Cross-Domain Robustness of Retrieval-Augmented 70B and 7B Models on Religious and Biomedical Benchmarks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the cross-domain robustness of retrieval-augmented 70B models compare to smaller (e.g., 7B) retrieval-augmented models when evaluated on religious (QuranQA) and non-religious (BioASQ). 9 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. 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 cross-domain robustness of retrieval-augmented 70B models compare to smaller (e.g., 7B) retrieval-augmented models when evaluated on religious (QuranQA) and non-religious (BioASQ) benchmarks, as measured by F1-score and hallucination detection metrics?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.2/10.

3 Results

12 papers retrieved. 9 claims extracted; 2 independently verified. Quality review score: 5.2/10.

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 study investigates 13 open-source Large Language Models in the context of Quranic studies.	✓	0.25
The system employs a Retrieval-Augmented Generation (RAG) architecture combining retrieval-based and generative methodol	×	0.07
Semantic search is performed over a vectorized dataset obtained from Qur'anic surah descriptions.	×	0.03
Generated responses include references to original dataset entries, such as surah descriptions or specific virtues.	×	0.05
Human evaluators assessed response quality based on three dimensions: Context Relevance, Answer Faithfulness, and Answer	✓	0.16
Context Relevance is calculated using the precision@k metric, where k represents the number of top retrieved results.	×	0.05
The dataset selection criteria included Authenticity, Descriptive Richness, Clarity and Accessibility, and Relevance.	×	0.03
The dataset source underwent a review to confirm compliance with recognized Islamic scholarship and the absence of specu	×	0.02
The evaluation platform logged and stored data, including scores and comments, for research purposes.	×	0.04

References

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

- http://arxiv.org/abs/2106.16020v1
- http://arxiv.org/abs/2503.08890v4