SOVEREIGN: How does dynamic iterative retrieval with varying passage counts per hop affect the efficiency-accuracy trade-

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Multi-Hop Question Answering (MHQA) tasks permeate real-world applications, posing challenges in orchestrating multi-step reasoning across diverse knowledge domains. While existing approaches have been improved with iterative retrieval, they still struggle to identify and organize dynamic knowledge. To address this, we propose DualRAG, a synergistic dual-process framework that seamlessly integrates reasoning and retrieval. DualRAG operates through two tightly coupled processes: Reasoning-augmented Querying (RaQ) and progressive Knowledge Aggregation (pKA). They work in concert: as RaQ navigate

1 Introduction

Analysis of: DualRAG: A Dual-Process Approach to Integrate Reasoning and Retrieval for Multi-Hop Question Answering. Research goal: How does dynamic iterative retrieval with varying passage counts per hop affect the efficiency-accuracy trade-off on HotPotQA compared to single-retrieval methods?.

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

6 papers retrieved. 9 claims extracted, 2 verified. Tribunal: 4.0/10 → REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

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
DualRAG achieves substantial performance improvements across a variety of datasets.	×	0.03
Incorporating external knowledge consistently outperforms relying solely on model parameters in multi-hop tasks.	×	0.06
DualRAG-FT, with fine-tuning, approaches and in some cases surpasses the performance achieved with oracle knowledge acce	✓	0.16
DualRAG is a dual-process framework where RaQ guides reasoning and retrieval, while pKA organizes retrieved knowledge to	✓	0.21
By identifying key entities, RaQ dynamically generates targeted queries, while pKA structures and integrates relevant in	×	0.12
A fine-tuned version of DualRAG enhances proficiency of LLMs in retrieval and generation, significantly reducing computa	×	0.03
Extensive experiments on multiple multi-hop question answering datasets validate the effectiveness and robustness of Dua	×	0.13
Most iterative RAG systems lack the ability to proactively identify emerging knowledge gaps.	×	0.03
Noise in retrieved documents, stemming from both the documents themselves and retrieval tools, is a common issue in iter	×	0.03

References

- https://arxiv.org/abs/2504.18243
- https://arxiv.org/abs/2404.14464
- http://arxiv.org/abs/2404.14464v1