Adaptive Diversity-Weight Tuning in Vendi-RAG for Throughput and Accuracy Trade-offs

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does adaptive diversity-weight tuning in Vendi-RAG affect throughput on the TriviaQA benchmark compared to fixed-weight retrieval for FLAN-T5-xxl, and what is the optimal efficiency-accuracy. 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. 8 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. 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: How does adaptive diversity-weight tuning in Vendi-RAG affect throughput on the TriviaQA benchmark compared to fixed-weight retrieval for FLAN-T5-xxl, and what is the optimal efficiency-accuracy trade-off when scaling from 1 to 10 retrieval rounds?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.2/10.

3 Results

14 papers retrieved. 8 claims extracted; 2 independently verified. Quality review score: 5.2/10.

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 uses a retrieval approach based on the Vendi Score (VS) to explicitly quantify semantic diversity in a set of	✓	0.20
The Vendi Score (VS) reaches its maximum value when all documents in the set are orthogonal (fully diverse).	×	0.08
In the sensitivity analysis of the VSR process, as the parameter s increases from 0.0 to 1.0, both Kendall's τ and Spear	×	0.03
Setting s = 0.0 in the VSR process represents a pure similarity search scenario without any emphasis on diversity.	×	0.03
Kendall's τ and Spearman's ρ are used as ranking comparison metrics to quantify deviations from the baseline in the sens	×	0.02
The sensitivity analysis was performed using 100 randomly sampled queries from the dataset.	×	0.04
Lower values of Spearman's Rank Correlation ρ indicate increased diversity through higher s values.	×	0.03
Experiments are conducted on three multi-hop QA benchmark datasets: MuSiQue, HotpotQA, and 2WikiMultiHopQA.	✓	0.17

References

- http://arxiv.org/abs/2503.16581v1
- http://arxiv.org/abs/2502.11228v2
- http://arxiv.org/abs/2605.02623v1