Benchmarking Open-Source LLMs in RAG Systems with Diploma Abstracts

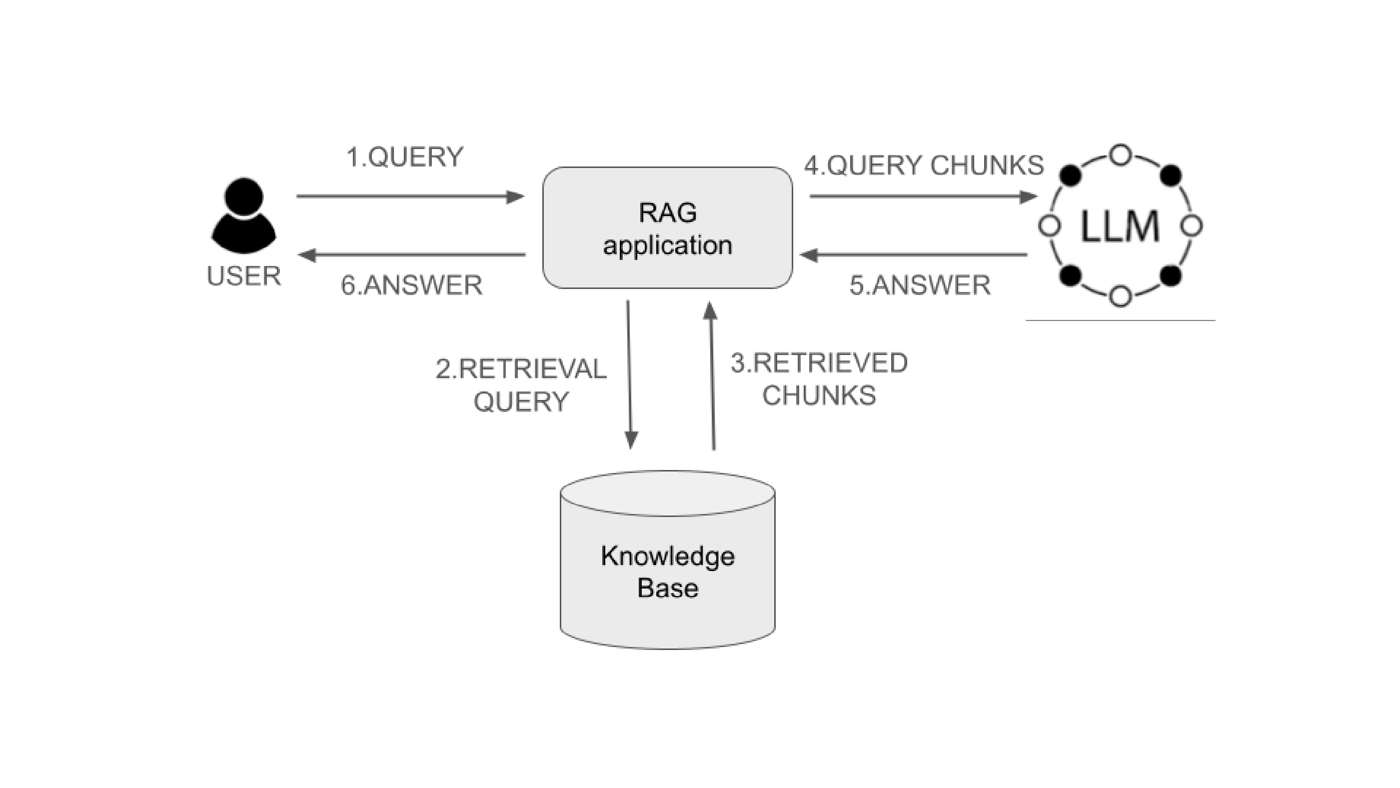
Authors:

Margit ANTAL (Sapientia University, Romania)
Krisztian BUZA (Budapest University of Economics and Business, Budapest, Hungary)

MathInfo
September 8-12, 2025

RAG System Architecture

Architecture



Introduction to RAG Systems

- **Retrieval-Augmented Generation (RAG) systems** enhance Language Model (LLM) performance.
- They **ground LLMs in external knowledge sources**, such as vector databases.
- **Challenge**: Evaluating the effectiveness of RAG systems, as both **retrieval** and **generation** components must be assessed.

