SOVEREIGN: How do different retrieval evaluation strategies (e.g., recall-based vs relevance-based) affect the downstream

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Large Language Models (LLMs) showcase impressive capabilities but encounter challenges like hallucination, outdated knowledge, and non-transparent, untraceable reasoning processes.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 do different retrieval evaluation strategies (e.g., recall-based vs relevance-based) affect the downstream QA accuracy and robustness of LLM-based RAG systems on multi-hop queries from MuSiQue and HotPotQA?.

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. 6 claims extracted, 6 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
Large Language Models (LLMs) encounter challenges like hallucination, outdated knowledge, and non-transparent, untraceab	✓	0.35
Retrieval-Augmented Generation (RAG) enhances the accuracy and credibility of generation, particularly for knowledge-int	✓	0.28
RAG allows for continuous knowledge updates and integration of domain-specific information	✓	0.23
RAG synergistically merges LLMs' intrinsic knowledge with external databases	✓	0.28
RAG paradigms include Naive RAG, Advanced RAG, and Modular RAG	✓	0.18
RAG frameworks have three foundational components: retrieval, generation, and augmentation techniques	✓	0.17

References

- https://doi.org/10.18653/v1/2025.findings-acl.253
- https://doi.org/10.3390/bdcc9120320
- https://doi.org/10.48550/arxiv.2312.10997