Scaling Multimodal Retrieval Index for MUST-RAG in Zero-Shot Music 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: What is the impact of scaling the size of the multimodal retrieval index on the performance of MUST-RAG for zero-shot QA in music-related domains, measured by response relevance and factual accuracy?.

2 Methodology

Systematic literature search across multiple databases yielded 10 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

10 papers retrieved. 27 claims extracted; 20 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	✓	0.16
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
GPT-4o was evaluated as an API-based zero-shot baseline.	✓	0.16
Llama 3.1 8B Instruct was evaluated as an open-source zero-shot baseline.	✓	0.17
MuLLaMA is a music-specific model designed to handle audio-based question answering.	✓	0.16
ChatMusician is a music-specific model that specializes in music understanding and generation with ABC notation.	✓	0.16
The QA fine-tuning process used Llama 3.1 8B Instruct trained on 8K multiple-choice QA pairs generated from MusWikiDB.	✓	0.23
The RAG Inference approach uses Llama 3.1 8B Instruct as the base model with MusWikiDB as the retrieval database.	✓	0.18
RAG fine-tuning was performed using a dataset augmented with additional context in the form of (context, question, answe	✓	0.18
Models were trained for one epoch using LoRA with 8-bit quantization.	✓	0.25
The training batch size was 2 with 4 gradient accumulation steps.	✓	0.24
The training learning rate was 3e-5 with a weight decay of 0.005.	✓	0.26
The training configuration used a warmup ratio of 0.1, a cosine scheduler, and the AdamW optimizer.	✓	0.21
The LoRA hyperparameters used were r=16, alpha=16, and dropout=0.1.	✓	0.18
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

References

- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2404.07220v2
- http://arxiv.org/abs/2404.14464v1