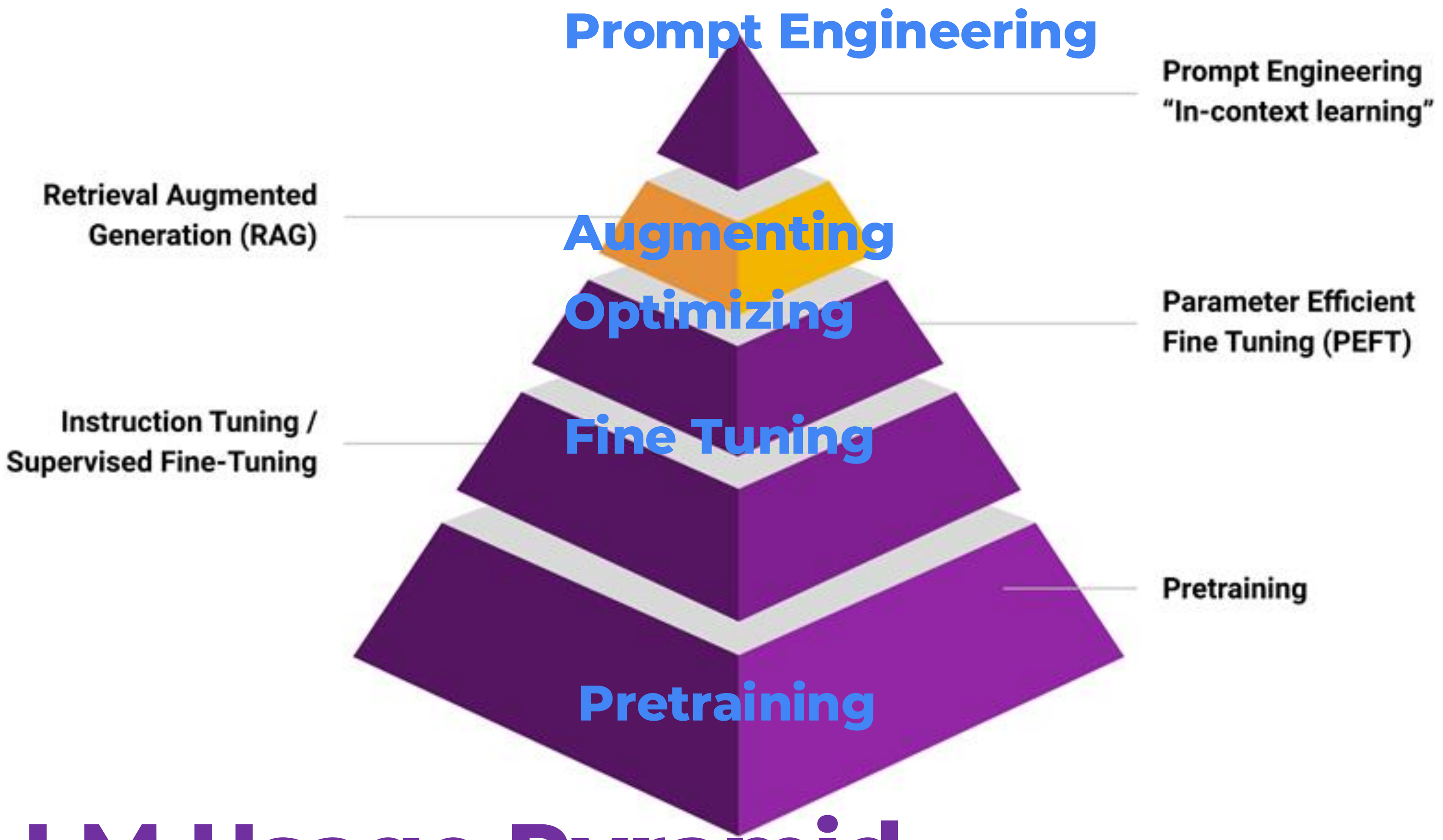
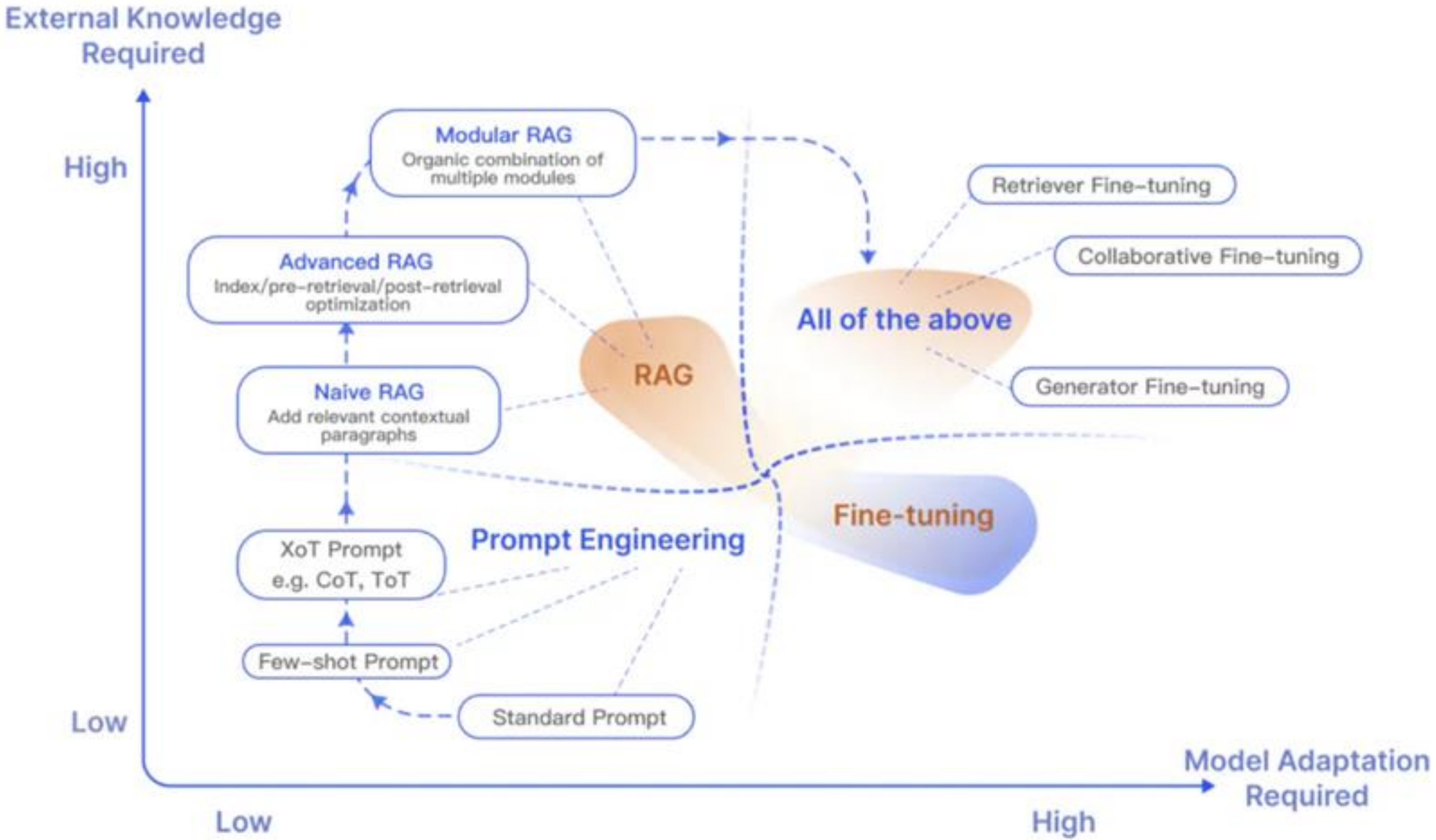
RAG: Retrieval Augmented Generation

