SOVEREIGN: How does FAIR-RAG's iterative refinement process scale in terms of inference latency and token-level processin

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

While Retrieval-Augmented Generation (RAG) mitigates hallucination and knowledge staleness in Large Language Models (LLMs), existing frameworks often falter on complex, multi-hop queries that require synthesizing information from disparate sources. Current advanced RAG methods, employing iterative or adaptive strategies, lack a robust mechanism to systematically identify and fill evidence gaps, often propagating noise or failing to gather a comprehensive context. We introduce FAIR-RAG, a novel agentic framework that transforms the standard RAG pipeline into a dynamic, evidence-driven reasoning

1 Introduction

Analysis of: FAIR-RAG: Faithful Adaptive Iterative Refinement for Retrieval-Augmented Generation. Research goal: How does FAIR-RAG's iterative refinement process scale in terms of inference latency and token-level processing time when evaluated on HotpotQA compared to baseline RAG approaches across 7B, 13B, and 70B parameter models?.

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. 0 claims extracted, 0 verified. Tribunal: 1.8/10 → REJECT (revision_round=0). 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.

References

- http://arxiv.org/abs/2510.22344v1
- http://arxiv.org/abs/2310.06825v1
- http://arxiv.org/abs/2510.25621v1