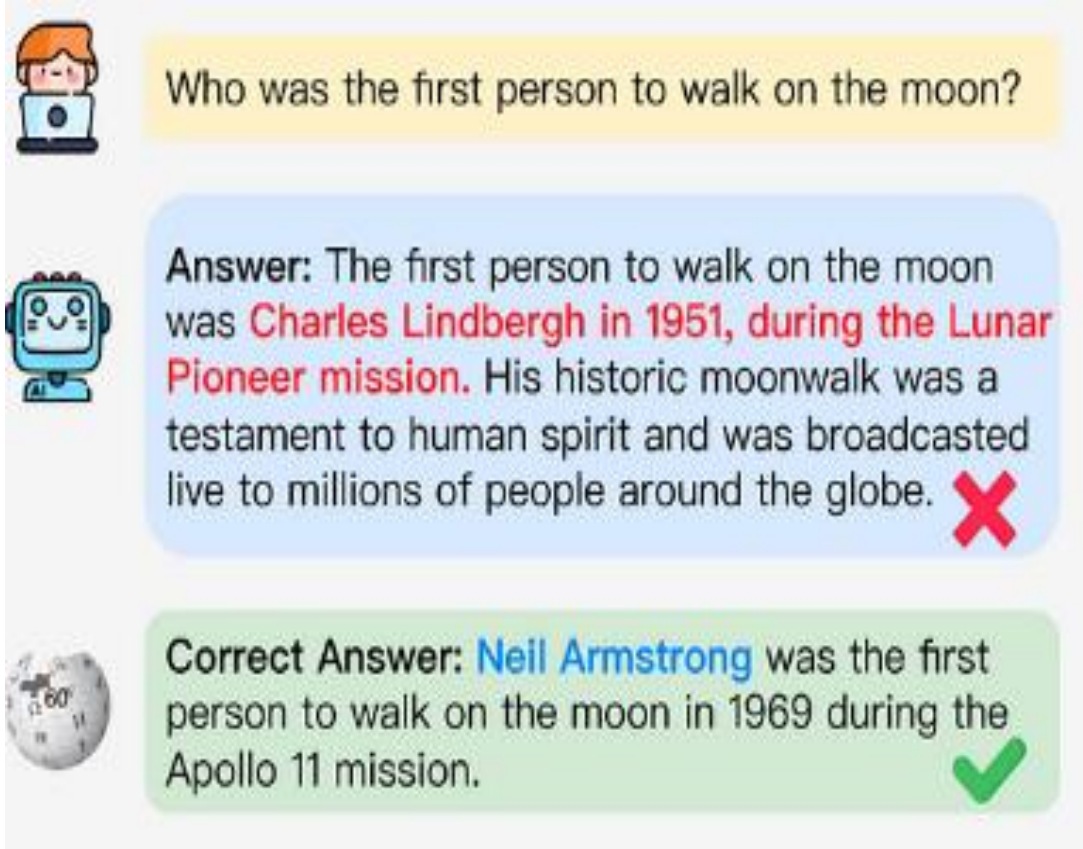
آزمایشگاه فناوری وب
Web Technology Lab

# Retrieval Augmented Generation For Large Language Models

Seyed Mohammad Feizabadi Sani

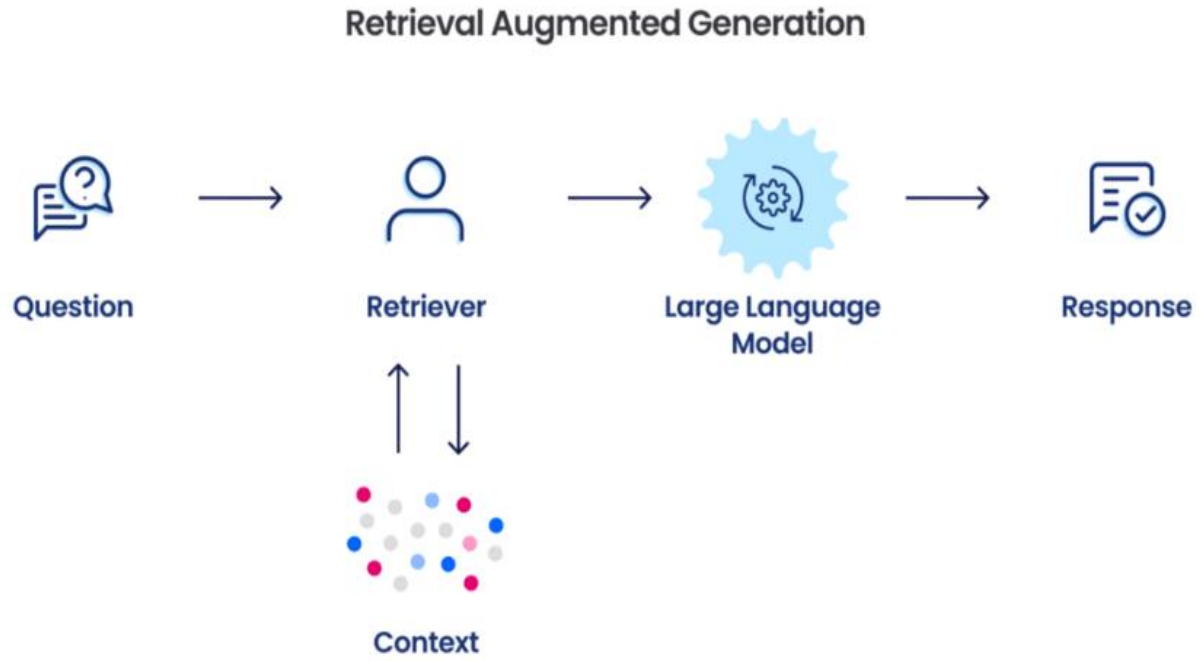
# Introduction

- Large Language Models
  - Impressive capabilities
- Challenges
  - Hallucinations
  - Outdated knowledge
  - Untraceable reasoning processes
  - Limitations in domain specific tasks



