Artificial intelligence (AI)

Types of RAG

Based on how they integrate retrieval and generation components.

Sourav Verma
@srgrace

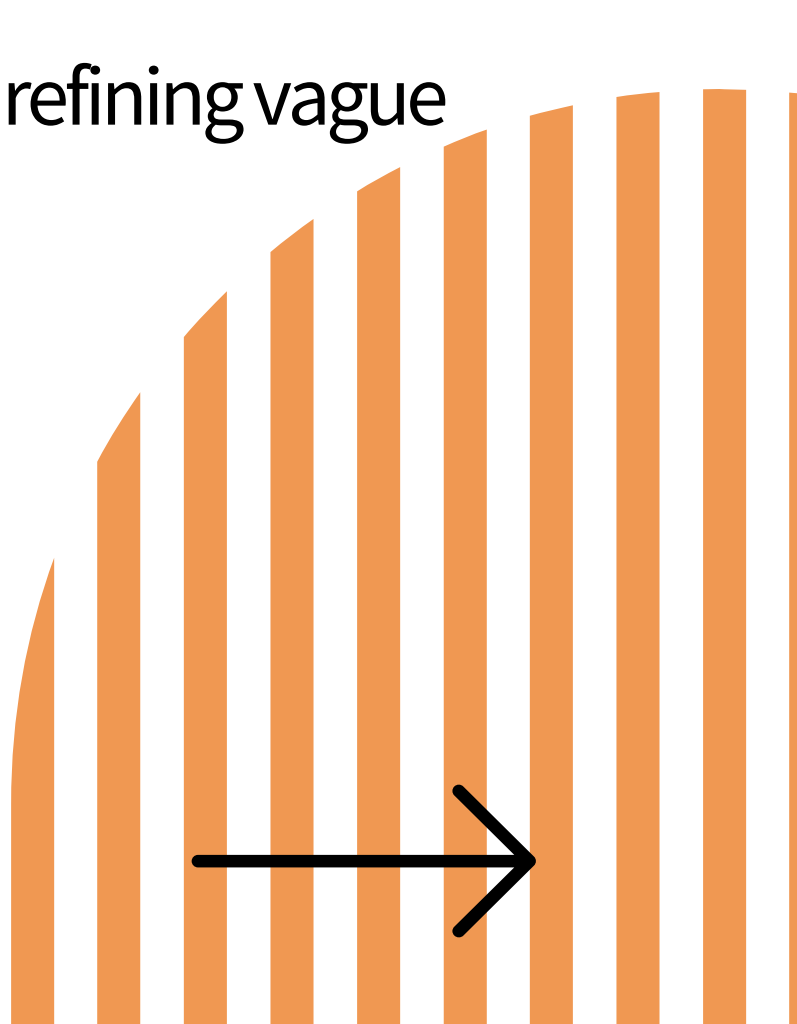
1. Based on Retrieval Timing

Pre-Retrieval Generation (Pre-RAG)

- **Mechanism:** The system retrieves documents or knowledge first, then generates a response based on the retrieved information.
- **Use Case:** Ideal for answering domain-specific queries or providing citations.
- **Example:** Search-enhanced Q&A systems.

Post-Retrieval Generation (Post-RAG)

- **Mechanism:** The system generates initial hypotheses or questions, uses them to retrieve relevant documents, and then refines its response.
- **Use Case:** Effective for exploratory tasks or refining vague queries.
- **Example:** Research assistant models.



Sourav Verma

2. Based on Retrieval Integration

Hard RAG (Pipeline RAG)

- **Mechanism:** Retrieval and generation are separate stages in a pipeline. The retrieved data is treated as fixed input for the generator.
- **Advantages:** Simplicity, interpretability.
- **Challenges:** Limited flexibility if the retrieved content is incomplete.
- **Example:** OpenAI's GPT with external plugin APIs.

Soft RAG (Joint RAG)

- **Mechanism:** Retrieval and generation are integrated; the model jointly optimizes retrieval relevance and response generation.
- **Advantages:** More seamless, can adapt retrieval dynamically based on generation needs.
- **Challenges:** Computationally more intensive.
- **Example:** Retrieval-augmented transformers (e.g., DPR + T5-based setups).