The Challenge of RAG Evaluation

- **Evaluation benchmarks** often lack complexity and domain specificity needed for comprehensive RAG assessment.
- There is a necessity for **robust datasets** and **metrics** that accurately reflect real-world applications.
- Evaluation must assess both **Retrieval** and **Generation** independently and in **combination**.

Our Contributions

1. Creation of a novel, **domain-specific dataset** for RAG evaluation using diploma thesis abstracts.
2. **Categorization of questions** into `summary`, `single fact`, and `reasoning` types.
3. Comprehensive **evaluation** of a RAG pipeline:
    - Retriever performance: **lexical** and **semantic search**.
    - Generation efficacy: **faithfulness** and **answer correctness**

The *SapiTheses* Dataset

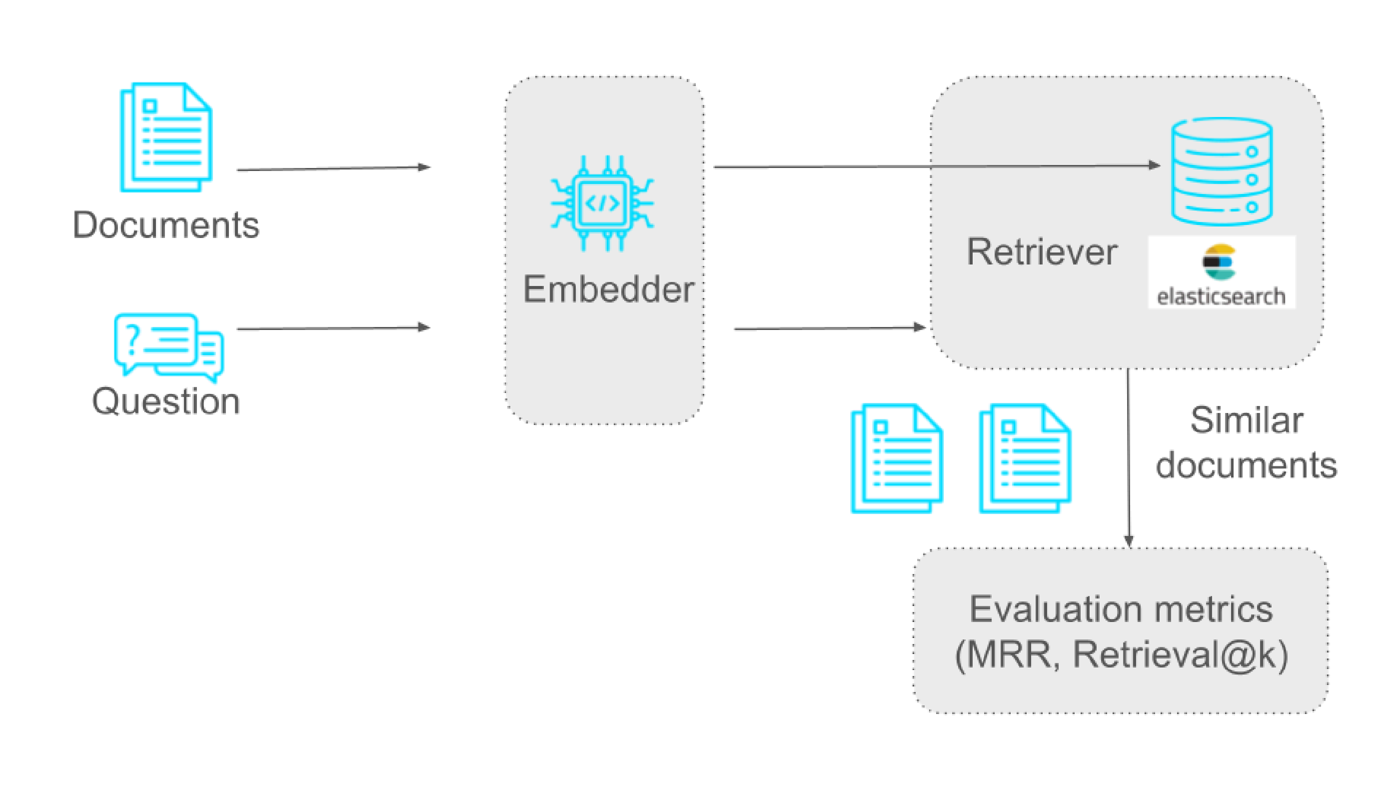
- **Source documents**: **227** English abstracts of diploma theses
- **Question-Answer pair generation**:
  - Used **Ragas** to generate questions
  - Each question relates to a single document (**single hop**)
  - **GPT-4o** was used by Ragas for question and answer generation.
  - **122 human-reviewed questions** remained after removing overly simple or general ones.