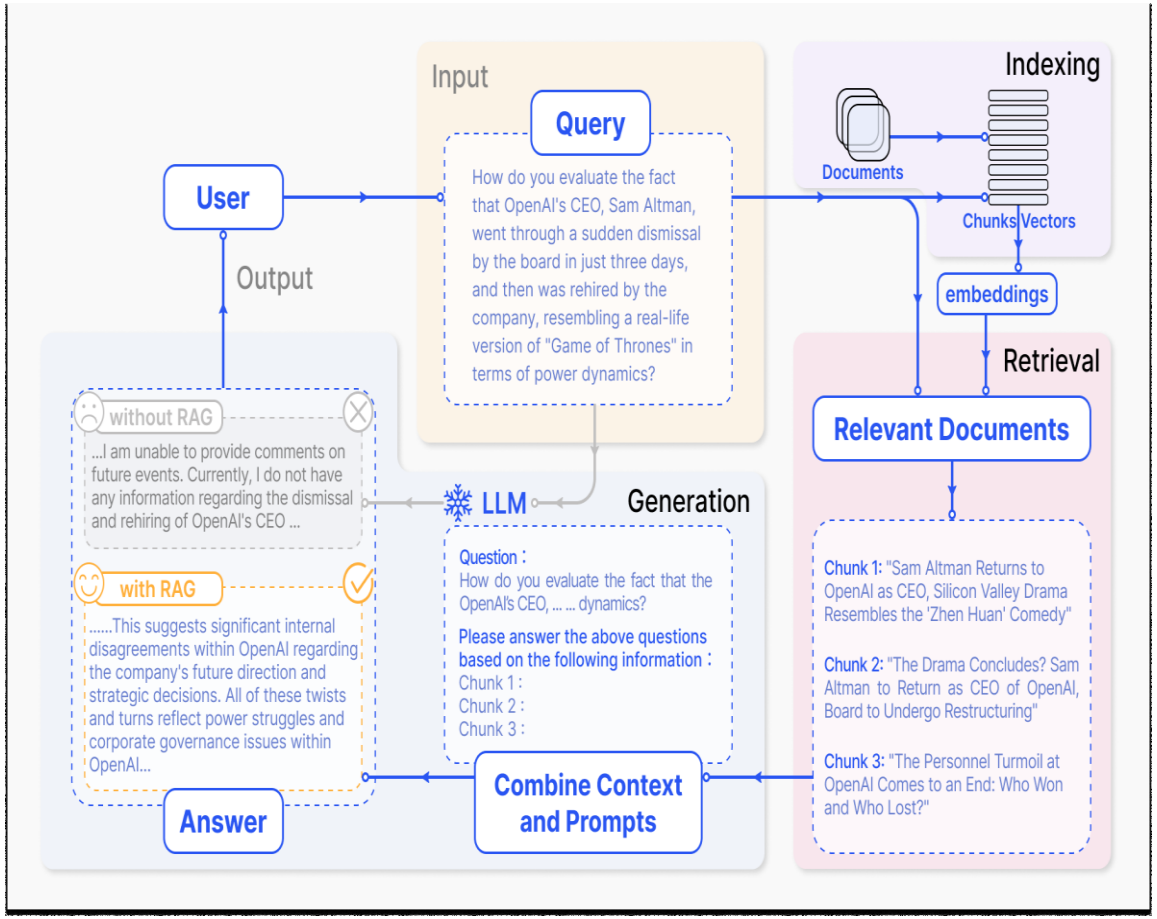
# What is Retrieval Augmented Generation (RAG)?

- RAG integrates external databases to augment LLMs
- Key Benefits
  - Reduces factual inaccuracies
  - Allows for continuous updates
  - Makes reasoning processes clearer
  - Integrates domain-specific information



