Vendi-RAG Inference Efficiency Across Model Sizes and Noisy Retrieval Contexts

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: How does the inference efficiency (measured in tokens per second) of Vendi-RAG with different model sizes (7B vs. 70B) vary when processing noisy retrieval contexts on TriviaQA and WebQuestions,. 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 the inference efficiency (measured in tokens per second) of Vendi-RAG with different model sizes (7B vs. 70B) vary when processing noisy retrieval contexts on TriviaQA and WebQuestions, while maintaining a fixed retrieval accuracy threshold?.

2 Methodology

Systematic literature search across multiple databases yielded 9 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

9 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 was evaluated on three multi-hop QA benchmark datasets: MuSiQue, HotpotQA, and 2WikiMultiHopQA.	✓	0.23
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.	×	0.02
Kendall's τ and Spearman's ρ were used to quantify deviations from the baseline in the sensitivity analysis.	×	0.02
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.	✓	0.21
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/2505.21439v1
- http://arxiv.org/abs/2310.06825v1
- http://arxiv.org/abs/2502.11228v2