HIERARCHICAL RETRIEVAL-AUGMENTED GENERATION FOR MULTIMODAL MUSIC QUESTION–ANSWERING USING VECTOR-BASED SEMANTIC RETRIEVAL

Anna-Maria CHRISTODOULOU (annammc@uio.no) [1, 3], **Kyrre GLETTE** [2, 3], and **Alexander Refsum JENSENIUS** [1, 3]

[1] *Department of Musicology*, **University of Oslo**, Norway
[2] *Department of Informatics*, **University of Oslo**, Norway
[3] *RITMO Center for Interdisciplinary Studies in Rhythm, Time, and Motion*, **University of Oslo**, Norway

ABSTRACT

This paper explores how hierarchical Retrieval-Augmented Generation (RAG) can support natural language interaction with multimodal music datasets. Such datasets consist of heterogeneous data formats that are difficult to explore with conventional file-based or metadata-driven tools. We present a lightweight prototype that combines multimodal feature extraction with vector-based semantic retrieval and a locally deployed Llama model (via Jan) to enable semantic querying and question–answering across data sources. We evaluate the system using crossmodal retrieval and dataset summarization, demonstrating its ability to integrate information across the dataset and generate contextually appropriate responses. Our results show that the system reduces interaction costs by over 50% while maintaining high retrieval accuracy. Results also suggest that the system can synthesize evidence from different data types to answer complex questions and uncover relationships within datasets that are otherwise difficult to access manually.

1. INTRODUCTION

Navigating large multimodal data collections remains a significant challenge, particularly in domains such as music research where datasets combine heterogeneous formats [1], including audio–video recordings, symbolic representations (e.g., MIDI), structured and unstructured text, and/or physiological measurements collected during a music performance. These datasets are often distributed across inconsistent folder structures, lack standardized naming conventions, and are frequently accompanied by limited or incomplete documentation [2]. As a result, effective exploration typically requires substantial technical expertise and prior familiarity with the dataset [3].

Previous approaches to improving dataset accessibility have focused on metadata [4], visualizations [5], or both [6, 7]. In fields such as music information retrieval (MIR) and computational musicology, these methods have supported analysis and navigation by structuring information or providing specialized data representations [8, 9]. However, such approaches are typically constrained to individual data types or require users to adopt multiple tools and workflows. Therefore, exploring a single musical piece may require the simultaneous use of specialized software for audio analysis, notation editing, and numerical data processing, thereby limiting the potential to support dataset exploration across heterogeneous musical materials.

Recent developments in machine learning have introduced new paradigms for interacting with complex data. In particular, natural language interfaces have emerged as a means of accessing and reasoning about information, reducing the need for explicit knowledge of underlying data structures [10, 11]. The application of these approaches to

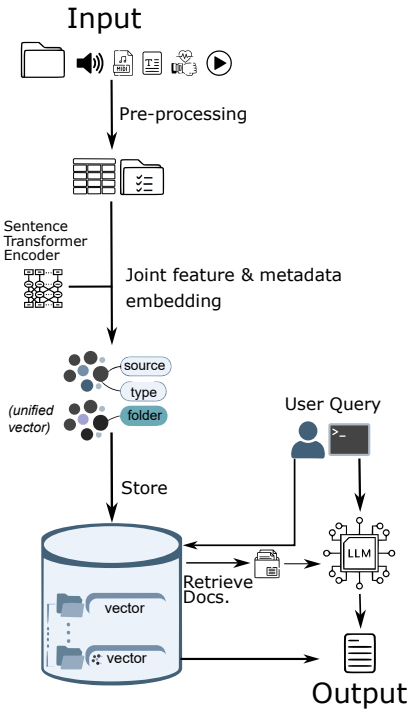


Figure 1. An overview of our developed system. Input data are pre-processed and mapped into a unified vector space using a Sentence-Transformer. The joint feature-metadata embeddings are stored in a vector database. In inference, a user query triggers the retrieval of relevant document segments, which are then synthesized by a Large Language Model (LLM) to produce a response.

Copyright: © 2026. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

multimodal music datasets remains limited.

A broader challenge is that multimodal music datasets are technically complex and difficult to navigate, as different data types are often distributed across files and folders. Enabling more intuitive access, therefore, requires approaches that bridge data types and expose these relationships.

