SOVEREIGN: Vendi-RAG: Adaptively Trading-Off Diversity And Quality Significantly Improves R

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 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 trade-off in inference efficiency (throughput in queries per second) when applying adversarial training on multi-hop queries to a dense retriever (e.g., Contriever) in a RAG system, evaluated on HotPotQA and SQuAD subsets?.

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

15 papers retrieved. 8 claims extracted, 7 verified. Tribunal: 7.5/10 → APPROVE (revision_round=0). Policy: AUTO_APPROVE.

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 accuracy increases of up to +4.2% on HotpotQA compared to Adaptive-RAG.	✓	0.18
Vendi-RAG achieves accuracy increases of up to +4.1% on 2WikiMultiHopQA compared to Adaptive-RAG.	✓	0.18
Vendi-RAG achieves accuracy increases of up to +1.3% on MuSiQue compared to Adaptive-RAG.	✓	0.20
Traditional RAG systems primarily focus on relevance-based retrieval and often struggle with redundancy.	✓	0.26
Vendi-RAG leverages the Vendi Score (VS) to promote semantic diversity in document retrieval.	✓	0.28
Vendi-RAG uses an LLM judge that evaluates candidate answers and outputs a score for balancing relevance and diversity.	✓	0.25
Experiments were conducted on HotpotQA, MuSiQue, and 2WikiMultiHopQA datasets.	×	0.12
The benefits of Vendi-RAG are more pronounced as the number of retrieved documents increases.	✓	0.27

References

- http://arxiv.org/abs/2502.11228v2
- http://arxiv.org/abs/2510.22344v1

- http://arxiv.org/abs/2401.15391v1