Scalability Trade-offs in Retrieval-Augmented Generation for Domain-Specific Question Answering

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 6 peer-reviewed papers addressing the following research question: How does the scalability of retrieval-augmented generation (e.g., index size, query latency) affect the response accuracy and faithfulness of 7B vs. 70B models on domain-specific QA benchmarks like. 9 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: Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models. Research question: How does the scalability of retrieval-augmented generation (e.g., index size, query latency) affect the response accuracy and faithfulness of 7B vs. 70B models on domain-specific QA benchmarks like QuranQA and BibleQA?.

2 Methodology

Systematic literature search across multiple databases yielded 6 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

6 papers retrieved. 9 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.

5 Extracted Claims

Claim	Verified	Confidence
The system employs a Retrieval-Augmented Generation (RAG) architecture, combining retrieval-based and generative methodo	×	0.10
The system executes semantic search and retrieval, response generation, and citations and contextualization tasks.	×	0.04
Context Relevance evaluates how precisely the retrieved and generated responses align with the user query while avoiding	×	0.07
Precision@k metric is used to calculate the relevance score, where k represents the number of top retrieved results cons	×	0.03
The dataset was chosen according to specific criteria: Authenticity, Descriptive Richness, Clarity and Accessibility, an	×	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 sea	×	0.03
The content needed to be created in a structured and clear manner, facilitating both manual review and computational pro	×	0.05
The dataset was meticulously curated to facilitate the process of addressing user concerns regarding Islamic doctrines.	×	0.01

References

- http://arxiv.org/abs/2402.12317v2

- http://arxiv.org/abs/2503.16581v1
- http://arxiv.org/abs/2508.05197v2