Retrieval Granularity Effects on Latency-Accuracy Trade-offs in 7B Models for MedQA

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the choice of retrieval granularity (sentence vs. paragraph) impact the trade-off between inference latency and accuracy in 7B models on the MedQA benchmark compared to smaller models. 15 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: MUST-RAG: MUSical Text Question Answering with Retrieval Augmented Generation. Research question: How does the choice of retrieval granularity (sentence vs. paragraph) impact the trade-off between inference latency and accuracy in 7B models on the MedQA benchmark compared to smaller models?.

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 3.5/10.

3 Results

14 papers retrieved. 15 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
The evaluation used two datasets: ArtistMus (in-domain) and TrustMus (out-of-domain).	×	0.09
Performance on factual and contextual questions was separately measured on the ArtistMus dataset.	×	0.03
TrustMus evaluation was conducted across four categories: People (Ppl), Instrument & Technology (IT), Genre, Forms, and	×	0.01
All evaluations use a multiple-choice QA format.	×	0.02
Zero-shot baselines evaluated include GPT-4o, Llama 3.1 8B Instruct, MuLLaMA, and ChatMusician.	×	0.05
MuLLaMA is designed to handle audio-based question answering.	×	0.10
ChatMusician specializes in music understanding and generation with ABC notation.	×	0.05
Llama 3.1 8B Instruct was fine-tuned on 8K multiple-choice QA pairs generated from MusWikiDB.	×	0.08
RAG inference was implemented using Llama 3.1 8B Instruct and MusWikiDB as the retrieval database.	×	0.08
RAG fine-tuning used a dataset in the form of (context, question, answer), augmenting the original QA fine-tuning datase	×	0.11
Models were trained for one epoch using LoRA with 8-bit quantization and specific hyperparameters.	×	0.04
For the ArtistMus dataset, half of the artists were included in the training data (Seen), while the other half were excl	×	0.02
MusWikiDB was developed by collecting music-related content from Wikipedia across seven categories: artists, genres, ins	×	0.06
MusWikiDB contains 31K pages, 629.2K passages, 65.5M total tokens, and a vocabulary size of 786K.	×	0.02
Wikipedia Corpus contains 3.2M pages, 21M passages, 2.1B total tokens, and a vocabulary size of 21.5M.	×	0.02

References

- http://arxiv.org/abs/2603.21389v1
- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2503.16581v1