Vendi-RAG Computational Overhead in Multi-Hop QA Benchmark Performance

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the computational overhead of Vendi-RAG's iterative joint optimization process compared to traditional RAG, measured in terms of latency and throughput on the MS MARCO passage ranking. 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. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Vendi-RAG: Adaptively Trading-Off Diversity And Quality Significantly Improves Retrieval Augmented Generation With LLMs. Research question: What is the computational overhead of Vendi-RAG's iterative joint optimization process compared to traditional RAG, measured in terms of latency and throughput on the MS MARCO passage ranking benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.3/10.

3 Results

13 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 4.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Experiments were conducted on three multi-hop QA benchmark datasets: MuSiQue, HotpotQA, and 2WikiMultiHopQA.	✓	0.17
The sensitivity analysis of the VSR process used 100 randomly sampled queries from the dataset.	×	0.06
Setting the parameter s = 0.0 serves as a baseline representing a pure similarity search scenario relying exclusively on	×	0.02
At s = 0.0, both Kendall's τ and Spearman's ρ are equal to 1.00.	×	0.03
At s = 1.0, Kendall's τ is 0.074 and Spearman's ρ is 0.078.	×	0.00
As the parameter s increases from 0.0 to 1.0, both Kendall's τ and Spearman's ρ decrease progressively.	×	0.03
On the 2WikiMultiHopQA dataset, Vendi-RAG-4o achieved an F1-score of 69.9.	×	0.10
On the 2WikiMultiHopQA dataset, Adaptive-RAG-4o achieved an F1-score of 63.4.	×	0.07
On the HotpotQA dataset, Vendi-RAG-4o achieved an Exact Match score of 56.5.	×	0.09
On the HotpotQA dataset, Adaptive-RAG-4o achieved an Exact Match score of 52.1.	×	0.06
On the MuSiQue dataset, Vendi-RAG-4o achieved an Accuracy of 63.4.	×	0.12
On the MuSiQue dataset, Adaptive-RAG-4o achieved an Accuracy of 59.3.	×	0.08
The Vendi Score (VS) explicitly quantifies semantic diversity in a set of documents.	×	0.13
The Vendi Score attains its maximum value n when all documents in the set are orthogonal.	×	0.06
Similarity search (SS) often results in redundant documents with high similarity.	×	0.03
Maximal Marginal Relevance (MMR) struggles to capture global semantic diversity.	×	0.06

References

- `http://arxiv.org/abs/2204.12852v1`

- http://arxiv.org/abs/2502.11228v2
- http://arxiv.org/abs/2510.25518v1