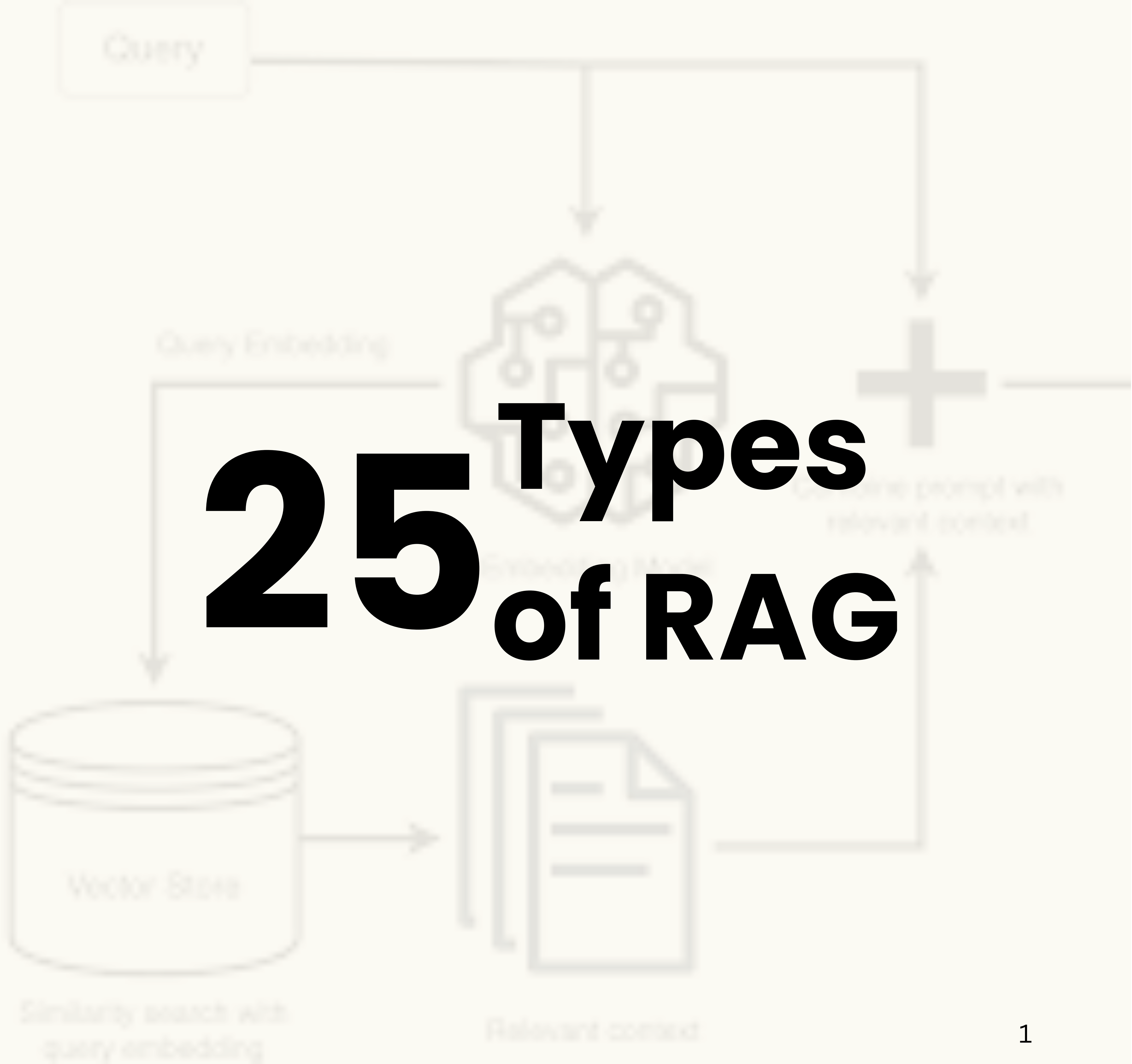
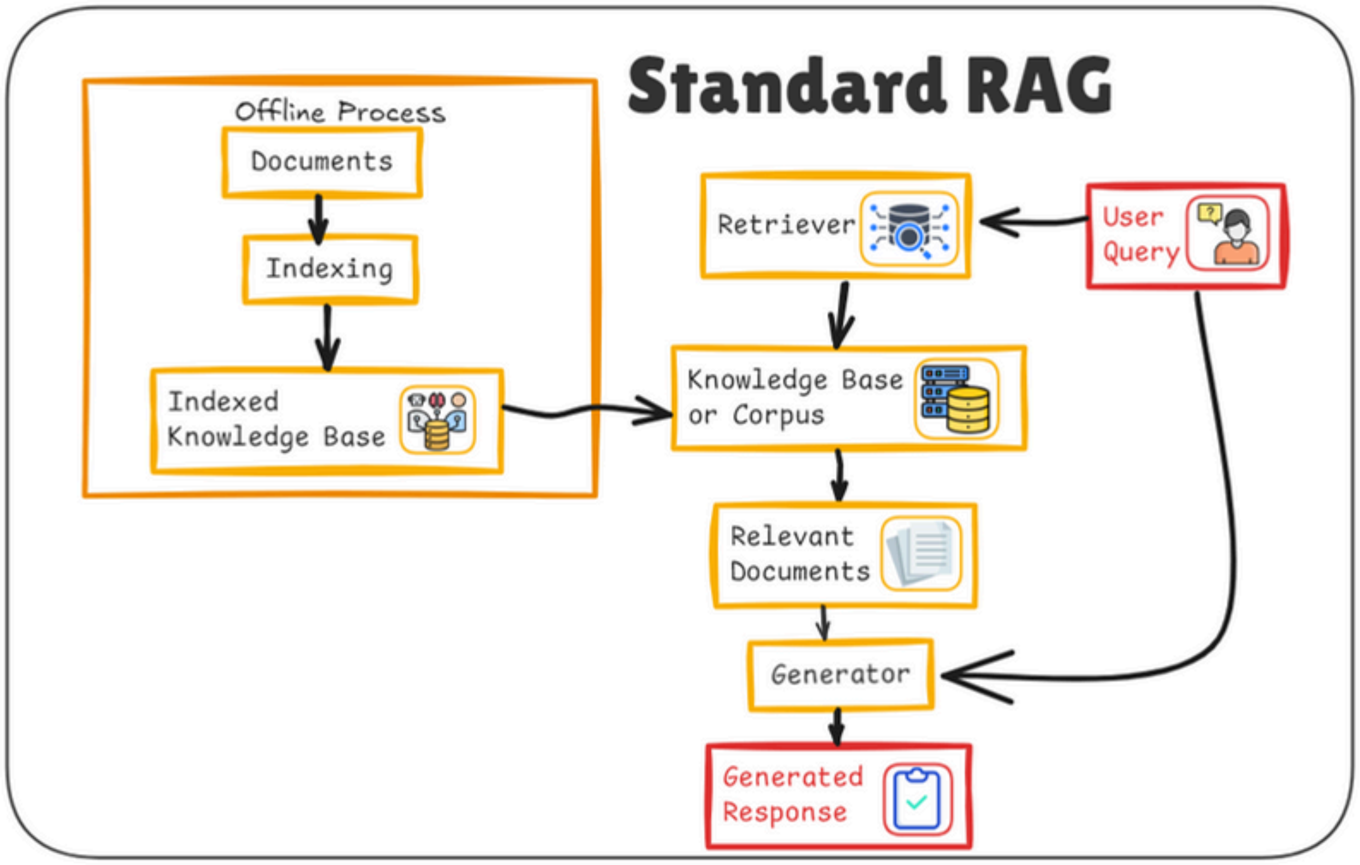