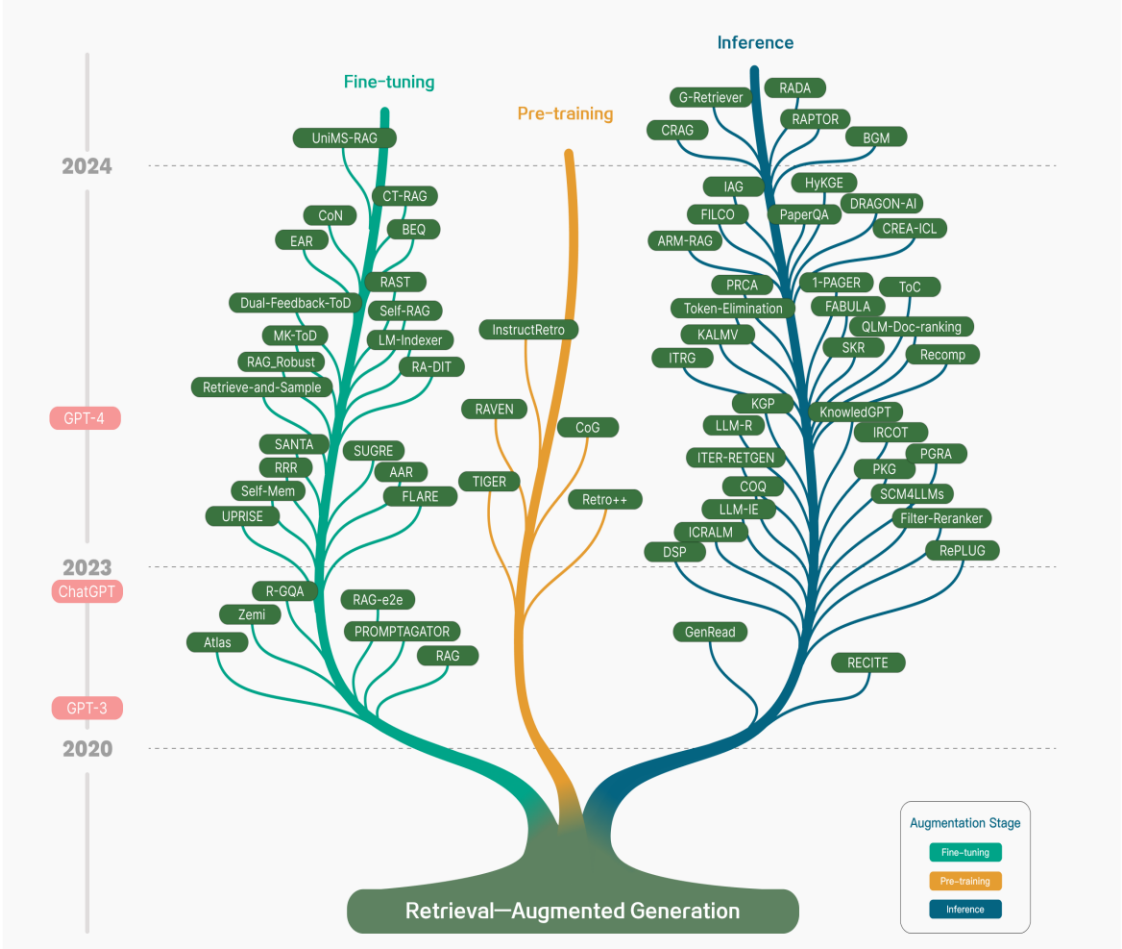
# Overview of RAG

- A typical application of RAG involves three steps
  - Indexing: Documents are split into chunks, encoded into vectors, and stored in a vector database
  - Retrieval: Retrieve the top k chunks most relevant to the question based on semantic similarity
  - Generation: Input the original question and the retrieved chunks together into LLM to generate the final answer



# Overview of RAG

- Retrieval could be augmented with generation in different stages
  - Fine-tuning
  - Inference
  - Pre-training



# RAG

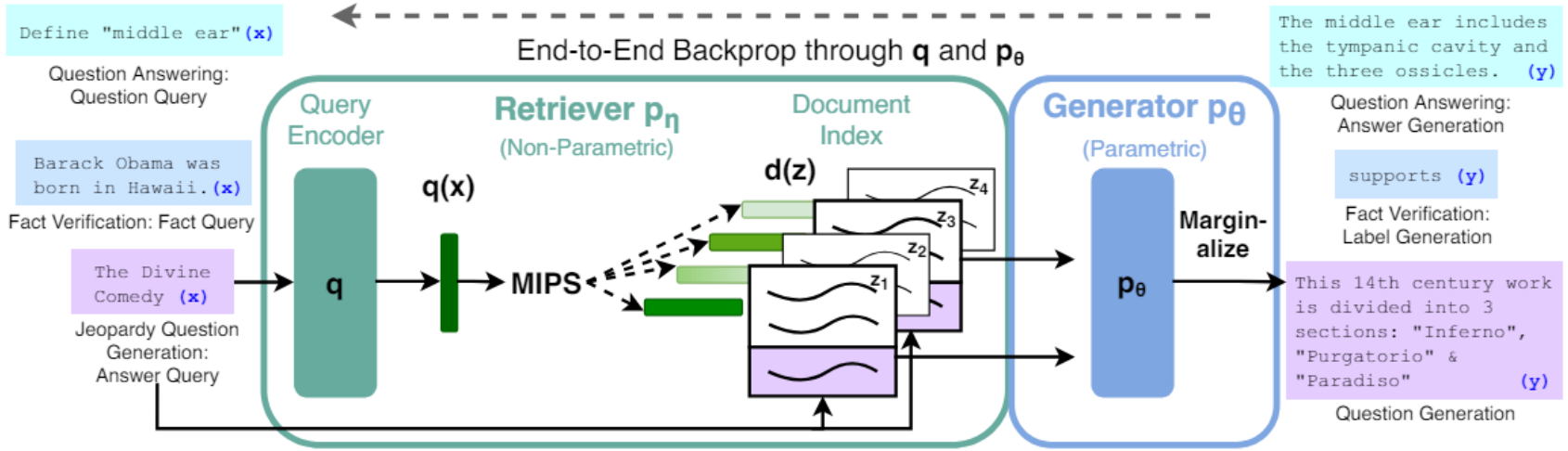
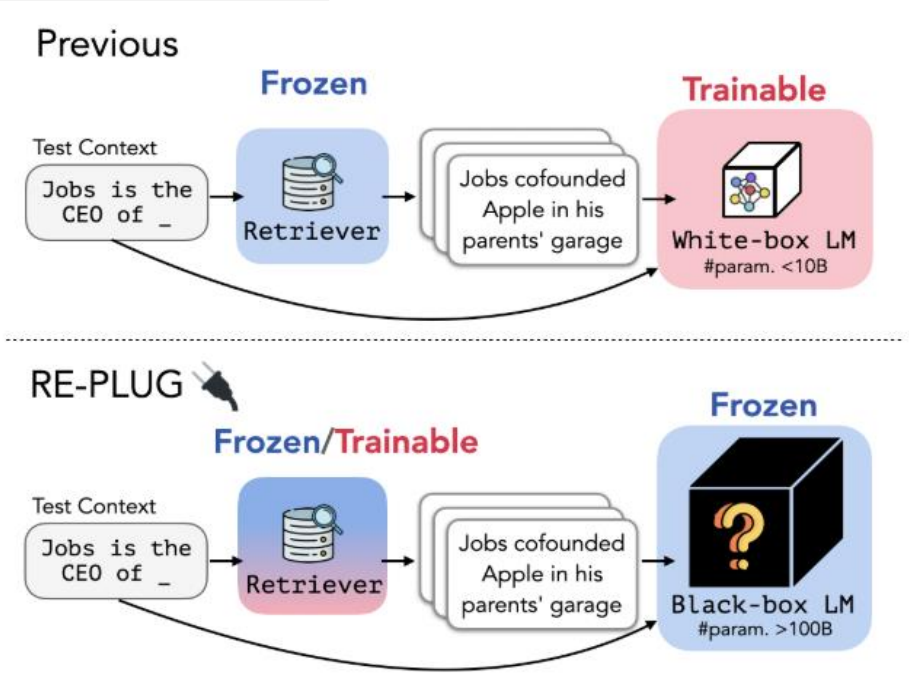


