SOVEREIGN: What is the impact of retrieval diversity optimization on inference efficiency and latency when scaling Vendi-

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 retrieval diversity optimization on inference efficiency and latency when scaling Vendi-RAG to handle 1000+ document corpora on the Natural Questions Open dataset?.

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

10 papers retrieved. 10 claims extracted, 2 verified. Tribunal: 2.8/10 → 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.

5 Extracted Claims

Claim	Verified	Confidence
Vendi-RAG achieves higher F1 scores than Adaptive-RAG on the 2WikiMultiHopQA, HotpotQA, and MuSiQue datasets.	✓	0.16
Vendi-RAG achieves higher Exact Match scores than Adaptive-RAG on the 2WikiMultiHopQA, HotpotQA, and MuSiQue datasets.	×	0.15
Vendi-RAG achieves higher Accuracy scores than Adaptive-RAG on the 2WikiMultiHopQA, HotpotQA, and MuSiQue datasets.	✓	0.19
As the parameter s increases from 0.0 to 1.0, both Kendall's τ and Spearman's ρ decrease progressively.	×	0.03
Setting s = 0.0 serves as a baseline representing a pure similarity search scenario.	×	0.02
Higher s values promote retrieval diversity by prioritizing documents that may be less similar.	×	0.08
The Vendi Score quantifies semantic diversity in a set of documents.	×	0.11
The Vendi Score attains its maximum value n when all documents are orthogonal (fully diverse).	×	0.05
Similarity search often results in redundant documents with high similarity.	×	0.03
MMR attempts to balance relevance and novelty using pairwise comparisons, but still struggles to capture global semantic	×	0.06

References

- http://arxiv.org/abs/2504.05181v2
- http://arxiv.org/abs/2502.11228v2
- http://arxiv.org/abs/2504.01346v4