Vendi-RAG Diversity-Weight Tuning for Latency-Accuracy Trade-offs in Multi-Hop QA

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 does the diversity-weight parameter in Vendi-RAG affect the trade-off between retrieval latency and EM score when evaluated on HotpotQA, and what is the optimal balance between efficiency and accuracy for FLAN-T5-xl on knowledge-intensive QA tasks?.

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 4.8/10.

3 Results

13 papers retrieved. 10 claims extracted; 2 independently verified. Quality review score: 4.8/10.

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.	×	0.04
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.	×	0.03
As s increases from 0.0 to 1.0, both Kendall's τ and Spearman's ρ decrease progressively.	×	0.03
Vendi-RAG variants and Adaptive-RAG variants were compared on three datasets using three evaluation metrics.	×	0.10
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.	×	0.06
The Vendi Score attains its maximum value n when all documents are orthogonal (fully diverse).	×	0.05

References

- http://arxiv.org/abs/2402.12317v2

- http://arxiv.org/abs/2502.11228v2
- http://arxiv.org/abs/2508.05197v2