RAG Frameworks

Shankar Kashamshetty

Introduction

Retrieval-Augmented Generation (RAG) is a machine learning framework that combines the advantages of both retrieval-based and generation-based models.

It leverages external data sources by retrieving relevant documents or facts and then generating an answer or output based on the retrieved information and the user query.

This blend of retrieval and generation leads to better-informed outputs that are more accurate and comprehensive than models that rely solely on generation.

RAG Frameworks

- Standard RAG
- Corrective RAG
- Speculative RAG
- Fusion RAG
- Agentic RAG
- Self RAG
- Graph RAG
- Modular RAG
- RadioRAG

Standard RAG

It is the foundational model of Retrieval-Augmented Generation

The model first retrieves relevant information from a large external dataset, such as a knowledge base or a document repository, and then generates a response using a language model

The retrieved documents serve as additional context to the input query, enhancing the language model’s capacity to create accurate and informative answers

When to Use:

- when the query requires precise and factual information

Example :

- QA systems or tasks that summarize large documents

Challenge/Issues:

- The retrieval step sometimes fails to identify the most relevant documents, leading to suboptimal or incorrect responses

Corrective RAG

It builds upon Standard RAG’s foundations but adds a layer designed to correct potential errors or inconsistencies in the generated response

After the retrieval and generation stages, a corrective mechanism is employed to verify the accuracy of the generated output. This correction can involve further consultation of the retrieved documents, fine-tuning the language model, or implementing feedback loops where the model self-assesses its output against factual data

Corrective RAG enhances trust in the model’s responses

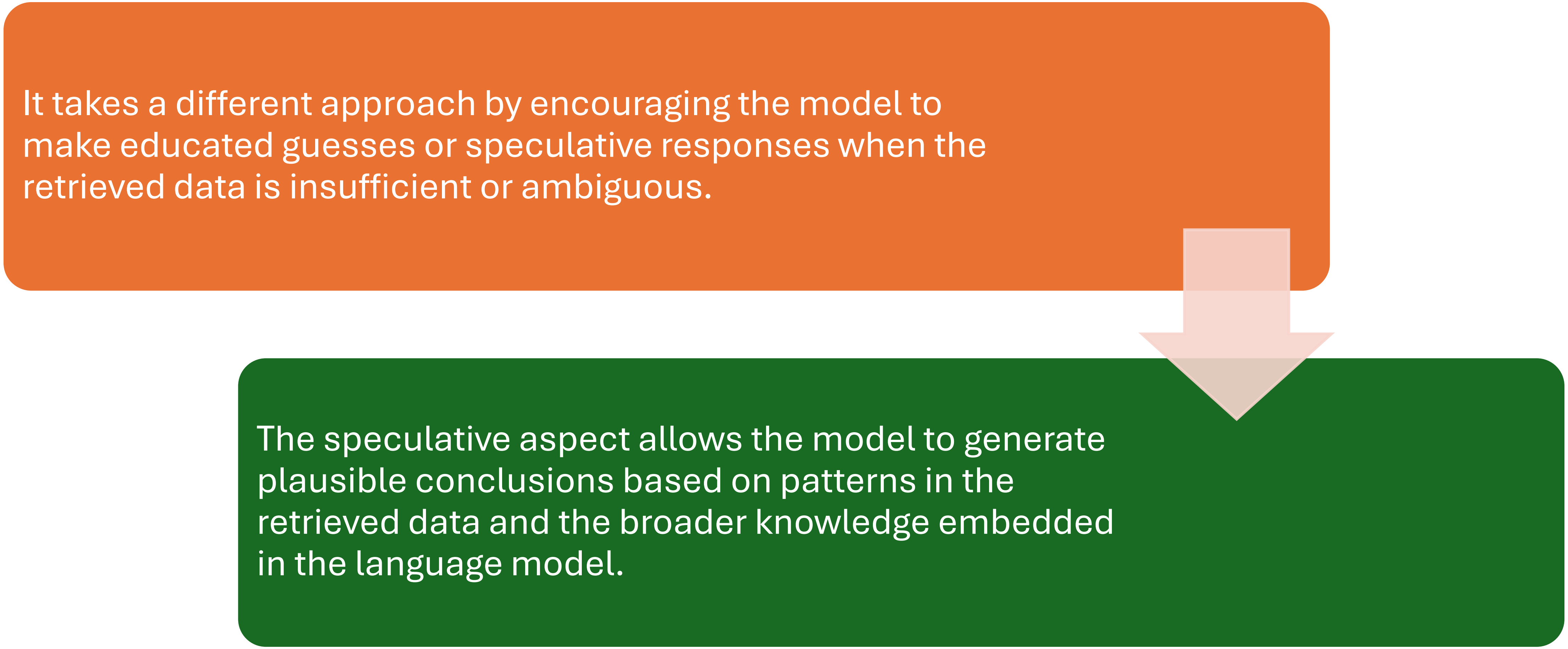
When to Use:

