Vendi-RAG Retrieval Rounds and Accuracy-Throughput Trade-offs on GSM8K with FLAN-T5-XXL

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of varying the number of retrieval rounds (1 to 10) in Vendi-RAG on the accuracy-throughput trade-off when applied to the GSM8K benchmark with FLAN-T5-xxl. 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. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. 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: What is the impact of varying the number of retrieval rounds (1 to 10) in Vendi-RAG on the accuracy-throughput trade-off when applied to the GSM8K benchmark with FLAN-T5-xxl?.

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 3.2/10.

3 Results

13 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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.

References

- http://arxiv.org/abs/2502.11228v2
- http://arxiv.org/abs/2312.17080v4
- http://arxiv.org/abs/2411.19443v1