Vendi-RAG Retrieval Diversity Trade-offs in Multi-Hop QA Latency and Accuracy

Assignee Research

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

1 Introduction

This paper examines: Vendi-RAG: Adaptively Trading-Off Diversity And Quality Significantly Improves Retrieval Augmented Generation With LLMs. Research question: How do different retrieval diversity parameters in Vendi-RAG impact the latency and EM score trade-offs when using FLAN-T5-xl generator on the HotpotQA benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.0/10.

3 Results

11 papers retrieved. 8 claims extracted; 2 independently verified. Quality review score: 5.0/10.

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
Vendi-RAG was evaluated on three multi-hop QA benchmark datasets: MuSiQue, HotpotQA, and 2WikiMultiHopQA.	✓	0.21
The sensitivity analysis of the VSR process was conducted using 100 randomly sampled queries from the dataset.	×	0.05
The sensitivity analysis evaluated the retrieval pipeline across multiple s values ranging from 0.0 to 1.0.	×	0.03
Setting s = 0.0 serves as a baseline representing a pure similarity search scenario.	×	0.03
Kendall's τ and Spearman's ρ were used to quantify deviations from the baseline in the sensitivity analysis.	×	0.01
As s increases from 0.0 to 1.0, both Kendall's τ and Spearman's ρ decrease progressively.	×	0.03
Vendi-RAG uses a retrieval approach based on the Vendi Score (VS) to quantify semantic diversity in a set of documents.	✓	0.19
The Vendi Score (VSk(D)) reflects the effective number of unique documents in D, attaining its maximum value n when all	×	0.06

References

- http://arxiv.org/abs/1805.05737v1
- http://arxiv.org/abs/2502.11228v2
- http://arxiv.org/abs/2411.16965v2