Vasudeva Varma



LLM Usage Pyramid


External Knowledge Required
High
Low
Modular RAG
Organic combination of multiple modules
Advanced RAG
Index/pre-retrieval/post-retrieval optimization
Naive RAG
Add relevant contextual paragraphs
XoT Prompt
e.g. CoT, ToT
Few-shot Prompt
Standard Prompt
Prompt Engineering
RAG
All of the above
Retriever Fine-tuning
Collaborative Fine-tuning
Generator Fine-tuning
Fine-tuning
Model Adaptation Required
Low
High

Augmenting

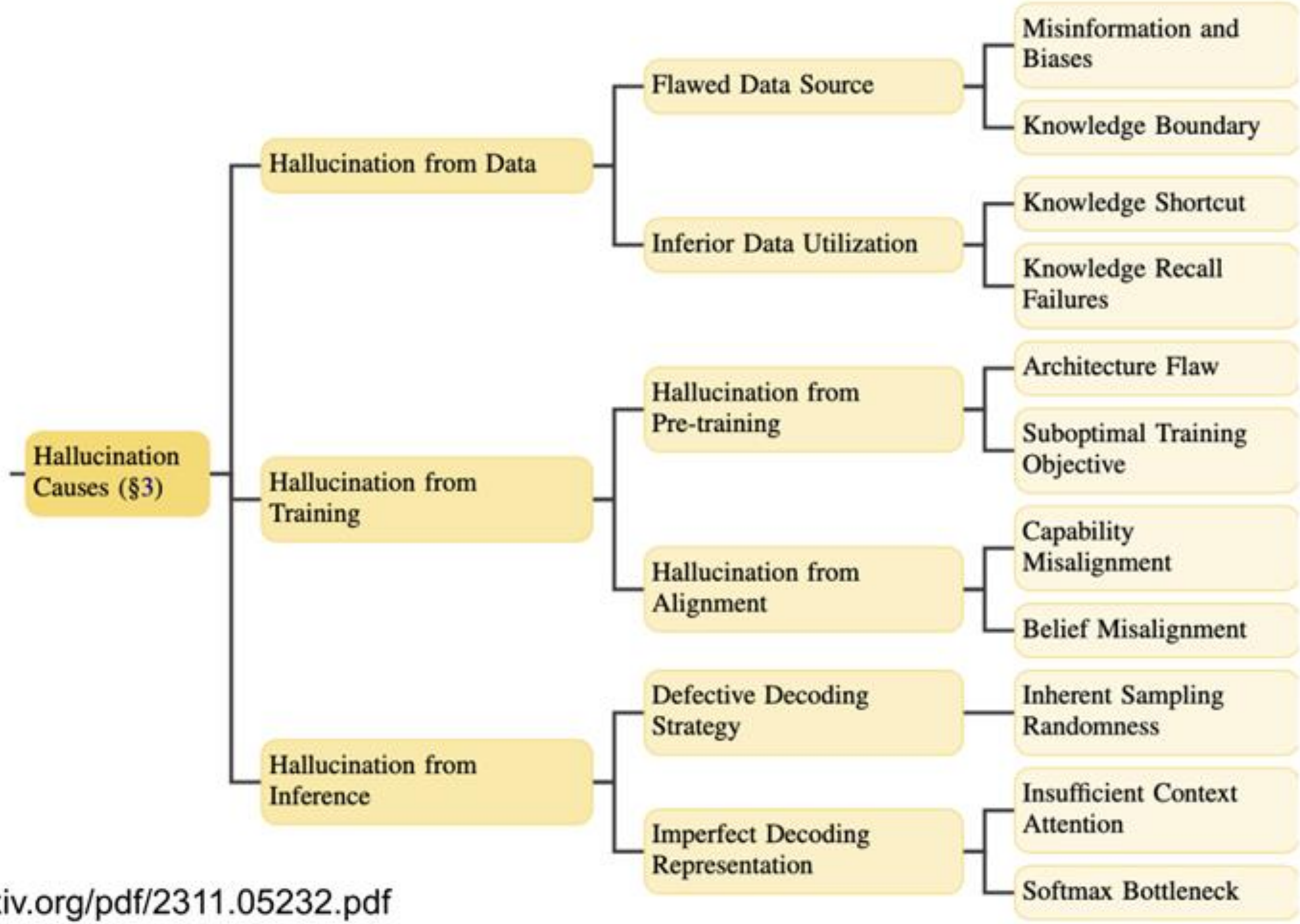
Retrieval-Augmented Generation (RAG) *combines retrieval systems with LLMs*: RAG integrates vector-based information retrieval with generative models to dynamically include external, task-specific knowledge.

How It Works?

1. **Query Understanding**: The LLM ***interprets*** the user's question.
2. **Information Retrieval**: The query is matched against a ***Document Repository***
3. **Contextual Generation**: The retrieved data is fed into the LLM, which ***generates a grounded, factual response.***

