VILNIUS TECH
Faculty of Fundamental Sciences

DAMSS
DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

Evaluating Information Retrieval Effectiveness of RAG Architectures Across Diverse LLM Scales Using HotpotQA and MuSiQue

Natalija Matyjaškoit · Pavel Stefanovič · Simona Ramanauskaitė

RESEARCH OBJECTIVE

To improve multi-hop question answering by keeping the best documents and finding missing information.

Different search methods are compared to test this approach.

BACKGROUND & RESEARCH

Multi-hop questions need information from several documents. One search may not find all the needed information. A second search can help, but it may replace useful documents from the first search. The proposed Top-3 method keeps the three best documents and searches for missing information to improve LLM answers.

DATASETS & EXPERIMENTAL SETUP

HotpotQA
500 questions · 4,951 passages
402 bridge · 98 comparison

MuSiQue
500 questions · 7,011 passages
266 two-hop · 150 three-hop · 84 four-hop

Six search methods and six LLMs were tested. Qwen was used to convert text into vectors, FAISS to search for similar passages, and a reranker to select the best results. Each LLM received five passages to generate an answer.

PROPOSED TOP-3 MULTI-HOP STRATEGY

First, the system retrieves the most relevant documents. Then, the top three documents are kept fixed. Next, information from the best document is used to search for missing evidence. Finally, the remaining documents are reranked, and the final Top-5 context is sent to the LLM.

01 INITIAL SEARCH
Dense retrieval + reranking

→

02 PROTECT TOP-3
Keep the three best passages

→

03 SECOND-HOP SEARCH
Use Top-1 evidence to find missing facts

→

04 RERANK OPEN
Assess new and remaining passages

→

05 TOP-5 TO THE LLM
3 preserved + 2 replaceable slots

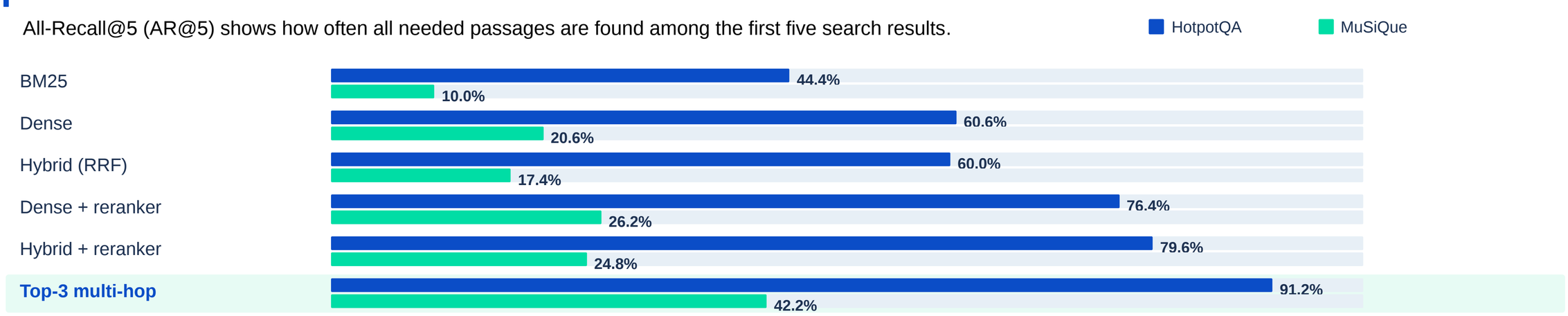
Final context: 1 2 3 4 5

Positions 1–3 stay fixed; positions 4–5 can be replaced by new evidence.

The additional search uses the same indexed corpus. No live web search is performed.

DOCUMENT RETRIEVAL

All-Recall@5 (AR@5) shows how often all needed passages are found among the first five search results.

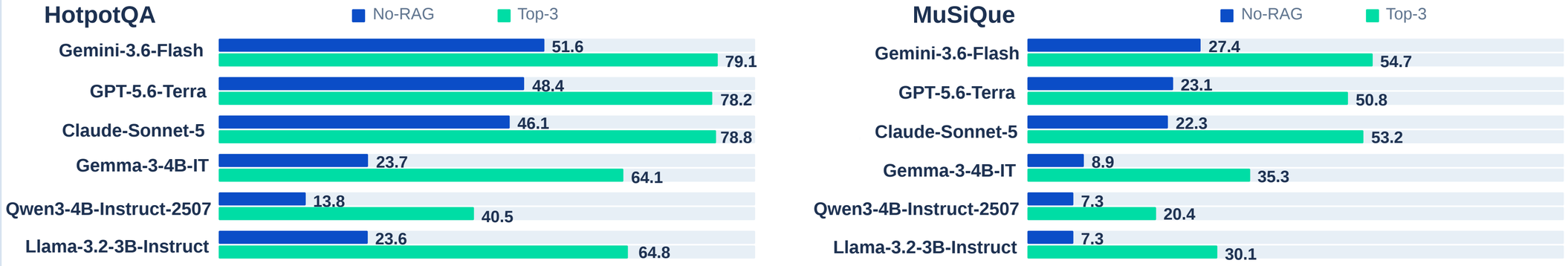


Top-3 vs Dense + reranker: +14.8 percentage points (HotpotQA) | +16.0 percentage points (MuSiQue).

These results show document search quality, not LLM answer quality..

LLM ANSWER QUALITY

The same questions were given to six language models in two ways: without extra documents (No-RAG) and with five documents found by Top-3. Token F1 measures how well the answers match the correct answers. A higher score is better.



All six models achieved better results with Top-3 on both datasets. This shows that additional information helps improve LLM answers.

MAIN FINDINGS & LIMITATIONS

01 IMPROVED COVERAGE
Top-3 achieved the best results among six search methods: 91.2% on HotpotQA and 42.2% on MuSiQue. It found more of the information needed to answer the questions.

02 BETTER MODEL ANSWERS
All six language models achieved higher EM, Token F1, and BERTScore F1 with Top-3 than without RAG. Results were lower on MuSiQue.

03 LIMITATIONS
The search used selected document collections, not the full Wikipedia. The method needs more testing with larger document collections.

Test larger document collections, measure search time and costs, and study how many documents should be kept..