Question Categorization Taxonomy

- **Fact Single Questions**: Seek **direct factual information** explicitly present in the abstract.
- **Reasoning Questions**: Require **logical inference or multi-step reasoning** based on the abstract; the answer is inferred, not explicit.
- **Summary Questions**: Ask for a **condensed version or key points** of the abstract.
- **Dataset Distribution**: **25** fact single, **73** reasoning, **24** summary questions.

RAG System Evaluation

Retriever Evaluation



Retrieval Subsystem

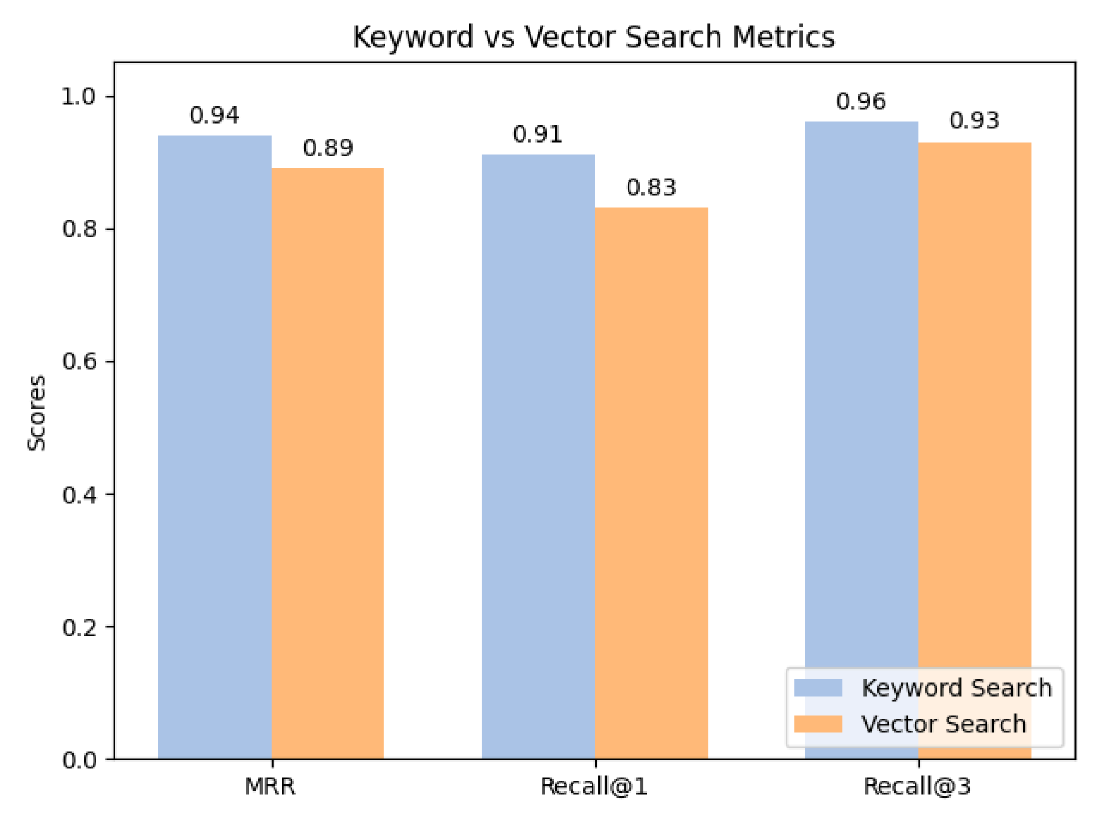
- **Tool**: *Elasticsearch*
- **Search Strategies Employed**:
  - **Lexical Search**:
    - **BM25** ranking algorithm.
  - **Semantic Search**:
    - **Embeddings** generated by the `all-mpnet-base-v2` model --> **768-dimensional** embeddings.

