Robustness of Llama-3-8B-128K Retrieval-Augmented Generation Across Music Question Types

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How robust is the retrieval-augmented generation of Llama-3-8B-128K across different music-related question types (fact-based, interpretive, comparative) on MuSiQue when evaluated using. Recent work on music question answering (Music-QA) has primarily focused on single-track understanding, where models answer questions about an individual audio clip using its tags, captions, or metadata. However, listeners often describe music in comparative terms, and existing. 10 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Jamendo-MT-QA: A Benchmark for Multi-Track Comparative Music Question Answering. Research question: How robust is the retrieval-augmented generation of Llama-3-8B-128K across different music-related question types (fact-based, interpretive, comparative) on MuSiQue when evaluated using human-annotated accuracy versus automated metric scores?.

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 6.3/10.

3 Results

10 papers retrieved. 10 claims extracted; 3 independently verified. Quality review score: 6.3/10.

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
Jamendo-MT-QA specifically targets comparative reasoning between music tracks, requiring models to integrate perceptual	×	0.15
Comparative question answering has been extensively studied in the NLP domain, particularly in multi-hop and relational	×	0.11
In Music-QA, text-only models can achieve strong results even without access to audio inputs, suggesting dataset biases	×	0.07
Existing analyses in the music and audio domain have primarily focused on single-track understanding or perceptual groun	✓	0.17
Jamendo-MT-QA draws inspiration from both multi-hop QA and recent diagnostic benchmark analyses by formulating comparati	✓	0.18
GPT-5 mini was used as the primary generator for dataset construction in Stage 3 of Jamendo-MT-QA.	×	0.11
For each track pair, the model produces exactly three comparative questions corresponding to yes/no, short-answer, and s	✓	0.19
All questions generated in Stage 3 are required to explicitly reference both tracks.	×	0.04
Several alternative LLMs were experimented with for Stage 3 generation, including Qwen3-32B, InternLM3-8B, Gemma 3 12B,	×	0.04
GPT-5 mini consistently produces step-by-step reasoning, which improves interpretability and makes the reasoning field m	×	0.02

References

- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2604.09721v1
- http://arxiv.org/abs/2105.14013v1