- Corrective RAG is especially useful in highly precise domains

Example:

- Medical Diagnosis, Legal advice, or Scientific research

Speculative RAG



It takes a different approach by encouraging the model to make educated guesses or speculative responses when the retrieved data is insufficient or ambiguous.

The speculative aspect allows the model to generate plausible conclusions based on patterns in the retrieved data and the broader knowledge embedded in the language model.

When to use:

- This model is designed to handle scenarios where complete information may not be available, yet the system still needs to provide a useful response

Example:

- In exploratory research or initial consultations in finance, marketing, or product development, Speculative RAG offers potential solutions or insights to guide further investigation or refinement

Challenge:

- Ensuring that users know the speculative nature of the responses. Since the model is designed to generate hypotheses rather than factual conclusions, the speculative nature must be communicated clearly to avoid misleading users.

Fusion RAG

It is an advanced model that merges information from multiple sources or perspectives to create a synthesized response.

Fusion RAG retrieves data from several sources and then uses the generation model to integrate these diverse inputs into a cohesive, well-rounded output.

When to Use:

- This approach is particularly useful when different datasets or documents offer complementary or contrasting information.

Example:

- This model is beneficial in complex decision-making processes, such as business strategy or policy formulation, where different viewpoints and datasets must be considered

Challenge:

- The risk of information overload or conflicting data points. The model needs to balance and reconcile diverse inputs without compromising the coherence or accuracy of the generated output.

Agentic RAG

It introduces autonomy into the RAG framework by allowing the model to act more independently in determining what information is needed and how to retrieve it.

Unlike traditional RAG models, which are typically limited to predefined retrieval mechanisms, Agentic RAG incorporates a decision-making component that enables the system to identify additional sources, prioritize different types of information, or even initiate new queries based on the user’s input.

When to Use:

- This autonomous behavior makes Agentic RAG particularly useful in dynamic environments where the required information may evolve, or the retrieval process needs to adapt to new contexts

Example:

- Autonomous research systems, Customer service bots, and Intelligent assistants

Challenge:

- Overly autonomous systems may stray too far from the intended task or provide irrelevant information to the original query.

