Vendi-RAG Diversity-Weight Impact on FLAN-T5-XL Latency and Throughput in ANLI and HANS

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the effect of varying Vendi-RAG's diversity-weight on FLAN-T5-xl inference latency and token throughput when evaluated on ANLI and HANS datasets. LLM inference is still evaluated mainly as a model or software problem: accuracy, latency, throughput, and hardware utilization. This is incomplete. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Position: LLM Inference Should Be Evaluated as Energy-to-Token Production. Research question: What is the effect of varying Vendi-RAG's diversity-weight on FLAN-T5-xl inference latency and token throughput when evaluated on ANLI and HANS datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

4 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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/2605.11733v1
- http://arxiv.org/abs/1210.1185v3
- http://arxiv.org/abs/2403.02310v3