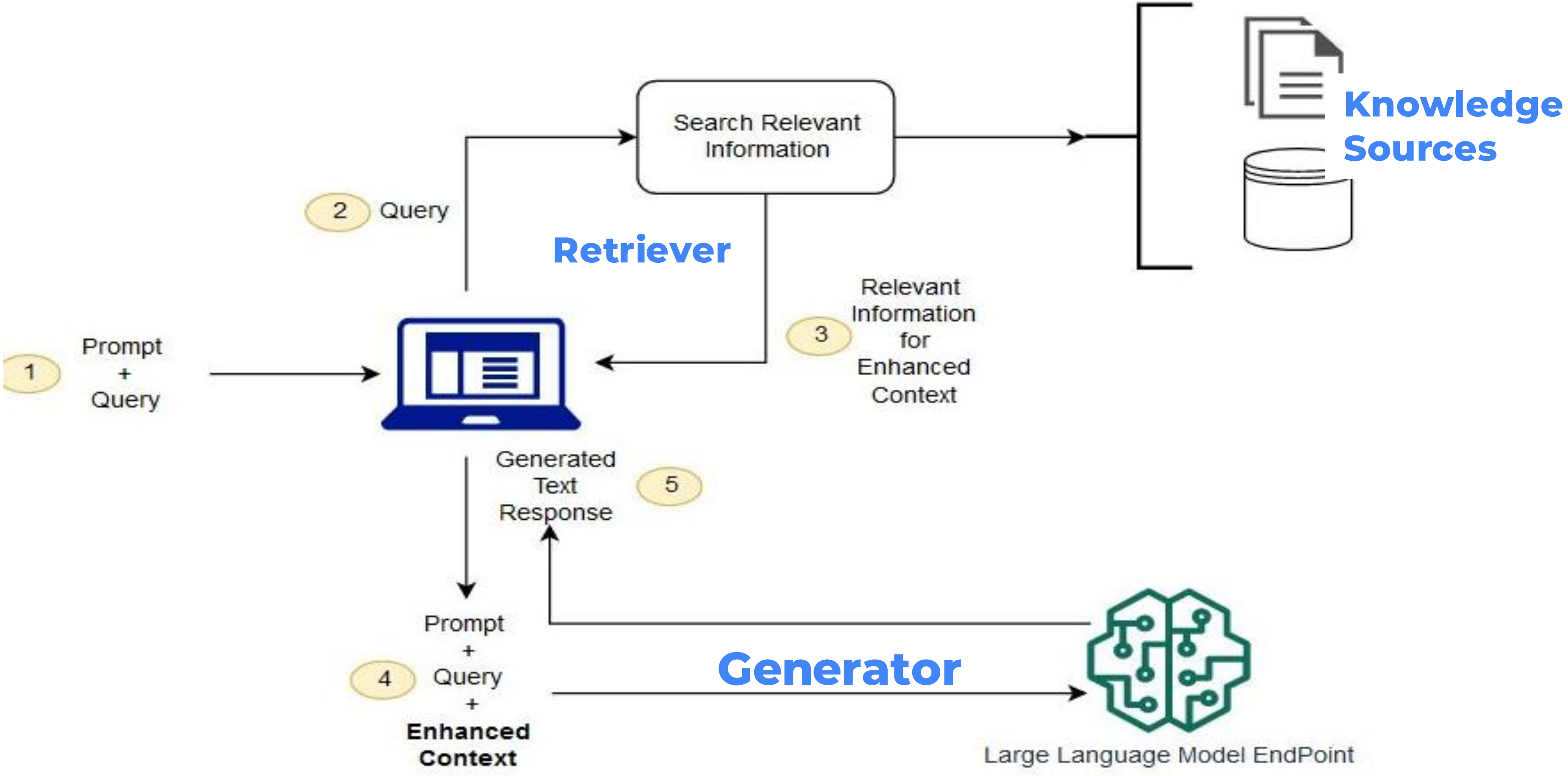
Key Advantages

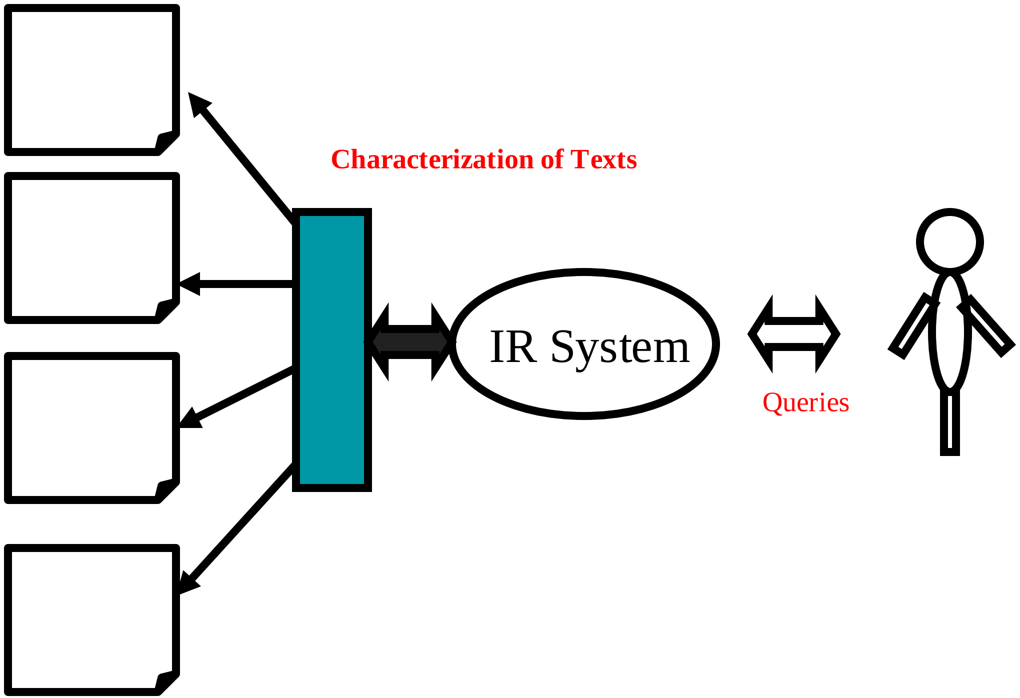
- **Enhanced Accuracy**: Incorporates real-world, dynamic data, generating **up-to-date, accurate information**
- **Domain-Specific Adaptability**: Tailored to specialized datasets (e.g., healthcare, legal); **Private data**
- **Reduced Hallucinations**: Limits reliance on outdated or inferred knowledge.



