Retrieval-Augmented Generation Performance and Latency Trade-offs in 7B vs 70B Models on Religious Datasets

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the retrieval-augmented generation (RAG) performance of a 7B model compare to a 70B model on specialized religious datasets when tested on benchmarks like HELM or MT-Bench, and what is the. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 1.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Six Llamas: Comparative Religious Ethics Through LoRA-Adapted Language Models. Research question: How does the retrieval-augmented generation (RAG) performance of a 7B model compare to a 70B model on specialized religious datasets when tested on benchmarks like HELM or MT-Bench, and what is the trade-off in response latency under 500ms per query in a batch size of 1?.

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 1.5/10.

3 Results

12 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 1.5/10.

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.

References

- http://arxiv.org/abs/2409.03708v2
- http://arxiv.org/abs/2506.06962v3
- http://arxiv.org/abs/2604.18404v1