Retrieval Quality Metrics

- **Mean Reciprocal Rank (MRR)**: Ranks the first relevant document's position.
- **Recall@k**: Frequency of the accurate context being found within the **top k** result.

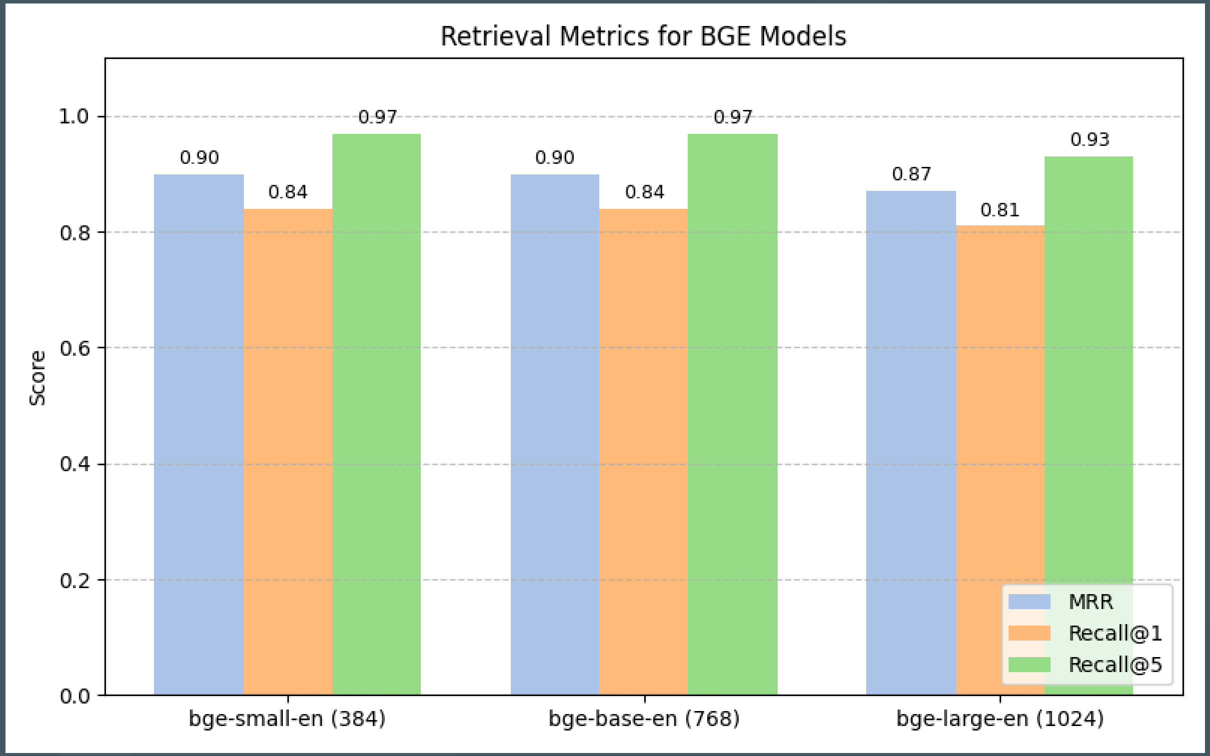
Retriever Performance Insights

Embedding model: `all-mpnet-base-v2 (768)`



Retriever Performance - BGE models

Embedding models: `small (384)`, `base (768)`, `large(1024)`



Answer Generation Subsystem

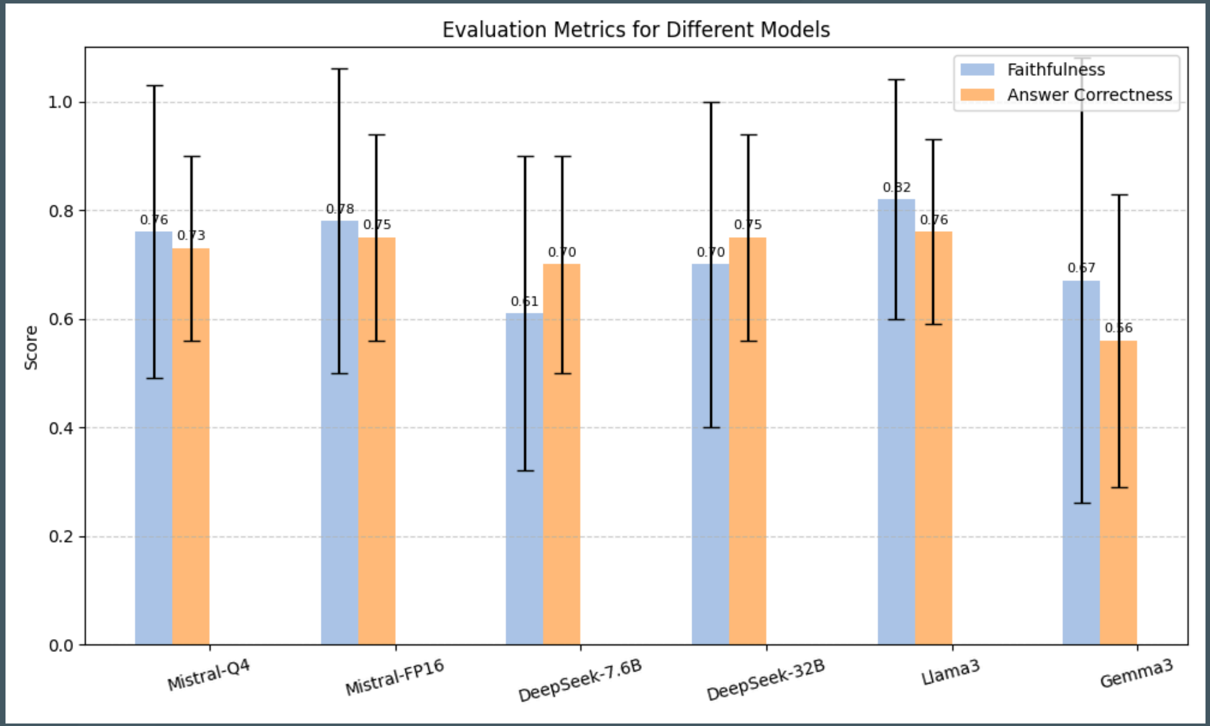
- **Open-Source LLMs tested**:
  - Mistral-Q4 (7.0B, 32K context)
  - Mistral-FP16 (7.0B, 32K context)
  - DeepSeek-r1-7.6B (7.6B, 128K context)
  - DeepSeek-r1-32B (32.0B, 128K context)
  - Llama3 (8.0B, 8K context)
  - Gemma3 (7.0B, 8K context)

Generation Performance Metrics

- **Faithfulness**: Checks if the answer is **factually consistent with the retrieved context**, helping identify hallucination.
- **Answer Relevance**: Measures how well the **answer addresses the input question**.
- **Semantic Similarity**: Assesses **content overlap** between the generated and reference answers.
- **Answer Correctness**: Overall judgment of **accuracy**, considering factual content and semantic alignment.

