Impact Of Domain-Specific Fine-Tuning On The Alignment Of Retrieval-Augmented 7B Models Compared To 70B Models In

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the effect of domain-specific fine-tuning on the alignment of retrieval-augmented 7B models compared to 70B models in religious QA tasks, as measured by the FAITH metric. 19 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Aggregated Knowledge Model: Enhancing Domain-Specific QA with Fine-Tuned and Retrieval-Augmented Generation Models. Research question: What is the effect of domain-specific fine-tuning on the alignment of retrieval-augmented 7B models compared to 70B models in religious QA tasks, as measured by the FAITH metric?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

14 papers retrieved. 19 claims extracted; 2 independently verified. Quality review score: 4.2/10.

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
Approximately 2800 (context, question, answer) tuples were used, with 80% allocated for model fine-tuning and 20% for va	×	0.07
Approximately 560 ScienceIT domain knowledge questions were processed by the first seven models (two fine-tuned models a	×	0.12
The eighth model, AKM, aggregated the responses from these seven models to generate its answer.	×	0.13
This entire process was repeated 100 times, resulting in a total of 56000 samples for evaluation.	×	0.03
BLEU Scores were used to evaluate n-gram accuracy.	×	0.03
ROUGE Scores were used to assess recall, precision, and F1 metrics.	×	0.03
STS (Semantic Textual Similarity) was used to examine the semantic similarity between model-generated and reference answ	×	0.03
TF-IDF and Cosine Similarity were used to vectorize the answers and compute the cosine similarity between each pair of a	×	0.03
Embeddings and Mean Embedding (using BERT) were utilized to represent each answer and compute the mean embedding of all	×	0.04
Clustering (using K-means) was the most effective method for selecting the most representative answer.	×	0.07
BLEU and ROUGE scores indicated specific strengths in text alignment and recall capabilities.	×	0.06
STS scores highlighted semantic similarities effectively handled by the models.	×	0.02
Models with Retrieval-Augmented Generation (RAG) features showed significant performance improvements.	✓	0.17
The Aggregated Knowledge Model (AKM) was introduced and evaluated in the study.	✓	0.16
The study presents a comparative analysis of seven models in the LBL ScienceIT domain.	×	0.14
The QA system enhances user experience by providing rapid and accurate responses.	×	0.04
The QA system offers a sustainable solution for information dissemination.	×	0.05
The initiative promotes accessible, accurate information for all researchers, ensuring inclusivity and comprehensive kno	×	0.03
The evolution of QA systems began with Stanford's BASEBALL in 1961, a rule-based linguistic model interfaced with a data	×	0.05

References

- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2503.16581v1
- http://arxiv.org/abs/2410.18344v1