

Retrieval-Augmented Generation Effects on AVPR in Legal QA Benchmarks for 7B and 70B Models

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of retrieval-augmented generation on the AVPR metric for legal domain QA benchmarks when comparing 7B and 70B models in a zero-shot setting. 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 impact of retrieval-augmented generation on the AVPR metric for legal domain QA benchmarks when comparing 7B and 70B models in a zero-shot setting?.

2 Methodology

Systematic literature search across multiple databases yielded 13 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

13 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.	✓	0.24
The system employs a Retrieval-Augmented Generation (RAG) architecture combining retrieval-based and generative methods.	×	0.13
The system performs semantic similarity search 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, to a	×	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.	×	0.05
The dataset selection criteria included Authenticity, Descriptive Richness, Clarity and Accessibility, and Relevance.	×	0.03
The dataset source underwent a thorough review to confirm compliance with recognized Islamic scholarship and the absence	×	0.02
The evaluation platform logged and stored data, including scores and comments, for research purposes after responses were	×	0.04

References

- <http://arxiv.org/abs/2308.12060v3>

- <http://arxiv.org/abs/2505.03970v1>
- <http://arxiv.org/abs/2503.16581v1>