SOVEREIGN: What is the trade-off between retrieval diversity and answer accuracy when applying Vendi-RAG to a dense retri

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 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 R. Research goal: What is the trade-off between retrieval diversity and answer accuracy when applying Vendi-RAG to a dense retriever like Contriever on SQuAD subsets, measured via F1 score and redundancy metrics?.

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

8 papers retrieved. 2 claims extracted, 2 verified. Tribunal: 7.3/10 → REVISE (revision_round=1). Policy: SOFT_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, a framework based on an iterative process that jointly optimizes retrieval diversity and answer quality, lead	✓	0.48
Traditional RAG systems primarily focus on relevance-based retrieval and often struggle with redundancy, especially when	✓	0.39

References

- https://doi.org/10.5281/zenodo.20415647
- https://doi.org/10.5281/zenodo.20415648
- https://doi.org/10.48550/arxiv.2312.10997