

Bhavishya Pandit

# Understanding

## Speculative RAG

*Overcoming the inaccuracies of traditional LLMs*

Swipe for more



# The Challenges of Traditional RAG Systems



Large Language Models (LLMs) often struggle with factual inaccuracies and hallucinations.



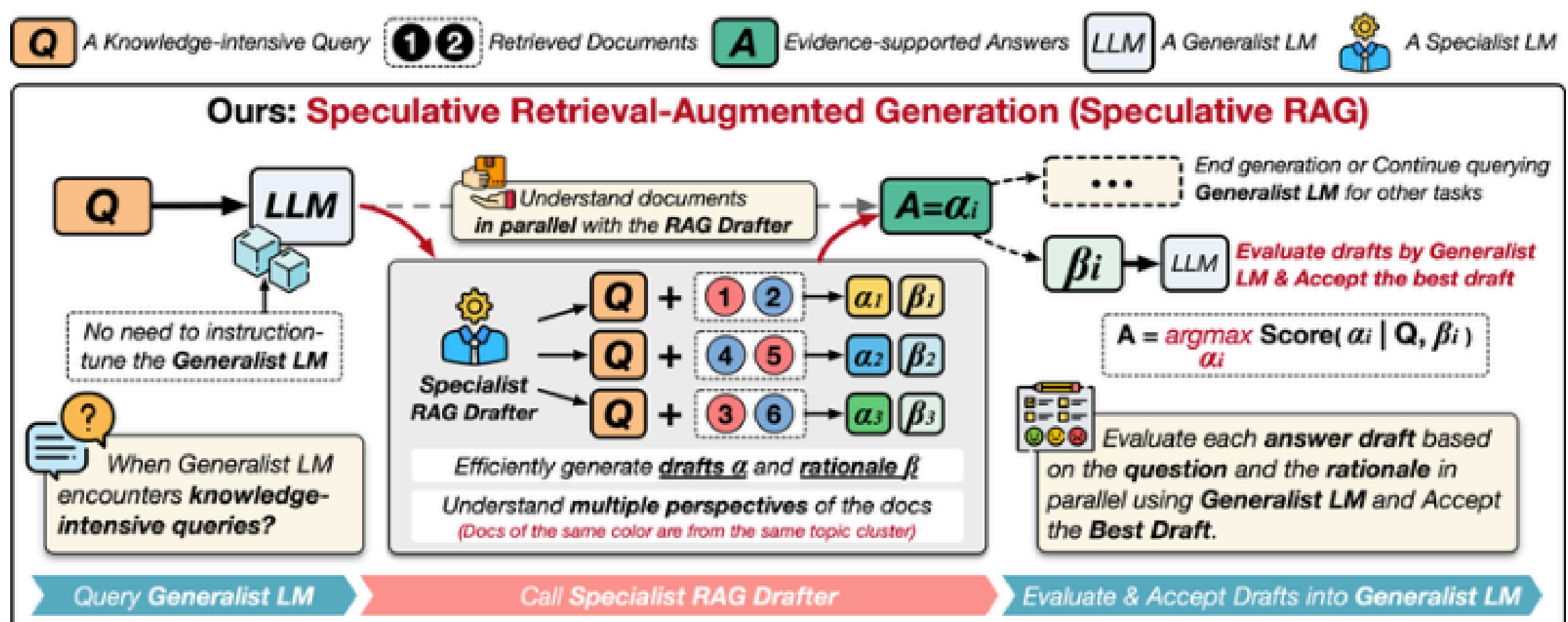
Standard RAG systems face latency issues when processing long, complex documents.



Striking a balance between accuracy and efficiency remains a key challenge.

# Speculative RAG: A Two-Step Framework

- Speculative RAG offloads the computational burden to a smaller specialist RAG drafter.
- A generalist RAG verifier validates drafts and selects the best one.
- Inspired by speculative decoding, it enhances both speed and accuracy.