25 Types of RAG

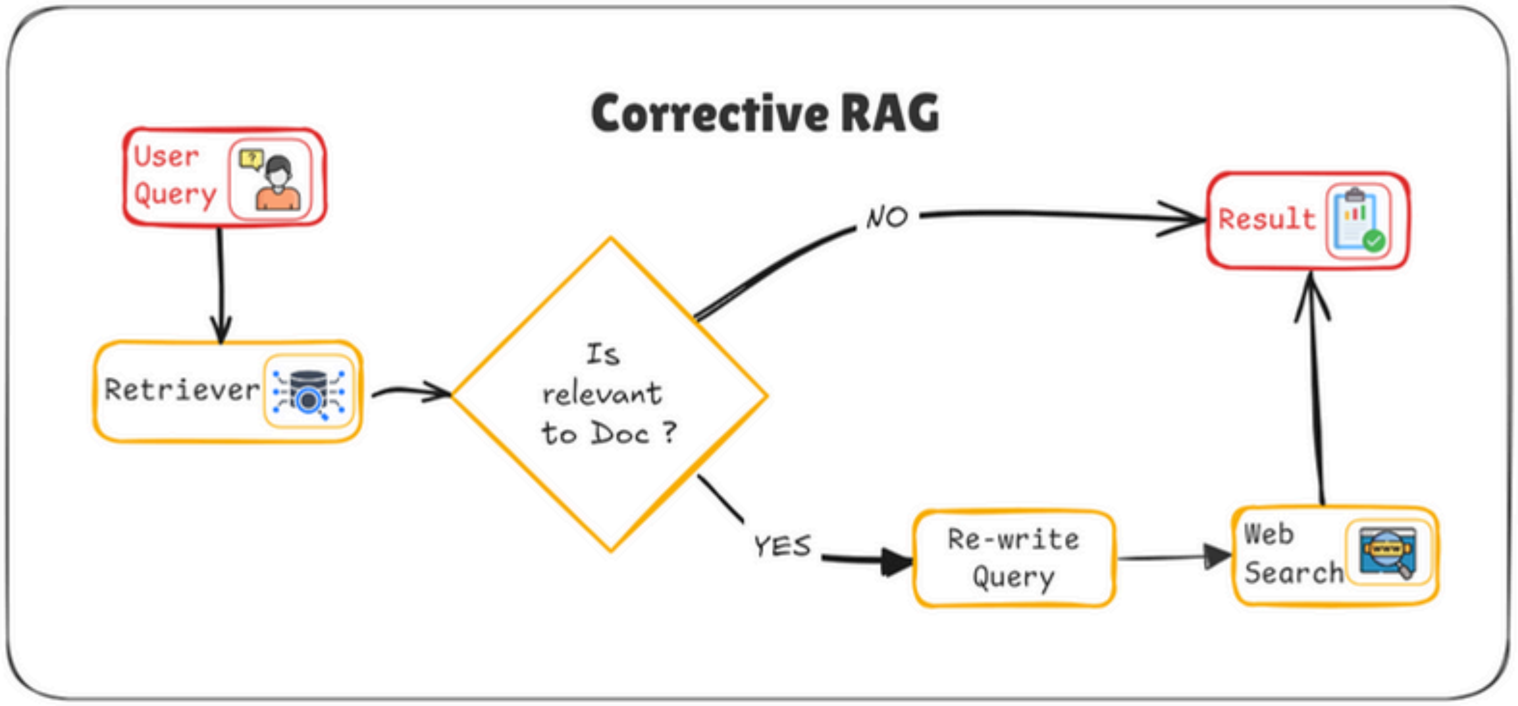
@Bhavishya Pandit

1. Standard RAG

- Combines **retrieval** with **large language models** for accurate, context-aware responses.
- Breaks **documents into chunks** for efficient information retrieval.
- Aims for **1-2 second response times** for real-time use.
- **Enhances answer quality** by leveraging external data sources.



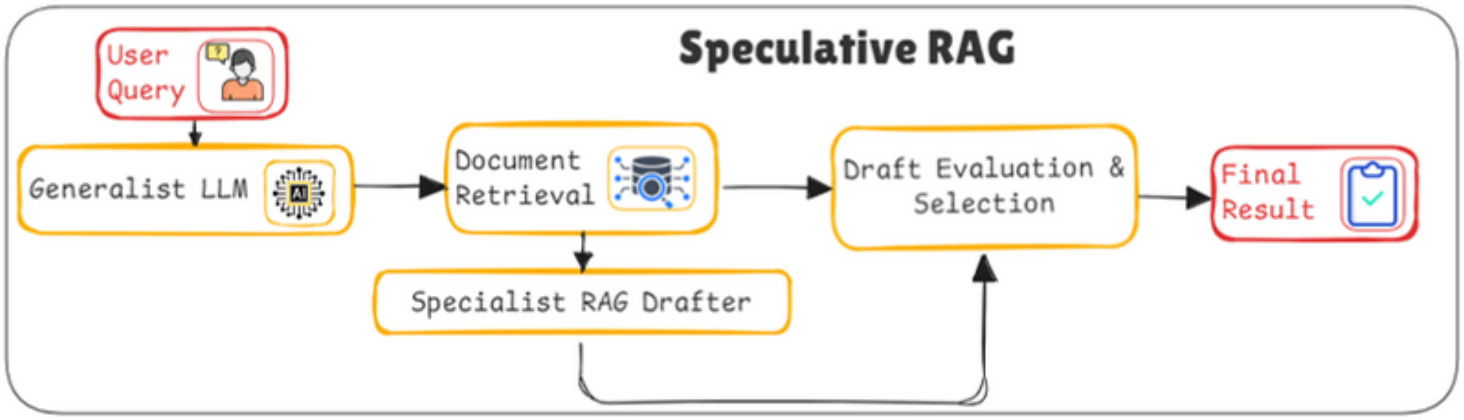
2. Corrective RAG



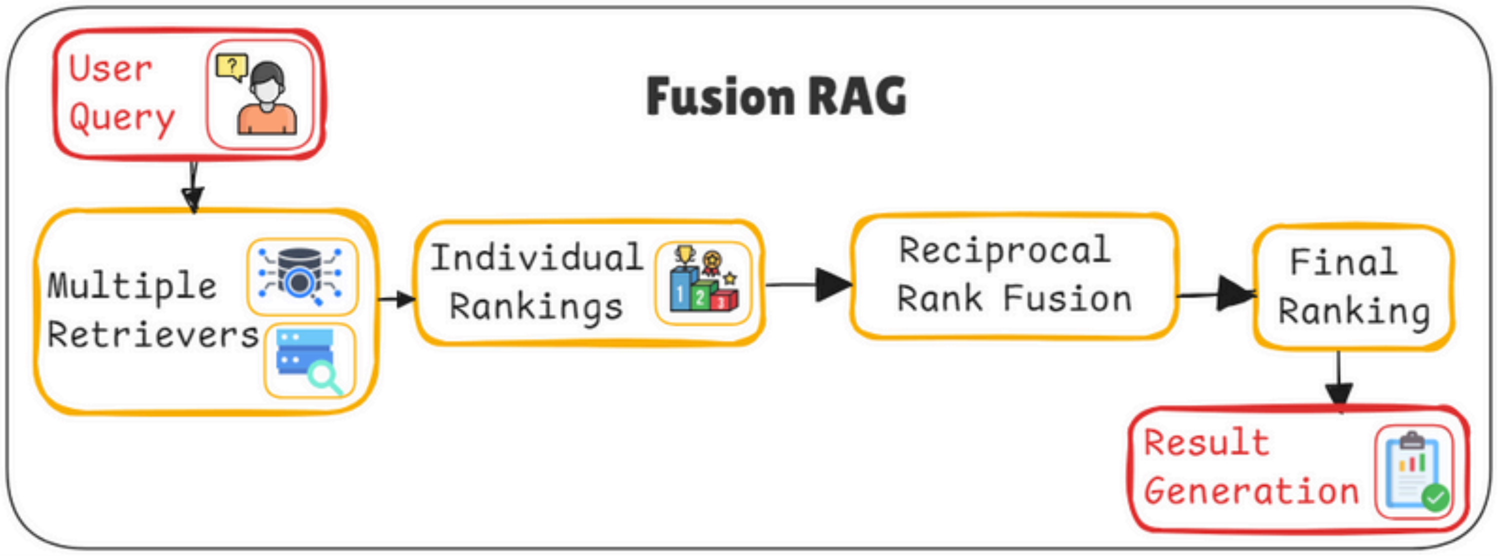
- Focuses on **identifying** and **fixing errors** in generated responses.
- Uses multiple passes to **improve outputs** based on feedback.
- Aims for **higher precision** and **user satisfaction** compared to standard RAG.
- Leverages user feedback to **enhance the correction** process .

3. Speculative RAG

- Uses a small **specialist model** for drafting and a larger **generalist model** for verification, ensuring efficiency and accuracy.
- **Parallel Drafting**: Speeds up responses by generating multiple drafts simultaneously.
- **Superior Accuracy**: Outperforms standard RAG systems.
- **Efficient Processing**: Offloads complex tasks to specialized models, reducing computational load.



