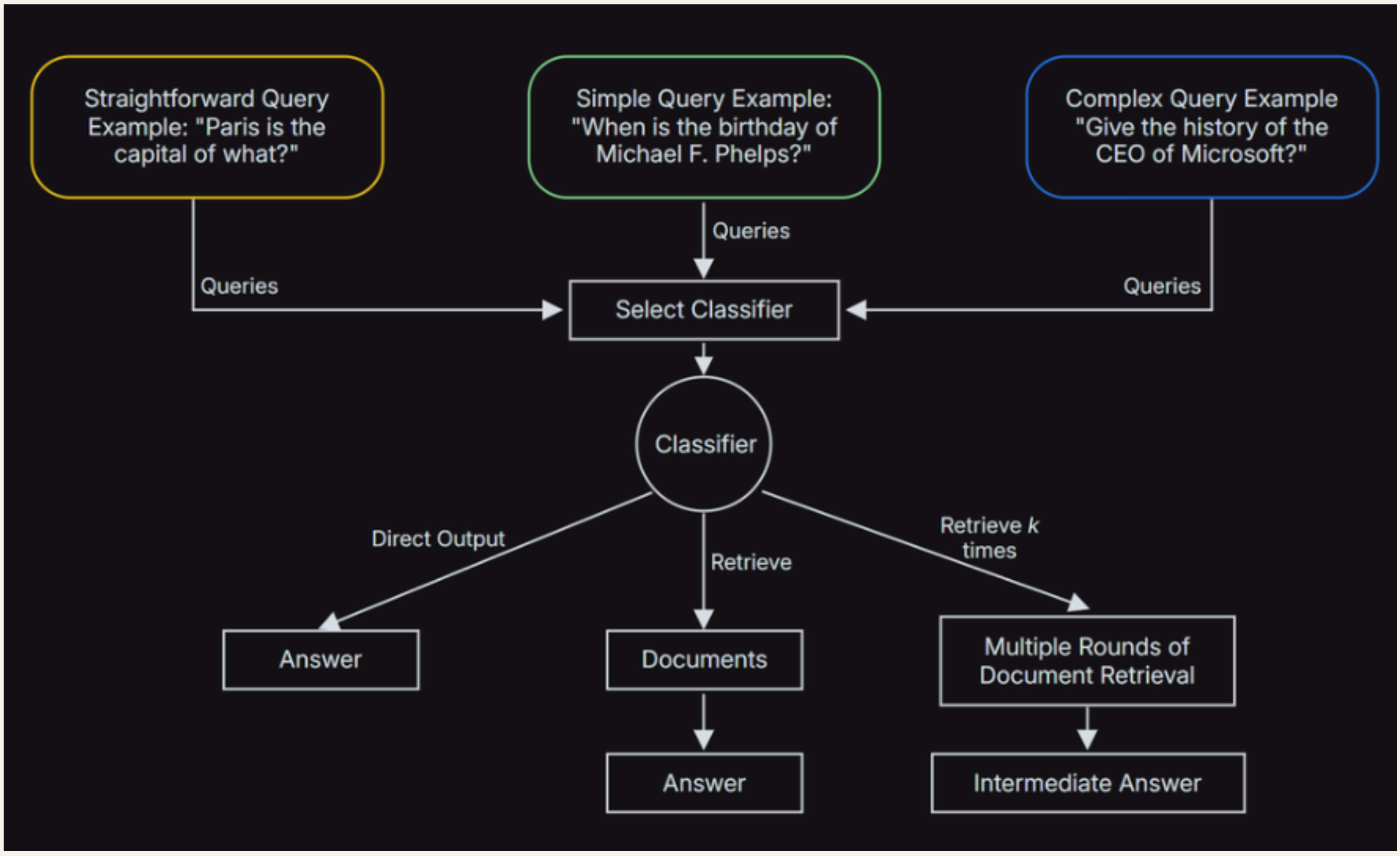
Guide to Agentic Adaptive RAG Systems



Dipanjan (DJ)

