Retrieval Augmentation Strategies for Robustness in Llama-3-8B Music Question Answering

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: To what extent do different retrieval augmentation strategies (e.g., multi-stage RAG, re-ranking) improve the robustness of Llama-3-8B on adversarial or ambiguous multi-track music QA benchmarks. 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. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. 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: To what extent do different retrieval augmentation strategies (e.g., multi-stage RAG, re-ranking) improve the robustness of Llama-3-8B on adversarial or ambiguous multi-track music QA benchmarks?.

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 3.8/10.

3 Results

11 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
ArtistMus and TrustMus were used as evaluation datasets.	×	0.04
TrustMus evaluation was conducted across four categories: People (Ppl), Instrument & Technology (IT), Genre, Forms, and	×	0.02
Each category in TrustMus comprises 100 questions.	×	0.03
All evaluations use a multiple-choice QA format.	×	0.05
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.	×	0.05
RAG fine-tuning was performed using a dataset in the form of (context, question, answer) by augmenting the original QA f	×	0.08
The models were trained for one epoch using LoRA with 8-bit quantization.	×	0.03
MusWikiDB contains 31K pages and 629.2K passages.	×	0.02
MusWikiDB has a vocabulary size of 786K and total of 65.5M tokens.	×	0.02
Wikipedia Corpus contains 3.2M pages and 21M passages.	×	0.02
Wikipedia Corpus has a vocabulary size of 21.5M and total of 2.1B tokens.	×	0.01

References

- http://arxiv.org/abs/2404.14464v1
- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2503.16581v1