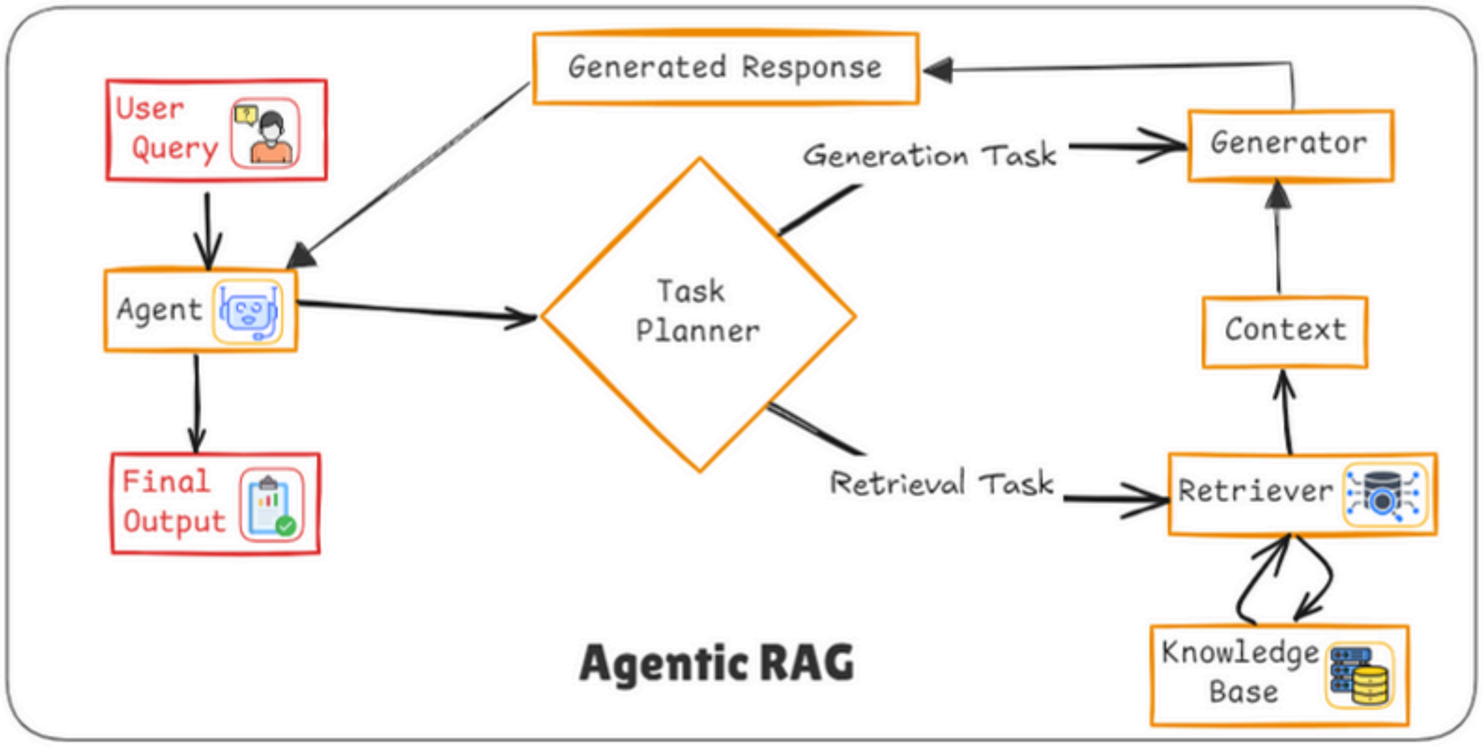
4. Fusion RAG



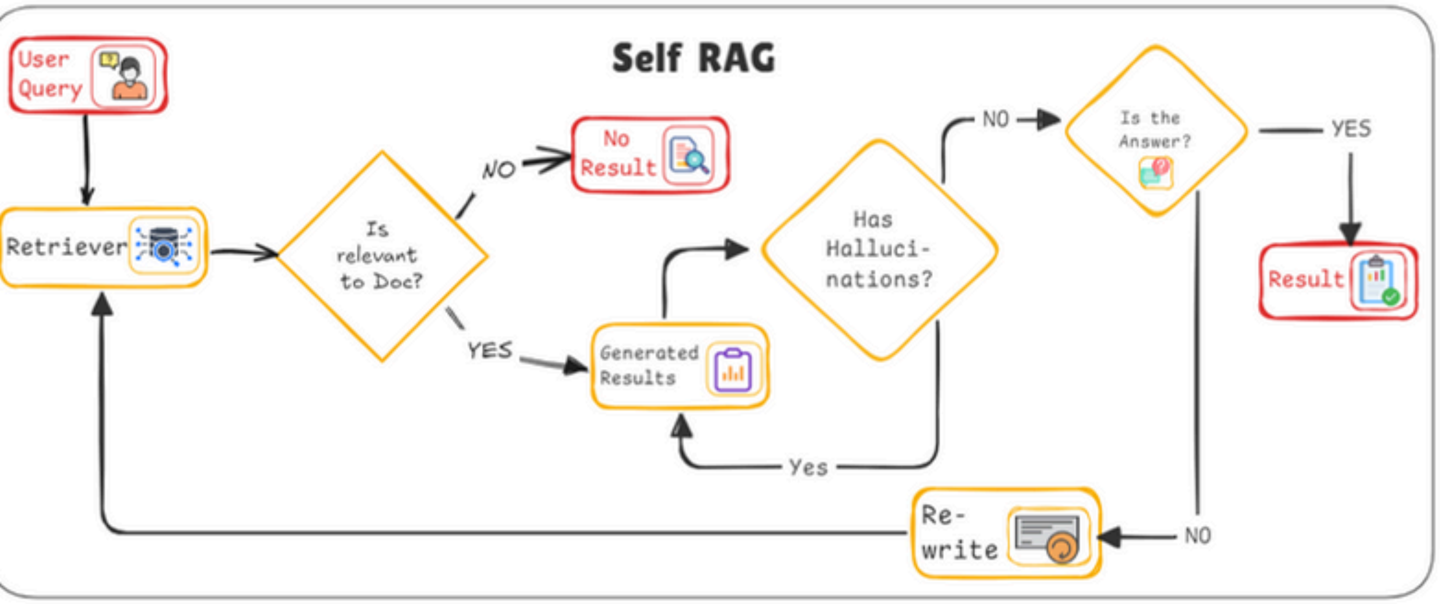
- Integrates **multiple retrieval** methods and data sources for enhanced response quality.
- Provides **comprehensive answers** by leveraging diverse data inputs.
- **Increases** system **resilience** by reducing dependence on a single source.
- Adapts retrieval **strategies dynamically** based on query context.

5. Agentic RAG

- Uses **adaptive agents** for real-time strategy adjustments in information retrieval.
- Accurately **interprets user intent** for relevant, trustworthy responses.
- **Modular design** enables easy integration of new data sources and features.
- Enhances **parallel processing** and **performance** on complex tasks by running agents concurrently.

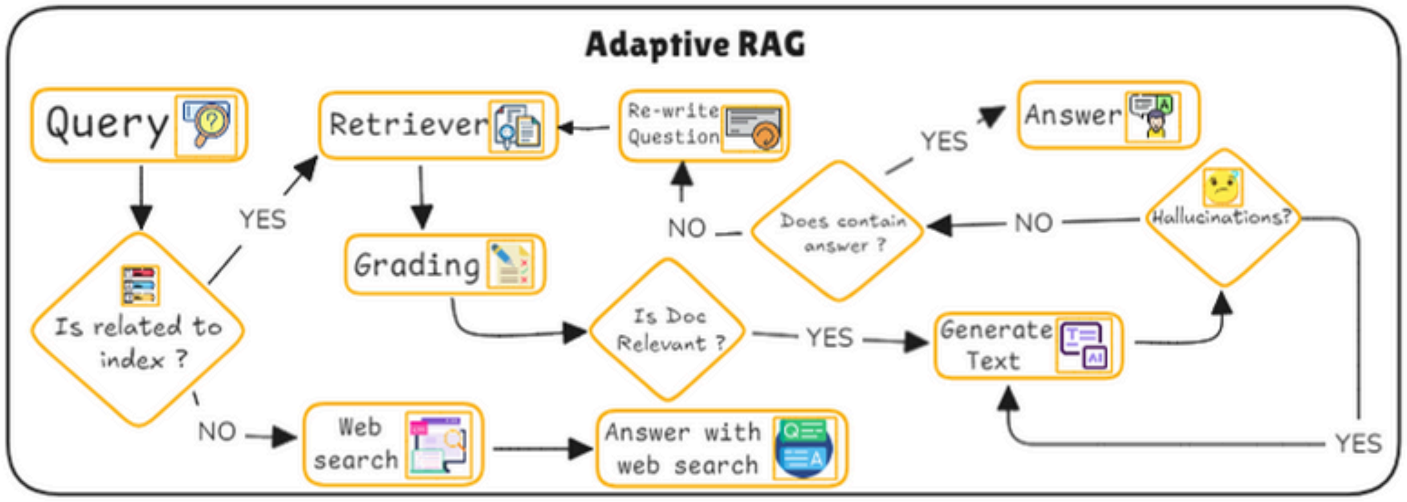


6. Self RAG



- Uses the model's own outputs as retrieval candidates for **better contextual relevance**.
- Refines responses iteratively, improving **consistency** and **coherence**.
- Grounds responses in prior outputs for **increased accuracy**.
- **Adapts retrieval** strategies based on the conversation's evolving context.