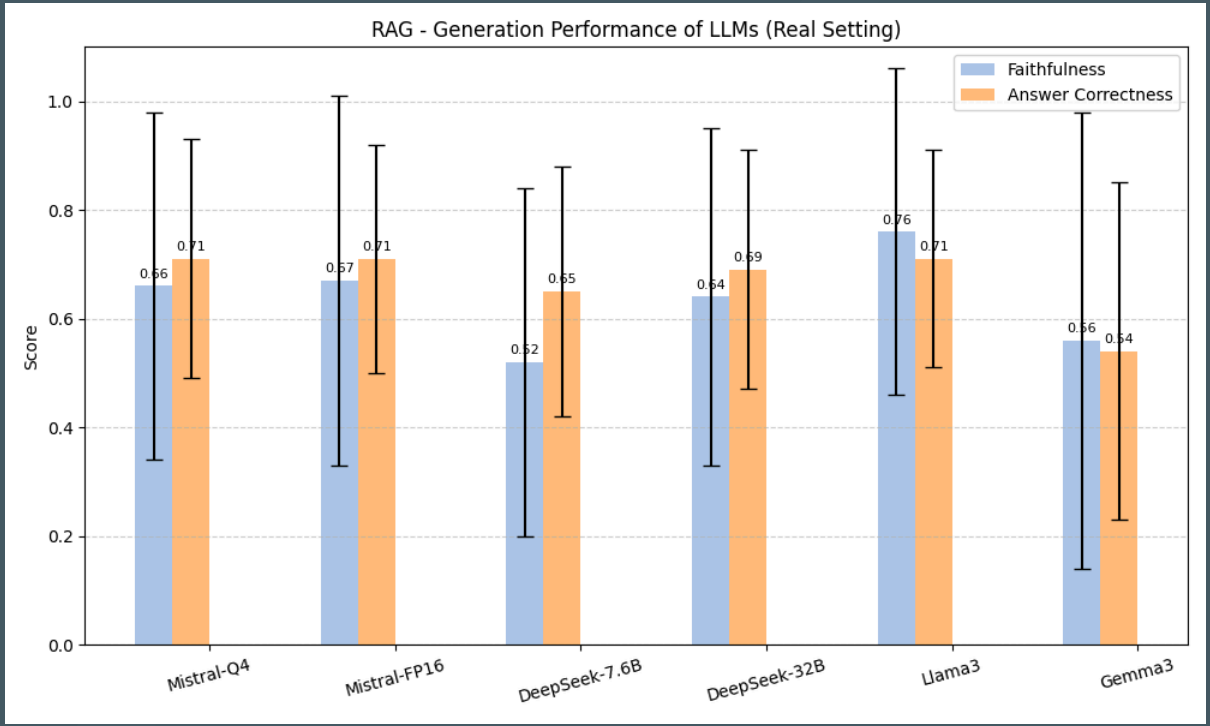
Generation Performance - Ideal Setting

LLM prompted with the original abstract (100% relevant)



