Retrieval-Augmentation Context Effects on Llama-3-8B-128K Accuracy in Jamendo-MT-QA

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the impact of varying retrieval-augmentation contexts (e.g., different music metadata sources, retrieval depths) on Llama-3-8B-128K's response accuracy for fact-based versus interpretive. 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. 11 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.0/10. 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: What is the impact of varying retrieval-augmentation contexts (e.g., different music metadata sources, retrieval depths) on Llama-3-8B-128K's response accuracy for fact-based versus interpretive questions in Jamendo-MT-QA, measured by both G-EVAL and GPT-4 scoring?.

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 6.0/10.

3 Results

15 papers retrieved. 11 claims extracted; 3 independently verified. Quality review score: 6.0/10.

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 is a benchmark for multi-track comparative music question answering.	✓	0.48
The MAE benchmark evaluates multi-audio processing capabilities across speech and general sound domains.	×	0.05
AIR-Bench introduces an instruction-following benchmark for audio-language models spanning speech, environmental sounds,	×	0.06
AudioBench provides a broad, task-diverse benchmark for audio-language models.	×	0.07
HotpotQA and WikiMultiHopQA are benchmarks that require models to aggregate evidence across multiple documents.	×	0.03
DROP emphasizes discrete and logical reasoning over textual contexts.	×	0.02
In Music-QA, text-only models can achieve strong results even without access to audio inputs.	×	0.07
Existing analyses in the music and audio domain have primarily focused on single-track understanding or perceptual groun	✓	0.15
Jamendo-MT-QA specifically targets comparative reasoning between music tracks.	✓	0.21
GPT-5 mini is used as the primary generator for dataset construction in Stage 3 of the methodology.	×	0.03
GPT-5 mini consistently produces step-by-step reasoning in the dataset construction process.	×	0.06

References

- http://arxiv.org/abs/2303.13375v2
- http://arxiv.org/abs/2404.14464v1
- http://arxiv.org/abs/2604.09721v1