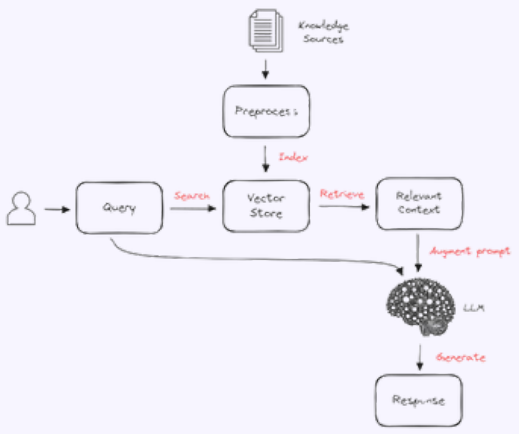
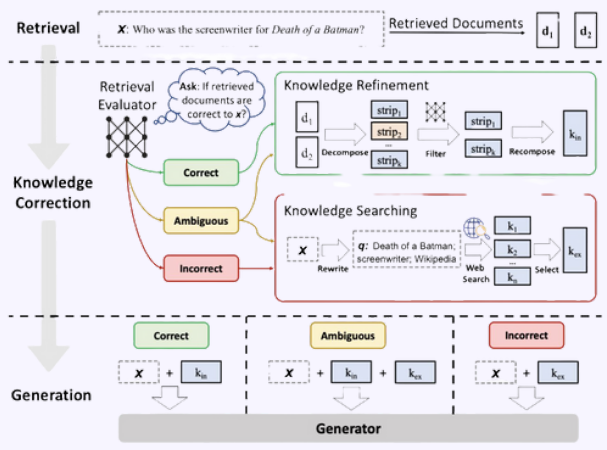
Different Types of RAG Techniques

