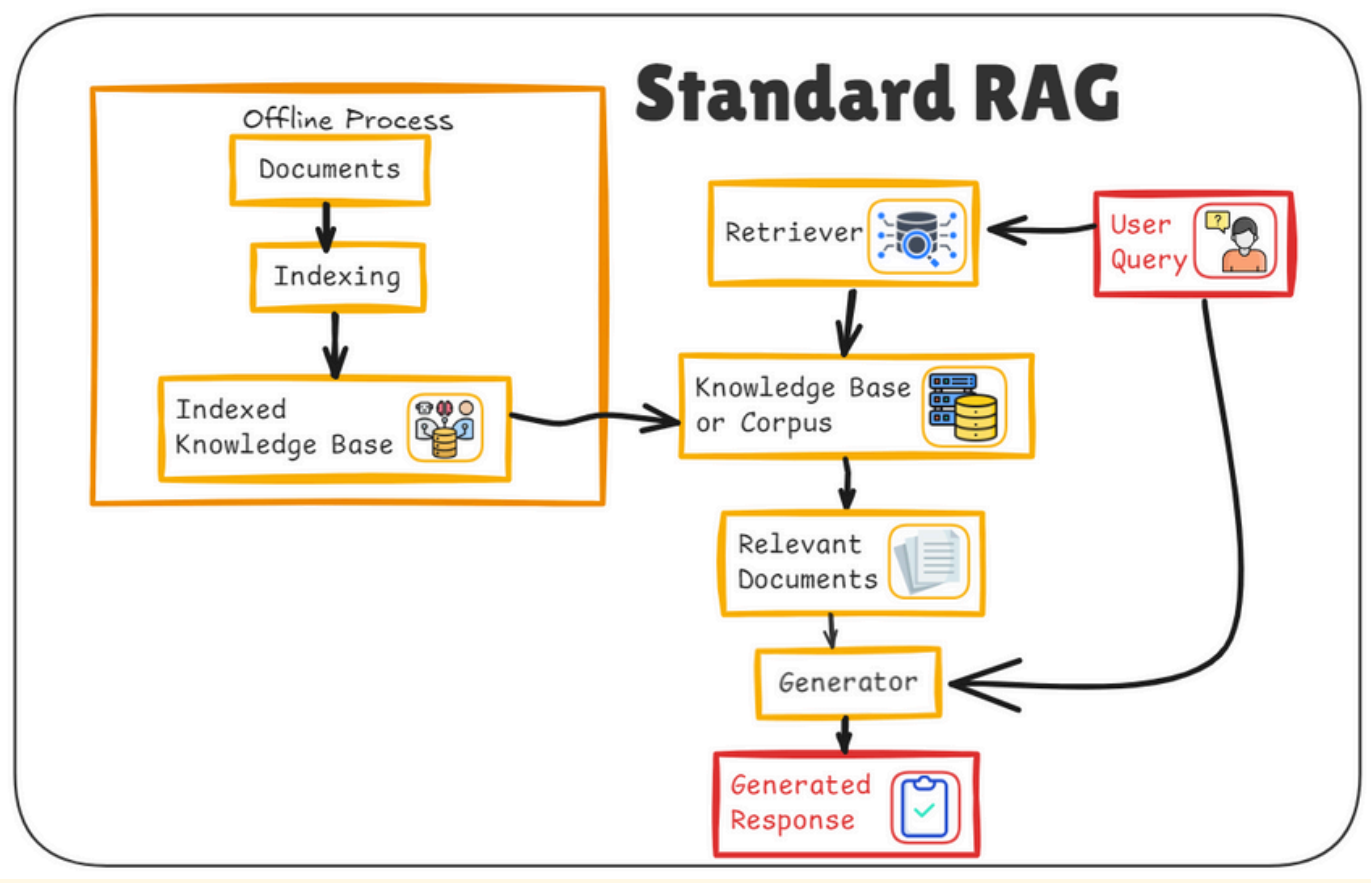
