Diversity-Weight Tuning in Vendi-RAG: Latency and EM Performance on HotpotQA

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: What is the impact of varying diversity-weight values on Vendi-RAG's retrieval latency and EM score when evaluated on the HotpotQA dataset with FLAN-T5-xl as the generator?.

2 Methodology

Systematic literature search across multiple databases yielded 5 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

3 Results

5 papers retrieved. 6 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
The Vendi Score explicitly quantifies semantic diversity in a set of documents.	×	0.10
Vendi-RAG uses a retrieval approach based on the Vendi Score to address limitations of similarity search and maximal mar	×	0.15
Setting s = 0.0 in the VSR process represents a pure similarity search scenario without any emphasis on diversity.	×	0.02
As s increases from 0.0 to 1.0 in the VSR process, both Kendall's τ and Spearman's ρ decrease progressively.	×	0.03
Lower values of Spearman's ρ indicate increased diversity through higher s values in the VSR process.	×	0.03
Higher s values in the VSR process promote retrieval diversity by prioritizing documents that may be less similar.	×	0.08

References

- http://arxiv.org/abs/2502.11228v2
- http://arxiv.org/abs/2504.05181v2
- http://arxiv.org/abs/2411.00744v2