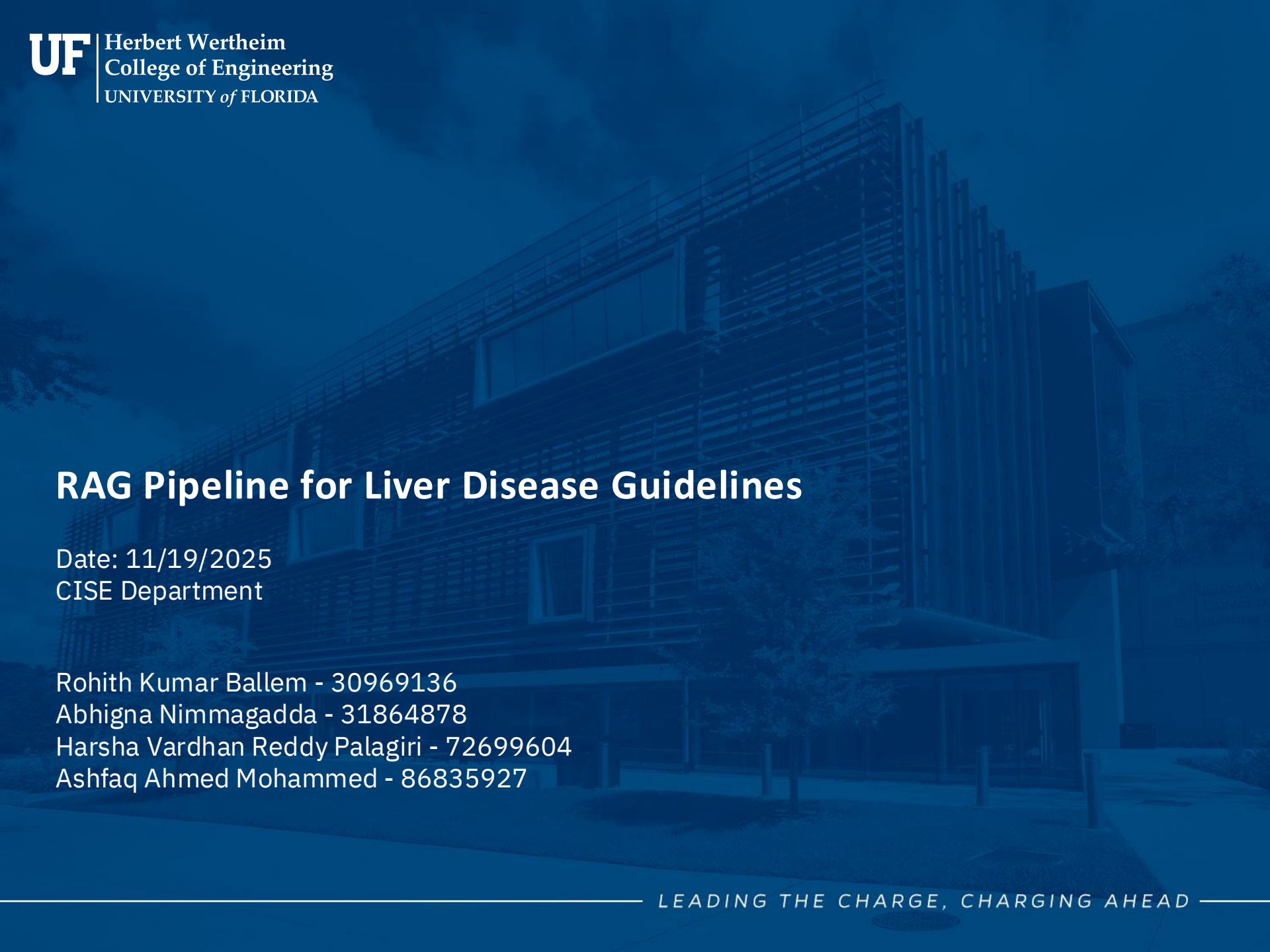


Herbert Wertheim College of Engineering
UNIVERSITY of FLORIDA

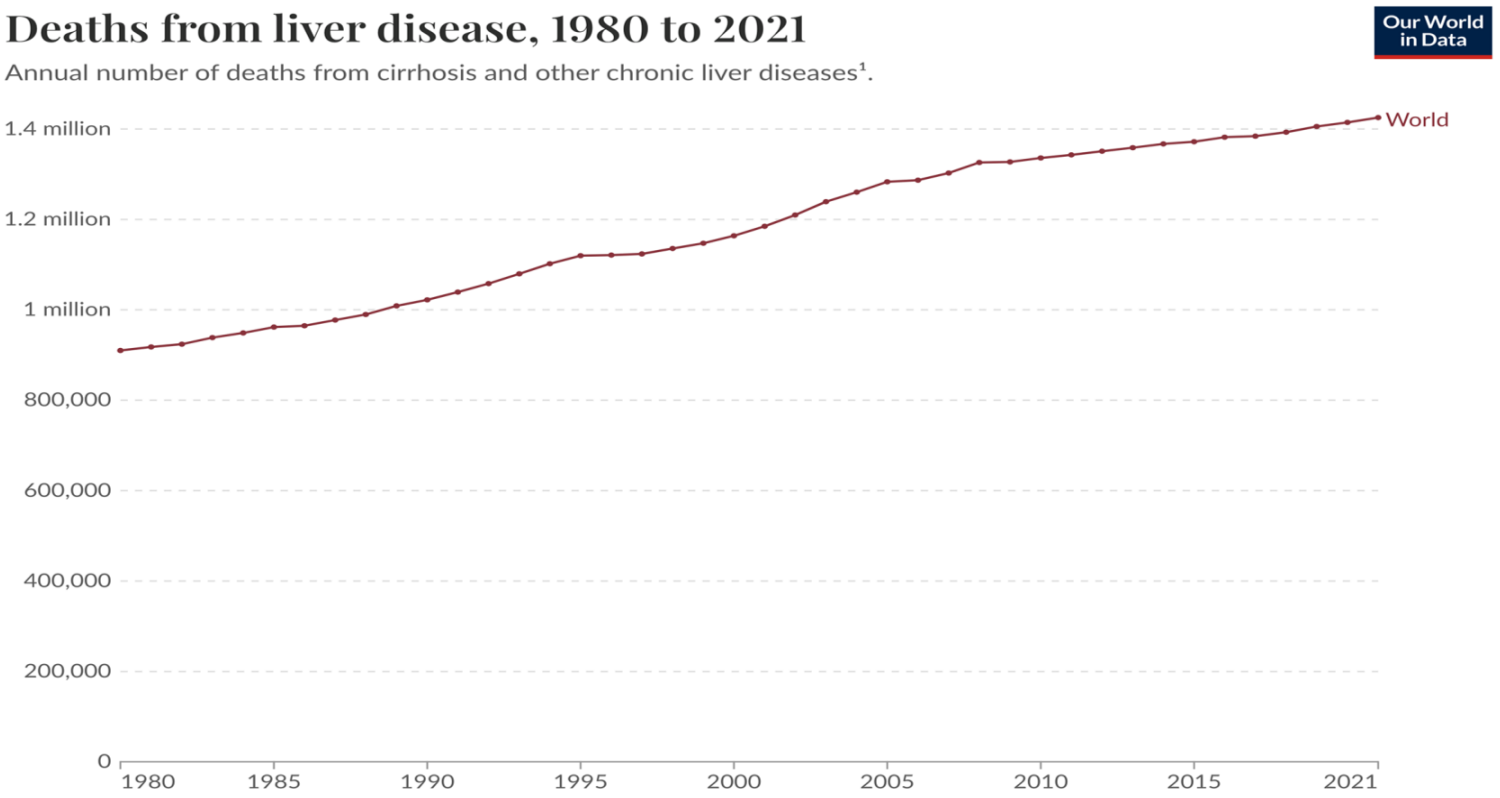
RAG Pipeline for Liver Disease Guidelines

Date: 11/19/2025
CISE Department

Rohith Kumar Ballem - 30969136
Abhigna Nimmagadda - 31864878
Harsha Vardhan Reddy Palagiri - 72699604
Ashfaq Ahmed Mohammed - 86835927

Why Liver Disease? - The Scale of the Problem

- Liver disease is a major global health crisis affecting billions of people worldwide.
- Liver disease accounts for 2 million deaths annually, that's 1 out of every 25 deaths globally. Yet it remains one of the most preventable and treatable conditions if managed correctly with current guidelines.



The Problem

The Core Problem

Large Language Models (like ChatGPT, Claude, Gemini) are trained on static data with a knowledge cutoff date.

They cannot access real-time medical guidelines and often provide confident answers that are incorrect or outdated.

Four Critical Failures

1. Hallucination Without Awareness

- LLMs cannot distinguish between facts they learned and information they "made up"
- No built-in mechanism to flag uncertainty in medical recommendations

