

SOVEREIGN: Evaluating Multi-Hop Reasoning in RAG Systems: A Comparison of LLM-Based Retrieval

SOVEREIGN Research Kernel

Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Retrieval-augmented generation (RAG) enhances large language models (LLMs) with external knowledge to answer questions more accurately. However, research on evaluating RAG systems—particularly the retriever component—remains limited, as most existing work focuses on single-context retrieval rather than multi-hop queries, where individual contexts may appear irrelevant in isolation but are essential when combined. In this research, we use the HotPotQA, MuSiQue, and SQuAD datasets to simulate a RAG system and compare three LLM-as-judge evaluation strategies, including our proposed Context-Awar

1 Introduction

Analysis of: Evaluating Multi-Hop Reasoning in RAG Systems: A Comparison of LLM-Based Retriever Evaluation Strategies. Research goal: How does the inference throughput of RAG retrievers degrade under adversarial perturbations in multi-hop vs. single-hop queries when using dense vs. sparse retrieval models on BEIR?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score < 6.5.

3 Results

11 papers retrieved. 7 claims extracted, 7 verified. Tribunal: 7.8/10 → APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
CARE consistently outperforms existing methods for evaluating multi-hop reasoning in RAG systems.	✓	0.37
Performance gains of CARE are most pronounced in models with larger parameter counts and longer context windows.	✓	0.24
Single-hop queries show minimal sensitivity to context-aware evaluation.	✓	0.30
The complete data of the experiments is provided at https://github.com/lorenzbrehme/CARE .	✓	0.17
Most existing work on evaluating RAG systems focuses on single-context retrieval rather than multi-hop queries.	✓	0.33
The HotPotQA, MuSiQue, and SQuAD datasets were used to simulate a RAG system.	✓	0.16
Three LLM-as-judge evaluation strategies were compared, including CARE.	✓	0.17

References

- <http://arxiv.org/abs/1712.07113v2>
- <http://arxiv.org/abs/2604.18234v1>
- <http://arxiv.org/abs/2110.06918v3>