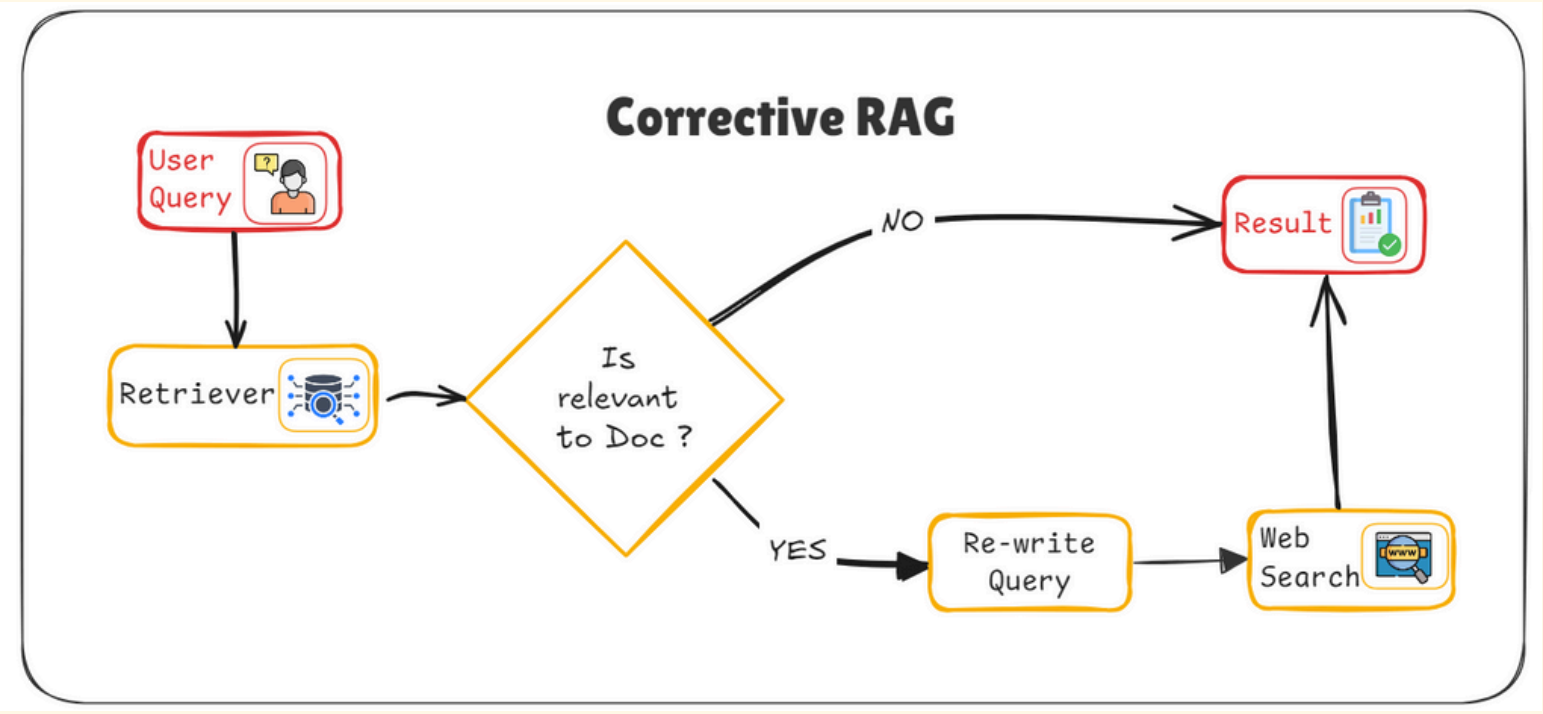
6 Different RAG Techniques

