Multimodal vs. Text-Only Models in Cross-Domain Factual Accuracy with RAG Integration

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How do multimodal models with domain-specific fine-tuning perform in terms of factual accuracy compared to text-only models when evaluated on cross-domain benchmarks like QA-Retrieval or TruthfulQA. 14 claims were extracted from source literature; 1 was 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: Aggregated Knowledge Model: Enhancing Domain-Specific QA with Fine-Tuned and Retrieval-Augmented Generation Models. Research question: How do multimodal models with domain-specific fine-tuning perform in terms of factual accuracy compared to text-only models when evaluated on cross-domain benchmarks like QA-Retrieval or TruthfulQA?.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 14 claims extracted; 1 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
Approximately 2800 (context, question, answer) tuples were used, with 80% allocated for model fine-tuning and 20% for va	×	0.08
Approximately 560 ScienceIT domain knowledge questions were processed by the first seven models (two fine-tuned models a	×	0.13
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.	×	0.04
BLEU Scores were used to evaluate n-gram accuracy, ROUGE Scores to assess recall, precision, and F1 metrics, and STS (Se	×	0.04
TF-IDF and Cosine Similarity did not yield the best performance for selecting the most representative answer.	×	0.02
Embeddings and Mean Embedding (using BERT) showed improvement over cosine similarity but still fell short in accurately	×	0.03
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, while STS scores highlight	×	0.04
Models with Retrieval-Augmented Generation (RAG) features showed significant performance improvements.	✓	0.18
The Aggregated Knowledge Model (AKM) outperformed other models in terms of BLEU, ROUGE, and STS scores.	×	0.12
The QA system enhances user experience by providing rapid and accurate responses, improving efficiency and focus on prim	×	0.03
The QA system offers a sustainable solution for information dissemination, evolving with the expanding needs of the Scie	×	0.06
The initiative promotes accessible, accurate information for all researchers, ensuring inclusivity and comprehensive kno	×	0.03

References

- http://arxiv.org/abs/2505.17058v1
- http://arxiv.org/abs/2508.05197v2
- http://arxiv.org/abs/2410.18344v1