Multimodal Capture Impact on VQA Accuracy in Expert Mind vs. Text-only RAG Baselines

Assignee Research

June 13, 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: How does the multimodal capture component in Expert Mind affect VQA accuracy on domain-specific datasets compared to text-only RAG baselines?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 9.2/10.

3 Results

16 papers retrieved. 15 claims extracted; 15 independently verified. Quality review score: 9.2/10.

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).	✓	0.18
Performance on factual and contextual questions was separately measured on the ArtistMus dataset.	✓	0.19
TrustMus evaluation was conducted across four categories: People (Ppl), Instrument & Technology (IT), Genre, Forms, and	✓	0.23
All evaluations use a multiple-choice QA format.	✓	0.21
Zero-shot baselines evaluated include GPT-4o, Llama 3.1 8B Instruct, MuLLaMA, and ChatMusician.	✓	0.20
MuLLaMA is designed to handle audio-based question answering.	✓	0.17
ChatMusician specializes in music understanding and generation with ABC notation.	✓	0.17
Llama 3.1 8B Instruct was fine-tuned on 8K multiple-choice QA pairs generated from MusWikiDB.	✓	0.24
RAG inference was implemented using Llama 3.1 8B Instruct and MusWikiDB as the retrieval database.	✓	0.18
RAG fine-tuning was performed using a dataset in the form of (context, question, answer).	✓	0.20
Models were trained for one epoch using LoRA with 8-bit quantization and specific hyperparameters.	✓	0.20
For the ArtistMus dataset, half of the artists were included in the training data (Seen), while the other half were excl	✓	0.36
MusWikiDB was developed by collecting music-related content from Wikipedia across seven categories: artists, genres, ins	✓	0.25
MusWikiDB contains 31K pages, 629.2K passages, 65.5M total tokens, and a vocabulary size of 786K.	✓	0.22
Wikipedia Corpus contains 3.2M pages, 21M passages, 2.1B total tokens, and a vocabulary size of 21.5M.	✓	0.22

References

- http://arxiv.org/abs/2004.05573v1
- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2408.07303v2