SOVEREIGN: What is the trade-off between inference throughput and multi-hop reasoning accuracy in LLM-based RAG systems w

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Retrieval-augmented generation has raise extensive attention as it is promising to address the limitations of large language models including outdated knowledge and hallucinations. However, retrievers struggle to capture relevance, especially for queries with complex information needs. Recent work has proposed to improve relevance modeling by having large language models actively involved in retrieval, i.e., to guide retrieval with generation. In this paper, we show that strong performance can be achieved by a method we call Iter-RetGen, which synergizes retrieval and generation in an iterativ

1 Introduction

Analysis of: Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. Research goal: What is the trade-off between inference throughput and multi-hop reasoning accuracy in LLM-based RAG systems when using compressed or distilled retriever models on HotpotQA and MuSiQue?.

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. 2 claims extracted, 2 verified. Tribunal: 8.0/10 → 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Retrieval-augmented generation can address the limitations of large language models including outdated knowledge and hal	✓	0.27
Iter-RetGen method can improve relevance modeling by having large language models actively involved in retrieval to guid	✓	0.35

References

- https://doi.org/10.18653/v1/2023.findings-emnlp.620
- https://doi.org/10.18653/v1/2023.acl-long.557
- https://doi.org/10.48550/arxiv.2312.10997