This paper addresses the following research question: *What are the capacities and limitations of a retrieval-augmented natural language interface for supporting the discovery of meaningful relationships within multimodal music datasets?* By framing multimodal datasets as spaces of interconnected information, this work aims to contribute to ongoing discussions in (creative) sound and music computing on new forms of data interaction. It explores the role of natural language as an interface for engaging with complex musical data and its potential to support more flexible and exploratory research practices. A sketch of our developed system is shown in Figure 1, and the code and implementation are publicly available online. [1]

2. RELATED WORK

Sophisticated computational tasks can integrate and reason over heterogeneous data. One example is question–answering, systems that answer questions posed by humans in a natural language. Such systems have evolved from unimodal (text-to-text) to bimodal (audio-to-text) [12] and are now explored in multimodal scenarios (audio, video, text, etc.) [13, 14].

Prior to large language model-based approaches (LLMs), multimodal music collections were commonly accessed through structured querying systems based on relational databases and ontology-based frameworks such as MusicBrainz [15] and the Music Ontology [16]. In addition, exploratory visual analytics tools such as repoVizz [17] support synchronized inspection of heterogeneous music-related data streams. However, these approaches generally rely on predefined schemas, explicit annotations, or interface-specific interaction paradigms, limiting their flexibility for ad-hoc natural language querying over heterogeneous and partially unstructured music datasets.

Although beneficial, general-purpose LLMs exhibit hallucinations, generating plausible yet ungrounded responses [18], while strong language biases often dominate multimodal reasoning, leading models to rely disproportionately on textual priors both due to the nature of training data and to architectural limitations of LLMs [19]. Krishna et al. [20] report a low baseline accuracy in question–answering without retrieval, while Sharma et al. [21] note that general-purpose pre-trained LLMs often lack the domain-specific knowledge and terminology required for accurate question–answering in specialized fields such as music.

To address these limitations, retrieval-augmented generation (RAG) systems have been proposed as a principled framework for grounding generation in external knowledge [22]. RAG combines an information retrieval component with a generative model, allowing the system to retrieve relevant information from large external data collections in response to a user query. The retrieved information is then used as additional context for the generative model, enabling it to produce responses that are grounded in external knowledge.

Recent work has begun to explore RAG in the context of music. For instance, [23] provide a survey of multimodal RAG systems, including audio, video, and text (in the form of code or metadata), highlighting how retrieval augmentation helps in music-related tasks such as captioning, music audio generation, and speech processing. Similarly, [24] introduce RAG into multimodal music synthesis, proposing a dual-track retrieval framework that combines broad thematic retrieval with targeted attribute-based retrieval, effectively forming a hierarchical retrieval mechanism for conditioning music generation. Their work also leverages multimodal datasets comprising audio, video, and text annotations. In a related direction, [25] propose a hybrid framework that integrates RAG with diffusion models for generating musically coherent dance sequences, incorporating optimized motion graph systems. However, these systems typically rely on pre-aligned data, which may not be available in archival collections.

While these approaches demonstrate the potential of retrieval augmentation and hierarchical reasoning in music-related generation tasks, they primarily focus on generation and synthesis, rather than question–answering and dataset exploration. Crucially, existing music RAG frameworks often overlook the role of natural language interfaces in abstracting the technical complexity of heterogeneous file formats (e.g., .mat, .mscz), a significant barrier for non-technical researchers. Moreover, many of these systems rely on either data-specific retrieval strategies or explicitly structured representations, without fully addressing the challenge of autonomous crossmodal inference and interpretable, hierarchical retrieval over multimodal music datasets.

3. METHOD

We employ a vector-based (RAG) architecture, in which both user queries and dataset files are mapped into a dense embedding space using the *all-MiniLM-L6-v2* Sentence-Transformer model [26]. Retrieval is performed via cosine similarity search over a persistent ChromaDB vector database [27], enabling semantically rich matching between queries and dataset descriptions, even when there is no exact keyword or lexical overlap between the query and the stored text representations.

Our system extends the standard RAG paradigm by incorporating a hierarchical retrieval strategy that accesses both high-level dataset representations, such as folder-level summaries, and fine-grained content segments. This allows the LLM to establish a global understanding of the dataset before integrating detailed evidence. The system uses a locally hosted API (*Jan-v3-4b-base-instruct*) [28] to ensure data privacy and persistent storage for the embedding collections. We focus on CPU-only execution, with query-time latency typically ranging between 1–4 seconds.

[1] https://github.com/MuTecEn/hierarchical-rag-multimodal-music-qa/tree/main

Memory usage is primarily determined by the embedding model and quantized LLM, and the system prioritizes reproducibility and accessibility on standard hardware.