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.



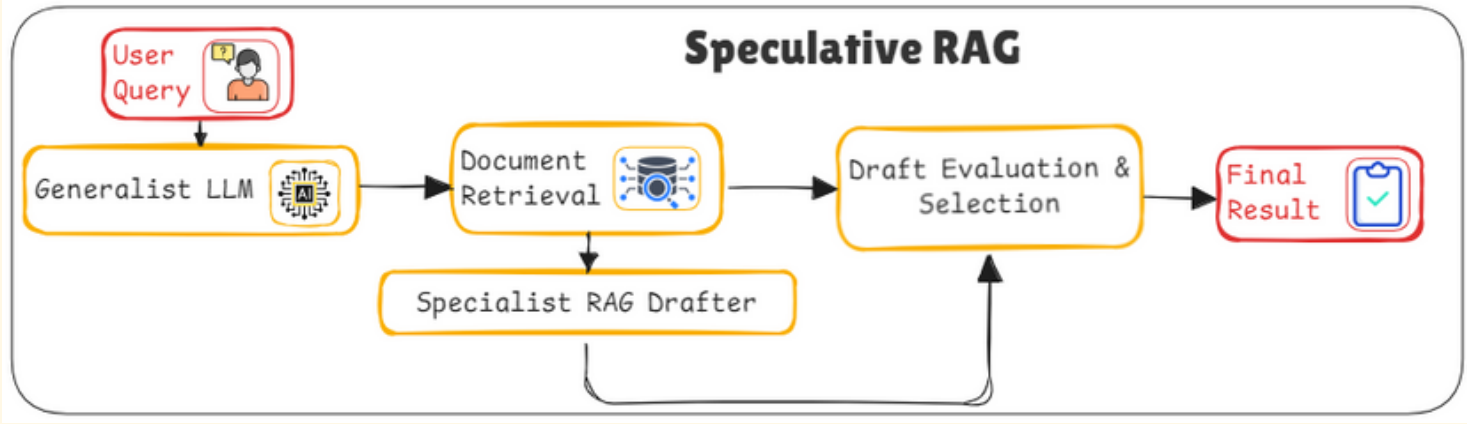
Corrective RAG



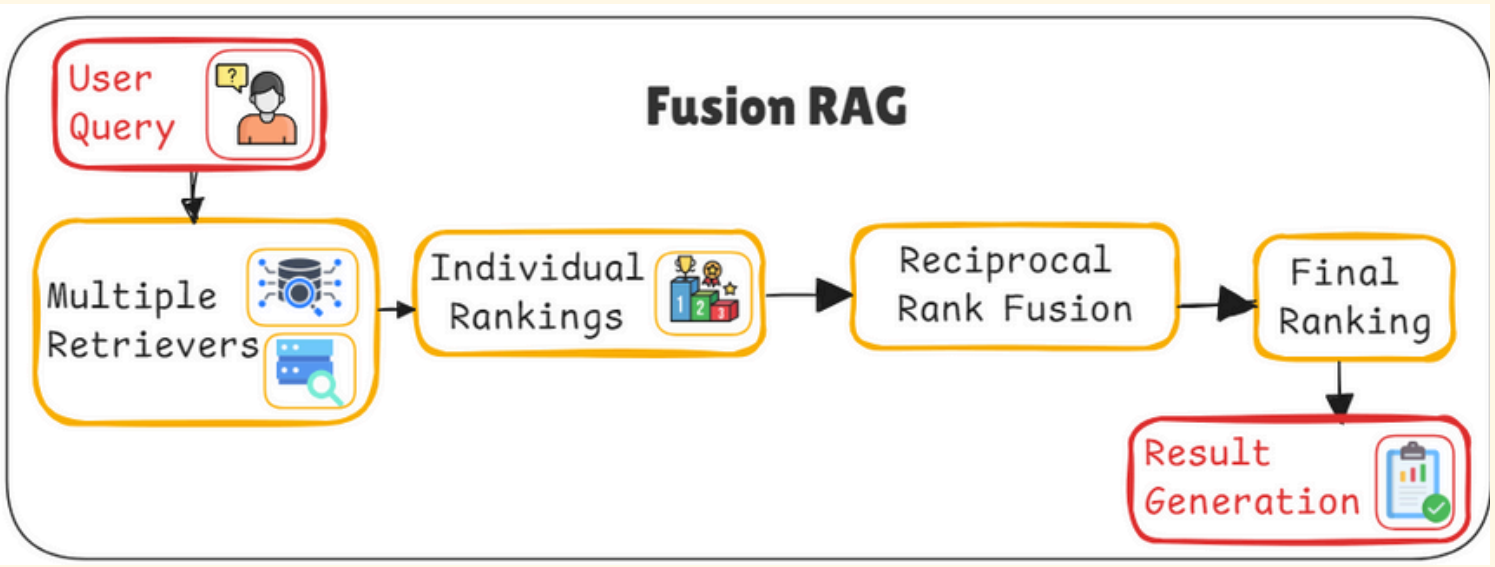
- Focuses on **identifying** and **fixing errors** in generated responses.
- Uses multiple passes to **improve outputs** based on feedback.
- Aims for **higher precision** and **user satisfaction** compared to standard RAG.
- Leverages user feedback to **enhance the correction** process .