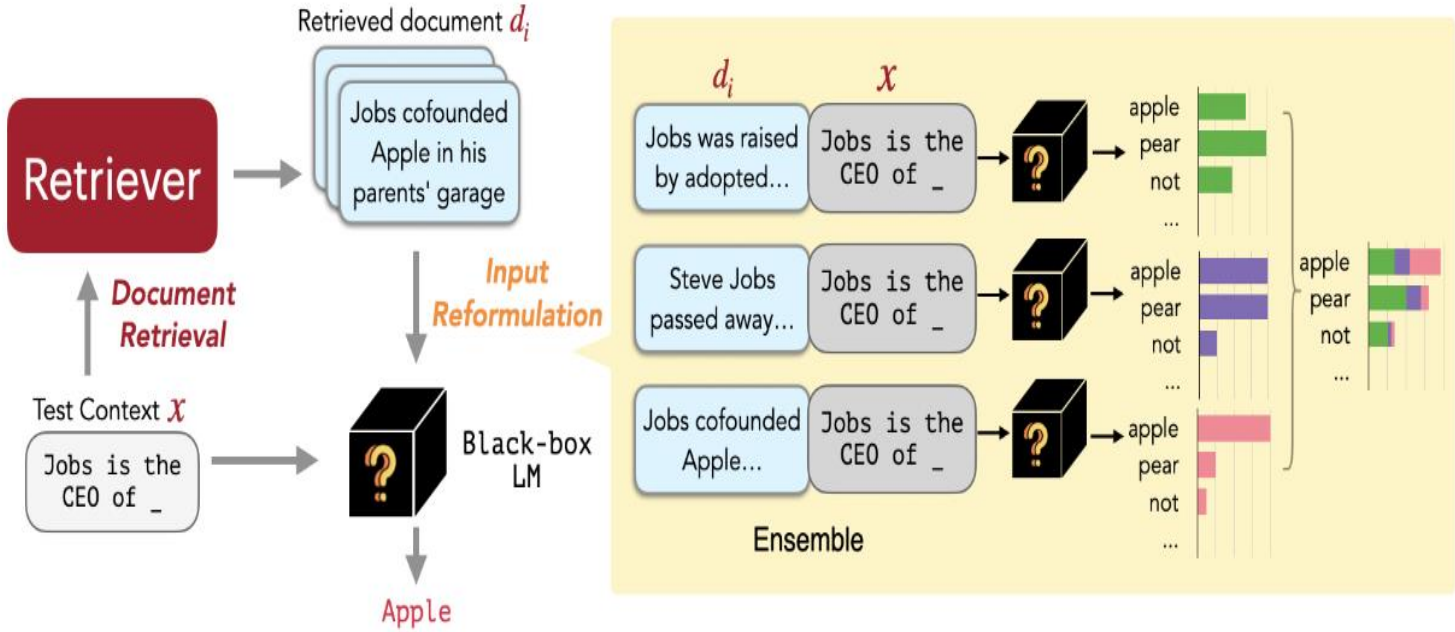

Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder* + *Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query $x$, we use Maximum Inner Product Search (MIPS) to find the top-K documents $z_i$. For final prediction $y$, we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