3. Based on Retrieval Methodology

Sparse Retrieval (Traditional IR)

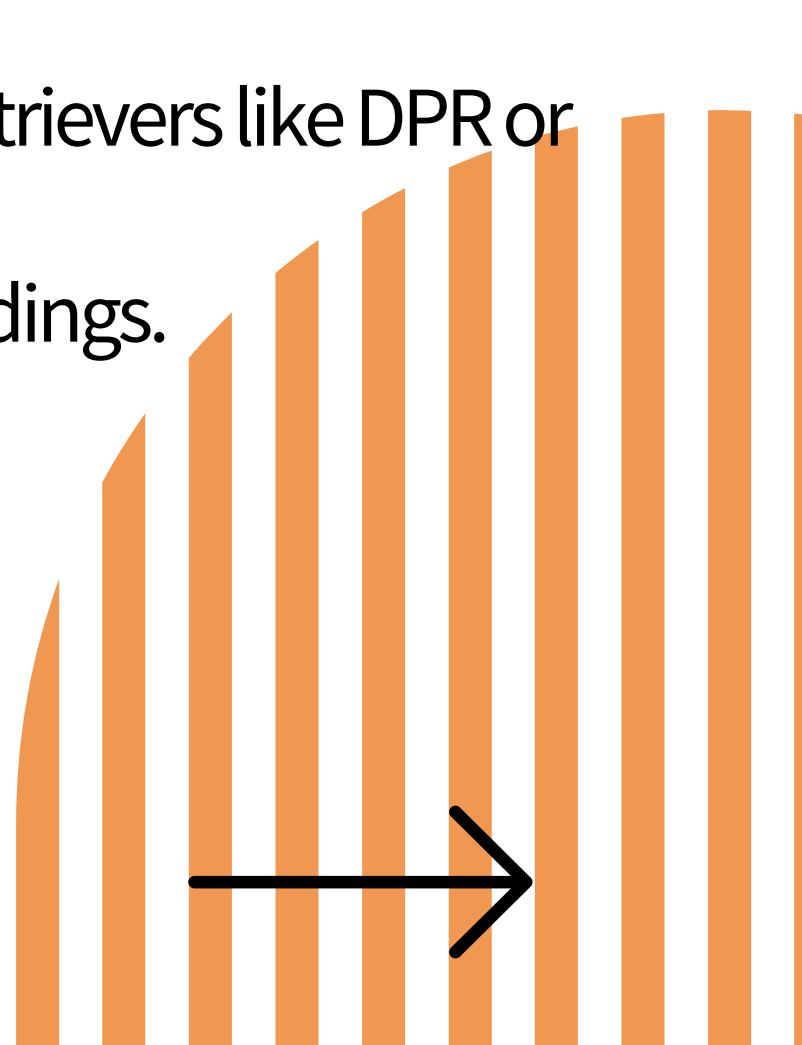
- **Mechanism:** Uses keyword matching and scoring techniques like TF-IDF or BM25 to retrieve content.
- **Advantages**: Lightweight and efficient for large datasets.
- **Challenges**: Limited to exact matches, less semantic understanding.
- **Example**: Elasticsearch, Lucene.

Dense Retrieval

- **Mechanism**: Uses embeddings from neural networks to retrieve semantically similar content.
- **Advantages**: Captures semantic relationships between queries and documents.
- **Challenges**: Requires pre-trained dense retrievers like DPR or sentence transformers.
- **Example**: FAISS with dense vector embeddings.

Hybrid Retrieval (Sparse + Dense)

Sourav Verma



4. Based on Retrieval Source

Closed-Domain RAG

- **Mechanism**: Retrieval is limited to a pre-defined, static knowledge base.
- **Advantages**: Reliable for specific domains.
- **Challenges**: Cannot answer out-of-domain queries.
- **Example**: Enterprise knowledge base assistants.

Open-Domain RAG

- **Mechanism**: Retrieval occurs across vast external sources, such as the web.
- **Advantages**: Highly flexible and up-to-date information.
- **Challenges**: Risk of retrieving low-quality or irrelevant data.
- **Example**: Google Bard, Bing Chat with search plugins.

