

Vendi-RAG Latency-Accuracy Trade-offs on TriviaQA Across Model Scales

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the trade-off between retrieval latency and answer accuracy in Vendi-RAG when evaluated on the TriviaQA benchmark with different model sizes. Accurate and contextually faithful responses are critical when applying large language models (LLMs) to sensitive and domain-specific tasks, such as answering queries related to quranic studies. General-purpose LLMs often struggle with hallucinations, where generated responses. 7 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: Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models. Research question: What is the trade-off between retrieval latency and answer accuracy in Vendi-RAG when evaluated on the TriviaQA benchmark with different model sizes?.

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 3.7/10.

3 Results

15 papers retrieved. 7 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.

5 Extracted Claims

Claim	Verified	Confidence
The system employs a Retrieval-Augmented Generation (RAG) architecture, combining retrieval-based and generative methods	×	0.09
The system executes semantic search and retrieval, response generation, and citations and contextualization tasks.	×	0.04
Context Relevance is evaluated using the precision@k metric, where k represents the number of top retrieved results	×	0.08
The dataset was chosen based on authenticity, descriptive richness, clarity and accessibility, and relevance.	×	0.04
The dataset underwent a thorough review to confirm its compliance with recognized Islamic scholarship and the absence of	×	0.02
The dataset must deliver comprehensive, contextually rich descriptions that can be effectively employed for semantic search	×	0.04
The content needed to be created in a structured and clear manner, facilitating both manual review and computational processing	×	0.05

References

- <http://arxiv.org/abs/2411.00907v3>
- <http://arxiv.org/abs/2502.11228v2>
- <http://arxiv.org/abs/2503.16581v1>