Can retrieval-augmented 3B models achieve comparable robustness to distractor contexts in multi-hop QA tasks a

Assignee Research

June 10, 2026

Abstract

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 deviate from authoritative sources, raising concerns about their reliability in religious contexts. This challenge highlights the need for systems that can integrate domain-specific knowledge while maintaining response accuracy, relevance, and faithfulness. In this study, we investigate 13 open-source LLMs categorized in

1 Introduction

This paper examines: Investigating Retrieval-Augmented Generation in Quranic Studies: A Study of 13 Open-Source Large Language Models. Research question: Can retrieval-augmented 3B models achieve comparable robustness to distractor contexts in multi-hop QA tasks as larger models while maintaining inference efficiency metrics like latency and throughput?.

2 Methodology

Systematic literature search across multiple databases yielded 11 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

11 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 methodo	×	0.08
The system executes semantic search and retrieval, response generation, and citations and contextualization tasks.	×	0.06
Context Relevance is evaluated using the precision@k metric, where k represents the number of top retrieved results cons	×	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 sea	×	0.03
The content needed to be created in a structured and clear manner, facilitating both manual review and computational pro	×	0.05

References

- http://arxiv.org/abs/2503.16581v1
- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2510.22344v1