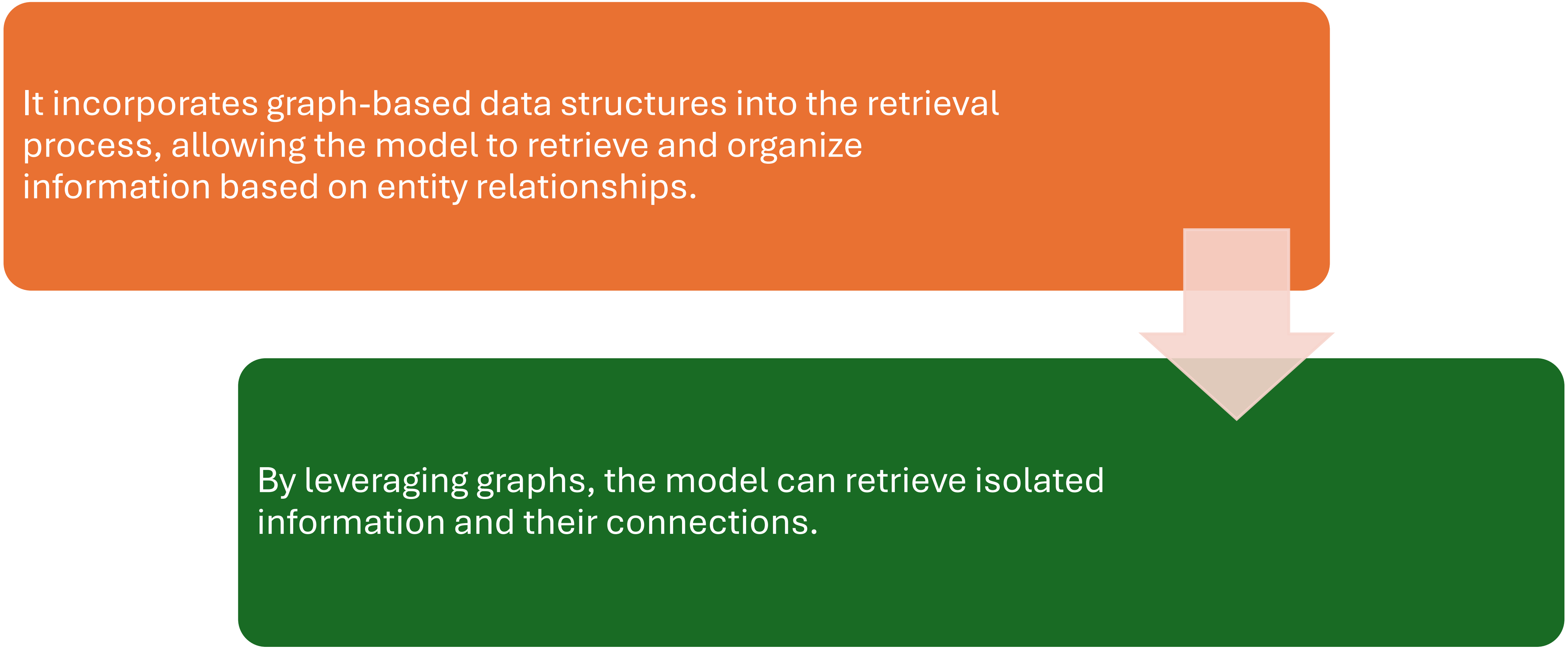
Self RAG

It is a more reflective variation of the model that emphasizes the system’s ability to evaluate its performance.

Model generates answers based on retrieved data and assesses the quality of its responses. This self-evaluation can occur through internal feedback loops, where the model checks the consistency of its output against the retrieved documents, or through external feedback mechanisms, such as user ratings or corrections.

When to Use:	Continuous improvement and accuracy are essential
Example:	Educational and training applications
Challenge:	Model’s ability to self-evaluate depends on the accuracy and comprehensiveness of the retrieved documents. If the retrieval process returns incomplete or incorrect data, the self-evaluation mechanisms may reinforce these inaccuracies.

Graphic RAG



It incorporates graph-based data structures into the retrieval process, allowing the model to retrieve and organize information based on entity relationships.

By leveraging graphs, the model can retrieve isolated information and their connections.

When to Use:

- Where the data structure is crucial for understanding and deep relational understanding

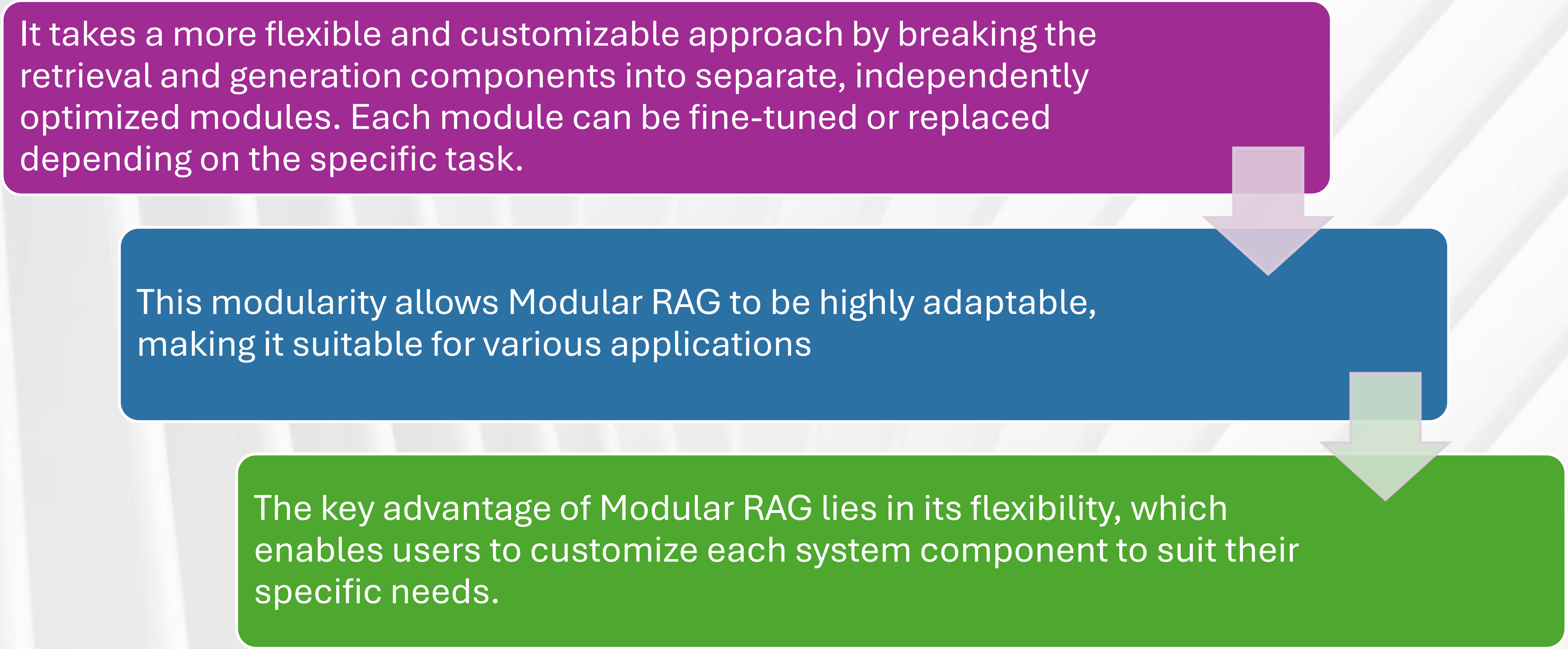
Example:

