Retrieval-Augmented vs. Non-Retrieval 7B Models on Medical Benchmarks: Accuracy and Latency Trade-offs

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the relative performance gain in accuracy and inference latency when comparing retrieval-augmented vs. non-retrieval-augmented 7B models on medical domain benchmarks like MedQA or MeSH,. 6 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Systematic Study of Retrieval Pipeline Design for Retrieval-Augmented Medical Question Answering. Research question: What is the relative performance gain in accuracy and inference latency when comparing retrieval-augmented vs. non-retrieval-augmented 7B models on medical domain benchmarks like MedQA or MeSH, measured via exact match and hallucination detection metrics?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

15 papers retrieved. 6 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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 MedQA USMLE benchmark contains 1273 clinical examination questions.	×	0.07
Performance was measured using exact-match accuracy, while computational efficiency was evaluated using total runtime an	×	0.04
Zero-shot performance was similar for the two models, with Gemma3 achieving slightly higher accuracy than LLaMA-Med42.	×	0.06
The full set of evaluated configurations is summarized in Appendix Table A1.	×	0.02
40 experimental configurations were evaluated.	×	0.04
Two instruction-tuned language models were evaluated: LLaMA3-Med42-8B and Gemma3.	×	0.15

References

- http://arxiv.org/abs/2604.07274v1
- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2506.00448v1