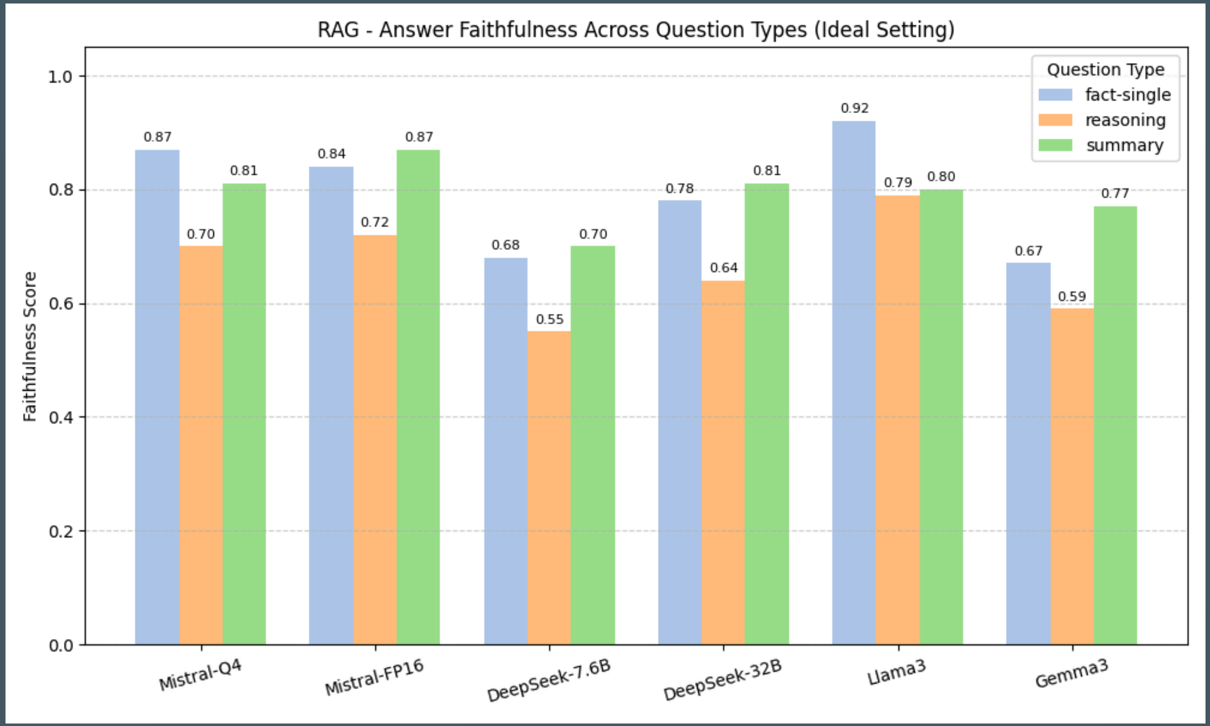
Generation Performance - Real Setting

LLM prompted with TOP 3 abstracts (reranked)



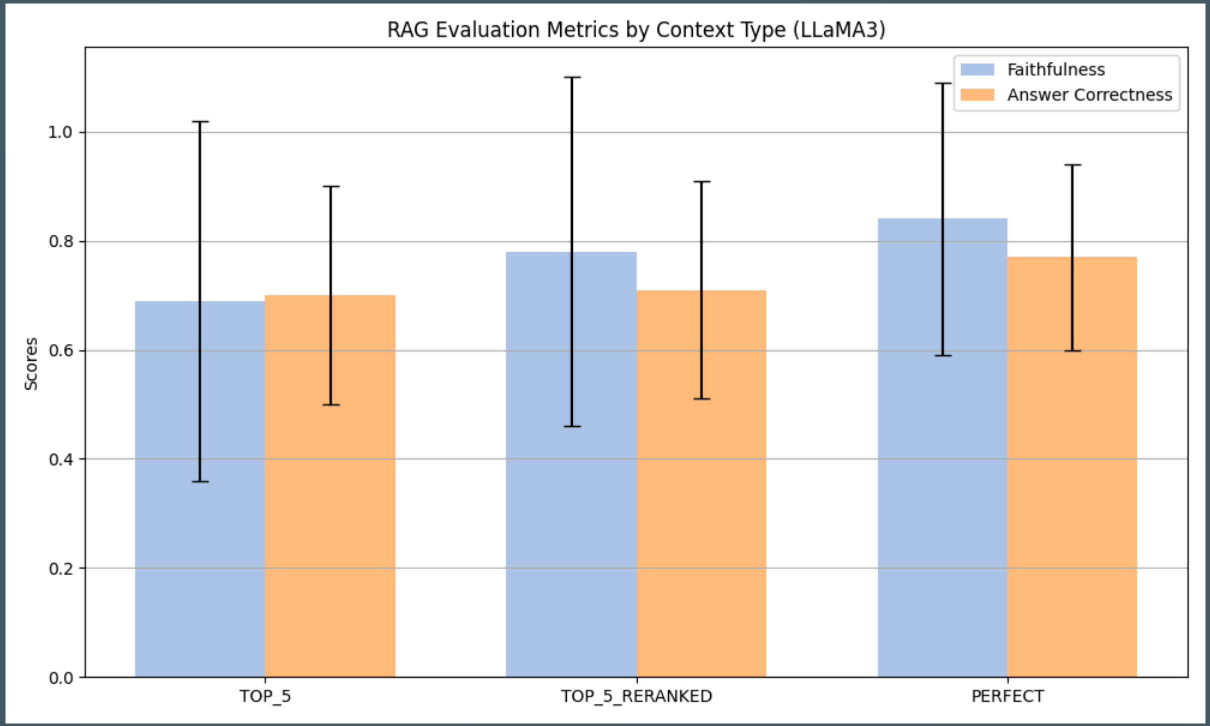
Performance Across Question Types

LLM prompted with the original abstract (PERFECT)



Reranking effect

Llama3 prompted with: PERFECT, TOP 5, TOP 5 reranked



Conclusions

- **Lexical and semantic search methods have distinct strengths**: Consider Hybrid search
- **Embedding choice is critical**: Mid-sized embeddings often outperform larger ones in retrieval tasks.
- **Llama3 and Mistral models offer balanced performance** in faithfulness and correctness.
- **Reranking** improves **generation** quality.
- **Reasoning-heavy questions remain a challenge** for RAG systems.