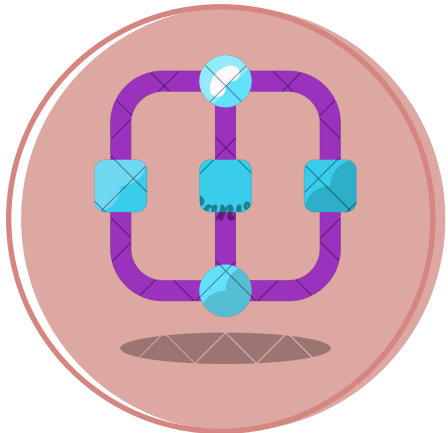


# How does it work?



## **Specialized Drafting**

A smaller, specialized LM quickly generates multiple answer drafts from subsets of retrieved documents.



## **Parallel Processing**

Drafts are created simultaneously, leveraging diverse perspectives to enhance answer quality.



## **Generalist Verification**

A larger generalist LM verifies each draft, focusing on selecting the most accurate answer.

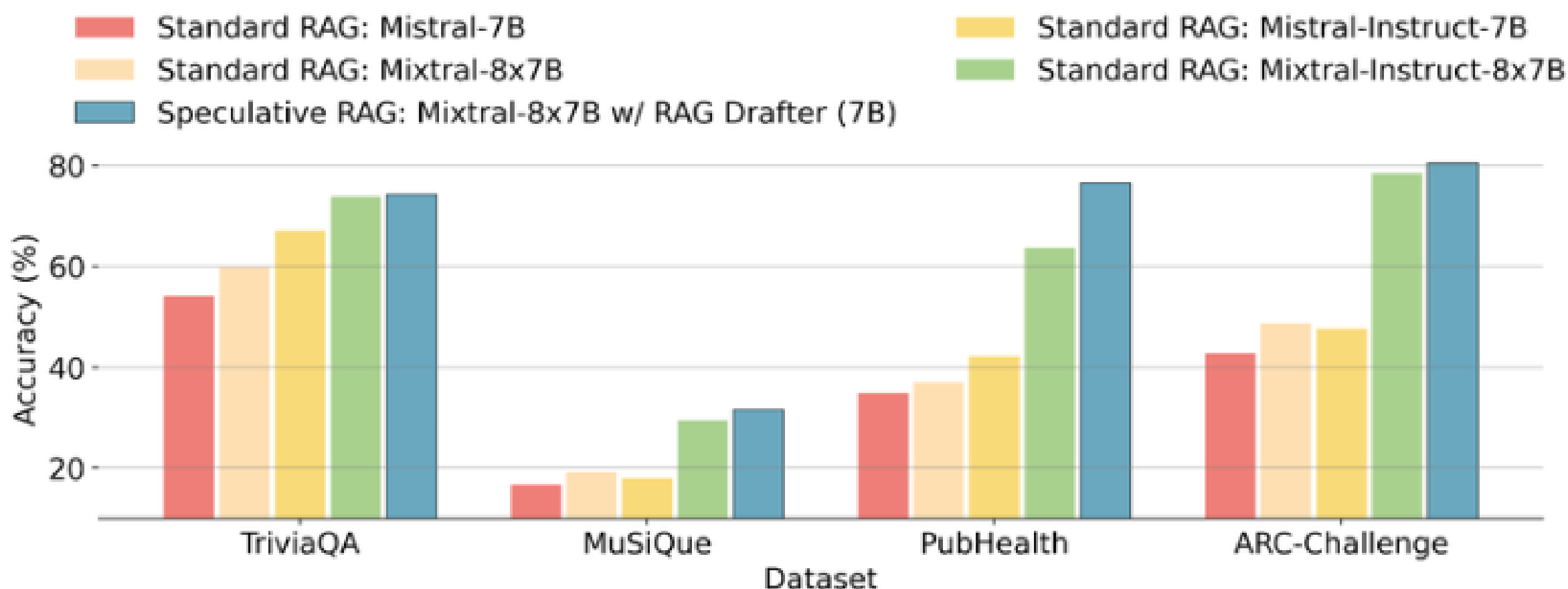


## **Final Selection**

The draft with the highest confidence score is chosen as the final output, ensuring both accuracy and efficiency.

# Why Speculative RAG?

- **Accuracy:** Outperforms traditional RAG systems by up to 12.97% on benchmarks.
- **Lower Latency:** Reduces latency by 51% due to efficient document processing.
- **Efficiency:** Efficiently handles complex documents with reduced computational load.
- **Scalability:** Easily adaptable to various tasks without additional tuning.