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis, etc

# REPLUG

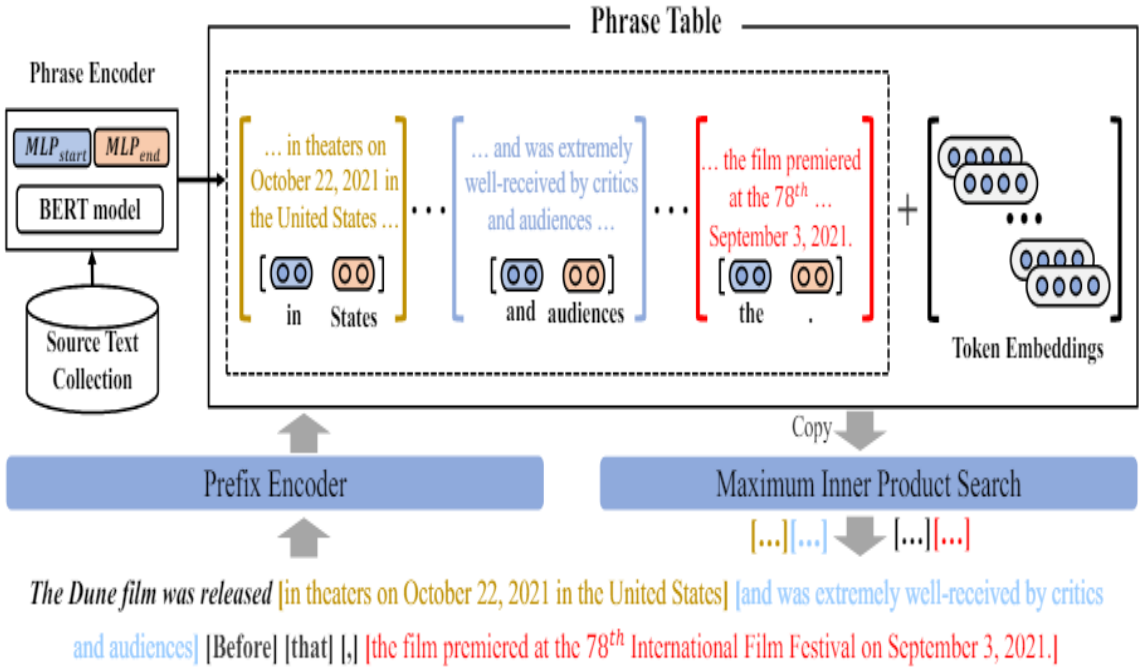


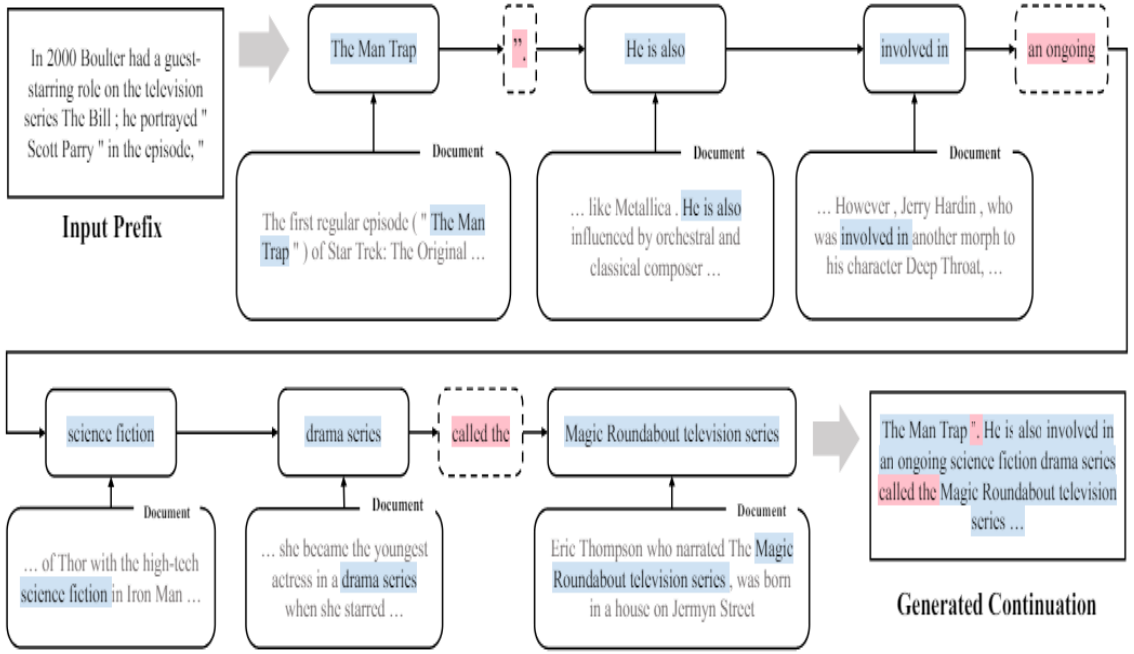




REPLUG: Retrieval-Augmented Black-Box Language Models

Weijia Shi, etc

# CoG



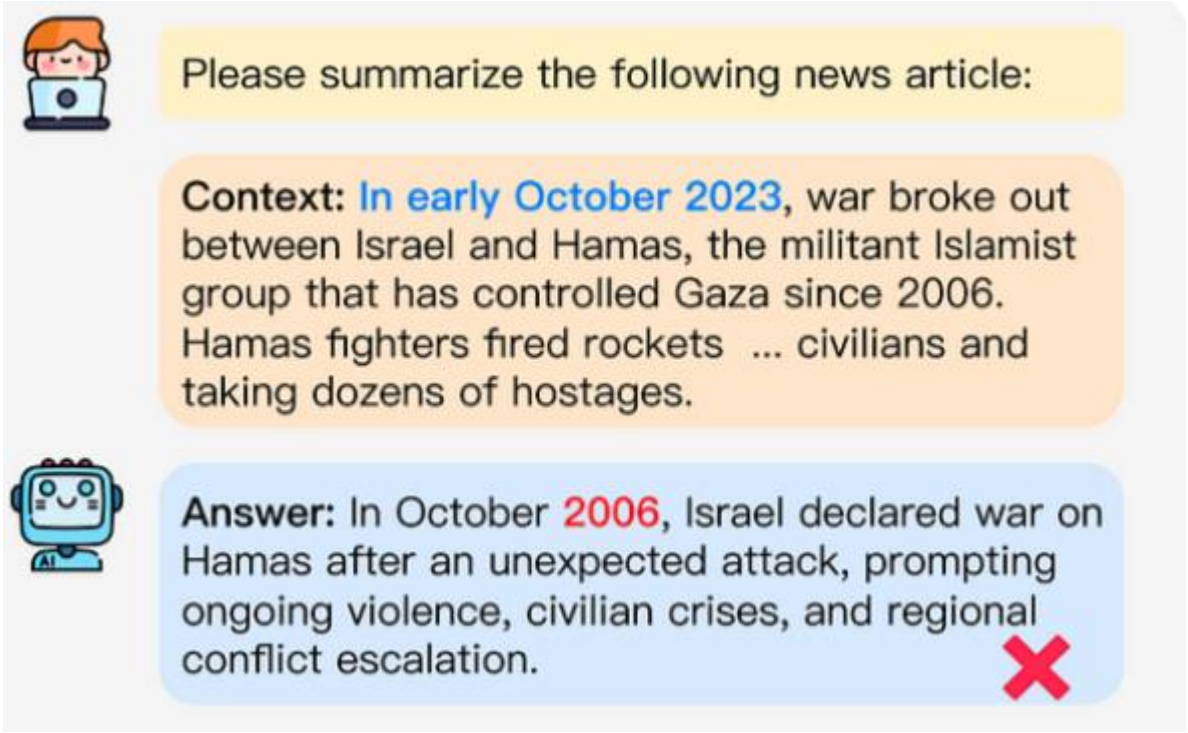




Copy is all you need

Tian Lan, etc

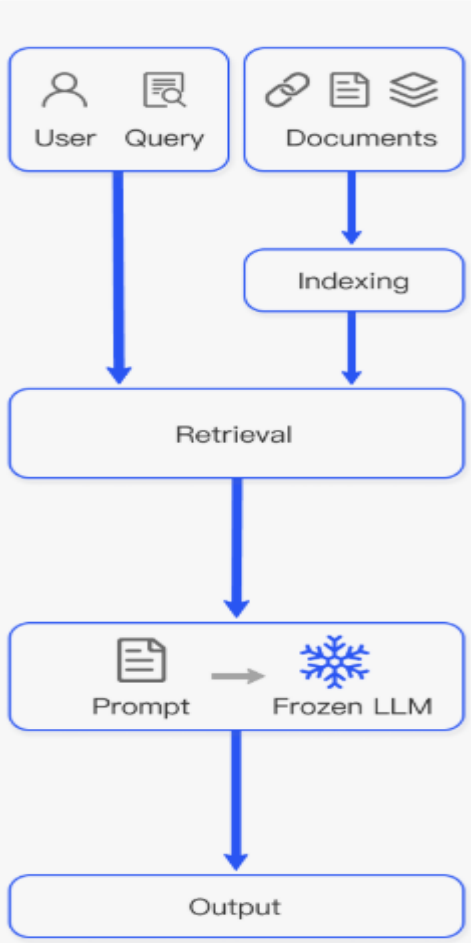
# Evolution of RAG Paradigms

- Naive RAG
- Advanced RAG
  - Optimizes indexing and retrieval
  - Introduces pre-retrieval and post-retrieval strategies
- Modular RAG
  - Adds specialized modules (e.g., Search)
  - Flexible and adaptive



