SOVEREIGN: How does the number of reasoning hops in multi-hop QA benchmarks (2-hop vs 3-hop in HotPotQA) affect the relat

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Large Language Models (LLMs) showcase impressive capabilities but encounter challenges like hallucination, outdated knowledge, and non-transparent, untraceable reasoning processes. Retrieval-Augmented Generation (RAG) has emerged as a promising solution by incorporating knowledge from external databases. This enhances the accuracy and credibility of the generation, particularly for knowledge-intensive tasks, and allows for continuous knowledge updates and integration of domain-specific information. RAG synergistically merges LLMs' intrinsic knowledge with the vast, dynamic repositories of exte

1 Introduction

Analysis of: Retrieval-Augmented Generation for Large Language Models: A Survey. Research goal: How does the number of reasoning hops in multi-hop QA benchmarks (2-hop vs 3-hop in HotPotQA) affect the relative MRR@10 improvement from adversarial training on retriever robustness in domain-adaptive RAG systems?.

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: 7.5/10 → 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
Large Language Models (LLMs) encounter challenges like hallucination, outdated knowledge, and non-transparent, untraceab	✓	0.34
Retrieval-Augmented Generation (RAG) has emerged as a promising solution by incorporating knowledge from external databa	✓	0.37
RAG enhances the accuracy and credibility of the generation, particularly for knowledge-intensive tasks.	✓	0.27
RAG allows for continuous knowledge updates and integration of domain-specific information.	✓	0.26
RAG synergistically merges LLMs' intrinsic knowledge with the vast, dynamic repositories of external databases.	✓	0.34
This comprehensive review paper offers a detailed examination of the progression of RAG paradigms, encompassing the Naiv	✓	0.36
The paper meticulously scrutinizes the tripartite foundation of RAG frameworks, which includes the retrieval, the genera	✓	0.32
The paper highlights the state-of-the-art technologies embedded in each of these critical components.	✓	0.25
This paper introduces up-to-date evaluation framework and benchmark.	✓	0.22
This article delineates the challenges currently faced and points out prospective avenues for research and development.	✓	0.28

References

- https://doi.org/10.18653/v1/p19-1613
- https://doi.org/10.48550/arxiv.2311.05232
- https://doi.org/10.48550/arxiv.2312.10997