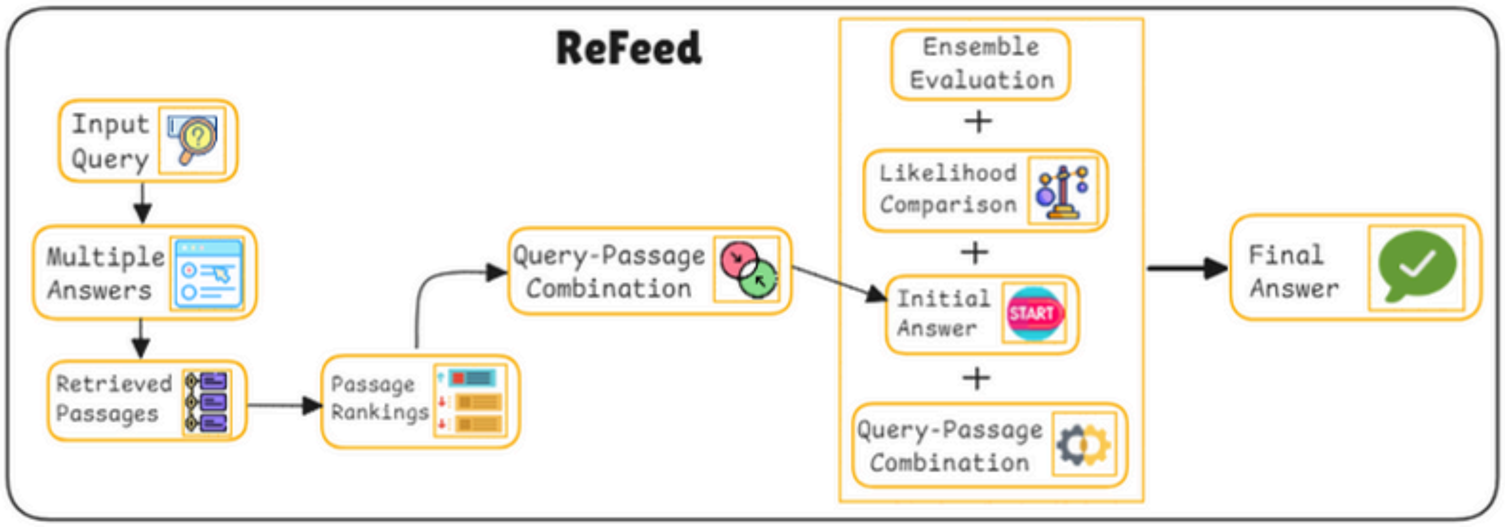
7. Adaptive RAG



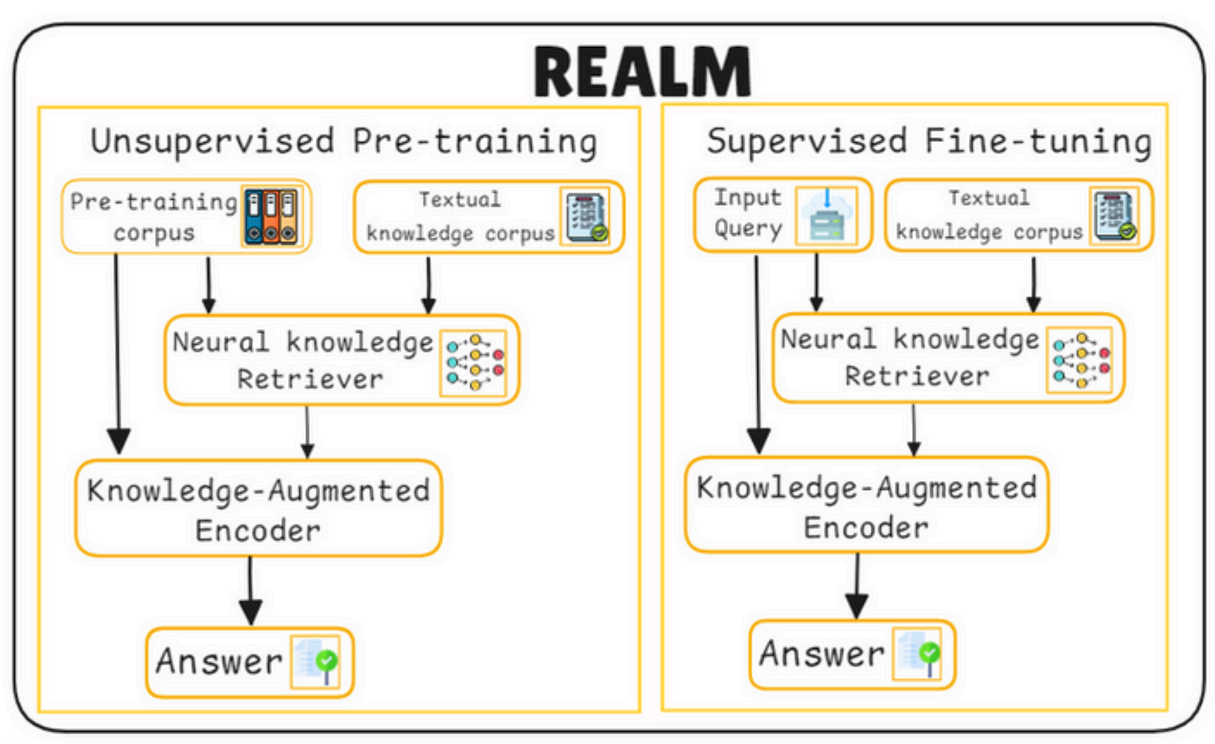
- It **dynamically** decides when to retrieve external knowledge, balancing internal and external knowledge.
- It uses **confidence scores** from the language model's internal states to assess retrieval necessity.
- An honesty probe helps the model **avoid hallucinations** by aligning its output with its actual knowledge.
- It **reduces unnecessary retrievals**, improving both efficiency and response accuracy.

8. REFEED Retrieval Feedback

- REFEED refines model outputs using **retrieval feedback without fine-tuning**.
- Initial answers are improved by retrieving relevant documents and adjusting the response based on the new information.
- Generates **multiple answers** to improve retrieval accuracy.
- Combines pre- and post-retrieval outputs using a **ranking system to enhance answer reliability**.



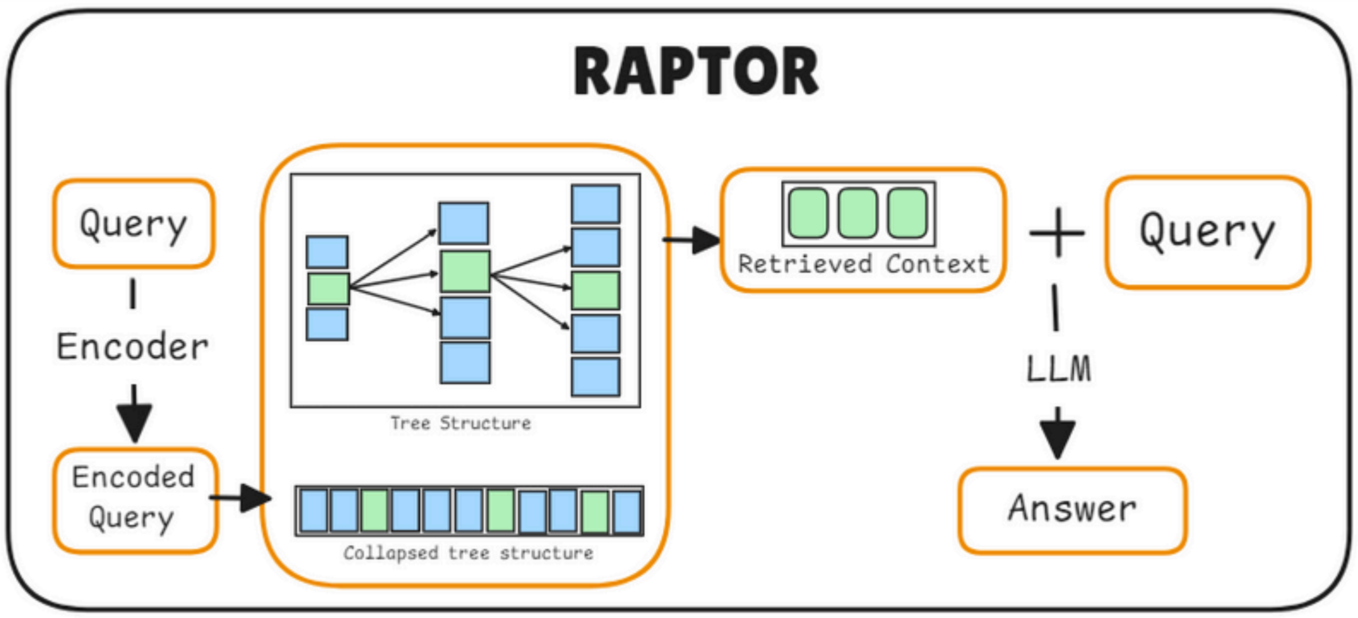
9. REALM



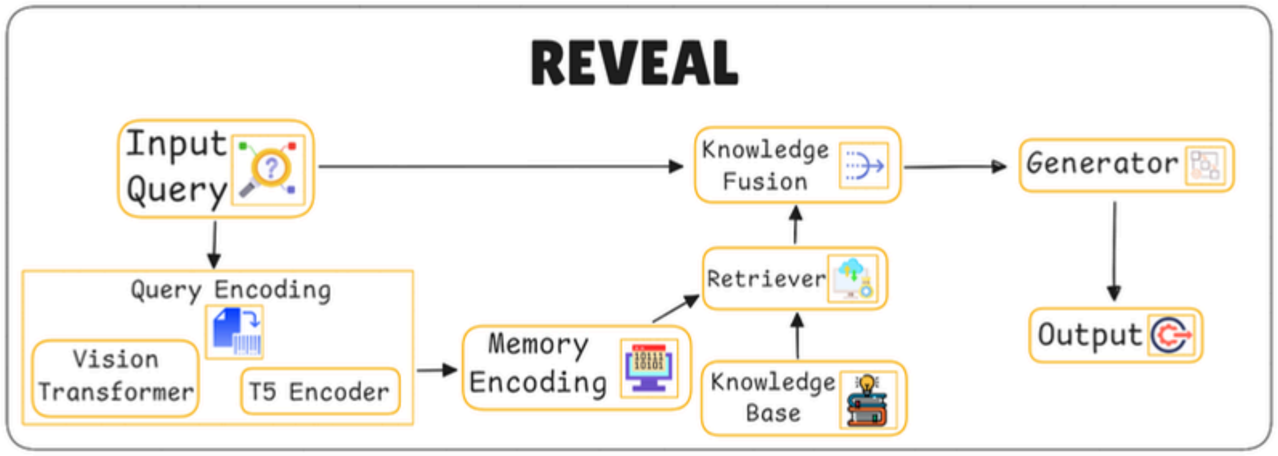
- REALM retrieves relevant documents from large corpora like Wikipedia to **enhance model predictions**.
- The retriever is trained with masked language modeling, optimizing retrieval to **improve** prediction **accuracy**.
- It uses **Maximum Inner Product Search** to efficiently find relevant documents from millions of candidates during training.
- REALM outperforms previous models in **Open-domain Question Answering** by integrating external knowledge.

10. RAPTOR- Tree-Organized Retrieval

- RAPTOR builds a **hierarchical tree** by **clustering** and **summarizing text recursively**.
- It enables retrieval at **different abstraction levels**, combining **broad themes** with specific details.
- RAPTOR **outperforms traditional methods** in complex question-answering tasks.
- Offers tree traversal and collapsed tree methods for **efficient information retrieval**.