2. Knowledge Has an Expiration Date

- The system confidently provides outdated recommendations

3. Frequent Contradiction of Guidelines

- What was standard care in 2020 may be harmful in 2025

4. Loss of Critical Clinical Details

- Exact dosages, lab thresholds, and timing are approximated or lost

Real Example

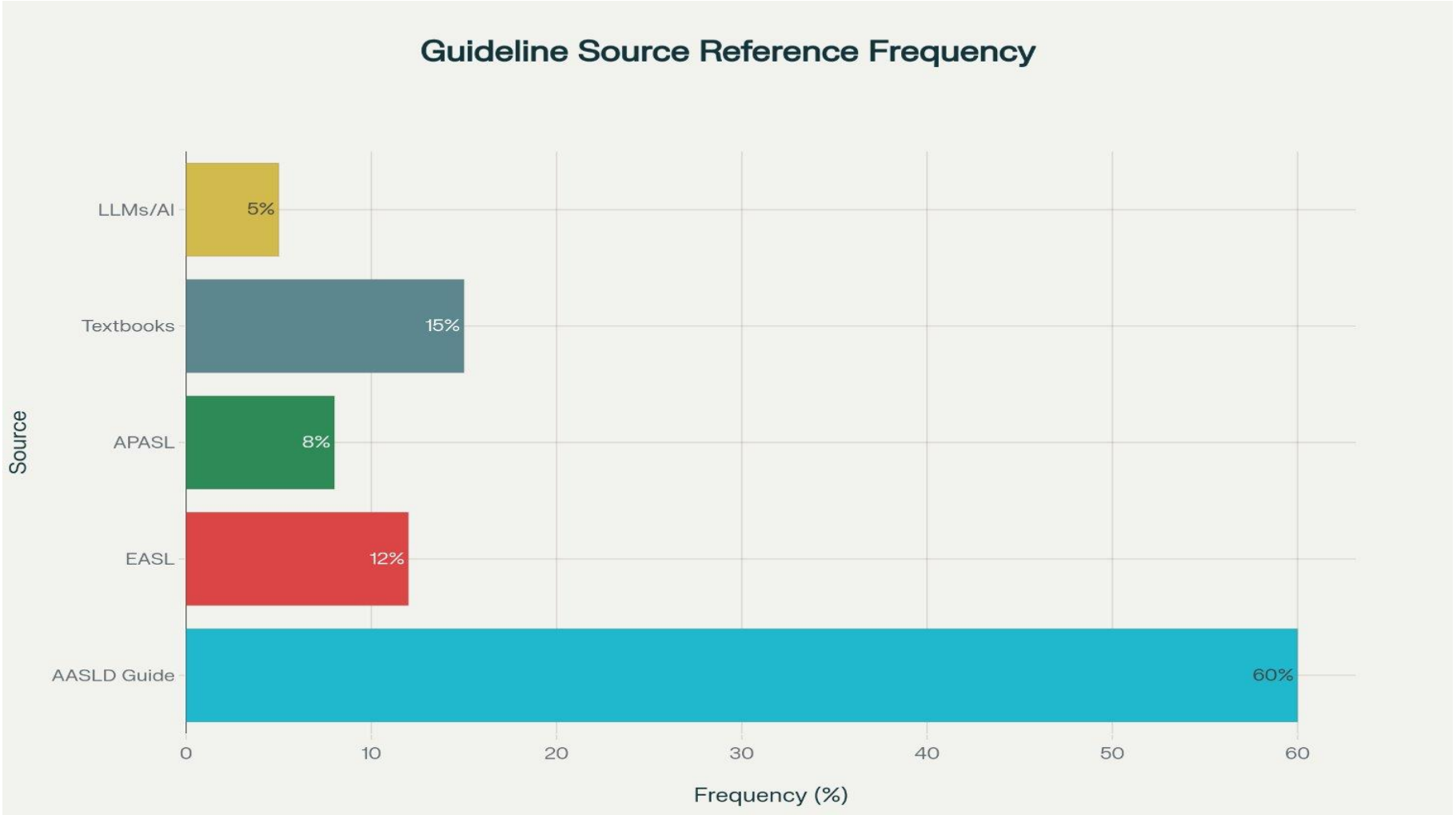
- Question: "What's the treatment for hepatitis C?"
- LLM (trained 2020): Interferon-based therapy
- AASLD 2024 Guideline: Direct-acting antivirals(DAAs) — better outcomes, fewer side effects
- Outcome if followed: Patient receives inferior, potentially harmful treatment

The Gap

- Clinicians need trustworthy, current, citable guidance grounded in official AASLD standards, not generic LLM responses.

What is AASLD?

The American Association for the Study of Liver Diseases is the leading authority on liver disease management. AASLD develops, publishes, and regularly updates evidence-based clinical practice guidelines that represent the gold standard for hepatology care.



Project motivation

The Problem We're Solving

Clinicians need rapid, reliable access to current AASLD guidelines, but the system is broken.

Our Solution

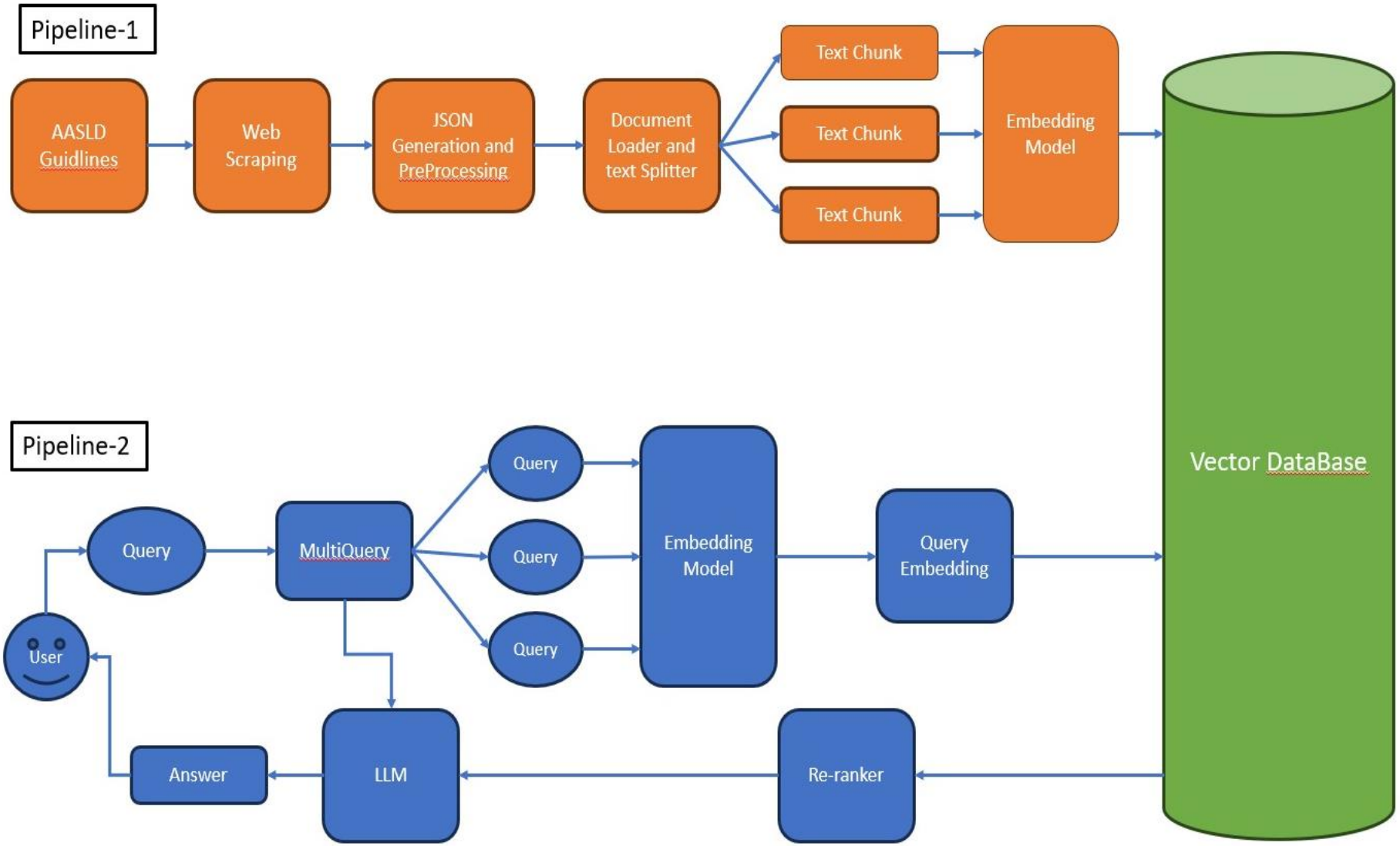
Build a RAG-Enhanced System that:

- Retrieves relevant AASLD guideline content in seconds
- Grounds answers directly in official guidelines
- Generates trustworthy recommendations with inline citations
- Preserves exact clinical details (dosages, thresholds, timing)
- Updates continuously as AASLD releases new guidelines

Impact

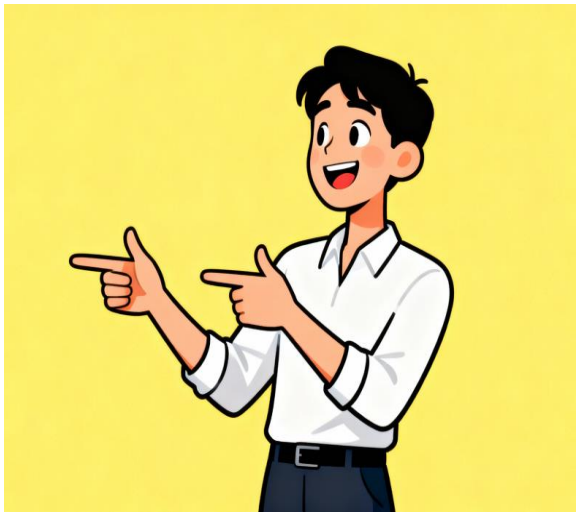
- Clinicians: Instant access to current, citable, evidence-based guidance
- Patients: Care based on latest clinical standards
- Healthcare Systems: Standardized, auditable decision support

RAG MODEL FLOWCHART



Steps for Data Extraction

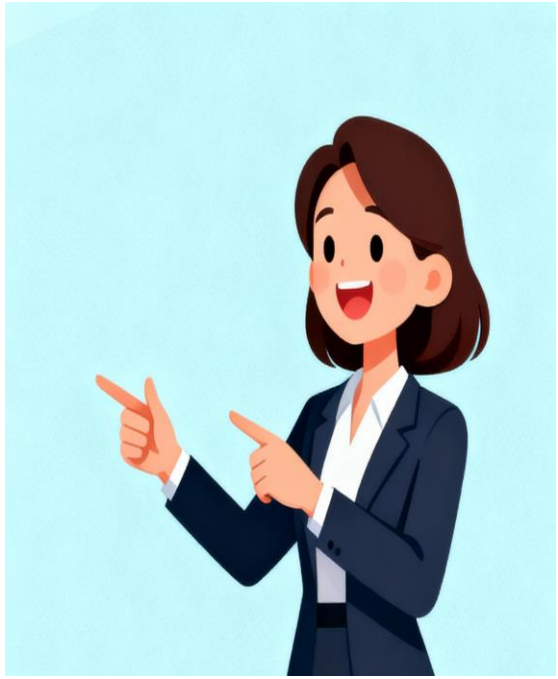
1. Extract Disease Links from AASLD Main Page
2. Extract Guideline Links from Disease Pages
3. Fetch HTML or PDF Content
4. Extract Text Sections, Tables, and Structured Data
5. Save to JSON and Text Files



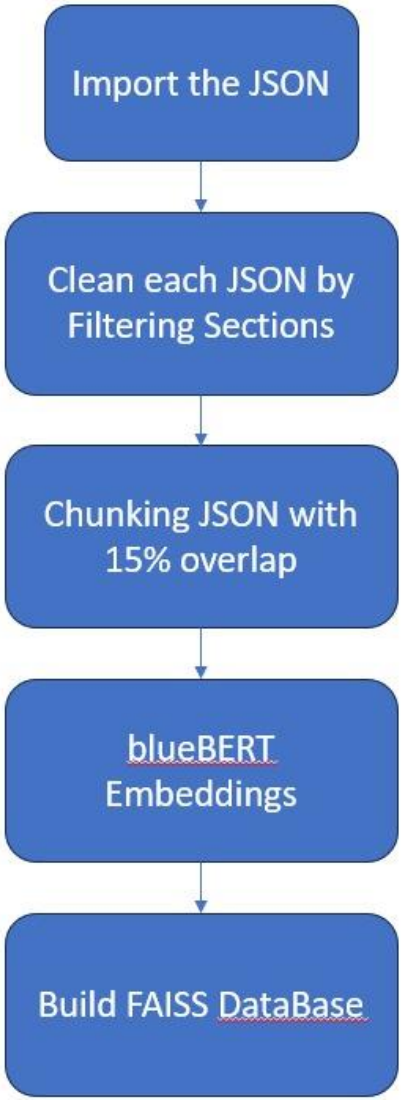
Steps For Data Cleaning

Steps For Data Cleaning

1. Load JSON Guideline ine Files
2. Clean and Normalize Text Content
3. Extract Recommendations and Clinical Thresholds
4. Save Cleaned Data and Generate Summary Report

