Retrieval Granularity Effects on Efficiency and Accuracy in Retrieval-Augmented Medical QA Models

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the retrieval granularity (e.g., sentences vs. paragraphs) affect the inference efficiency and accuracy trade-off in retrieval-augmented 7B models when benchmarked on medical QA datasets. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Systematic Study of Biomedical Retrieval Pipeline Trade-offs in Performance and Efficiency. Research question: How does the retrieval granularity (e.g., sentences vs. paragraphs) affect the inference efficiency and accuracy trade-off in retrieval-augmented 7B models when benchmarked on medical QA datasets like BioASQ or MedQA?.

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 3.5/10.

3 Results

12 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
Corpus choice is the dominant factor governing retrieval quality, outweighing chunking and indexing decisions.	×	0.07
The aggregated corpus consistently outperforms every individual dataset across all query types.	×	0.03
The aggregated corpus achieves a win rate above 63% against all single-corpus configurations.	×	0.07
MedRAG/textbooks dominate technical and exam-style queries.	×	0.06
Wikipedia is preferred for specialized conversational queries.	×	0.03
Chunking granularity has no universally optimal setting; its effect on retrieval quality is strongly conditioned on quer	×	0.07
When all queries are aggregated, retrieval quality is largely insensitive to chunk size.	×	0.07
Aggregate win rates remain close to chance (48.4%–51.6%) across granularities ranging from 64 to 2048 tokens.	×	0.01
High tie rates across chunk sizes indicate substantial semantic overlap between adjacent chunk sizes.	×	0.02
Exam-style and fact-heavy queries strongly favor small chunks, with performance peaking at 128 tokens.	×	0.04
Performance for exam-style and fact-heavy queries degrades monotonically as chunk size increases beyond 128 tokens.	×	0.05
Scientific queries benefit from larger chunks, with win rates increasing steadily up to 2048 tokens.	×	0.02
Conversational and keyword-based queries exhibit weaker and less consistent preferences across granularities, showing re	×	0.04

References

- http://arxiv.org/abs/2510.25621v1
- http://arxiv.org/abs/2604.20853v1

- http://arxiv.org/abs/2507.23334v2