Vendi-RAG Diversity-Weight Tuning and Its Effects on FLAN-T5-xl Code Generation Accuracy

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the impact of varying the diversity-weight parameter in Vendi-RAG on the accuracy of FLAN-T5-xl for code generation tasks in the HumanEval benchmark compared to BM25 retrieval. 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. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. 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 diversity-weight parameter in Vendi-RAG on the accuracy of FLAN-T5-xl for code generation tasks in the HumanEval benchmark compared to BM25 retrieval?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

15 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
Experiments were conducted on three multi-hop QA benchmark datasets: MuSiQue, HotpotQA, and 2WikiMultiHopQA.	✓	0.17
The sensitivity analysis of the VSR process used 100 randomly sampled queries from the dataset.	×	0.03
Setting the parameter s = 0.0 represents a pure similarity search baseline relying exclusively on cosine similarity or d	×	0.05
In the sensitivity analysis, increasing the parameter s from 0.0 to 1.0 causes both Kendall's τ and Spearman's ρ to decr	×	0.02
At parameter s = 0.2, the Kendall's τ value is 0.797 and the Spearman's ρ value is 0.828.	×	0.01
At parameter s = 1.0, the Kendall's τ value is 0.074 and the Spearman's ρ value is 0.078.	×	0.01
On the 2WikiMultiHopQA dataset, Vendi-RAG-4o achieved an F1-score of 69.9.	×	0.09
On the 2WikiMultiHopQA dataset, Adaptive-RAG-4o achieved an F1-score of 63.4.	×	0.07
On the HotpotQA dataset, Vendi-RAG-4o achieved an Exact Match score of 56.5.	×	0.09
On the HotpotQA dataset, Adaptive-RAG-4o achieved an Exact Match score of 52.1.	×	0.06
On the MuSiQue dataset, Vendi-RAG-4o achieved an Accuracy score of 63.4.	×	0.12
On the MuSiQue dataset, Adaptive-RAG-4o achieved an Accuracy score of 62.8.	×	0.10
The Vendi Score (VS) explicitly quantifies semantic diversity in a set of documents.	×	0.11
The Vendi Score attains its maximum value n when all documents in the set are orthogonal.	×	0.06
Similarity search (SS) retrieval often results in redundant documents with high similarity.	×	0.05
Maximal Marginal Relevance (MMR) struggles to capture global semantic diversity despite attempting to balance relevance	×	0.07

References

- http://arxiv.org/abs/2306.08568v2
- http://arxiv.org/abs/2210.05512v1
- http://arxiv.org/abs/2502.11228v2