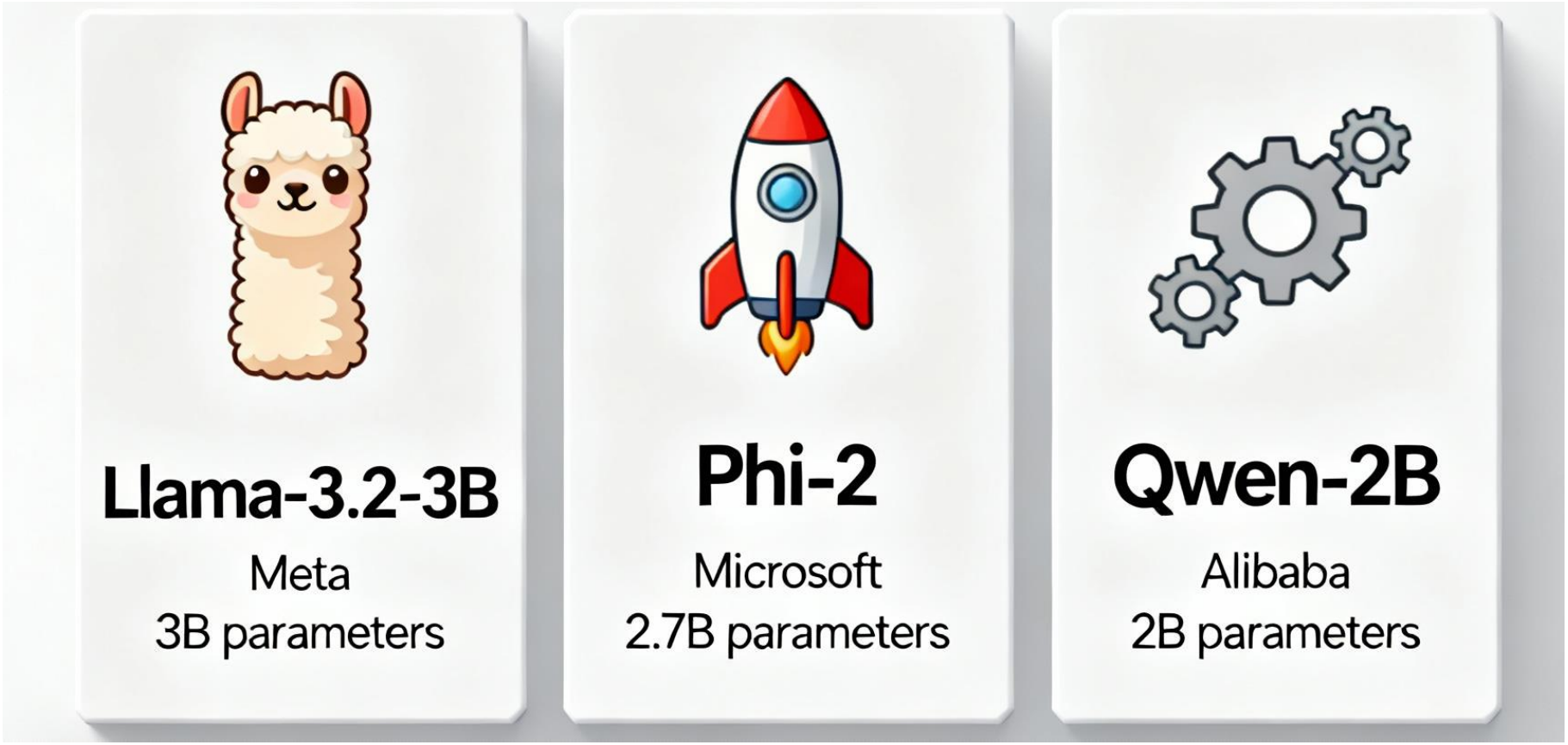


Data Processing & Embeddings



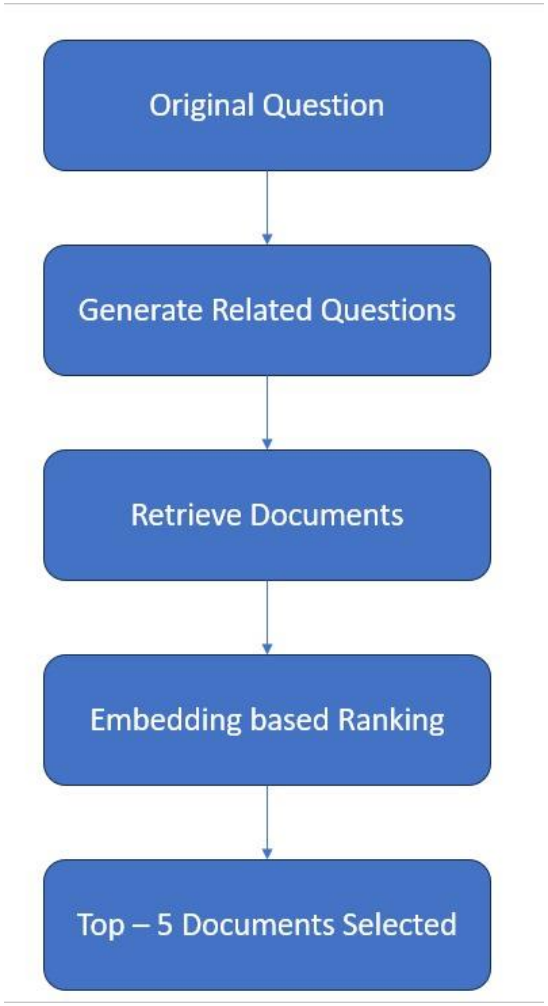
- Filtered Docs: 1507
- Chunk Size: 512 Token
- Total Chunks: 12782
- Vector Dimension: 768
- Indexing Type: indexFlatIP

