# SOVEREIGN: Does the Tree of Reviews iterative retrieval method improve robustness to irrelevant context in multi-hop QA c

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 Retrieval Augmented Generation With LLMs. Research goal: Does the Tree of Reviews iterative retrieval method improve robustness to irrelevant context in multi-hop QA compared to single-step retrieval on the 2WikiMultihop dataset, measured by F1 score and precision?.

## 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

2 papers retrieved. 6 claims extracted, 5 verified. Tribunal: 8.2/10 → 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 improvements of up to +4.2% on HotpotQA, +4.1% on 2WikiMultiHopQA, and +1.3% on MuSiQue comp | ✓ | 0.23 |
| Vendi-RAG jointly optimizes retrieval diversity and answer quality through an iterative process. | ✓ | 0.23 |
| Vendi-RAG uses the Vendi Score (VS) to promote semantic diversity in document retrieval. | ✓ | 0.23 |
| Vendi-RAG uses an LLM judge to evaluate candidate answers and output a score for balancing relevance and diversity. | ✓ | 0.17 |
| Experiments were conducted on HotpotQA, MuSiQue, and 2WikiMultiHopQA datasets. | × | 0.11 |
| Traditional RAG systems primarily focus on relevance-based retrieval and often struggle with redundancy. | ✓ | 0.25 |

## References


- https://www.semanticscholar.org/paper/014ca63a6ebdb9d4f8dc24b23c32f4b0eb04909a
- https://www.semanticscholar.org/paper/c64681d1b8668b2c98331448691afe4ff46db1ec