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.



Fusion RAG

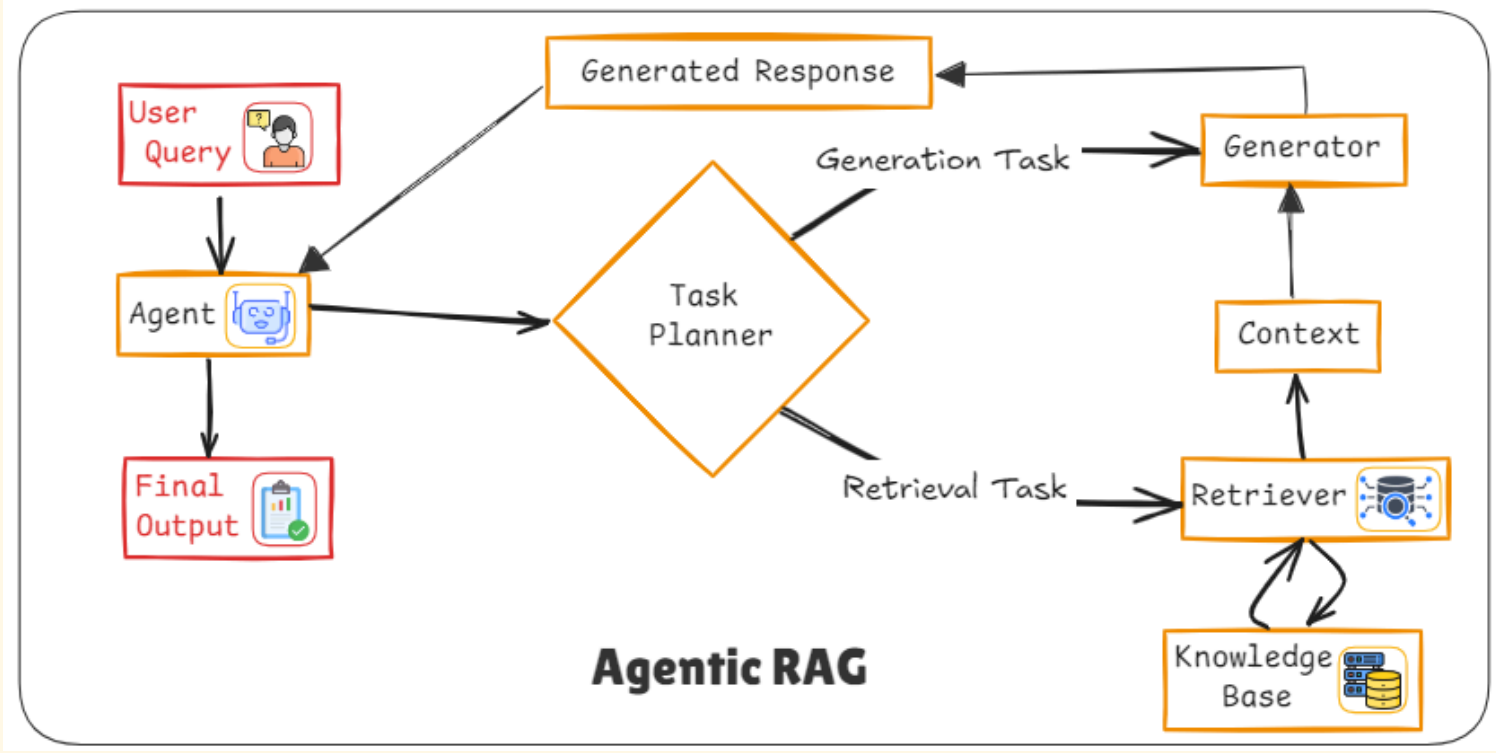


- Integrates **multiple retrieval** methods and data sources for enhanced response quality.
- Provides **comprehensive answers** by leveraging diverse data inputs.
- **Increases** system **resilience** by reducing dependence on a single source.