Sourav Verma

5. Based on Retrieval Feedback Loop

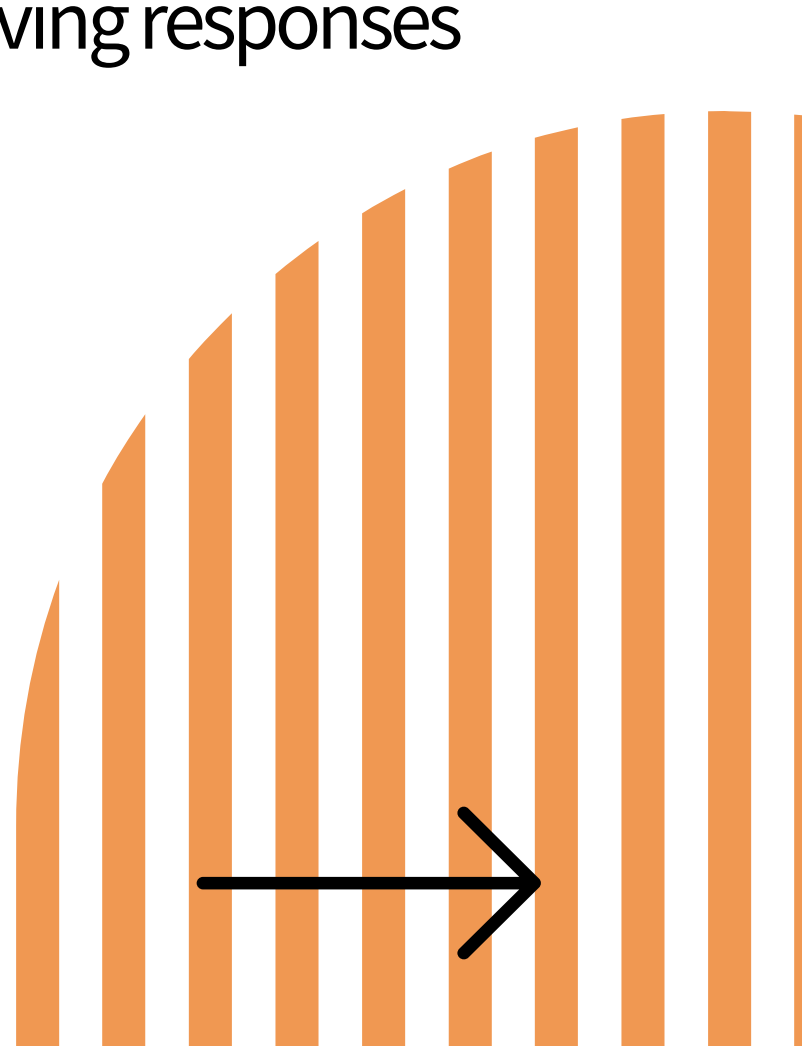
Single-Pass RAG

- **Mechanism**: Retrieves information once before generating the response.
- **Advantages**: Simple and fast.
- **Challenges**: May fail if initial retrieval is insufficient.

Iterative RAG

- **Mechanism**: The system retrieves additional content iteratively based on the progress of the generation process.
- **Advantages**: Handles complex or ambiguous queries effectively.
- **Challenges**: Slower due to multiple retrieval rounds.
- **Example**: Conversational assistants improving responses iteratively.

Sourav Verma



6. Based on Generative Model Usage

Decoder-Only RAG

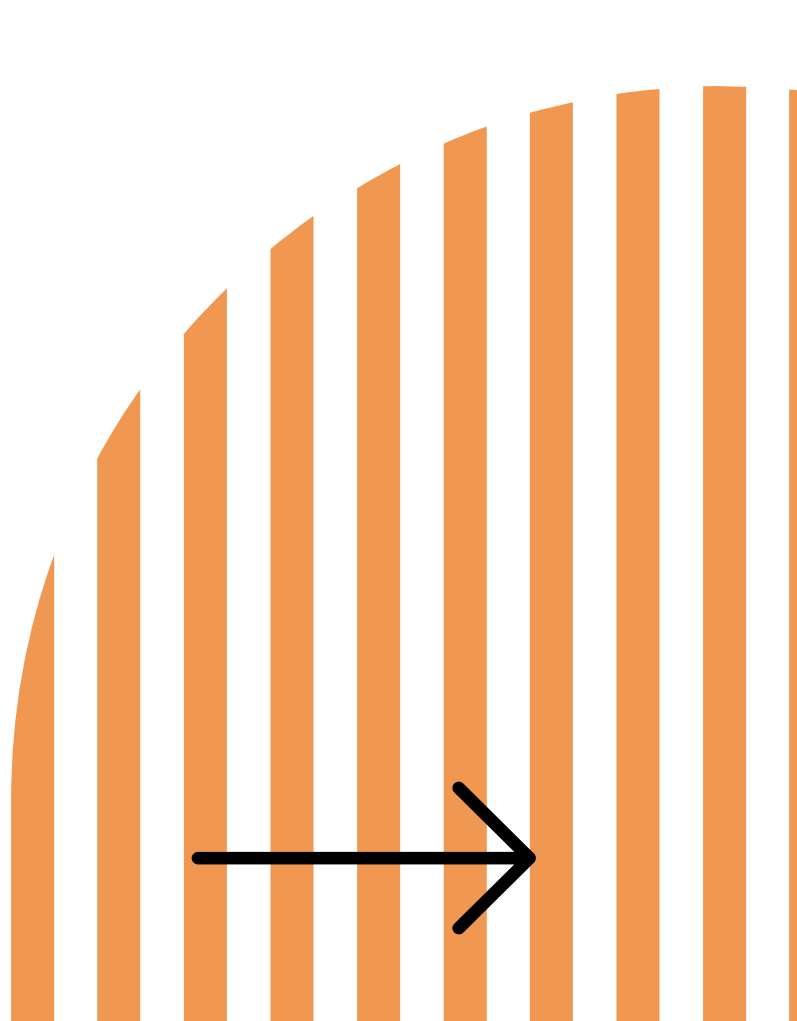
- **Mechanism**: Uses decoder-only transformers like GPT for generation based on retrieved content.
- **Advantages**: Strong generative capabilities.
- **Challenges**: May lack retrieval-specific optimizations.
- **Example**: GPT models integrated with retrieval APIs.

