SOVEREIGN: How does the diversity-weight parameter in Vendi-RAG affect retrieval throughput (queries/second) versus Exact

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: How does the diversity-weight parameter in Vendi-RAG affect retrieval throughput (queries/second) versus Exact Match score on the TriviaQA dataset using FLAN-T5-xl?.

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

5 papers retrieved. 6 claims extracted, 1 verified. Tribunal: 5.0/10 $\rightarrow$ 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
The sensitivity analysis was performed using 100 randomly sampled queries from the dataset	×	0.05
Setting s = 0.0 serves as a baseline representing a pure similarity search scenario	×	0.02
Kendall's τ measures the rank order similarity between two lists by evaluating concordant and discordant pairs	×	0.01
Spearman's Rank Correlation ρ assesses the monotonic relationship between two rankings	×	0.00
Vendi-RAG is evaluated on three benchmark datasets: MuSiQue, HotpotQA, and 2WikiMultiHopQA	✓	0.17
The Vendi Score (VS) explicitly quantifies semantic diversity in a set of documents	×	0.14

References

- http://arxiv.org/abs/1302.1134v1
- http://arxiv.org/abs/2402.03801v1
- http://arxiv.org/abs/2502.11228v2