# Applications



## Knowledge-Intensive QA

answering complex questions



## Real-Time Information Retrieval

up-to-date info like news aggregators



## Medical and Legal Text Analysis

Provides precise and timely insights

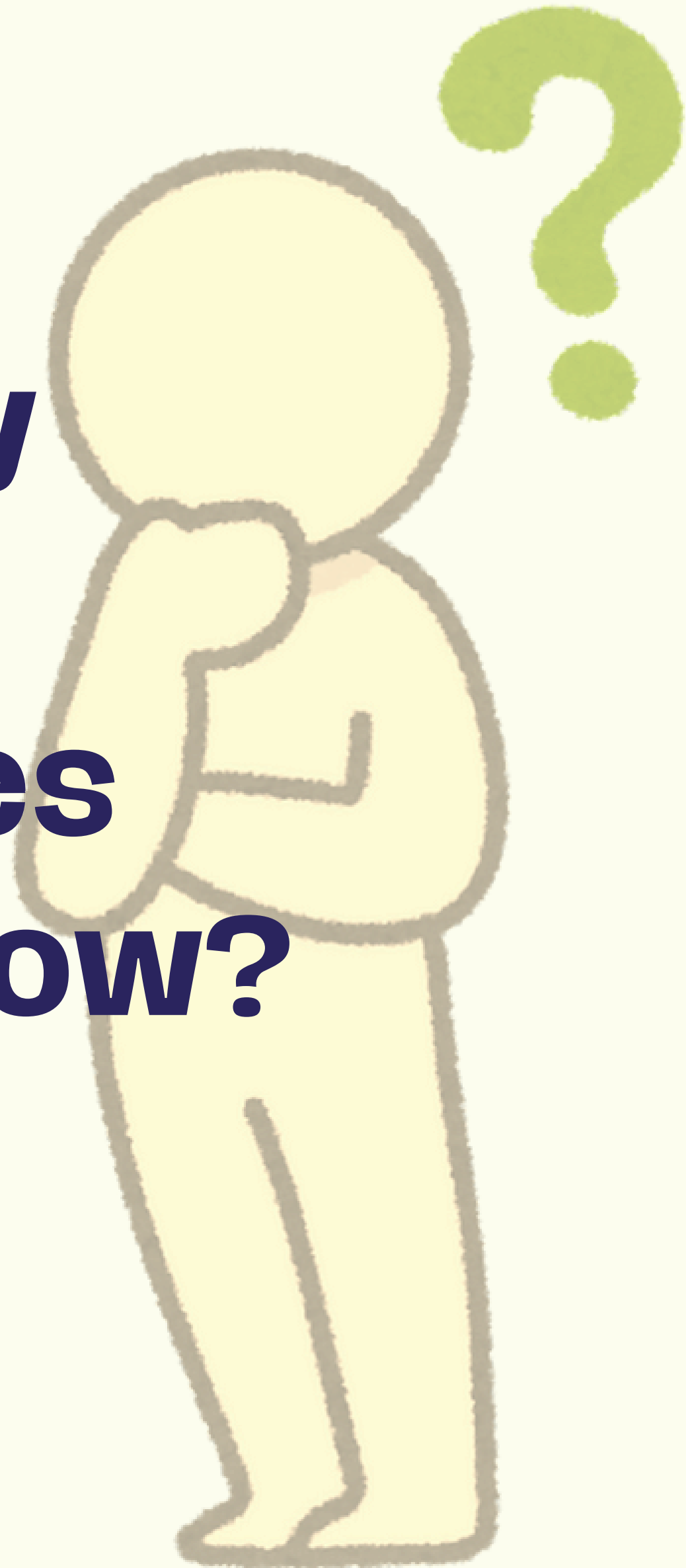


## Educational Tools

context-rich content for e-learning platforms

Bhavishya Pandit

**How many  
RAG  
techniques  
do you know?**

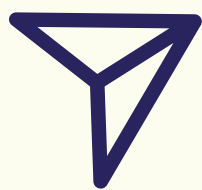




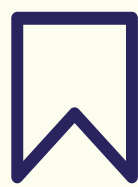
Bhavishya Pandit



**Follow for more  
AI/ML posts**



**Share your  
thoughts**



**Save for  
later**



**Like this  
Post**