Adaptive RAG System

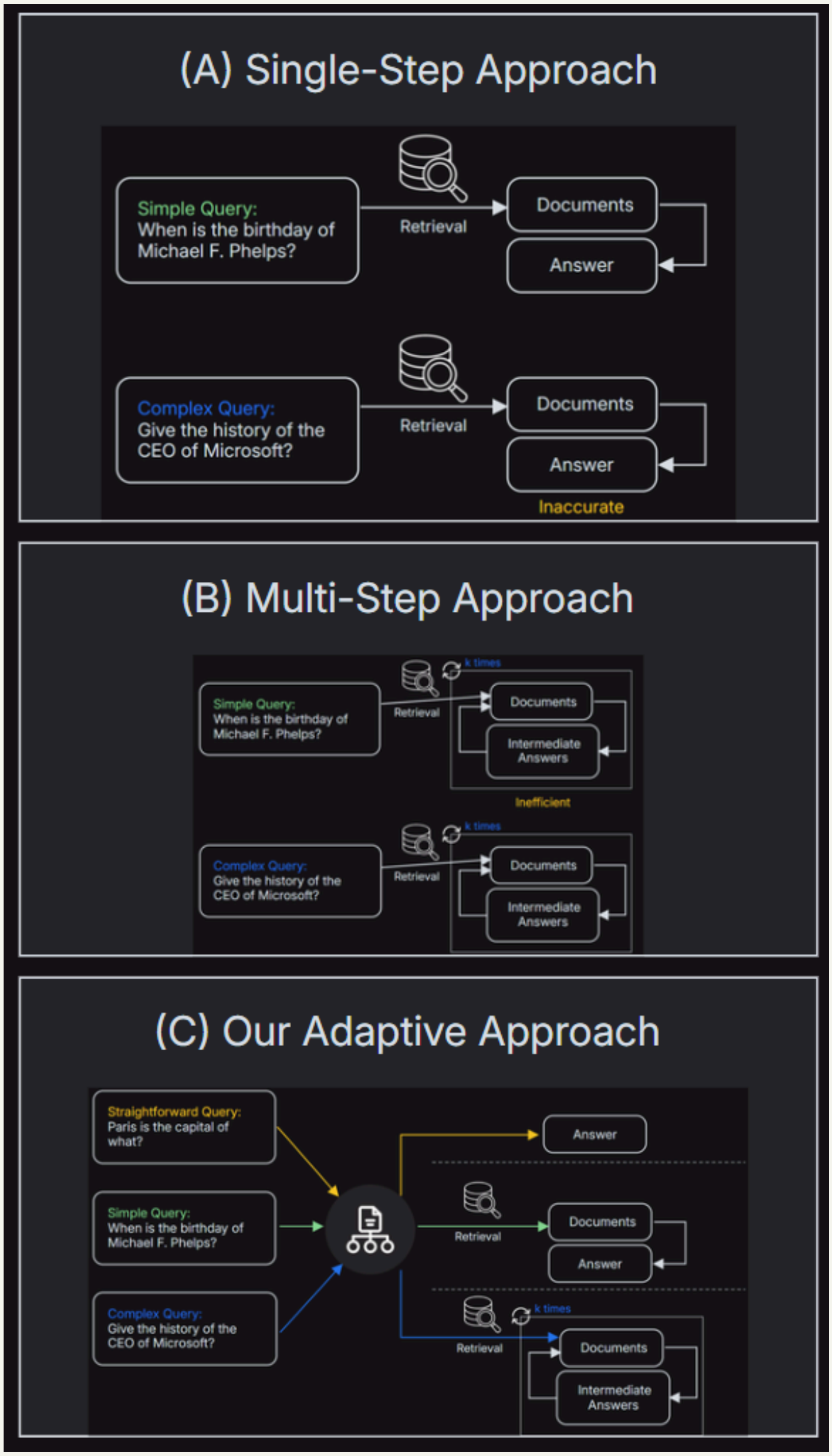


- **Step 1: Use an LLM to classify the user query into a certain category (direct, RAG, multi-hop RAG, etc.)**
- **Step 2: Route the query to a specific workflow depending on its category**
- **Step 3: Execute the relevant RAG workflow to answer the user query**

Overall Adaptive RAG can contain multiple workflows like regular RAG, self-reflective RAG, and more which can be executed based on the query complexity. It is not limited only to the depicted workflows.