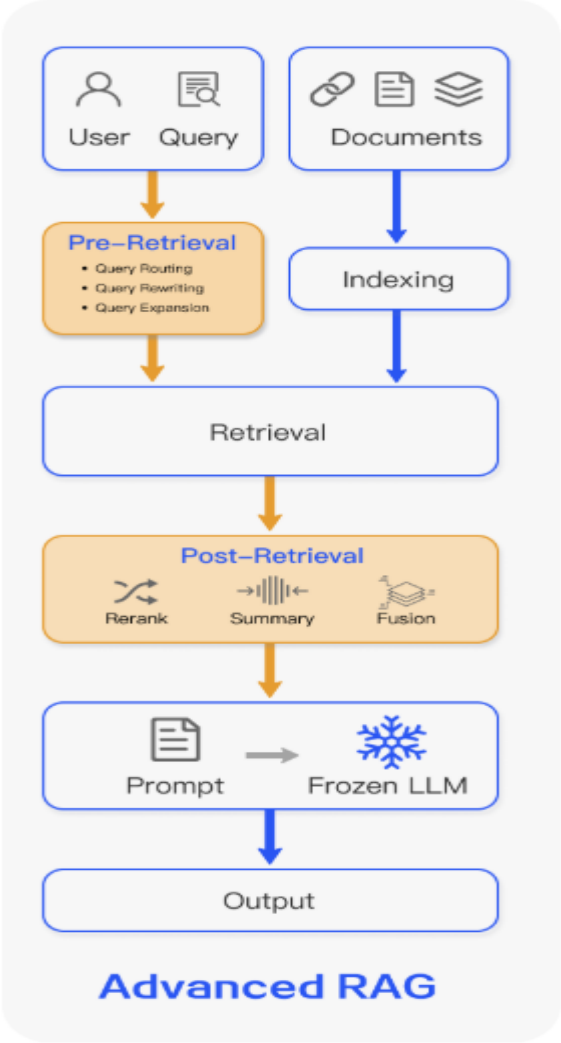
# Naive RAG

- The Naive RAG research paradigm represents the earliest methodology, which includes indexing, retrieval, and generation
  - Basic retrieval-read framework
  - Limitations: precision, hallucinations, integration challenges

