SOVEREIGN: What is the impact of faithfulness constraints on LLM generation throughput measured in tokens per second acro

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Retrieval-augmented generation (RAG) enhances large language models (LLMs) for domain-specific question-answering (QA) tasks by leveraging external knowledge sources. However, traditional RAG systems primarily focus on relevance-based retrieval and often struggle with redundancy, especially when reasoning requires connecting information from multiple sources. This paper introduces Vendi-RAG, a framework based on an iterative process that jointly optimizes retrieval diversity and answer quality. This joint optimization leads to significantly higher accuracy for multi-hop QA tasks. Vendi-RAG lev

1 Introduction

Analysis of: Vendi-RAG: Adaptively Trading-Off Diversity And Quality Significantly Improves Retrieval Augmented Generation With LLMs. Research goal: What is the impact of faithfulness constraints on LLM generation throughput measured in tokens per second across different RAG architectures on 2WikiMultihopQA exact match accuracy?.

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. 4 claims extracted, 1 verified. Tribunal: 6.1/10 $\rightarrow$ 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
Vendi-RAG improves retrieval diversity for multi-hop QA tasks	✓	0.26
Higher values of the parameter s in the VSR process introduce greater diversity in document ranking	×	0.05
Vendi-RAG outperforms baseline methods on multi-hop QA datasets	×	0.15
Kendall's τ and Spearman's ρ decrease as the parameter s increases from 0.0 to 1.0, indicating increased retrieval diver	×	0.05

References

- http://arxiv.org/abs/2502.11228v2
- http://arxiv.org/abs/2510.25621v1
- http://arxiv.org/abs/2401.15391v1