Scaling RAG Latency-Accuracy Trade-offs in 7B and 70B Models on Religious Datasets

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the trade-off between response latency and RAG accuracy for 7B vs. 70B models when scaling batch size from 1 to 8 on religious datasets, as measured by MT-Bench score degradation per query. 9 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. 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 response latency and RAG accuracy for 7B vs. 70B models when scaling batch size from 1 to 8 on religious datasets, as measured by MT-Bench score degradation per query?.

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 4.5/10.

3 Results

15 papers retrieved. 9 claims extracted; 2 independently verified. Quality review score: 4.5/10.

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 study investigates 13 open-source Large Language Models in the context of Quranic studies.	✓	0.23
The system employs a Retrieval-Augmented Generation (RAG) architecture combining retrieval-based and generative methodol	×	0.09
The system performs semantic similarity searches over a vectorized dataset obtained from Qur'anic surah descriptions.	×	0.02
Generated responses include references to original dataset entries, such as surah descriptions or specific virtues.	×	0.06
Human evaluators assessed response quality based on three dimensions: Context Relevance, Answer Faithfulness, and Answer	✓	0.17
Context Relevance is calculated using the precision@k metric, where k represents the number of top retrieved results.	×	0.06
The dataset selection criteria included authenticity, descriptive richness, clarity and accessibility, and relevance.	×	0.03
The dataset source was reviewed to confirm compliance with recognized Islamic scholarship and the absence of speculative	×	0.02
The evaluation platform logged and stored data, including scores and comments, for research purposes.	×	0.04

References

- http://arxiv.org/abs/2510.21459v1

- http://arxiv.org/abs/2503.16581v1
- http://arxiv.org/abs/2604.00715v1