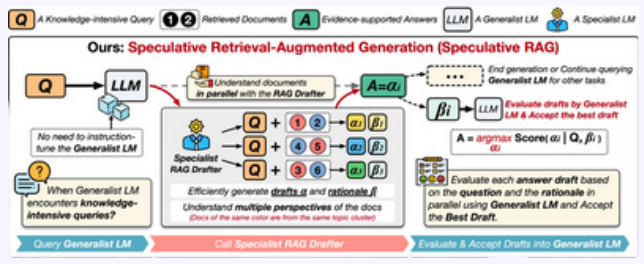
Standard RAG



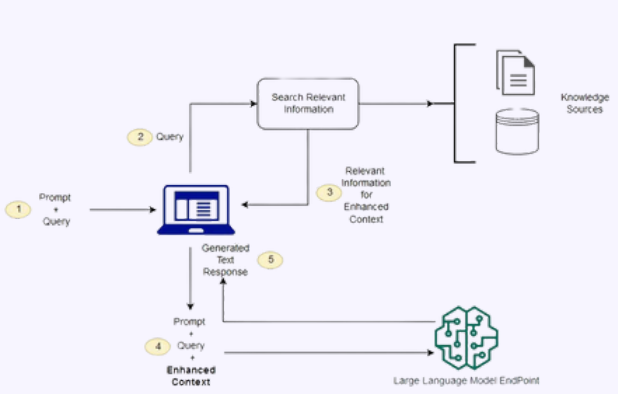
Corrective RAG



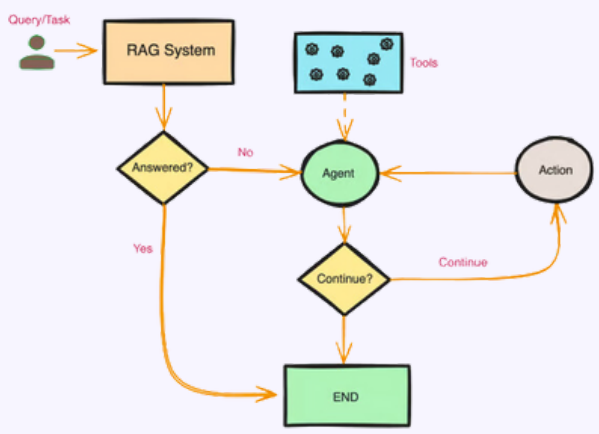
Speculative RAG



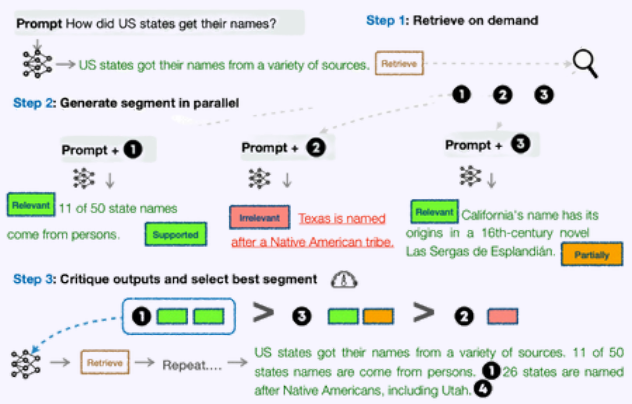
Fusion RAG



Agentic RAG

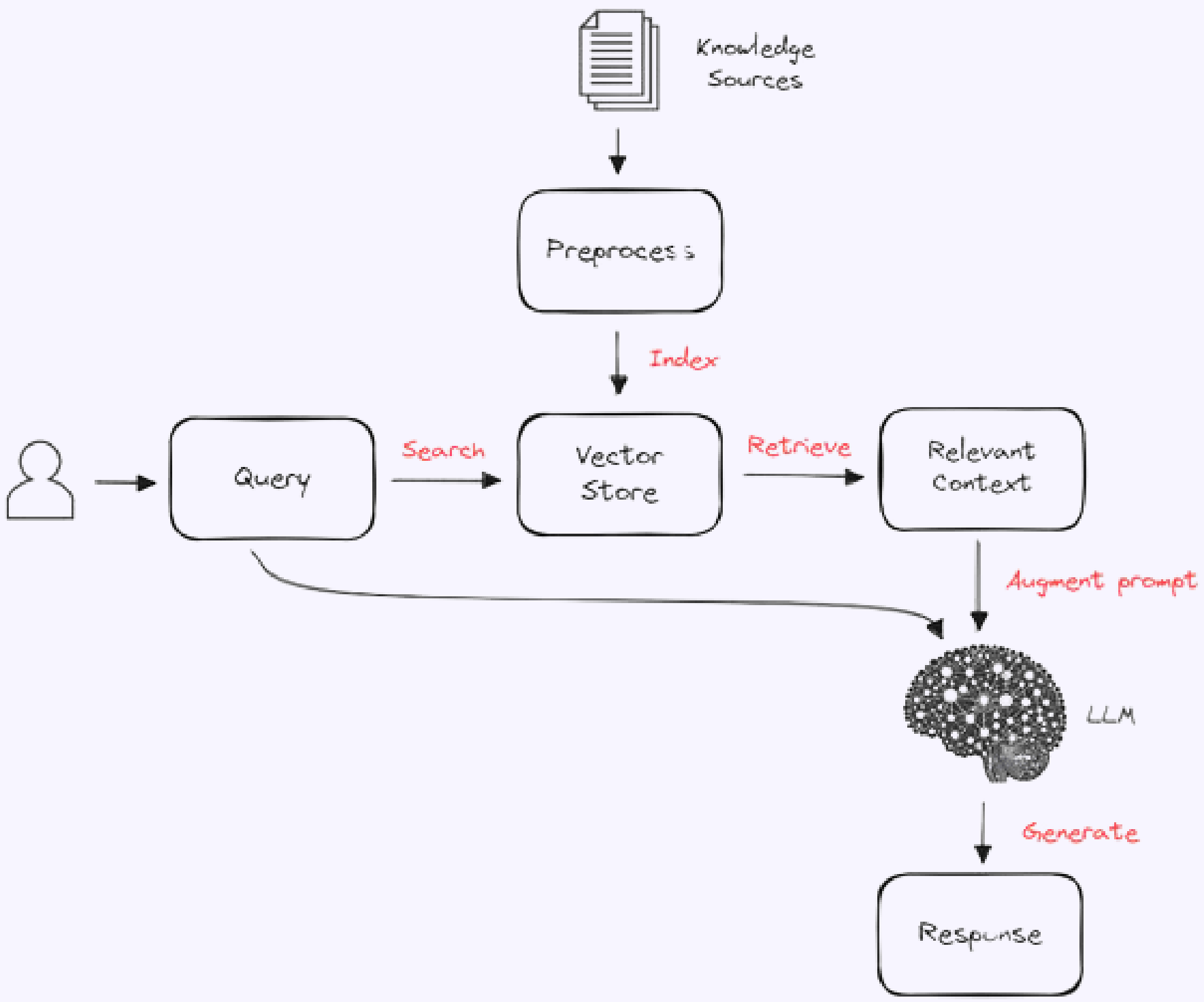


Self RAG



Standard RAG

- Standard RAG combines a retrieval model (like a search engine) with a generative model (like GPT). The retrieval