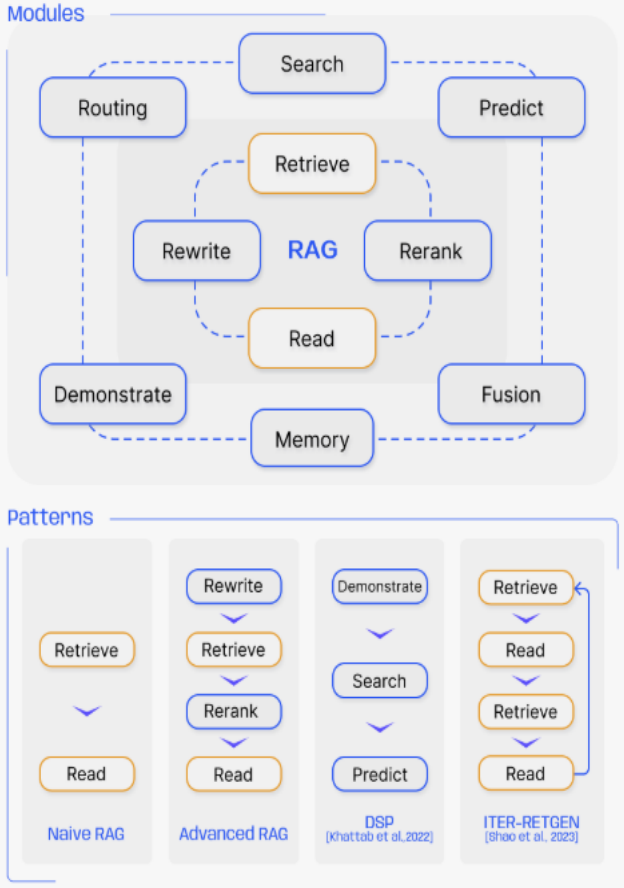


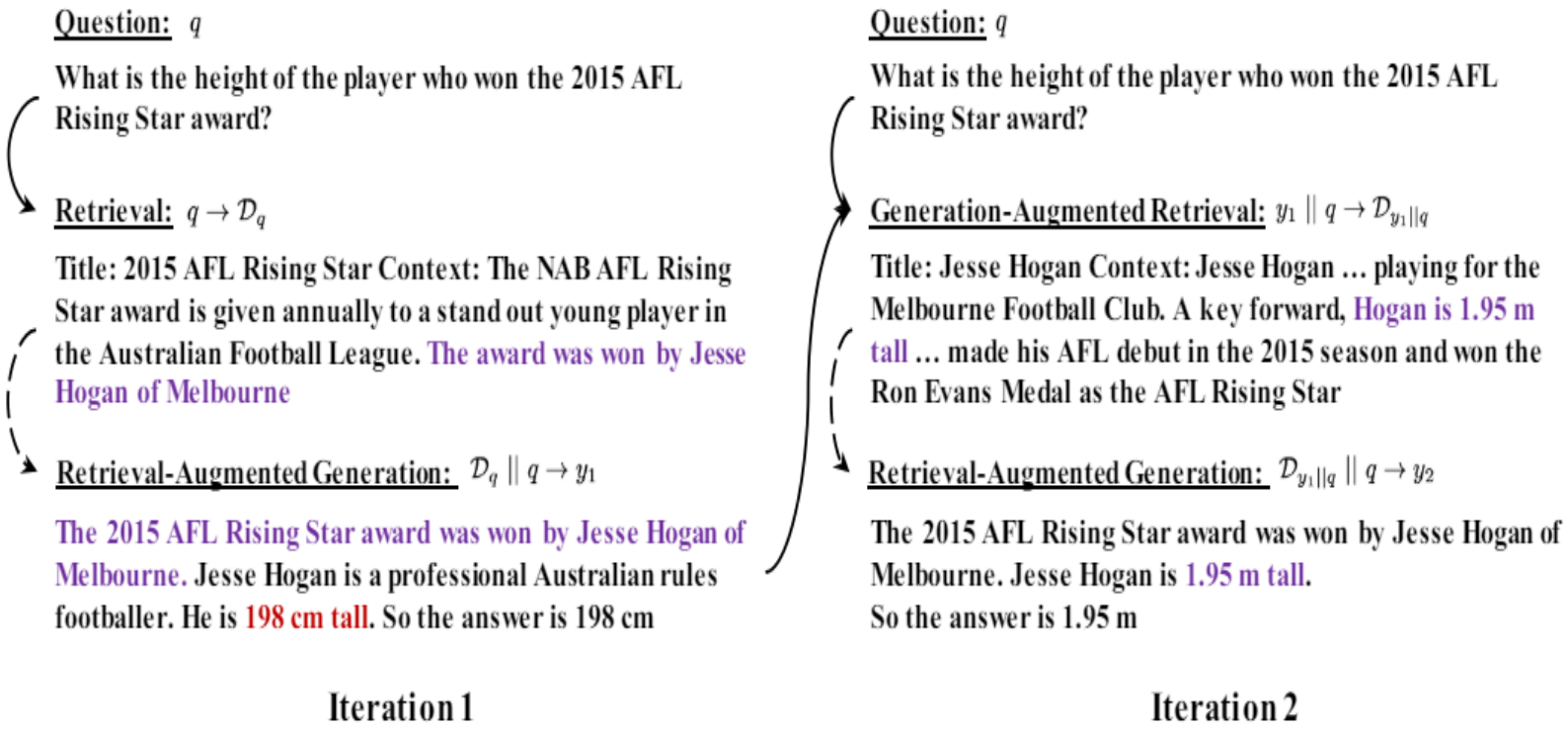
# Advanced RAG

- Advanced RAG introduces specific improvements to overcome the limitations of Naive RAG. It employs pre-retrieval and post-retrieval strategies, such as:
  - 1) Pre-retrieval: Optimizing indexing structure and original query to enhance content quality and retrieval accuracy.
  - 2) Post-retrieval: Integrating retrieved context with the query, re-ranking chunks, and compressing context to improve the relevance and coherence of the final response.

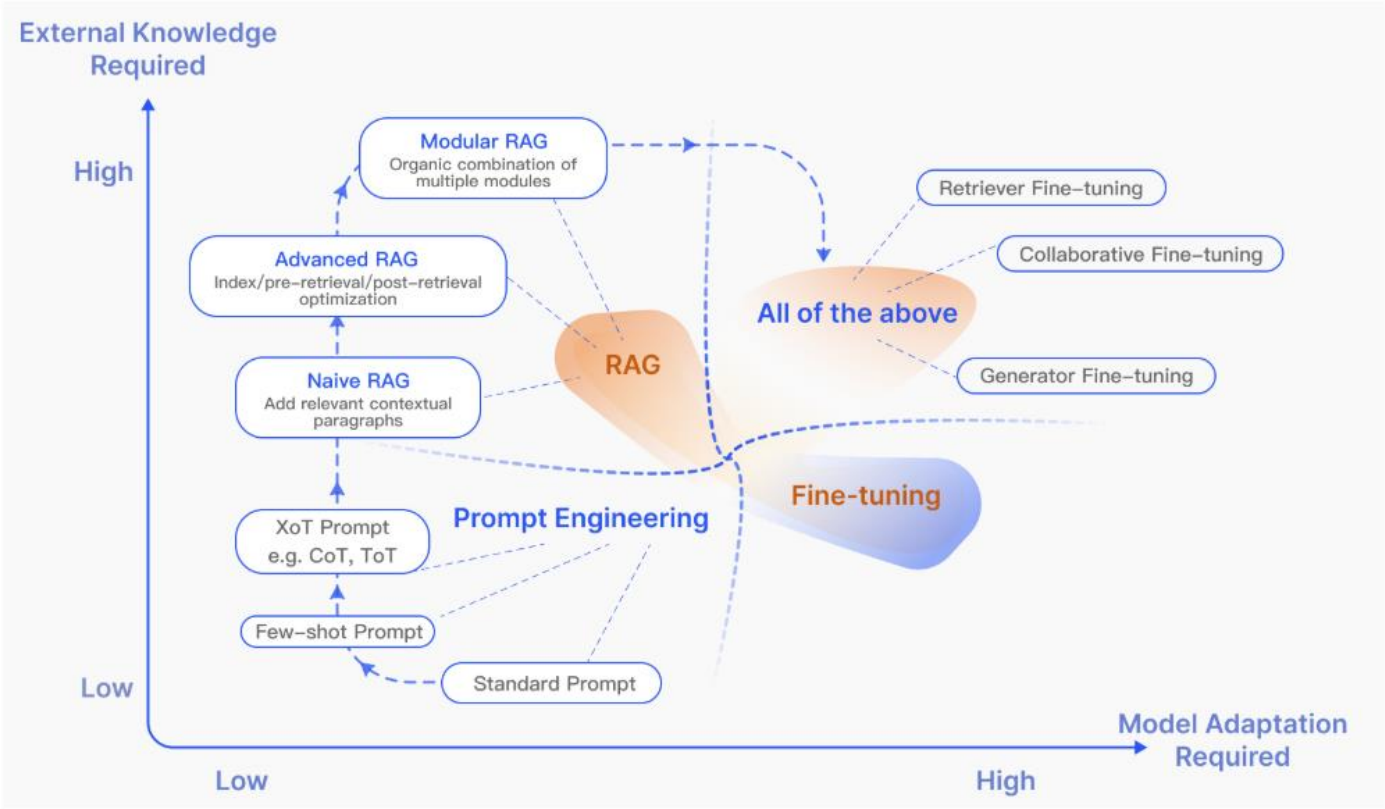


# Modular RAG


Modules
Search
Routing
Predict
Retrieve
Rewrite
RAG
Rerank
Read
Demonstrate
Fusion
Memory
Patterns
Retrieve
Read
Naive RAG
Rewrite
Retrieve
Rerank
Read
Advanced RAG
Demonstrate
Search
Predict
DSP
[Khattab et al.,2022]
Retrieve
Read
Retrieve
Read
ITER-RETGEN
[Shao et al., 2023]
Modular RAG