LLM Models Used



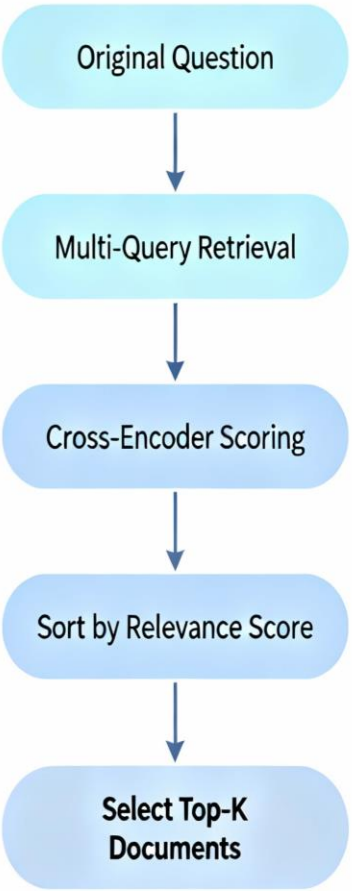
MultiQuery

- Generates multiple related questions
- Retrieves documents for each question
- Combines results (embedding-based ranking)
- Fast, simple, lightweight

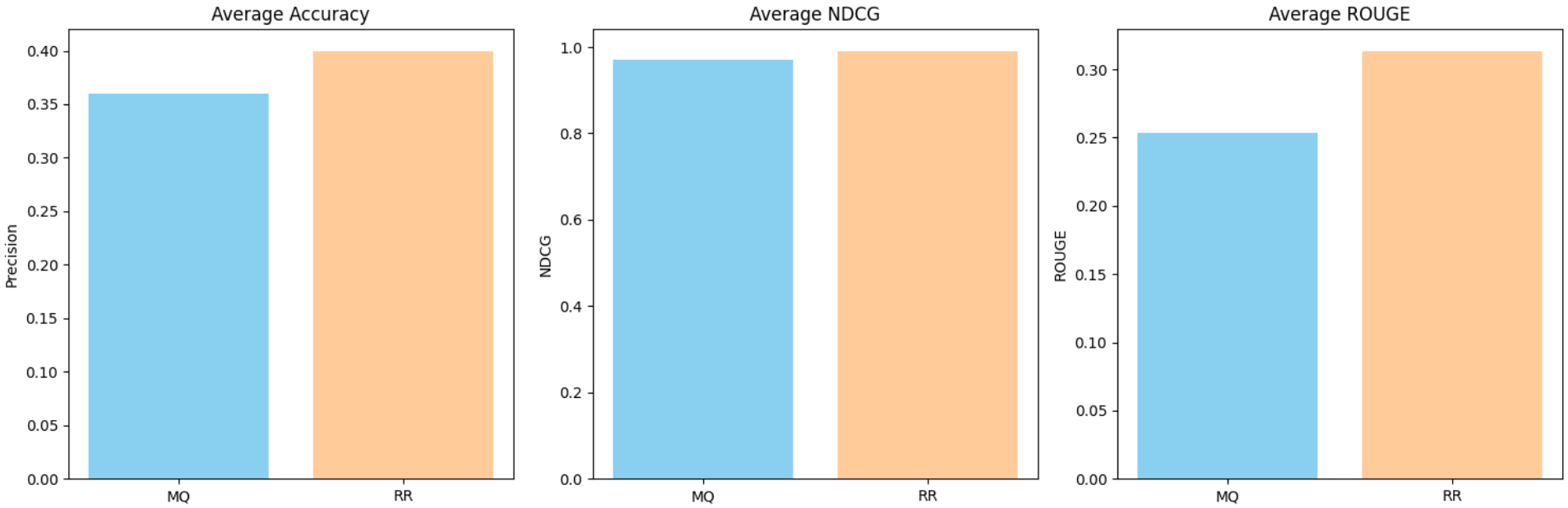


Re-ranker with Multiquery

- Perform Multiquery retrieval
- Uses Cross-Encoder to re-rank documents
- Scores (question, document) pairs directly
- More accurate, slower, heavier