model fetches relevant documents or pieces of information from a large database, and the generative model uses this information to generate coherent and contextually accurate responses or content.



How it works?

- **Step 1**: Query Input - A user query or input is provided to the retrieval component of the system.

- **Step 2**: Retrieval Process - The retriever searches a large corpus or database for documents or text passages that are most relevant to the query. This is often done using vector search or dense retrieval methods, where both the query and documents are encoded into high-dimensional vectors.

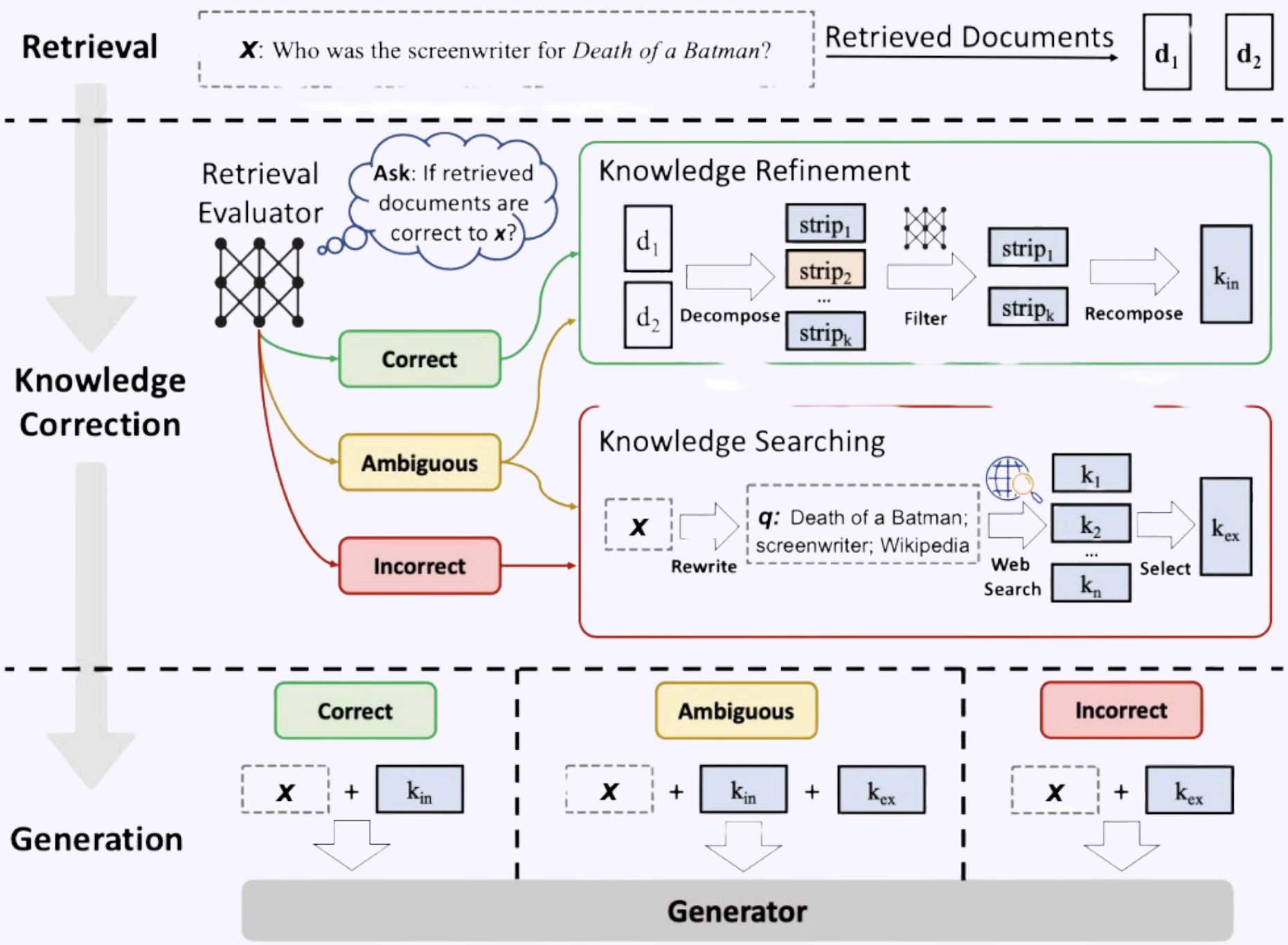
- **Step 3**: Selection of Top Documents - The retriever ranks the documents based on their relevance to the query and selects the top-k documents (e.g., the top 5 most relevant passages).