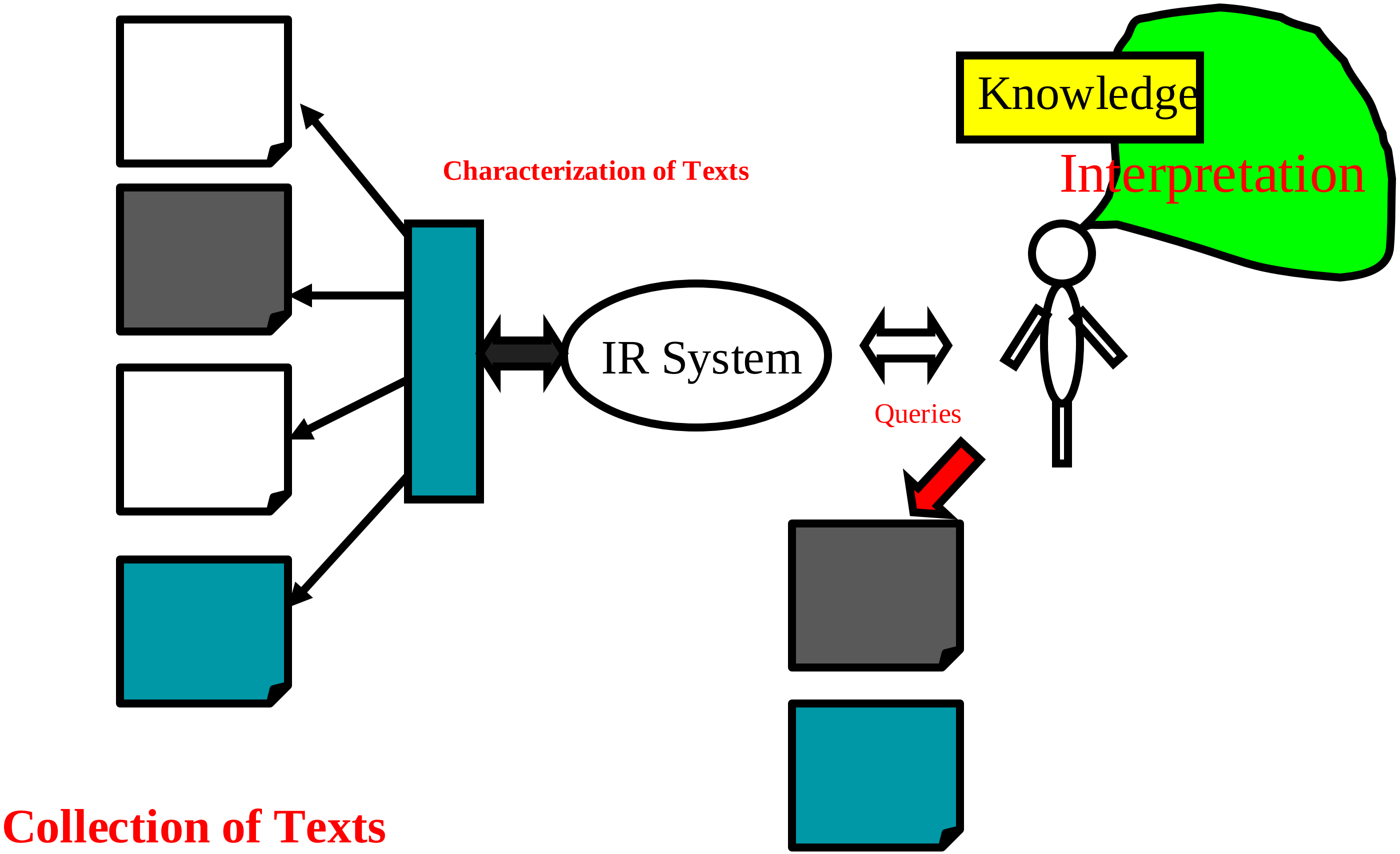
https://arxiv.org/pdf/2311.05232.pdf

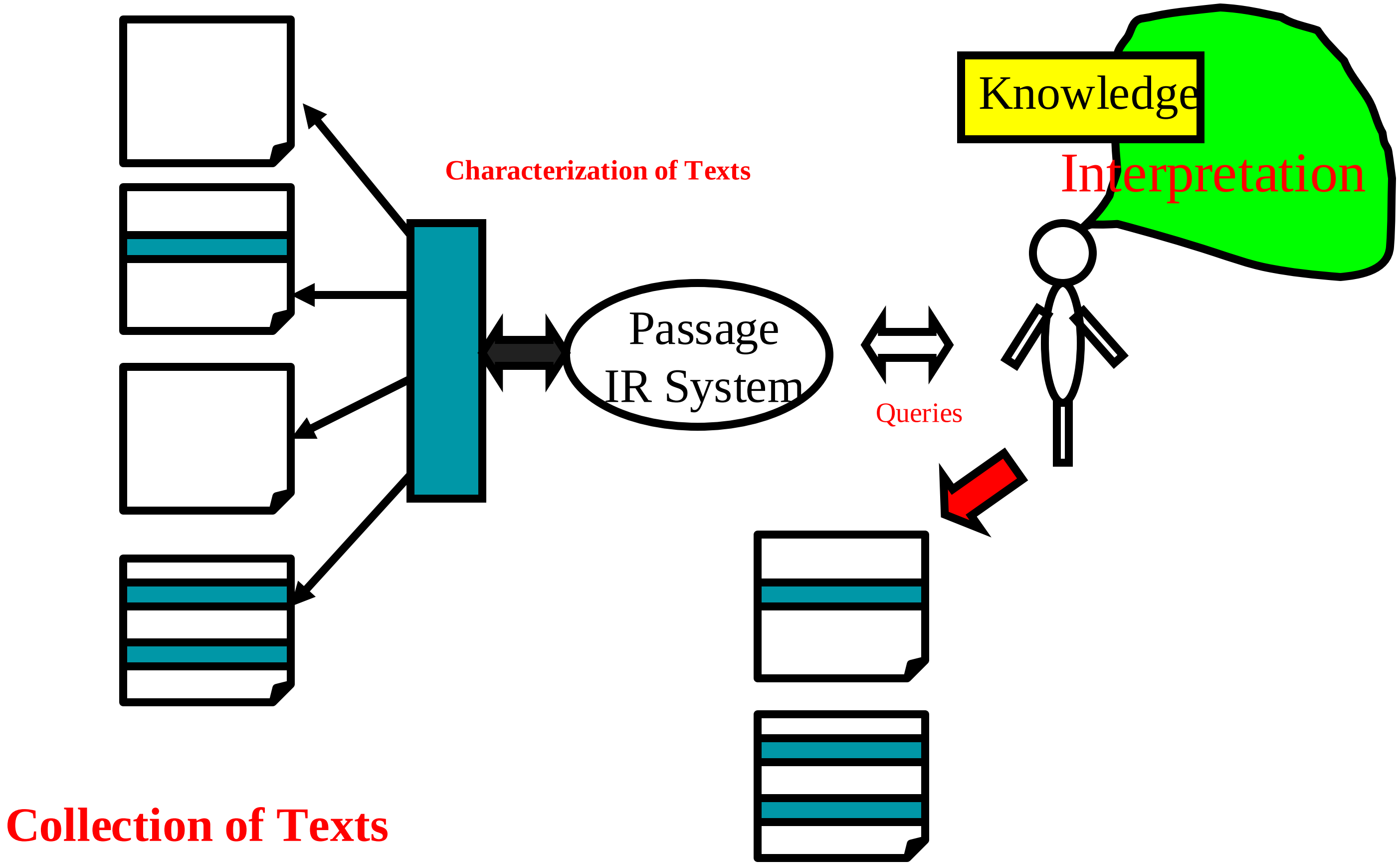
Key Components of RAG: Retriever, Generator and Knowledge Sources