Llama

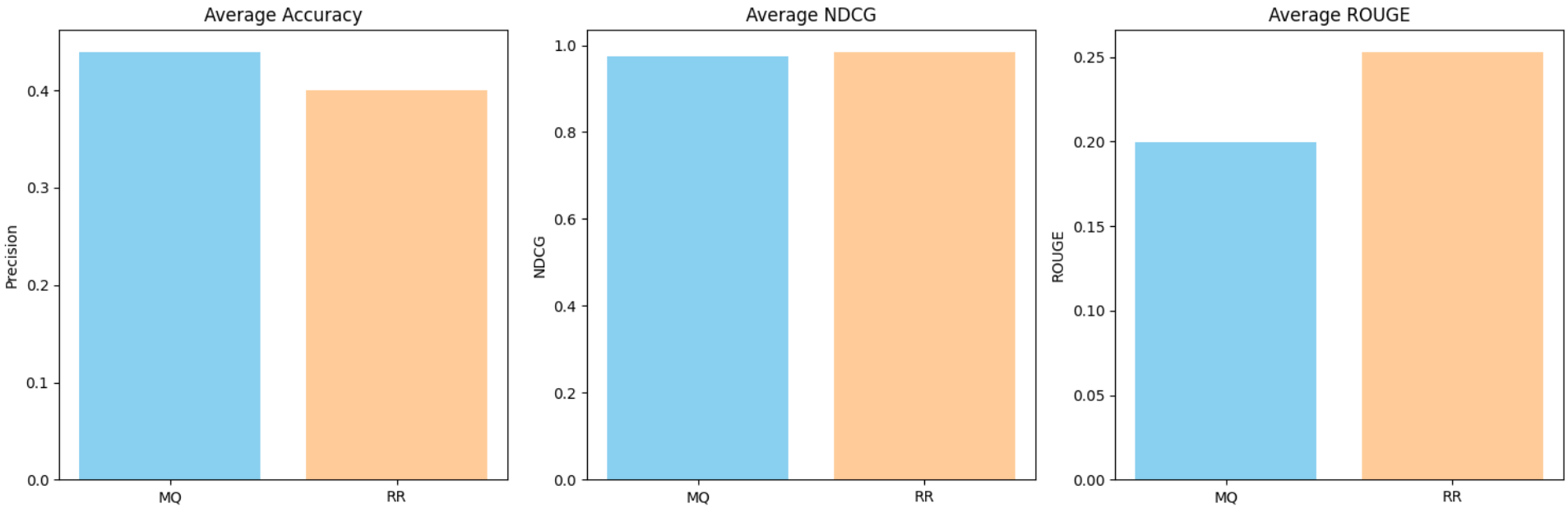


The graphs indicate that **ReRank (RR)** outperforms **MultiQuery (MQ)** across all three evaluation metrics, showing higher accuracy (precision), stronger ranking quality (NDCG), and better textual alignment with ground-truth answers (ROUGE).

For each individual metric, RR consistently produces more relevant retrieved chunks, ranks them in a more optimal order, and generates responses that better match the expected content.

Overall, these trends suggest that RR enhances both retrieval relevance and final answer quality, making it a more effective approach than MQ for this task.

Qwen

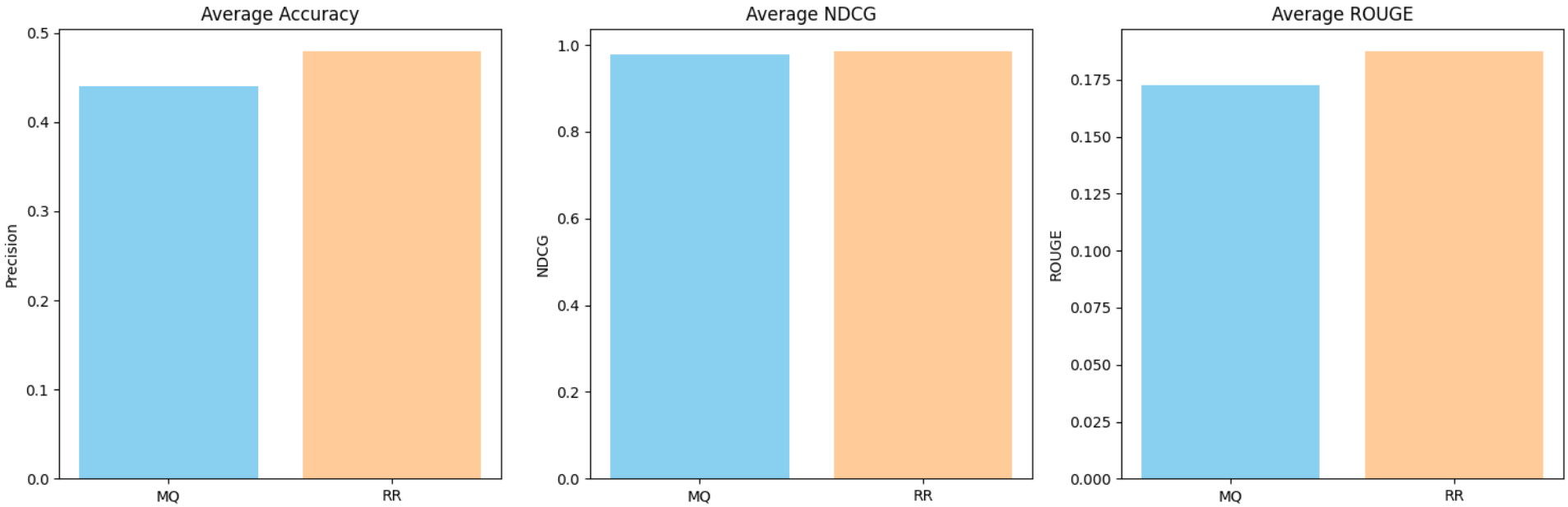


The graph shows that MultiQuery (MQ) achieves higher accuracy, while ReRank (RR) performs better in NDCG and ROUGE, indicating stronger ranking quality and better textual alignment with reference answers.

Although MQ retrieves slightly more relevant items on average, RR organizes and expresses information more effectively, resulting in higher-quality ranked outputs and summaries.