- **Step 4**: Generative Response - The selected documents are then passed to the generative model (like GPT). The model uses this context to generate a coherent response that directly answers the query while incorporating the retrieved information.

- **Step 5**: Output - The final response is presented to the user, leveraging the retrieved content to enhance accuracy and detail.

Corrective RAG

- Standard RAG combines a retrieval model (like a search engine) with a generative model (like GPT). The retrieval model fetches relevant documents or pieces of information from a large database, and the generative model uses this information to generate coherent and contextually accurate responses or content.



How it works?

- **Step 1**: Initial Retrieval and Generation - The process begins like Standard RAG, where the retriever fetches relevant information and the generative model creates a response.

- **Step 2**: Validation Process - The generated response is then validated against a trusted dataset or source. This could involve comparing the generated content with data from authoritative sources (like medical databases, academic papers, or trusted news outlets).

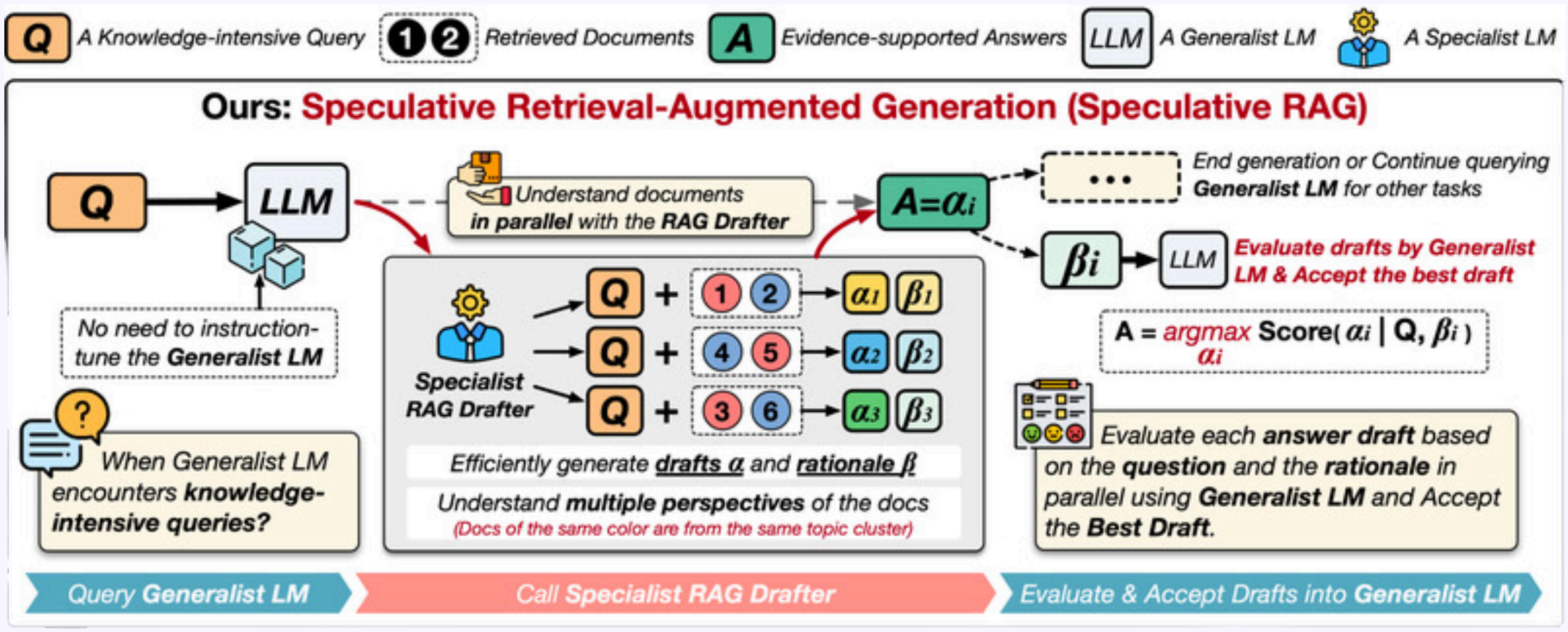
- **Step 3**: Correction Mechanism - If discrepancies or errors are detected during validation, the model uses the feedback to correct the response. This might involve generating a new response or refining the existing one.