Source: Author

Adaptive RAG System

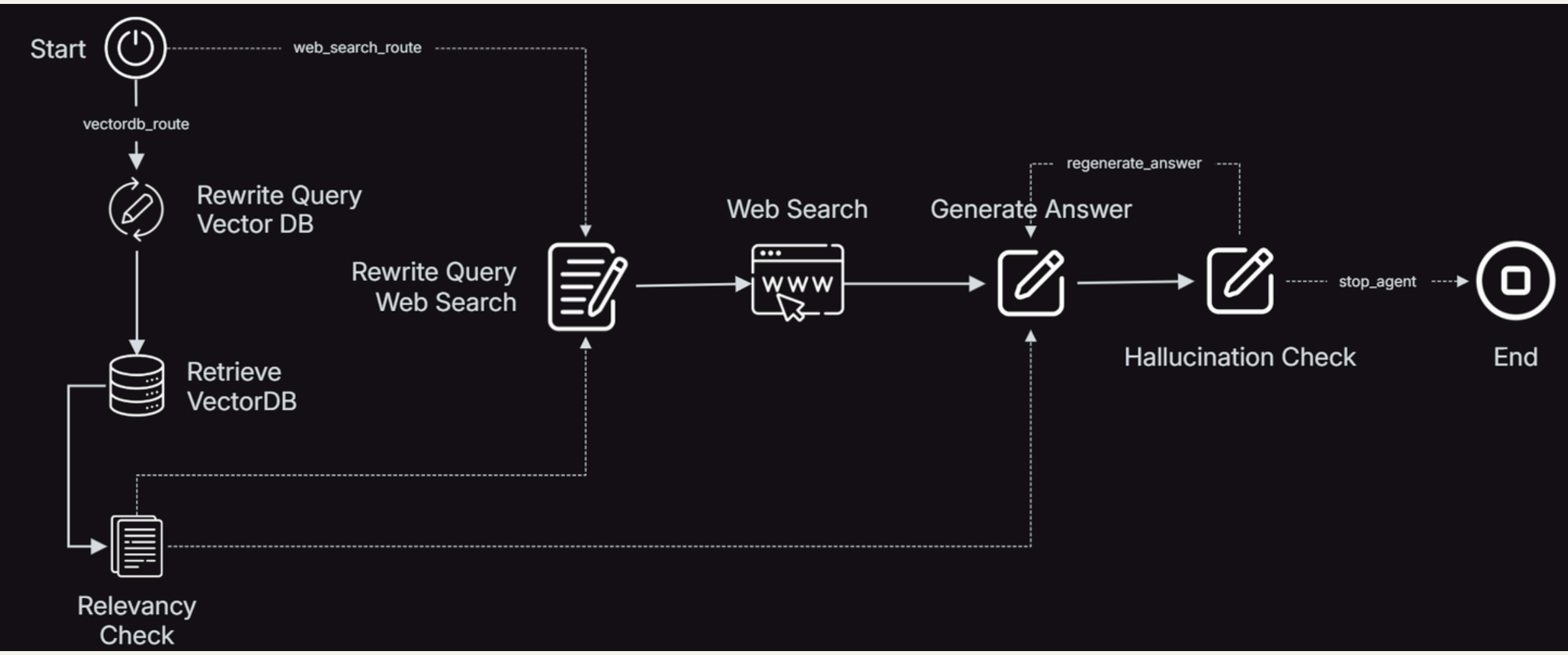


- **This system was proposed in the paper, Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity, Jeong et al., March 2024, where they propose a workflow as depicted in the figure here to enhance a regular RAG system**

- **Adaptive RAG dynamically chooses between different strategies for answering questions - ranging from simple single-step approaches to more complex multi-step or even no-retrieval processes, purely based on the complexity of the query**