- Knowledge graphs, social networks, or semantic web applications
- Legal context, it could retrieve relevant case law and the precedents that connect those cases, providing a more nuanced understanding of the topic
- Biological research, where understanding the relationships between genes, proteins, and diseases is crucial

Challenge:

- One of the main challenges with Graph RAG is ensuring that the graph structures are updated and maintained accurately, as outdated or incomplete graphs could lead to incorrect or incomplete responses.

Modular RAG



It takes a more flexible and customizable approach by breaking the retrieval and generation components into separate, independently optimized modules. Each module can be fine-tuned or replaced depending on the specific task.

This modularity allows Modular RAG to be highly adaptable, making it suitable for various applications

The key advantage of Modular RAG lies in its flexibility, which enables users to customize each system component to suit their specific needs.

When to Use:

- Different retrieval engines could be used for different datasets or domains, while the generative model could be tailored for particular types of responses (e.g., factual, speculative, or creative).

Example:

- In a hybrid customer support system, one module might focus on retrieving information from a technical manual, while another could retrieve FAQs. The generation module would then tailor the response to the specific query type, ensuring that technical queries receive detailed, factual answers. At the same time, more general inquiries are met with broader, user-friendly responses

Challenges:

- Ensuring that the various modules work seamlessly together can be challenging, particularly when dealing with highly specialized retrieval systems or combining different generative models.

Radio RAG

It is a specialized implementation of RAG developed to address the challenges of integrating real-time, domain-specific information into LLMs for radiology

Traditional LLMs, while powerful, are often limited by their static training data, which can lead to outdated or inaccurate responses, particularly in dynamic fields like medicine.

RadioRAG has been rigorously tested using a dedicated dataset, RadioQA, composed of radiologic questions from various subspecialties, including breast imaging and emergency radiology.

By retrieving precise radiological information in real time, RadioRAG enhances the diagnostic capabilities of LLMs, particularly in scenarios where detailed and current medical knowledge is crucial

When to Use:

- Retrieving up-to-date information from sources in real-time, enhancing the accuracy & relevance of the model’s responses

Example:

- Medicine

References

- https://arxiv.org/abs/2407.13193
- https://arxiv.org/abs/2407.16833
- https://arxiv.org/abs/2408.08921
- https://arxiv.org/abs/2407.21059
- https://arxiv.org/abs/2407.15621
- https://arxiv.org/pdf/2407.21059v1
- https://arxiv.org/abs/2310.11511
- https://arxiv.org/abs/2408.14484
- https://arxiv.org/abs/2402.03367
- https://arxiv.org/pdf/2407.08223v1
- https://arxiv.org/pdf/2312.10997v5
- https://arxiv.org/abs/2401.15884