- **Step 4**: Iteration and Feedback Loop - The system iterates this process, continuously refining the response until it aligns with the correct information or falls within an acceptable error margin.

- **Step 5**: Final Output - The validated and corrected response is provided to the user.

Speculative RAG

- Speculative Retrieval-Augmented Generation (Speculative RAG) is an approach that involves generating multiple possible responses or outputs for a given input query, using a retrieval model to provide relevant information. The generated responses are then evaluated using a feedback mechanism to select the most plausible or relevant one. The goal is to enhance the model's ability to produce accurate and contextually appropriate answers, especially when there is ambiguity or multiple potential interpretations of a query.



How it works?

- **Step 1**: Retrieval - Similar to Standard RAG, it starts by retrieving multiple documents relevant to the query.

- **Step 2**: Generative Speculation - The generative model creates multiple speculative responses based on the retrieved documents. Instead of producing a single answer, it explores several possible outputs.

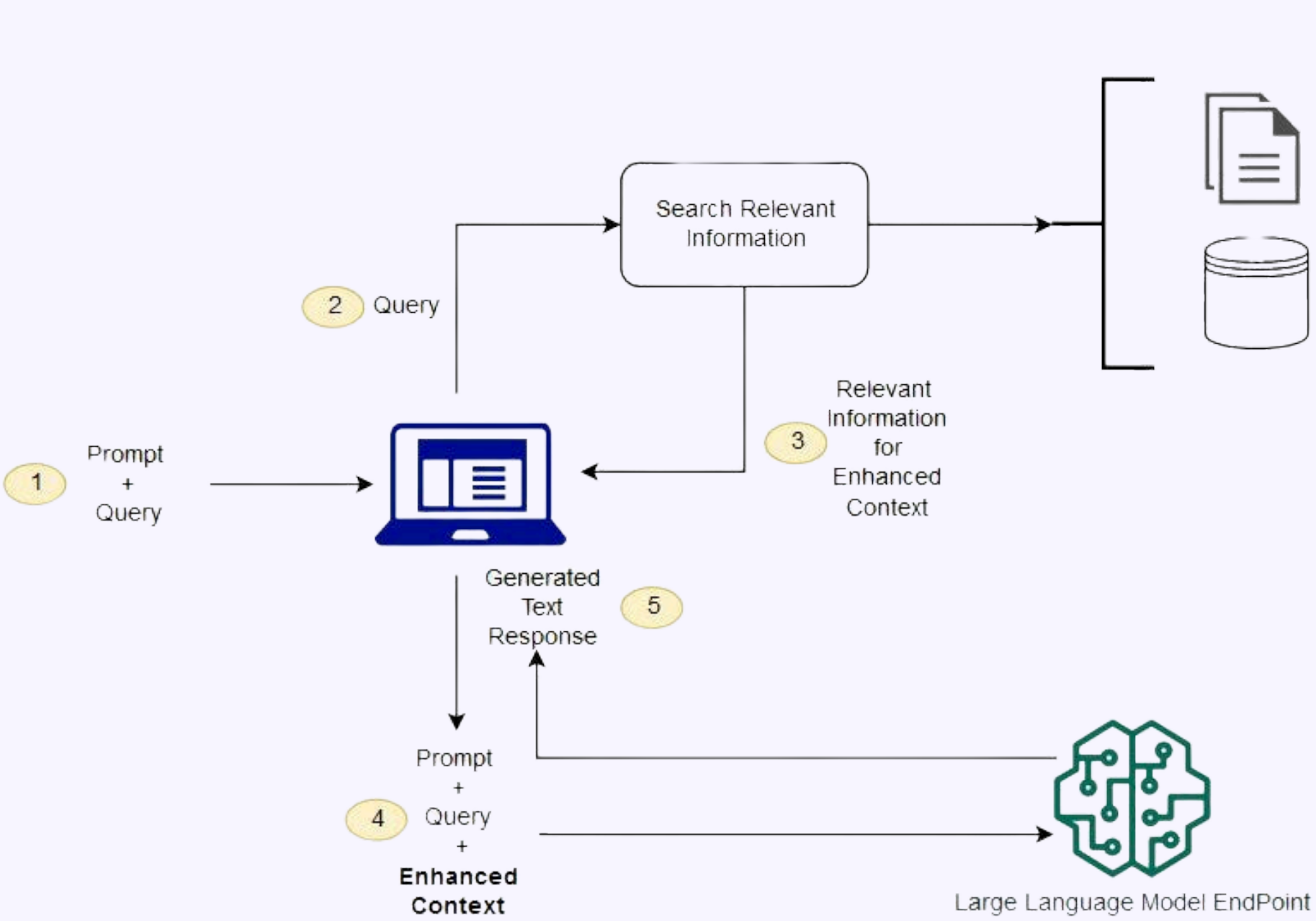
- **Step 3**: Feedback and Ranking - Each of the generated responses is evaluated using a feedback mechanism that scores them based on various criteria like relevance, coherence, completeness, and factual accuracy. This could involve comparing the responses against additional retrieved documents or using scoring models.

