Scaling Performance of Retrieval-Augmented 3B Models with Domain-Specific Knowledge Base Size in Multi-Hop QA

Assignee Research

June 12, 2026

Abstract

Recent advancements in Large language models (LLMs) have demonstrated remarkable capabilities across diverse domains. While they exhibit strong zero-shot performance on various tasks, LLMs' effectiveness in music-related applications remains limited due to the relatively small proportion of music-specific knowledge in their training data. To address this limitation, we propose MusT-RAG, a comprehensive framework based on Retrieval Augmented Generation (RAG) to adapt general-purpose LLMs for text-only music question answering (MQA) tasks. RAG is a technique that provides external knowledge to L

1 Introduction

This paper examines: MUST-RAG: MUSical Text Question Answering with Retrieval Augmented Generation. Research question: How does the performance of retrieval-augmented 3B models scale with increasing domain-specific knowledge base size in multi-hop QA tasks, measured by both accuracy and inference latency trade-offs?.

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 7.4/10.

3 Results

15 papers retrieved. 22 claims extracted; 16 independently verified. Quality review score: 7.4/10.

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 study used two datasets for evaluation: ArtistMus (in-domain) and TrustMus (out-of-domain).	×	0.13
Performance on factual and contextual questions was separately measured on the ArtistMus dataset.	✓	0.19
The TrustMus dataset evaluation covers four categories: People (Ppl), Instrument & Technology (IT), Genre, Forms, and Th	✓	0.15
Each category in the TrustMus dataset comprises 100 questions.	×	0.06
All evaluations in the study use a multiple-choice QA format.	✓	0.17
A response is considered incorrect if it deviates from the expected format.	×	0.12
The zero-shot baselines evaluated include GPT-4o, Llama 3.1 8B Instruct, MuLLaMA, and ChatMusician.	✓	0.21
MuLLaMA is designed to handle audio-based question answering.	✓	0.17
ChatMusician specializes in music understanding and generation with ABC notation.	✓	0.18
The QA fine-tuning dataset consists of 8K multiple-choice QA pairs generated from MusWikiDB.	✓	0.20
RAG inference was implemented using Llama 3.1 8B Instruct as the base model and MusWikiDB as the retrieval database.	✓	0.19
RAG fine-tuning was performed by augmenting the original QA fine-tuning dataset with additional context in the form of (	✓	0.24
Models were trained for one epoch using LoRA with 8-bit quantization.	✓	0.24
The training hyperparameters include a batch size of 2, gradient accumulation steps of 4, and a learning rate of 3e-5.	✓	0.26
For the ArtistMus dataset, half of the artists were included in the training data (Seen) and half were excluded (Unseen)	✓	0.36
MusWikiDB was developed by collecting music-related content from Wikipedia across seven categories: artists, genres, ins	✓	0.25
Data collection for MusWikiDB used a page depth of 2.	×	0.08
Sections shorter than 60 tokens were removed from the MusWikiDB dataset.	×	0.15
MusWikiDB contains 31K pages compared to 3.2M pages in the general Wikipedia corpus.	×	0.15
MusWikiDB has a vocabulary size of 786K compared to 21.5M in the general Wikipedia corpus.	✓	0.20

References

- http://arxiv.org/abs/2402.12317v2
- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2510.25621v1