Retrieval-Augmented Generation and Fine-Tuned 70B Models on Religious and Non-Religious QA Benchmarks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the integration of retrieval-augmented generation with fine-tuned 70B models compare to zero-shot performance when evaluated on religious (QuranQA) and non-religious (BioASQ) benchmarks, as. 0 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: Can Language Models Critique Themselves? Investigating Self-Feedback for Retrieval Augmented Generation at BioASQ 2025. Research question: How does the integration of retrieval-augmented generation with fine-tuned 70B models compare to zero-shot performance when evaluated on religious (QuranQA) and non-religious (BioASQ) benchmarks, as measured by ROUGE-L and BERTScore metrics?.

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 3.2/10.

3 Results

13 papers retrieved. 0 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.

References

- http://arxiv.org/abs/2307.12114v3
- http://arxiv.org/abs/2502.11228v2
- http://arxiv.org/abs/2508.05366v1