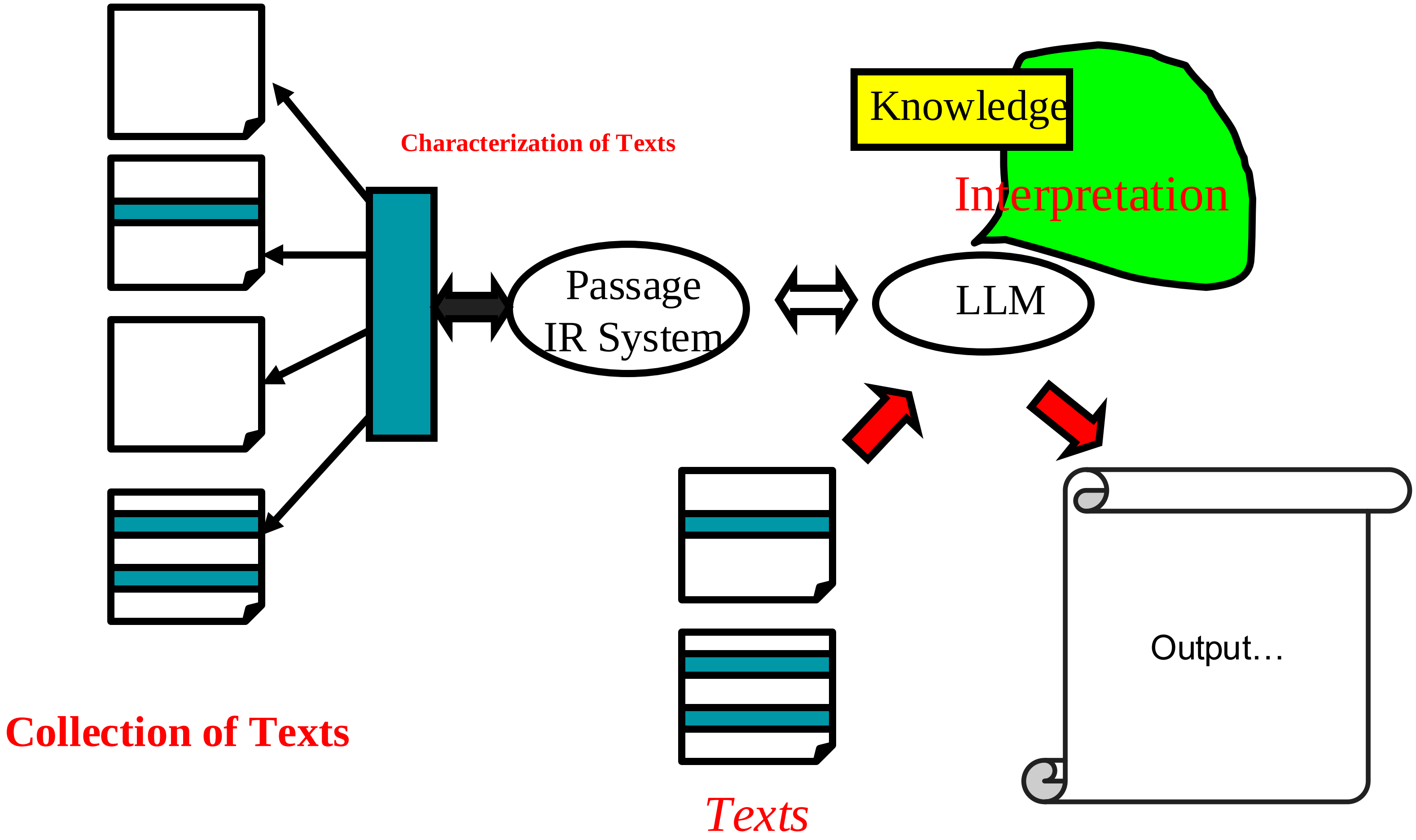


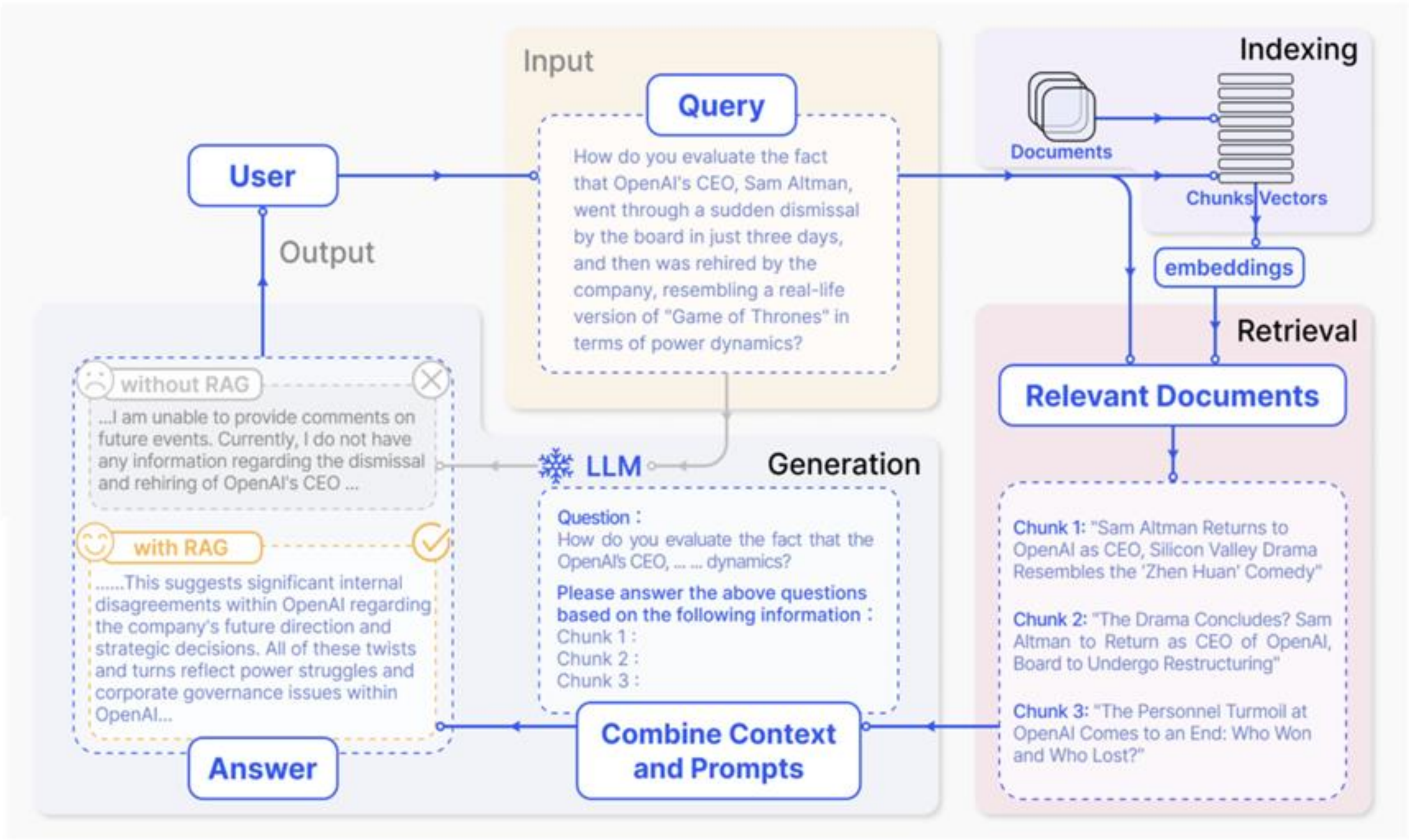


Collection of Texts

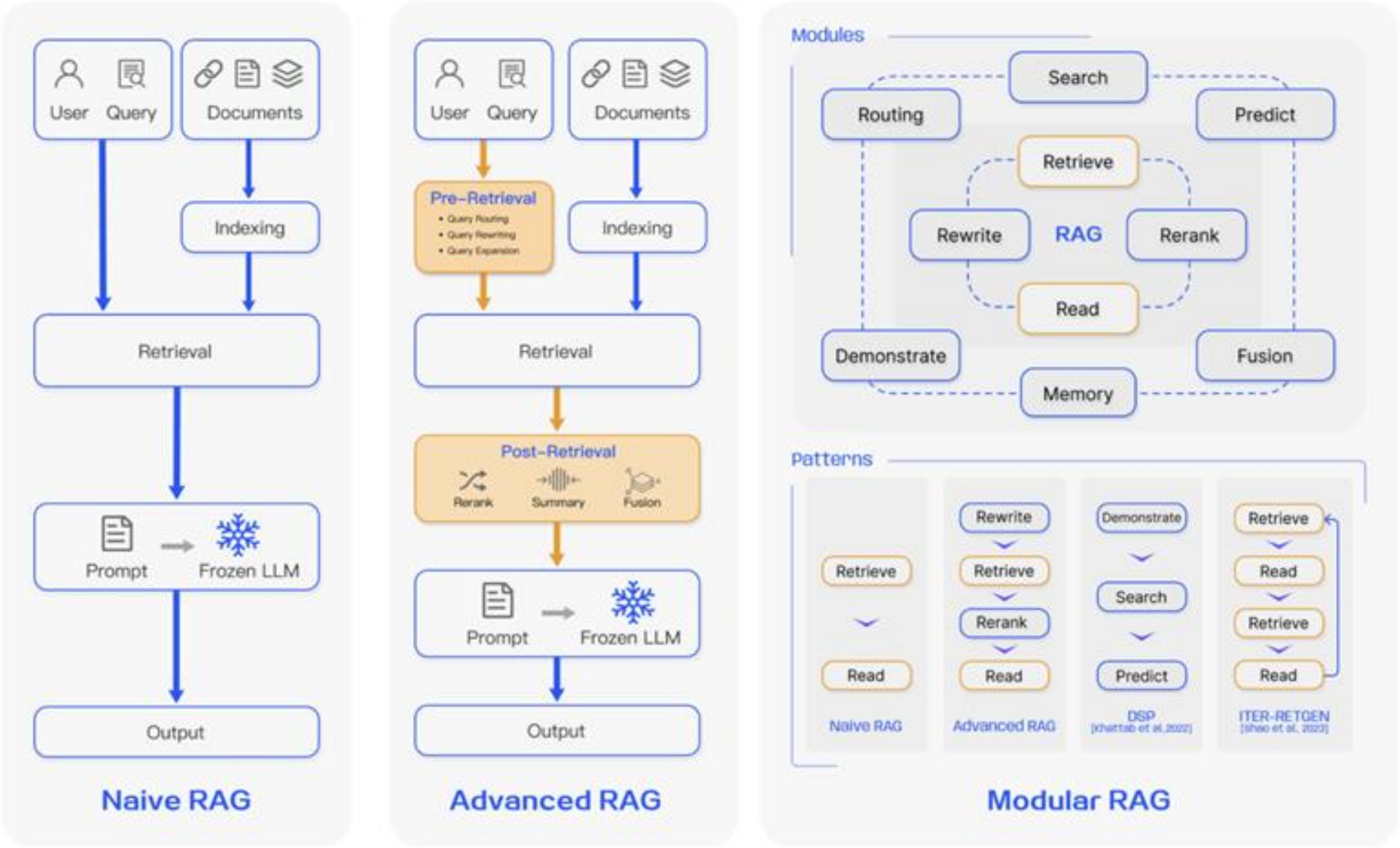

Knowledge
Interpretation
Characterization of Texts
IR System
Queries
Collection of Texts


Knowledge
Interpretation
Characterization of Texts
Passage
IR System
Queries
Collection of Texts


Characterization of Texts
Knowledge
Interpretation
Passage
IR System
LLM
Output…
Texts
Collection of Texts



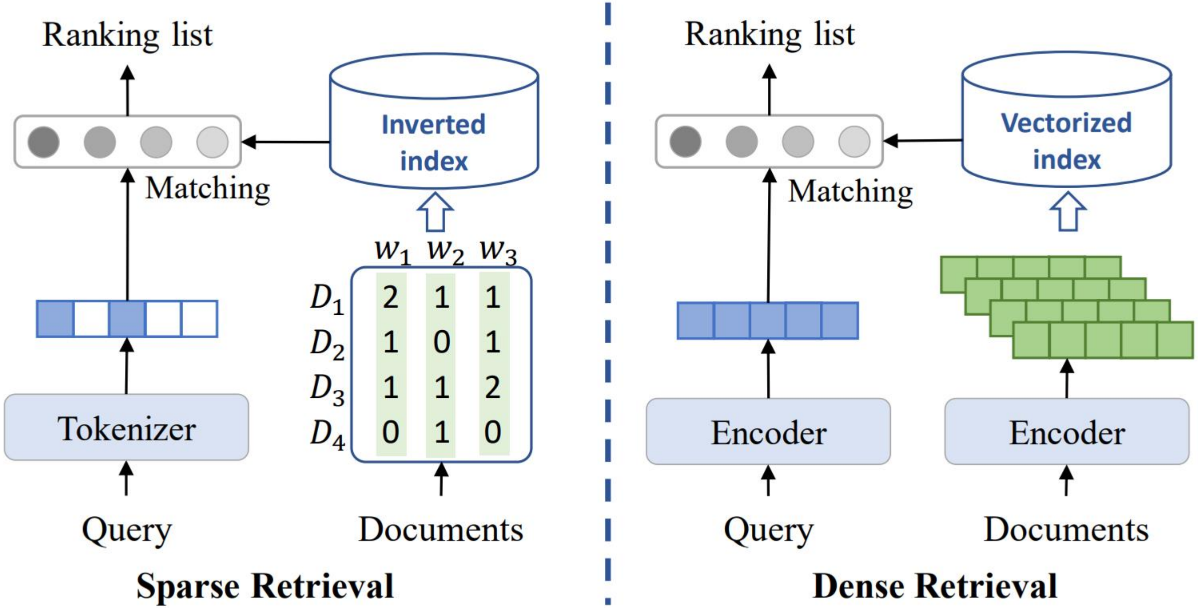
Input
Query
How do you evaluate the fact that OpenAI's CEO, Sam Altman, went through a sudden dismissal by the board in just three days, and then was rehired by the company, resembling a real-life version of "Game of Thrones" in terms of power dynamics?
Indexing
Documents
Chunks Vectors
embeddings
User
Output
Retrieval
Relevant Documents
without RAG
...I am unable to provide comments on future events. Currently, I do not have any information regarding the dismissal and rehiring of OpenAI's CEO ...
with RAG
......This suggests significant internal disagreements within OpenAI regarding the company's future direction and strategic decisions. All of these twists and turns reflect power struggles and corporate governance issues within OpenAI...
LLM
Generation
Question :
How do you evaluate the fact that the OpenAI's CEO, ... dynamics?
Please answer the above questions based on the following information :
Chunk 1 :
Chunk 2 :
Chunk 3 :
Chunk 1: "Sam Altman Returns to OpenAI as CEO, Silicon Valley Drama Resembles the 'Zhen Huan' Comedy"
Chunk 2: "The Drama Concludes? Sam Altman to Return as CEO of OpenAI, Board to Undergo Restructuring"
Chunk 3: "The Personnel Turmoil at OpenAI Comes to an End: Who Won and Who Lost?"
Answer
Combine Context and Prompts