Encoder-Decoder RAG

- **Mechanism**: Uses encoder-decoder models like T5 or BART to process and generate content.
- **Advantages**: Better at integrating and summarizing retrieved content.
- **Challenges**: More resource intensive.
- **Example**: T5-based RAG pipelines.

Hybrid Generative RAG

Sourav Verma



7. Specialized RAG Types

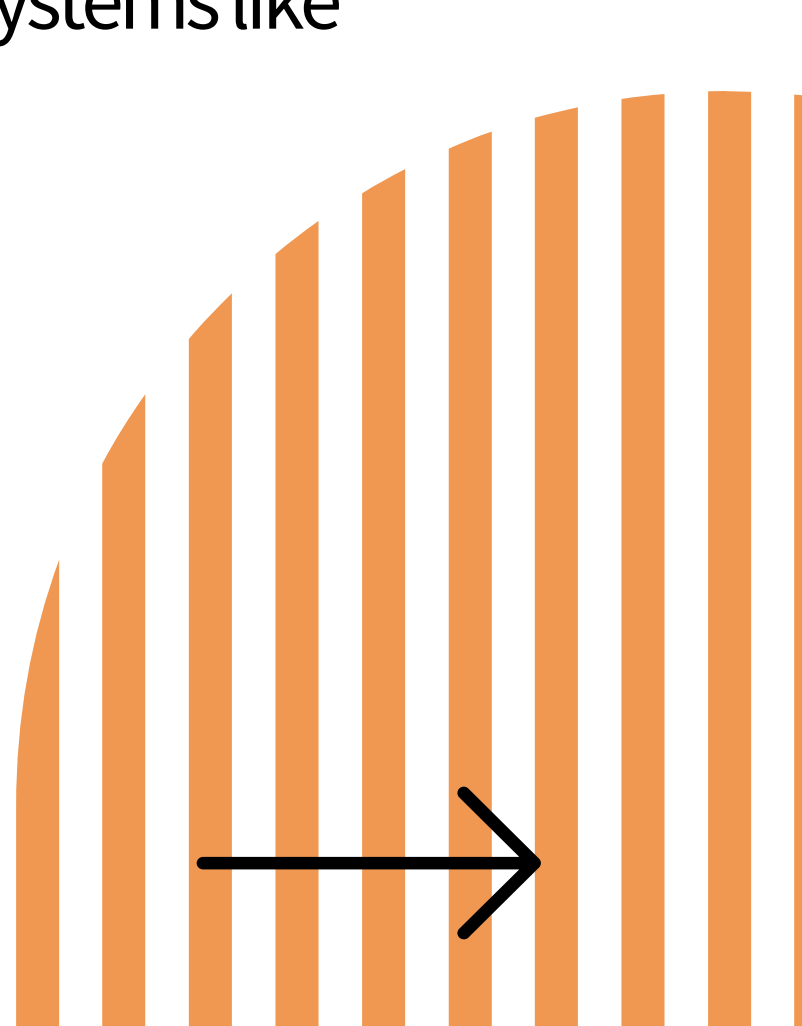
Multimodal RAG

- **Mechanism**: Retrieves information across text, images, or videos and generates multimodal outputs.
- **Use Case**: Applications in education, e-commerce, or medical imaging.
- **Example**: Systems combining CLIP for image retrieval with GPT for textual reasoning.

Conversational RAG

- **Mechanism**: Designed for dialogue systems, maintaining contextual retrieval across multi-turn interactions.
- **Use Case:** Chatbots, customer support.
- **Example**: Retrieval-augmented dialogue systems like ChatGPT with memory.

Sourav Verma



Artificial intelligence (AI)

Repost if you find the Information helpful and follow along

Sourav Verma
@srgrace