- Adapts retrieval **strategies dynamically** based on query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

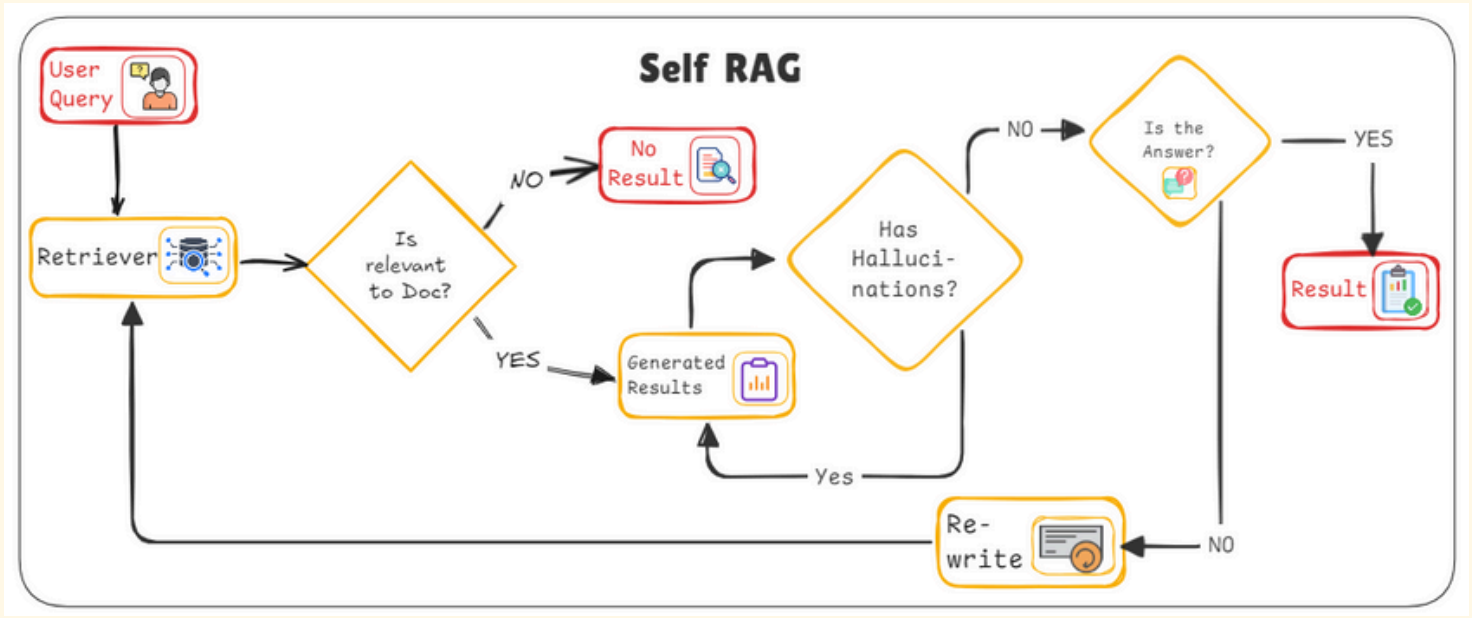
@Bhavishya Pandit

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.



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.

@Bhavishya Pandit

Do you know any other technique?

LET US KNOW IN THE COMMENTS

@Bhavishya Pandit

Bhavishya Pandit

Share your thoughts

Save for later

Like this post