https://arxiv.org/pdf/2312.10997.pdf

Types of Retrieval

- **Blackbox retrieval (ask Google/Bing)**
- **Sparse Retrieval**: Methods like BM25 and TF-IDF.
- **Dense Retrieval**: Techniques leveraging DPR or BERT embeddings.
    - Document level dense retrieval
    - Token level dense retrieval



Dense Retrieval in RAG: The Basics

- Query and document embeddings capture semantic meaning.
- Similarity comparison using metrics like cosine similarity.
- Key advantage: semantic matching beyond keyword overlap.

RAG Model Variants

- **Document Level Dense Retrieval: (RAG-Sequence)**: Sequential integration of retrieved documents.
- **Token Level Dense Retrieval: (RAG-Token)**: Token-level integration for fine-grained control.

Tools and Frameworks

- **Popular Tools**: LangChain, Haystack, FAISS. ScaNN

Dense Retrieval at the Document Level

- Data is split into chunks (e.g., paragraphs or sections).
- Chunks are encoded into single dense vectors.
- Retrieval is based on top-k similarity matches.
- Example: Amazon Alexa FAQs.

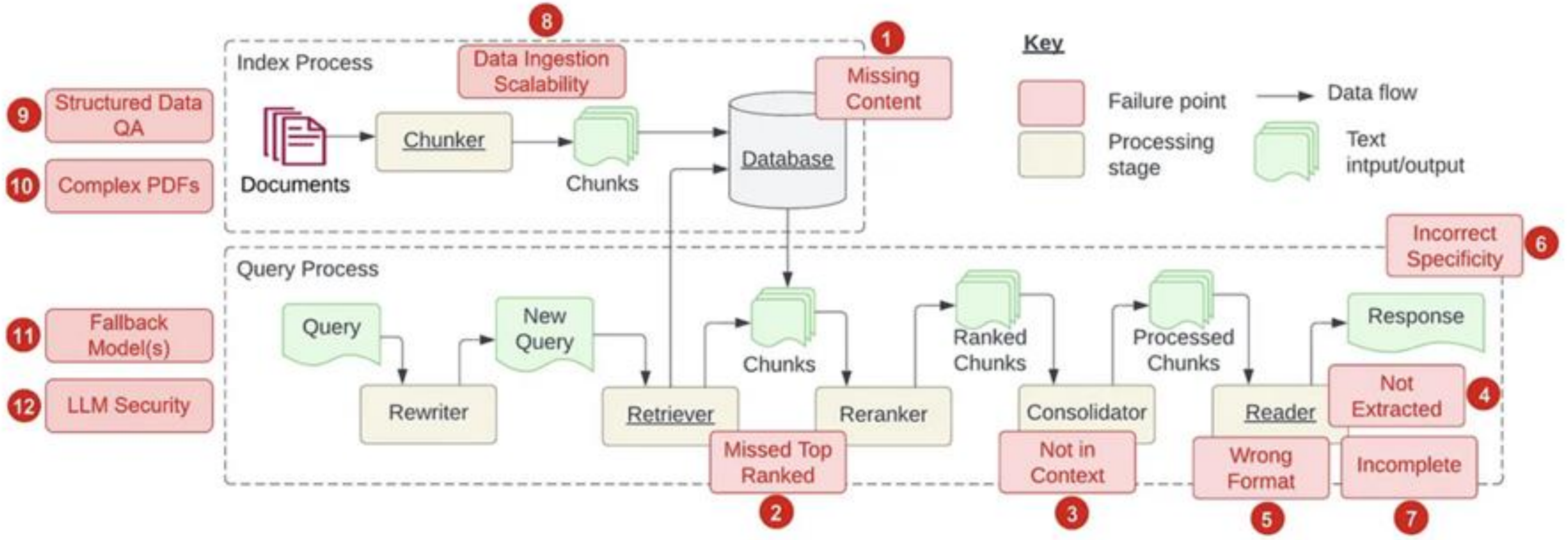
Dense Retrieval at the Token Level

- Token embeddings offer fine-grained precision.
- Matches specific tokens instead of entire chunks.
- Useful for open-domain QA (e.g., “Paris” for “What’s the capital of France?”).
- Applications: biomedical research, legal texts.

Dense Retrieval: Document level vs Token Level

Aspect	Document-Level Retrieval	Token-Level Retrieval
Granularity	Coarser (e.g., paragraphs)	Fine-grained (e.g., tokens)
Index Size	Smaller	Larger
Use Cases	General information retrieval	Precise question answering
Challenges	Chunking, embedding quality	High cost, contextual noise

- **Role of hybrid systems: Combine both approaches to leverage strengths.**
- **Open research areas: Adaptive indexing, context-aware embeddings, and efficient token retrieval.**



https://arxiv.org/pdf/2401.05856.pdf

Retrieval Granularity: When do we retrieve?

1

Once, at the beginning of generation

- Default method used by most systems (Lewis et al. 2020)

2

Several times during generation, as necessary

- Generate a search token (Schick et al. 2023)
- Search when the model is uncertain (Jiang et al. 2023)

3

Every token

- Find similar final embeddings (Khandelwal et al. 2019)
- Approximate attention with nearest neighbours (Bertsch et al. 2023)

Triggering Retrieval w/ Tokens

Toolformer (Schick et al.2023) generates tokens that trigger retrieval (or other tools)

Training is done in an iterative manner - generate and identify successful retrievals