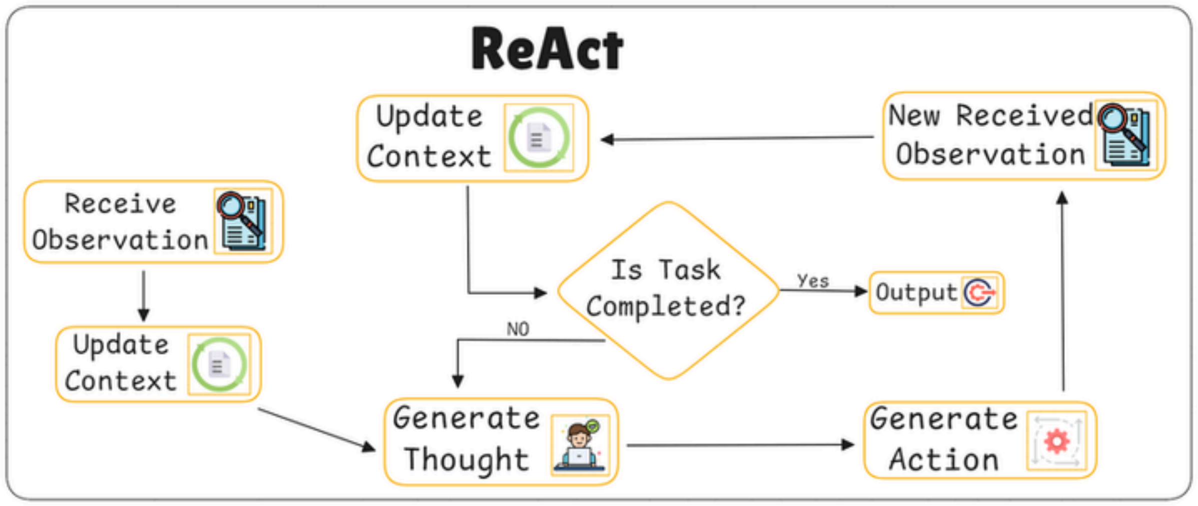
11. REVEAL for Visual-Language Model



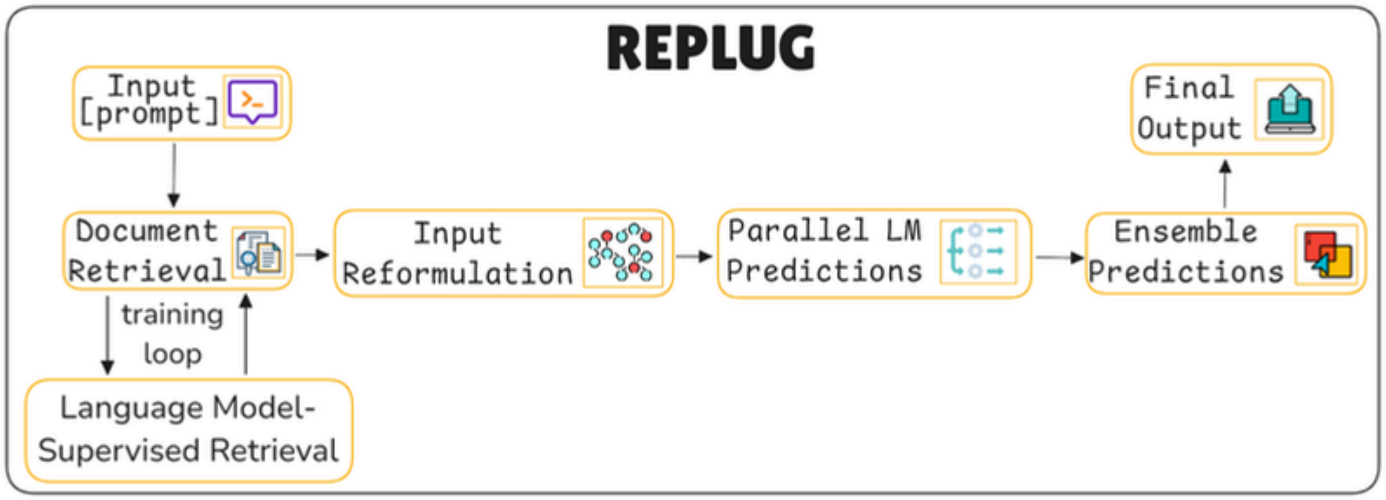
- This technique combines reasoning with **task-specific actions** and **external knowledge**, **improving decision-making**.
- It **minimizes errors** by grounding reasoning in real-world facts, reducing inaccuracies and hallucinations.
- The method offers clear, human-like task-solving steps, **enhancing transparency** and **interpretability**.
- REVEAL achieves **strong performance across tasks with fewer training** examples, making models efficient, adaptable, and responsive.

12. REACT

- The ReAct technique combines **reasoning and action**, allowing models to interact with their environment.
- It maintains situational **awareness by updating context** with past actions and thoughts.
- The model generates **task-aligned thoughts** to guide logical decision-making.
- **Real-time feedback** refines understanding, reducing errors and enhancing transparency and reliability.

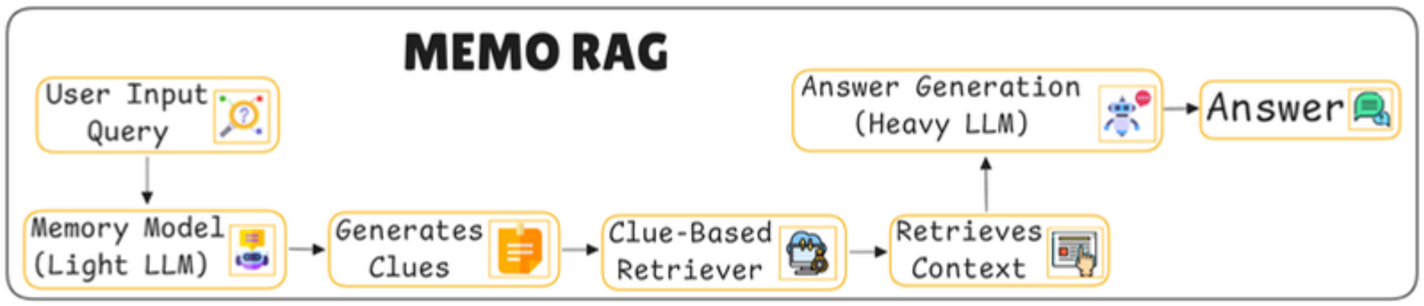


13. REPLUG Retrieval Plugin



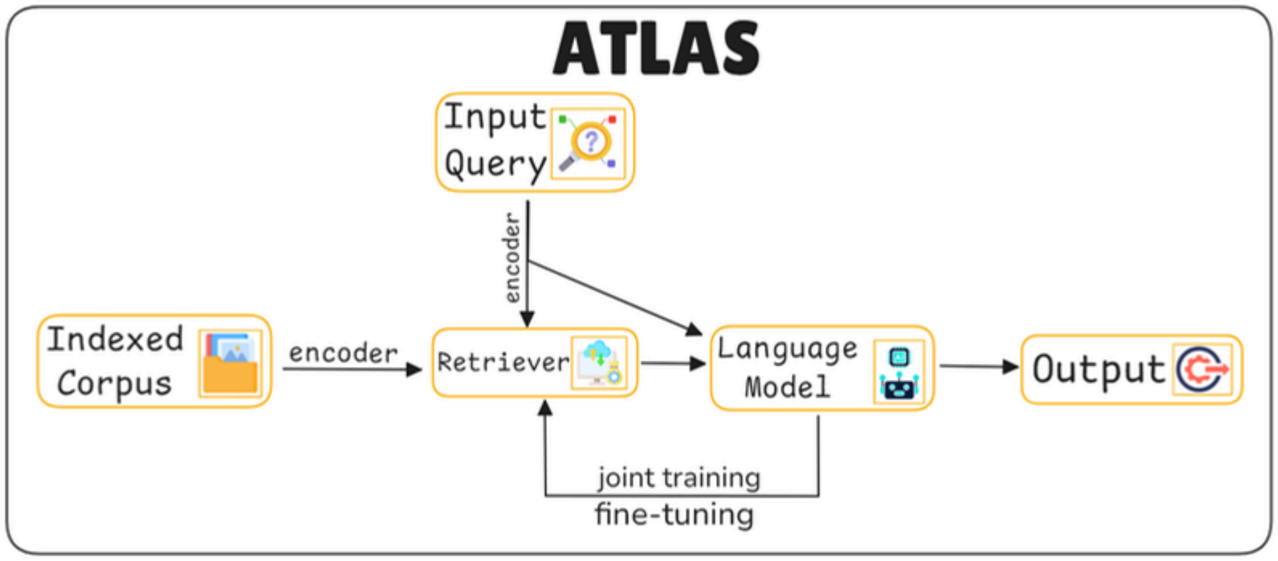
- REPLUG **enhances LLMs** by retrieving relevant external documents to **improve prediction accuracy**.
- It treats the language model as a fixed "**black box**", prepending retrieved information to the input.
- This **flexible design** works with existing models without modifications, integrating external knowledge to r**educe errors and hallucinations**.
- The retrieval component **can be fine-tuned with model feedback**, aligning better with the model's needs and expanding niche knowledge.

14. MEMO RAG



- Memo RAG **combines memory and retrieval** to handle complex queries.
- A memory model generates draft answers that guide the search for external information.
- The retriever then gathers relevant data from databases, which a more powerful language model uses to create a comprehensive final answer.
- This method helps Memo RAG **manage ambiguous queries** and **efficiently process large amounts of information** across various tasks.