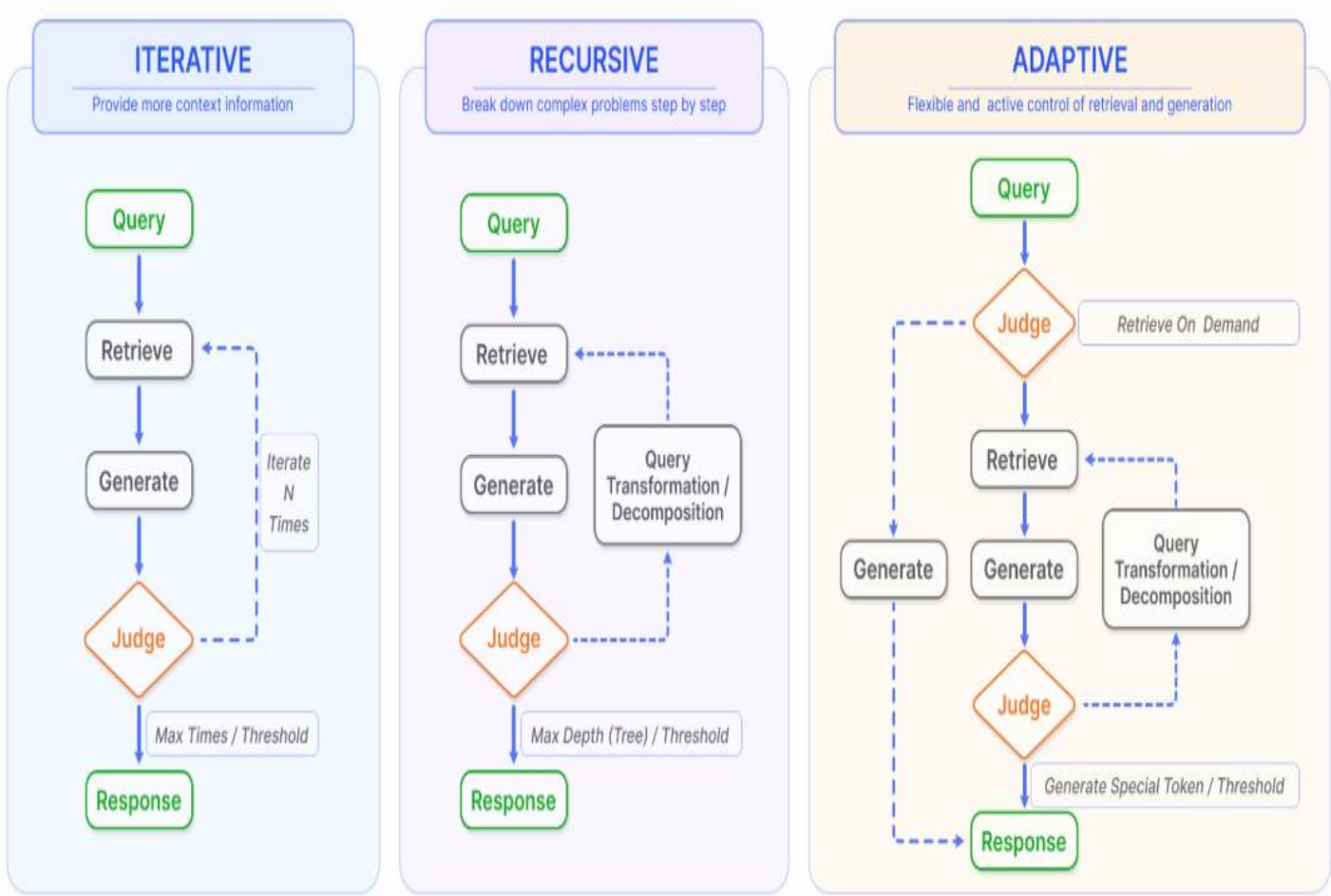

Question: $q$
What is the height of the player who won the 2015 AFL Rising Star award?
Retrieval: $q \rightarrow \mathcal{D}_q$
Title: 2015 AFL Rising Star Context: The NAB AFL Rising Star award is given annually to a stand out young player in the Australian Football League. The award was won by Jesse Hogan of Melbourne
Retrieval-Augmented Generation: $\mathcal{D}_q \parallel q \rightarrow y_1$
The 2015 AFL Rising Star award was won by Jesse Hogan of Melbourne. Jesse Hogan is a professional Australian rules footballer. He is 198 cm tall. So the answer is 198 cm
Iteration 1
Question: $q$
What is the height of the player who won the 2015 AFL Rising Star award?
Generation-Augmented Retrieval: $y_1 \parallel q \rightarrow \mathcal{D}_{y_1 \parallel q}$
Title: Jesse Hogan Context: Jesse Hogan ... playing for the Melbourne Football Club. A key forward, Hogan is 1.95 m tall ... made his AFL debut in the 2015 season and won the Ron Evans Medal as the AFL Rising Star
Retrieval-Augmented Generation: $\mathcal{D}_{y_1 \parallel q} \parallel q \rightarrow y_2$
The 2015 AFL Rising Star award was won by Jesse Hogan of Melbourne. Jesse Hogan is 1.95 m tall.
So the answer is 1.95 m
Iteration 2

# RAG and Other Model Optimization Methods



# Core Components of RAG

- Retrieval
  - Indexing documents
  - Semantic similarity calculation
- Generation
  - Synthesizing query and retrieved documents
  - Addressing hallucinations
- Augmentation
  - Integrating retrieved information
  - Ensuring relevance and coherence



# Challenges and Future Directions

- Current Challenges
  - Retrieval precision, generation quality, augmentation complexity
- Future Research
  - Enhancements in retrieval methods
  - Improved integration techniques
  - Exploration of new applications

# Q&A

L
LARGE
LARGE
LANGUAGE MODELS
LARGUGE
LANGUAGE
MODELS
LLMS
Thanks for
your attention.

# References

[1] Y. Gao et al., “Retrieval-Augmented Generation for Large Language Models: A Survey,” 2023, arXiv.

doi: 10.48550/ARXIV.2312.10997.