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

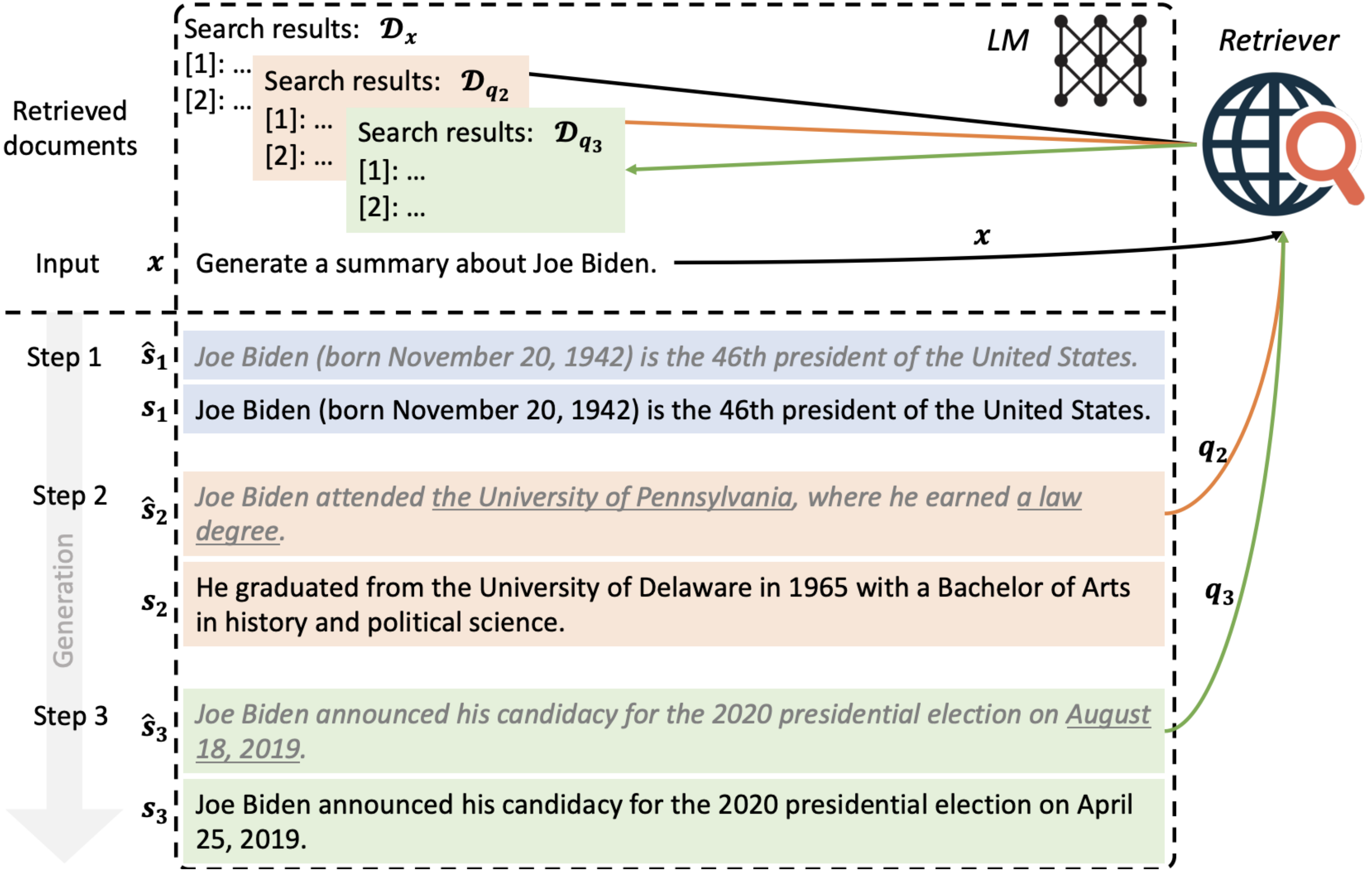
Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

Triggering Retrieval w/ Uncertainty

FLARE (Ji
to generat
retrieval if



General Applications of RAG

- Open-domain Q&A.
- Customer support automation.
- Content generation with contextual relevance.

RAG in Education

- Adaptive learning systems offering personalized resources.
- Interactive tutoring with real-time Q&A referencing textbooks.
- Knowledge discovery for lifelong learning.

RAG in Healthcare

- Clinical decision support with evidence-based recommendations
- Patient education by simplifying medical terminology.
- Research assistance through literature summarization.

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis[†‡], Ethan Perez[*],

Aleksandra Piktus[†], Fabio Petroni[†], Vladimir Karpukhin[†], Naman Goyal[†], Heinrich Küttler[†],

Mike Lewis[†], Wen-tau Yih[†], Tim Rocktäschel[†‡], Sebastian Riedel[†‡], Douwe Kiela[†]

[†]Facebook AI Research; [‡]University College London; [*]New York University;
`plewis@fb.com`

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

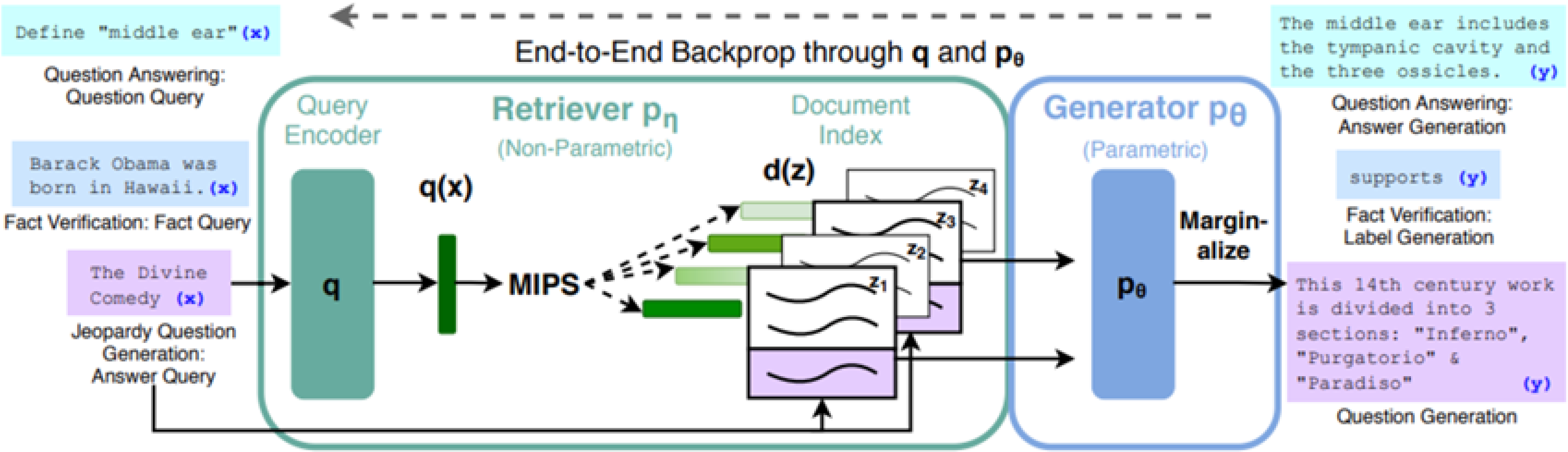


Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder* + *Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x, we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i. For final prediction y, we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

Key messages:

The last mile is the hardest; **RAG plays a critical role in making LLM implementations actually work.**

While technology tremendously enhances productivity, *domain knowledge is the cornerstone to success.*

Need to build middleware, responsibility layer, and domain specific foundation models

Further Reading and Resources

- **Suggested Materials**:
  - "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" (Lewis et al.).
  - Tutorials on LangChain and Haystack.

Thank you