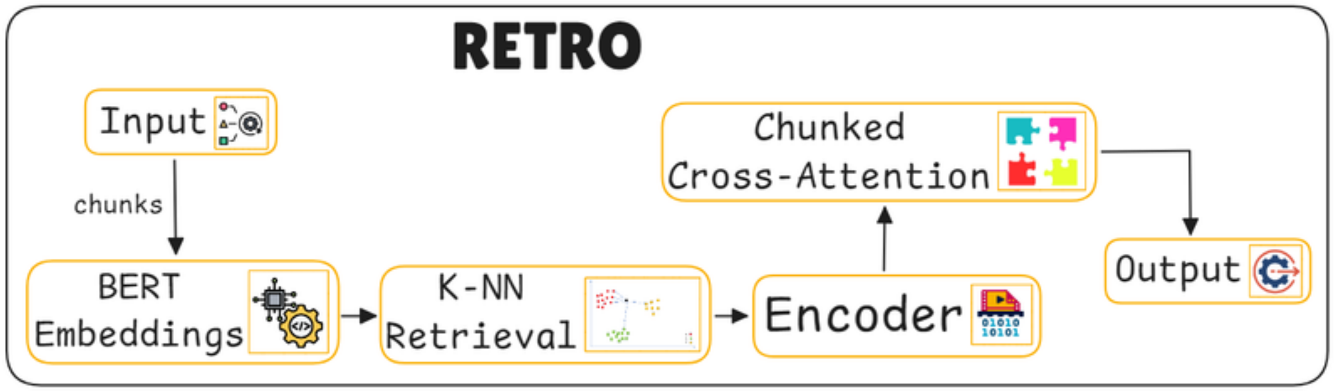
15. Attention-based RAG



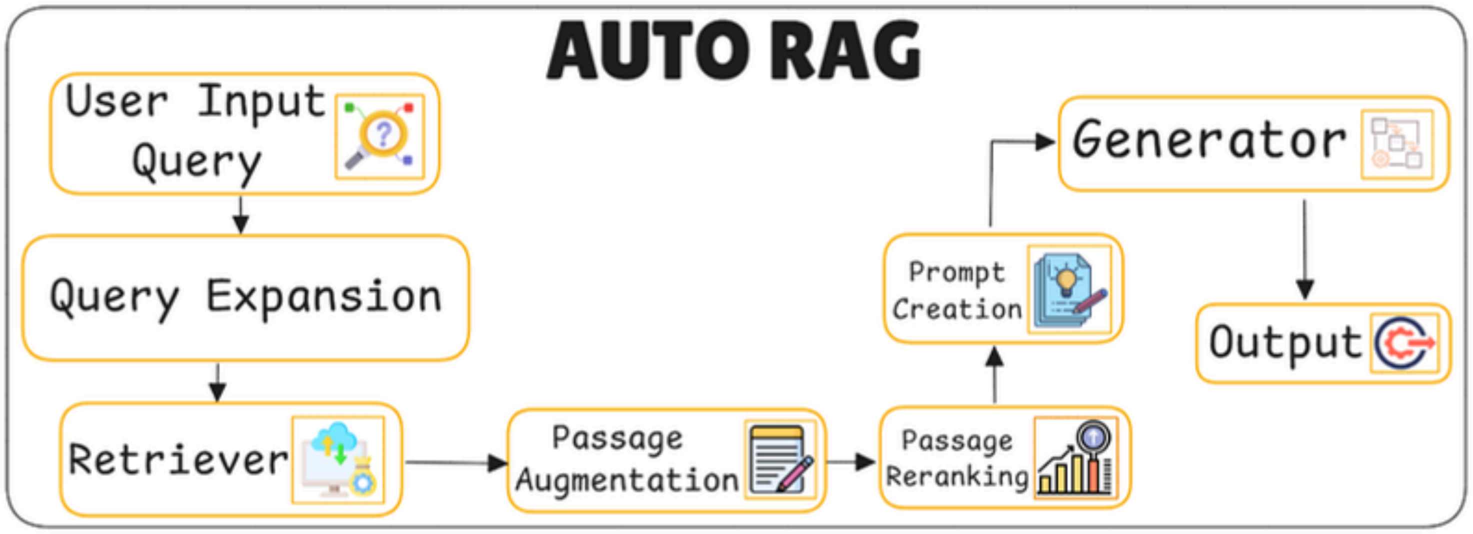
- ATLAS **improves language models** by retrieving external documents to enhance accuracy, especially in question-answering tasks.
- It uses a **dual-encoder retriever** to identify the top-K relevant documents from large text corpora.
- A **Fusion-in-Decoder model** integrates query and document information, generating accurate responses while reducing reliance on memorization.
- The **document index is updatable without retraining**, ensuring it remains current and effective for knowledge-intensive tasks.

16. RETRO

- RETRO splits input text into chunks and **retrieves similar information from a large text database** to enrich context.
- It uses **pre-trained BERT embeddings** to pull in relevant chunks from external data, enhancing context.
- **Chunked cross-attention** integrates these chunks, improving predictions without a major increase in model size.
- This approach **enhances tasks** like question answering and text generation efficiently, accessing extensive knowledge with lower computational demands than larger models.



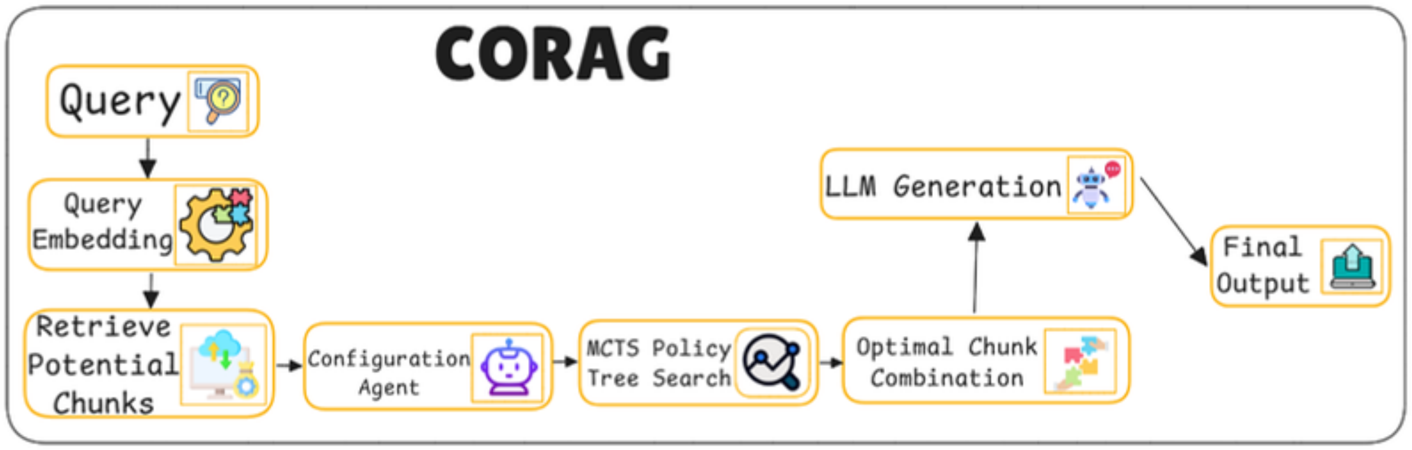
17. AUTO RAG

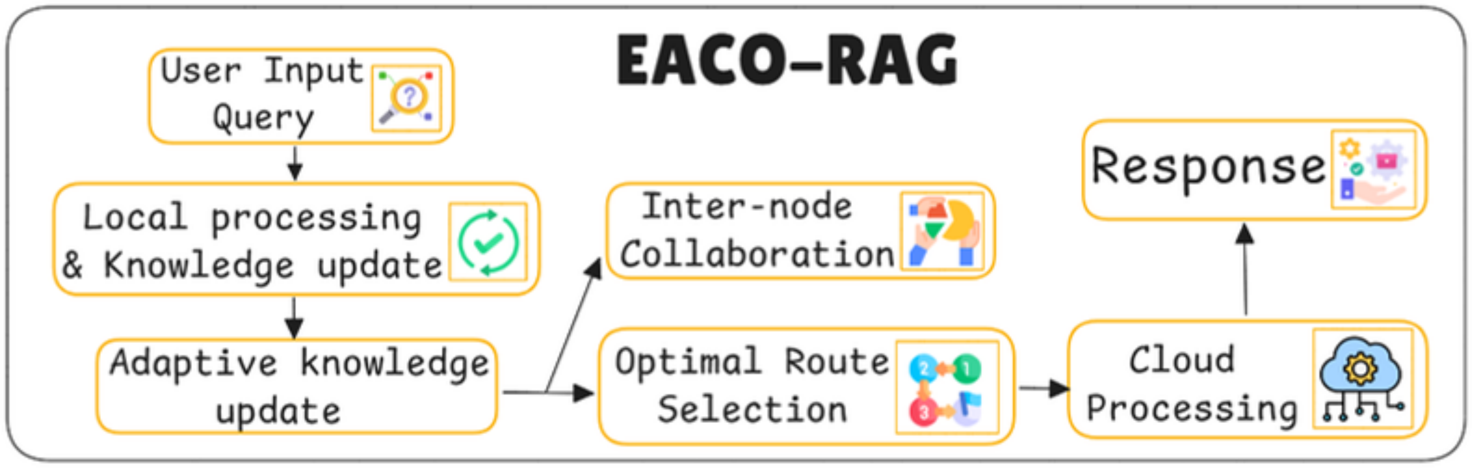


- AutoRAG automates optimization for Retrieval-Augmented Generation (RAG) systems.
- It evaluates modules like query expansion, retrieval, and reranking for best performance.
- The framework uses a modular, node-based structure to test various configurations.
- A greedy optimization approach enhances efficiency across different datasets.