- **Step 4**: Selection Process - The model ranks all possible responses and selects the highest-scoring one as the final output.

- **Step 5**: Presentation - The chosen response is then presented to the user as the final answer.

Fusion RAG

- Fusion RAG integrates information from multiple retrieved documents or sources to generate a comprehensive and well-rounded response.



How it works?

- **Step 1**: Retrieval of Diverse Documents - The retriever fetches multiple relevant documents, ensuring they represent diverse perspectives or cover different aspects of the query topic.

- **Step 2**: Information Integration - The generative model analyzes the content of these documents and identifies common themes, facts, or points of view. It weighs conflicting information, integrates consistent data, and balances different perspectives.

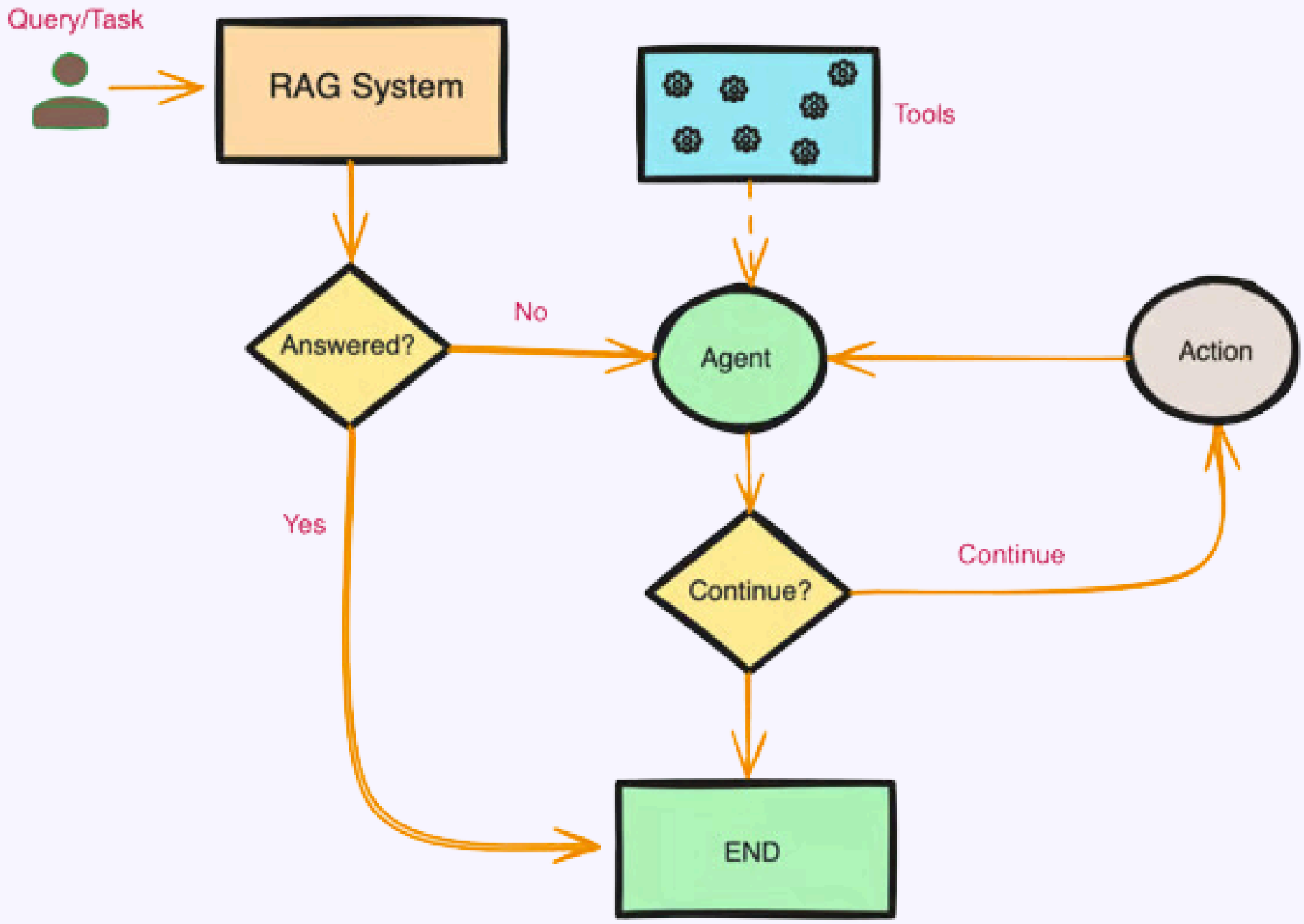
- **Step 3**: Synthesis of a Unified Response - The model synthesizes a single, coherent response that integrates the relevant information from all retrieved documents. It aims to present a balanced view, combining multiple pieces of evidence or viewpoints.

- **Step 4**: Conflict Resolution - When conflicting information is present, the model uses additional context, source credibility, or predefined rules to resolve discrepancies and form a unified answer.

- **Step 5**: Output - The integrated response is provided to the user, ensuring it covers all necessary aspects of the query.

Agentic RAG

- Agentic RAG involves the AI acting autonomously with a predefined goal, using the retrieval process to make decisions and guide its actions.



How it works?

- **Step 1**: Goal Definition - The model is provided with a specific objective or goal (e.g., explaining a concept, providing personalized tutoring, or solving a complex task).

- **Step 2**: Autonomous Action Planning - The generative model autonomously determines which information to retrieve and which actions to take to achieve the goal. It may ask itself questions, plan steps, or decide on intermediate objectives.

