Llama-3-8B with MusT-RAG vs. Atlas and REALM on MusWikiDB Robustness Benchmarks

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the performance of Llama-3-8B with MusT-RAG compare to other retrieval-augmented frameworks like Atlas or REALM on the robustness of multi-track music QA benchmarks under varying levels of. 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. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. 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 performance of Llama-3-8B with MusT-RAG compare to other retrieval-augmented frameworks like Atlas or REALM on the robustness of multi-track music QA benchmarks under varying levels of question ambiguity and adversarial perturbations?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

11 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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).	×	0.11
Performance on factual and contextual questions was separately measured on the ArtistMus dataset.	×	0.04
TrustMus evaluation was conducted across four categories: People (Ppl), Instrument & Technology (IT), Genre, Forms, and	×	0.02
All evaluations use a multiple-choice QA format.	×	0.05
Zero-shot baselines evaluated include GPT-4o, Llama 3.1 8B Instruct, MuLLaMA, and ChatMusician.	×	0.03
MuLLaMA is designed to handle audio-based question answering.	×	0.09
ChatMusician specializes in music understanding and generation with ABC notation.	×	0.04
Llama 3.1 8B Instruct was fine-tuned on 8K multiple-choice QA pairs generated from MusWikiDB.	×	0.06
RAG inference uses Llama 3.1 8B Instruct as the base model and MusWikiDB as the retrieval database.	×	0.07
RAG fine-tuning was performed using a dataset in the form of (context, question, answer).	×	0.10
Models were trained for one epoch using LoRA with 8-bit quantization and specific hyperparameters.	×	0.03
For the ArtistMus dataset, half of the artists were included in the training data (Seen), while the other half were excl	×	0.04
MusWikiDB was developed to address the lack of a music-specific vector database for RAG in MQA.	✓	0.16
MusWikiDB contains music-related content from Wikipedia across seven categories: artists, genres, instruments, history,	×	0.05
MusWikiDB has 31K pages, 629.2K passages, 65.5M total tokens, and a vocabulary size of 786K.	×	0.01
Wikipedia Corpus has 3.2M pages, 21M passages, 2.1B total tokens, and a vocabulary size of 21.5M.	×	0.01

References

- http://arxiv.org/abs/2404.14464v1
- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2604.09721v1