Overall, the metrics suggest a trade-off where MQ leads in raw precision, but RR provides more structured, contextually aligned, and semantically richer results.

Phi-2

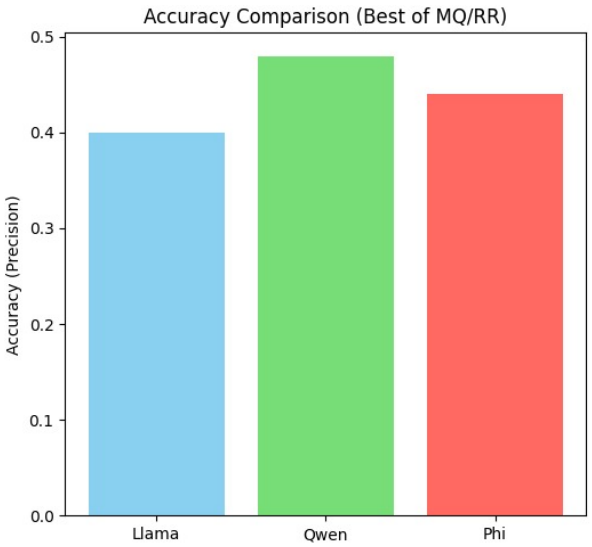


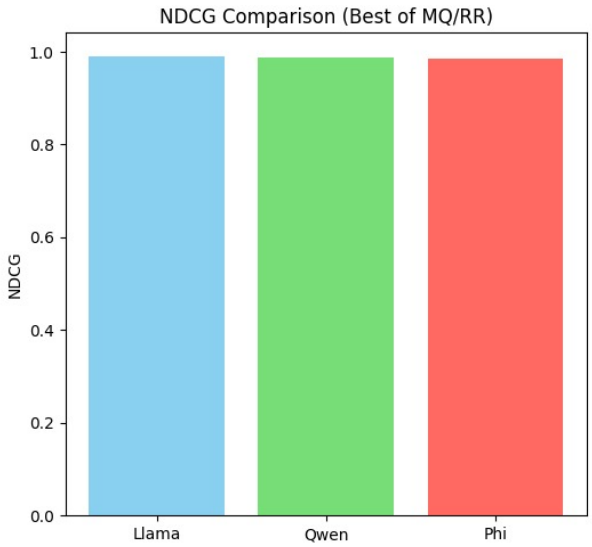
The graph shows that ReRank (RR) performs better than MultiQuery (MQ) across all metrics (Accuracy, NDCG, and ROUGE) indicating that RR retrieves more relevant information, ranks it more effectively, and produces responses with stronger textual similarity to the ground truth.

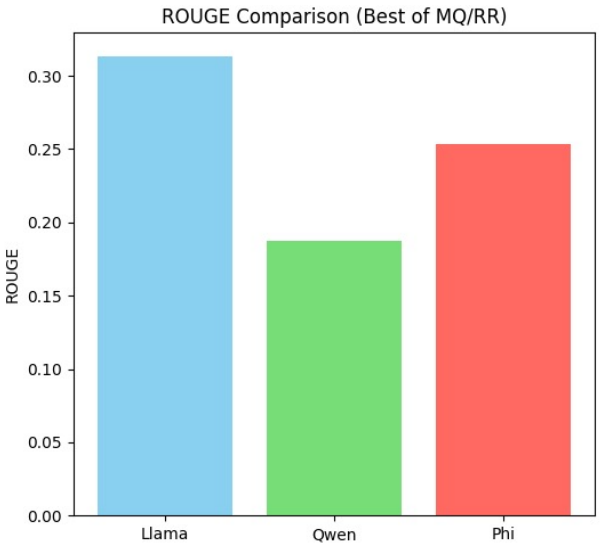
Individually, Accuracy and ROUGE show noticeable improvements with RR, while NDCG remains high for both methods but still slightly favors RR, reflecting better ordering of relevant results.

Overall, RR consistently enhances both retrieval quality and downstream response quality compared to MQ.

Comparison







The graphs show that Qwen achieves the highest accuracy, followed closely by Phi, while Llama trails behind, suggesting that Qwen retrieves the most relevant information overall. For NDCG, all three models perform almost identically, indicating that each model is strong at ranking relevant chunks near the top once they retrieve them.
However, in ROUGE, Llama leads, with Phi performing moderately and Qwen showing the weakest overlap with ground-truth text, highlighting that Llama generates responses with richer textual similarity despite its lower accuracy.

Future Scope

- Implement dynamic chunking to preserve context continuity in documents.
- Use a dynamic re-ranking strategy, enabling re-ranking only for complex queries where it actually improves retrieval quality.
- Host the RAG pipeline online and expose its functionality through FastAPI endpoints.
- Improve evaluation by testing with more varied and representative queries to better assess effectiveness.

Quiz

Q: Which option shows the correct sequence of steps in an RAG Workflow?

1. FAISS Database Creation -> Chunking - > MultiQuery - > ReRanker
2. Chunking - > FAISS Database Creation -> MultiQuery -> ReRanker
3. Chunking -> MultiQuery -> ReRanker -> FAISS Database Creation
4. Chunking -> FAISS Database Creation -> ReRanker -> MultiQuery

Answer: 2. Chunking - > FAISS Database Creation -> MultiQuery -> ReRanker

UF | Herbert Wertheim College of Engineering
UNIVERSITY *of* FLORIDA

LEADING THE CHARGE, CHARGING AHEAD