18. CORAG :Cost-Constrained RAG

- It enhances RAG by optimizing relevant chunk selection from databases.
- It tackles three challenges: correlating chunks efficiently, handling non-monotonic utility where adding chunks may reduce utility, and adapting to diverse query types.
- CORAG uses Monte Carlo Tree Search (MCTS) for optimal chunk combination while factoring in cost constraints, achieving up to a 30% improvement over baseline models.



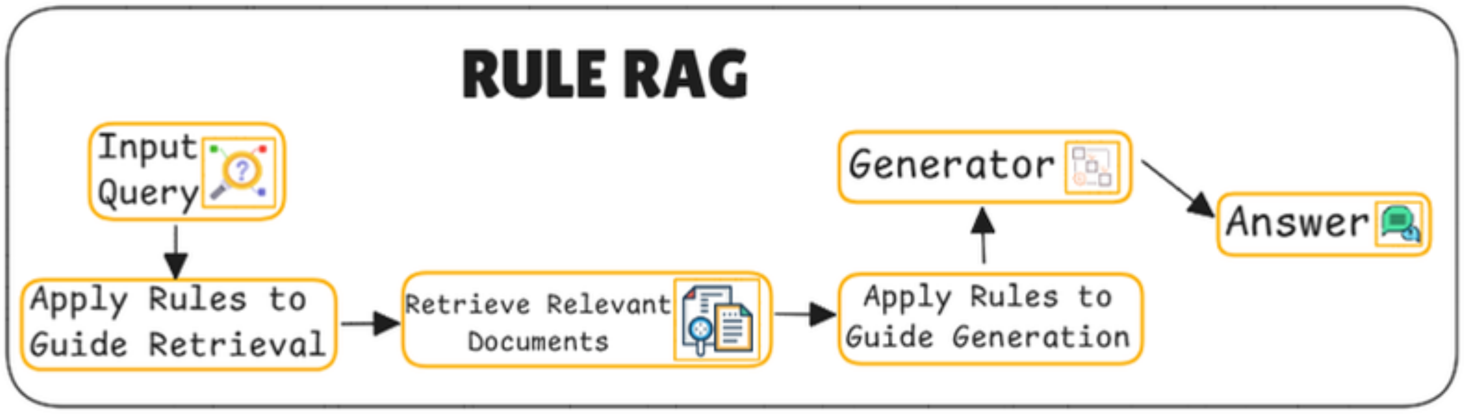


19. EACO-RAG

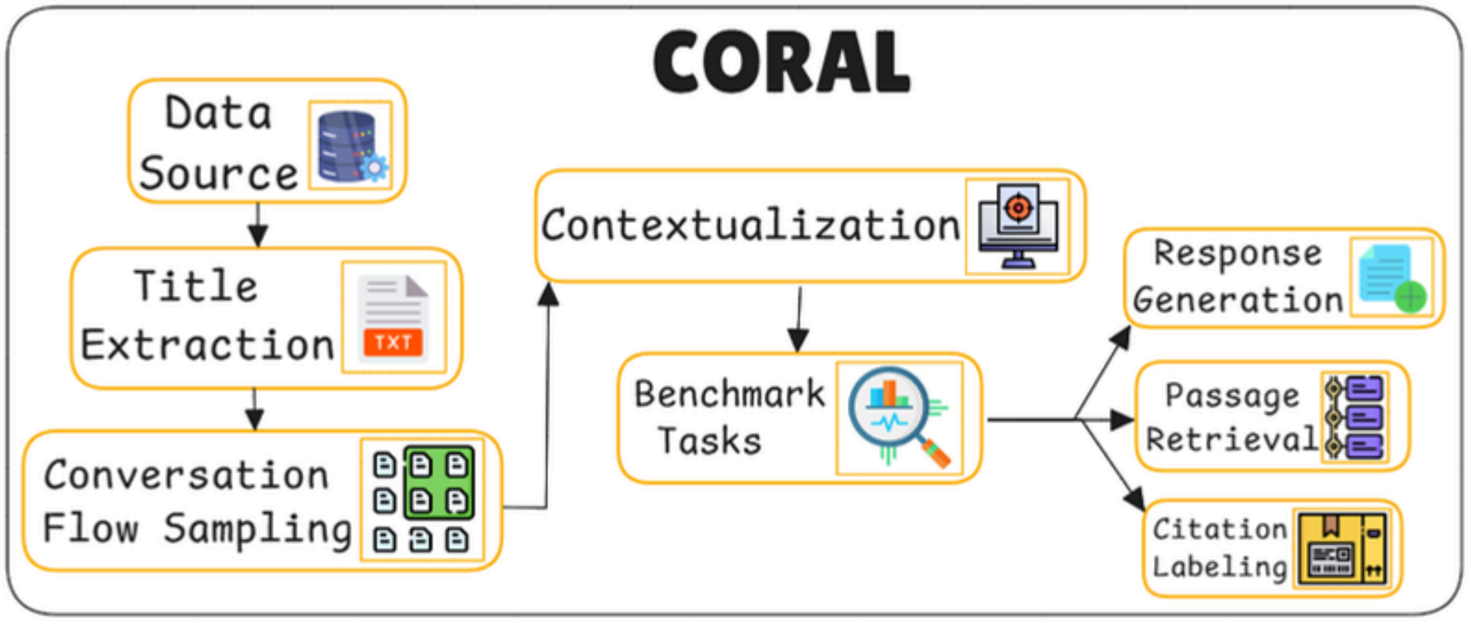
- EACO-RAG enhances RAG with edge computing for **faster, efficient responses**.
- Vector datasets are distributed across edge nodes, **reducing delays and resource use**.
- Adaptive knowledge updates and inter-node collaboration **improve response accuracy**.
- A multi-armed bandit approach **optimizes cost**, **accuracy**, and **delay in real-time**.

20. RULE RAG

- Rule-RAG **enhances question answering** by adding rule-based guidance to RAG.
- It retrieves documents **logically relevant** to queries using predefined rules.
- Rules are also used to guide answer generation for a**ccuracy and context**.
- It includes **in-context learning** (ICL) and a **fine-tuned** version (FT) for better retrieval and generation.



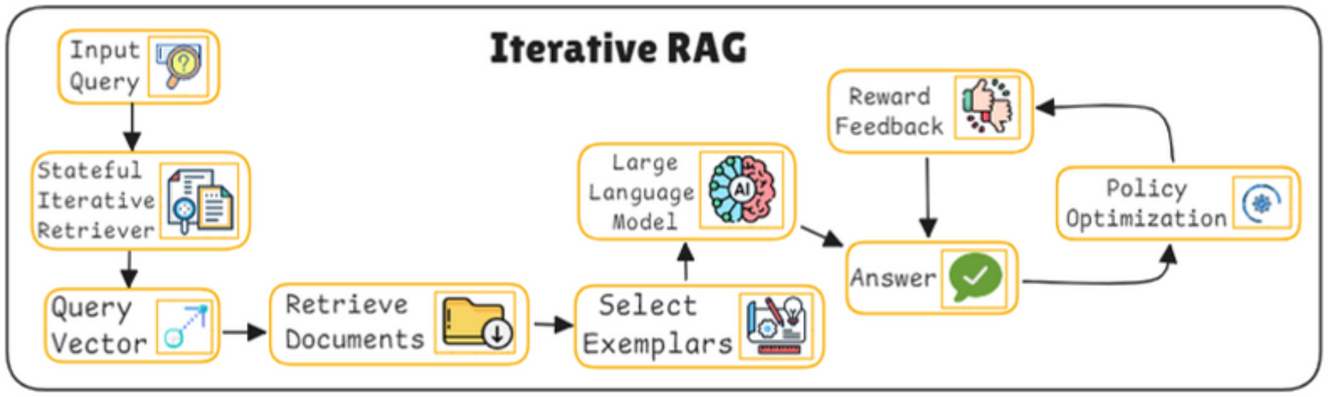
21. Conversational RAG



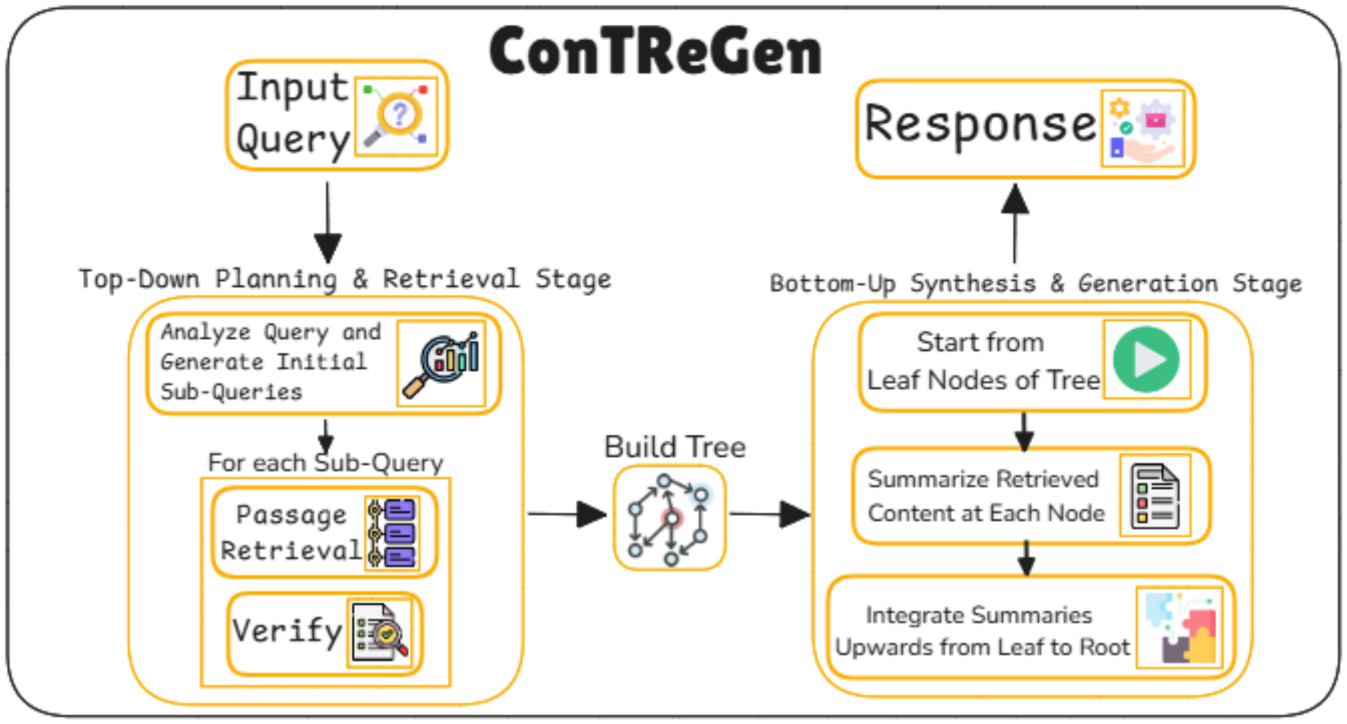
- CORAL benchmarks **multi-turn conversational** RAG using Wikipedia data.
- It evaluates passage retrieval, response generation, and citation labeling.
- CORAL **handles open-domain**, **realistic**, **multi-turn conversations**.
- It bridges single-turn RAG research and real-world multi-turn needs.