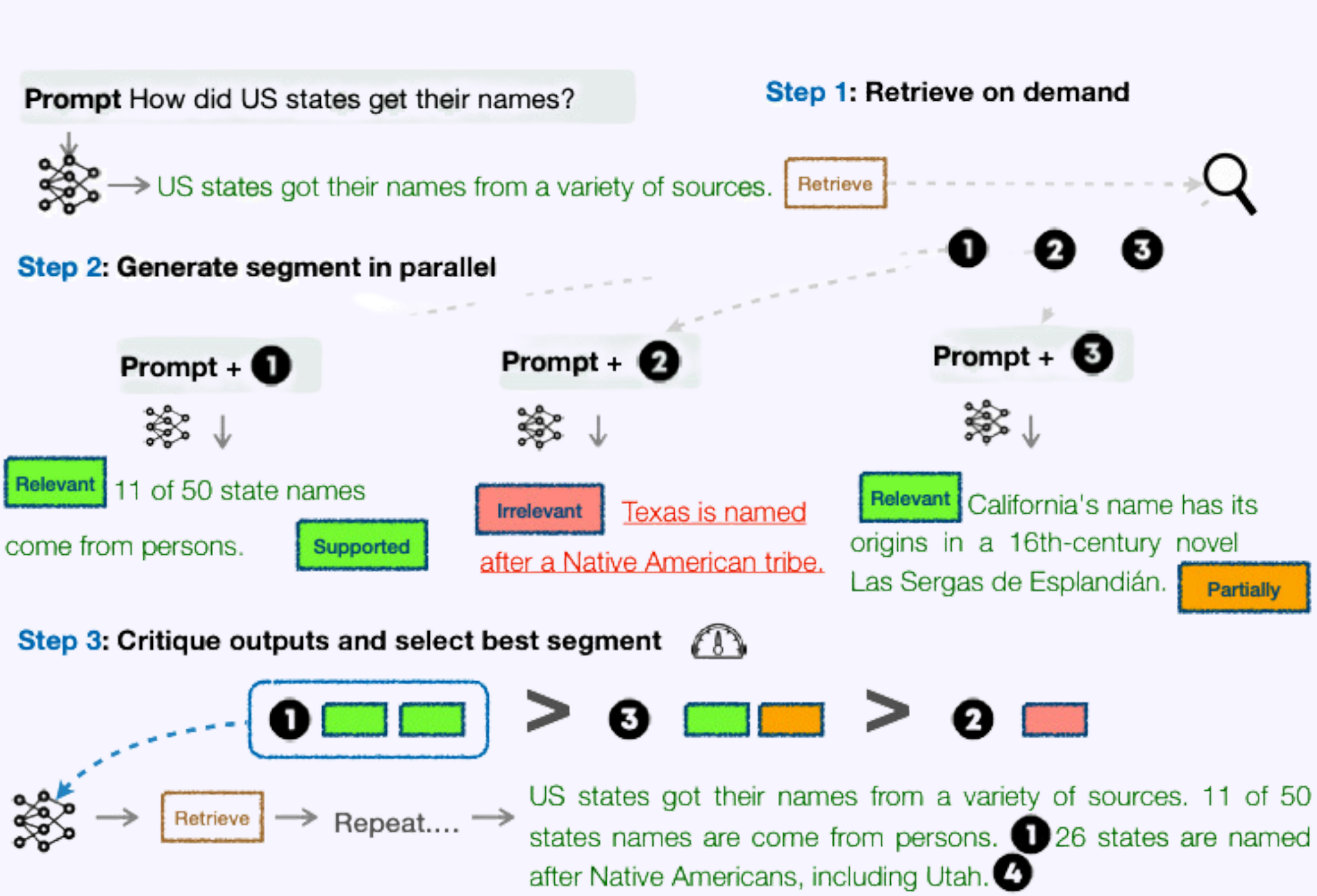
- **Step 3**: Iterative Retrieval and Generation - The model retrieves information in iterative cycles, each time refining its understanding of the task or goal. It may adapt its retrieval strategy based on new insights gained from previous iterations.

- **Step 4**: Dynamic Adjustment - The model continuously evaluates its progress toward the goal, adjusting its actions, retrieval strategies, or questions dynamically based on feedback or partial results.

- **Step 5**: Completion of Goal - Once the model determines that it has achieved the goal or completed the task, it generates a final output that reflects its findings or solution.

Self RAG

- Self RAG uses the model's own generated outputs as new data for future retrieval, enabling continuous self-improvement and learning over time.



How it works?

- **Step 1**: Initial Retrieval and Generation - The model begins by retrieving information and generating responses based on the input query.

- **Step 2**: Self-Storage - The generated outputs are stored in a dedicated repository as new data points.

- **Step 3**: Iterative Self-Retrieval - For future queries, the model retrieves not only from the original corpus but also from its own previously generated content. This allows it to build upon its past outputs, refining them further.

- **Step 4**: Self-Improvement Loop - As more content is generated and stored, the model learns from its past responses, identifying patterns, gaps, or inaccuracies, and continuously adjusting its retrieval and generation strategies to improve over time.

- **Step 5**: Enhanced Output - Over successive iterations, the model produces increasingly accurate and context-aware responses, leveraging the self-referential data it has accumulated.

Moreover,
we are offering a

Free Certification

on RAGs, check the link in the description

@Harshit Ahluwalia