SOVEREIGN: How does instruction-tuned retrieval performance on multi-hop queries from MuSiQue compare to single-context a

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 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 instruction-tuned retrieval performance on multi-hop queries from MuSiQue compare to single-context adversarial training when evaluated on out-of-domain BEIR subsets (SciFact, TREC-COVID) in terms of MRR@10?.

2 Methodology

Multi-query arXiv search (1 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

3 papers retrieved. 9 claims extracted, 3 verified. Tribunal: 6.2/10 → REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

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
The CARE method consistently outperforms existing methods for evaluating multi-hop reasoning in RAG systems.	✓	0.35
Experiments were conducted using HotPotQA, MuSiQue, and SQuAD datasets.	×	0.11
The indirect evaluation approach is derived from the eRAG method.	×	0.02
The direct evaluation method is based on the ARES framework.	×	0.05
On HotPotQA dataset, CARE achieved 0.827±0.02 accuracy compared to 0.642±0.03 for indirect and 0.720±0.03 for direct met	×	0.03
On MuSiQue dataset, CARE achieved 0.755±0.02 accuracy compared to 0.631±0.03 for indirect and 0.702±0.03 for direct meth	×	0.03
Performance gains of CARE are most pronounced in models with larger parameter counts and longer context windows.	✓	0.23
Single-hop queries show minimal sensitivity to context-aware evaluation.	✓	0.30
CARE consistently outperformed other approaches across all models except for the LLaMa 3.1-8b model.	×	0.05

References

- http://arxiv.org/abs/2604.18234v1
- http://arxiv.org/abs/2404.14464v1

- http://arxiv.org/abs/2005.04474v1