22. Iterative RAG

- Unlike traditional retrieval, iterative RAG performs **multiple retrieval steps**, refining its search based on feedback from previously selected documents.
- Retrieval decisions follow a **Markov decision process**.
- **Reinforcement learning** improves retrieval performance.
- The iterative retriever **maintains an internal state**, allowing it to adjust future retrieval steps based on the accumulated knowledge from previous iterations.



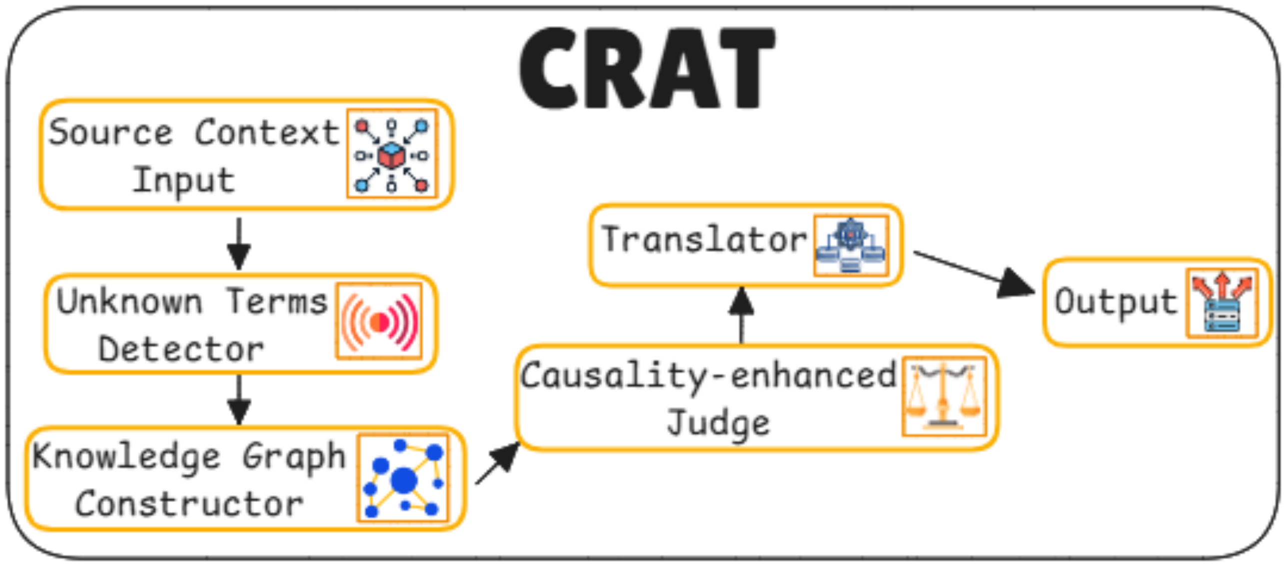
23. Context-driven Tree-structured Retrieval



- It is a context-driven, **tree-structured RAG approach** that decomposes complex queries into hierarchical sub-queries, enhancing retrieval depth.
- Its workflow has two stages: a **top-down exploration** of query facets, creating a tree of retrieved passages, followed by **bottom-up synthesis**, integrating summarized information to produce a coherent long-form response.
- This framework **reduces gaps in information** and **improves the quality of generated content**.

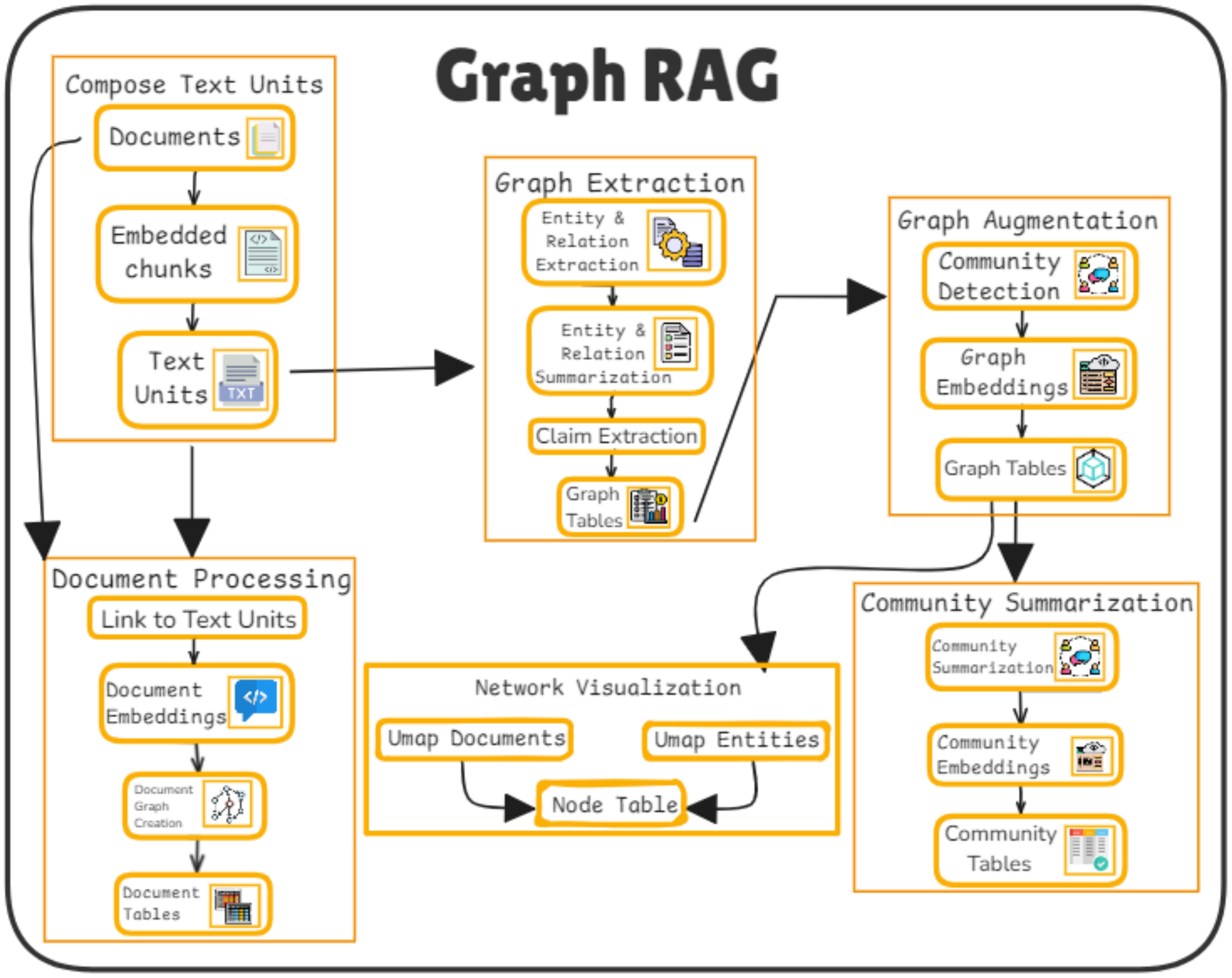
24. Causality-Enhanced Reflective and Retrieval-Augmented Translation

- **Multi-Agent Framework**: CRAT enhances translation by detecting, clarifying, and translating ambiguous terms.
- **Knowledge Graph** : Combines internal and external sources to capture context for accurate term use.
- **Causality Validation**: A judge agent validates information to ensure context-aligned translations.
- **Refined Output**: CRAT delivers precise, consistent translations by using validated knowledge.



25. Graph RAG

- Graph RAG constructs a **knowledge graph** on-the-fly, linking relevant entities during retrieval.
- It leverages **node relationships** to decide when and how much external knowledge to retrieve.
- **Confidence scores** from the graph guide expansion, avoiding irrelevant additions.
- This approach **improves efficiency** and **response accuracy** by keeping the knowledge graph compact and relevant.



Bhavishya Pandit

Save

Like

Repost