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

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 F1 score of LLMs on multi-hop HotPotQA with adversarial distractor insertion compare between 128K-token context windows and iterative retrieval with reranking, when using LLM-as-a-judge evaluation instead of standard exact match?.

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

9 papers retrieved. 8 claims extracted, 8 verified. Tribunal: 7.3/10 → 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
Most existing work on evaluating RAG systems focuses on single-context retrieval rather than multi-hop queries.	✓	0.35
The research uses the HotPotQA, MuSiQue, and SQuAD datasets to simulate a RAG system.	✓	0.18
The research compares three LLM-as-judge evaluation strategies, including the proposed Context-Aware Retriever Evaluatio	✓	0.32
Experiments with LLMs from OpenAI, Meta, and Google demonstrate that CARE consistently outperforms existing methods for	✓	0.43
The 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.	✓	0.31
The complete data of the experiments is provided at https://github.com/lorenzbrehme/CARE.	✓	0.19

References

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

- http://arxiv.org/abs/2504.11972v2