Source: Author

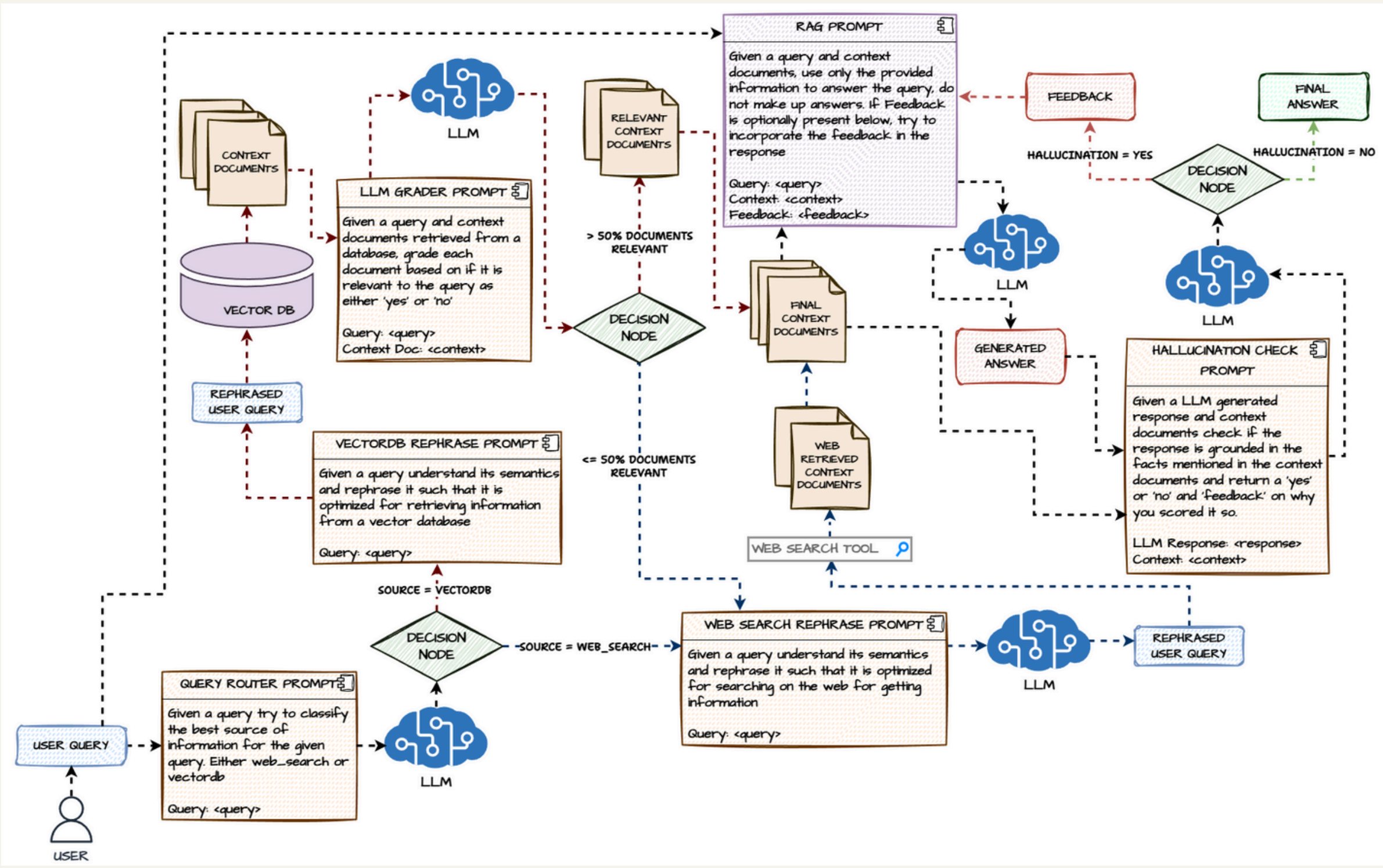
Adaptive RAG Workflow



- **Analyze a user query and adapt and dynamically route it to the best-suited workflow to answer it**
- **Leverages patterns and architectures like reflection, corrective, and self-reflective RAG to generate the best response to the user query**
- **Classifies the user query to route to the best possible workflow:**
  - **Web Search Route Workflow**
    - Rewrites query to make it more optimized for search
    - Gets live relevant context data from the web
    - Generates query response
    - Checks response for hallucinations and improves response using reflection if needed
  - **Vector DB Route Workflow**
    - Rewrites query to make it more optimized for vector db retrieval
    - Retrieves relevant context data from the vector db
    - Uses LLM as a Judge to grade each context document as relevant or not
    - If more than half the documents are irrelevant then augment additional context using web search (Corrective RAG)
    - Generates query response
    - Checks response for hallucinations and improves response using reflection if needed (Self-Reflective RAG)

Source: Author

Adaptive RAG Architecture



- **Analyze a user query and adapt and dynamically route it to the best-suited workflow to answer it**
- **Leverages patterns and architectures like reflection, corrective, and self-reflective RAG to generate the best response to the user query**
- **Classifies the user query to route to the best possible workflow:**
  - **Web Search Route Workflow**
    - Uses web search to get live context information
    - Checks response for hallucinations and improves response using reflection if needed (Self-Reflective RAG)
  - **Vector DB Route Workflow**
    - Gets context information from the vector db and if more than half the documents are irrelevant then augment additional context using web search (Corrective RAG)
    - Checks response for hallucinations and improves response using reflection if needed (Self-Reflective RAG)

Source: Author