

# 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 retrieval latency scale with the number of hops in multi-hop RAG queries when comparing dense retriever (e.g., DPR) vs. sparse retriever (e.g., BM25) on HotPotQA and MuSiQue under adversarial context perturbations?.

## 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

7 papers retrieved. 10 claims extracted, 10 verified. Tribunal: 7.7/10 → 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
Retrieval-augmented generation (RAG) enhances large language models (LLMs) with external knowledge to answer questions m	✓	0.32
Research on evaluating RAG systems, particularly the retriever component, remains limited, as most existing work focuses	✓	0.43
In multi-hop queries, individual contexts may appear irrelevant in isolation but are essential when combined.	✓	0.27
The study uses the HotPotQA, MuSiQue, and SQuAD datasets to simulate a RAG system.	✓	0.17
The study compares three LLM-as-judge evaluation strategies, including the proposed Context-Aware Retriever Evaluation (	✓	0.32
Experiments with LLMs from OpenAI, Meta, and Google demonstrate that CARE consistently outperforms existing methods for	✓	0.42
The performance gains of CARE are most pronounced in models with larger parameter counts and longer context windows.	✓	0.25
Single-hop queries show minimal sensitivity to context-aware evaluation.	✓	0.29
The results highlight the critical role of context-aware evaluation in improving the reliability and accuracy of retriev	✓	0.41
The complete data of the experiments is provided at <a href="https://github.com/lorenzbrehme/CARE">https://github.com/lorenzbrehme/CARE</a> .	✓	0.20

## References

- <http://arxiv.org/abs/2604.09019v2>
- <http://arxiv.org/abs/2604.18234v1>
- <http://arxiv.org/abs/2108.06279v2>