3.1 Pre-processing and Indexing

Two music research datasets were used to test the system:

- MusicLab Abels KORK (KORK): 147 audio files, 19 text files (for timing alignment and dataset description), 12 performance videos, and 126,204 Equivital vest CSV files with physiological measurements (accelerometry, respiration, ECG, and IBI) collected from musicians and audience members during a live concert setting [29].
- Greek Folk Music (GFMD): 7 audio files, 53 text files (lyrics and short analytical descriptions), 292 MIDI files, 700 plot images (in PNG), and 166 score images (in PDF) [30].

Data preprocessing and indexing were performed in Python using *Sentence-Transformer* for embeddings [26], *PyPDF2* for PDFs [31], *librosa* for audio [32], *pretty_midi* for MIDI files [33], and *Musical Gestures Toolbox* for video [34].

Both datasets were converted to text to compute embeddings via a single sentence transformer (*all-MiniLM-L6-v2*). Text and CSV files were read in full, keyword-summarized, and split into 150-word chunks. Non-text modalities were serialized into descriptive text strings containing key metadata and analytical metrics: audio files included duration, tempo, onset density, and spectral/chroma profiles; MIDI summaries logged duration, instruments, note counts, and pitch ranges; video files recorded a quantity of motion index. To support hierarchical reasoning, we additionally generated recursive folder-level summaries. In this context, folders correspond to the hierarchical directory structure of the dataset, where each folder groups related multimodal resources (e.g., audio recordings, symbolic scores, textual metadata, and paradata) belonging to a common organizational unit. For each directory, we aggregated file-type distributions, total item counts, and dominant keywords across all subdirectories. This process is applied recursively in a bottom-up manner, such that summaries are first computed for leaf-level folders and then progressively merged into higher-level parent directory summaries. The resulting folder descriptions are embedded as separate *folder_summary* documents, creating a vectorized map of the dataset's organizational structure.

3.2 Hierarchical Retrieval and Generation

The system combines hierarchical retrieval from ChromaDB with constrained generation from a local LLM (Llama) [35], which is served via an OpenAI-compatible API (model *Jan-v3-4b-base-instruct*). All query embeddings are computed with the same Sentence Transformer (*all-MiniLM-L6-v2*) used during indexing. The system first embeds a user's natural language query and performs two similarity searches in the vector database: (i) a folder-level search that retrieves the three most relevant folder summaries, and (ii) a content-level search that retrieves the five most relevant file-level documents. The textual contents of these documents, together with their metadata (source and type), are concatenated into a single context string with separate sections for dataset structure and detailed content. This context is injected into a fixed prompt template that (a) instructs the LLM to act as an assistant exploring a multimodal music research dataset and (b) enforces a strict JSON output format. The template specifies the exact JSON schema (question, file_types, description, musical_materials, structure_relationship, conclusion) and explicitly forbids any text outside valid JSON. The LLM is called once per query with this prompt. Its response is parsed as JSON, and if parsing succeeds, the resulting object is merged with the retrieval metadata to form the final structured output. If parsing fails, the raw model output is returned for inspection.

3.3 Question Types

The evaluation query set consists of 28 unique questions, each designed to probe a specific capability of the system (e.g., crossmodal retrieval, dataset summarization, or semantic reasoning). Each question was posed once for each of the two datasets (KORK and GFMD), yielding a total of 56 query–dataset combinations. The queries are organized into categories (crossmodal retrieval, content-based retrieval, semantic and cultural queries, dataset understanding and summarization, creative and exploratory queries, and complex reasoning). The full list of all 28 questions is provided in Appendix Table 2. Crossmodal retrieval queries require the system to connect information across different data types, while content-based retrieval queries focus on feature-based access within a single data type (such as audio or MIDI). Semantic and cultural queries probe the system's ability to recognize meaningful musical or contextual concepts. Dataset understanding and summarization queries examine its capacity to generate coherent overviews of the collection. Creative and exploratory queries test its potential to support insight or inspiration, and complex reasoning queries challenge the system to integrate information from multiple sources and produce structured, comparative outputs.

3.4 Evaluation

To evaluate our system, we have devised a hybrid evaluation framework that assesses both the system's efficiency and its higher-level reasoning. This framework combines quantitative and qualitative analyses.

3.4.1 Quantitative analysis

The quantitative evaluation of the system was designed to assess two primary dimensions: *efficiency* (the reduction of technical and interaction barriers) and *retrieval performance* (accuracy and completeness). Given the system's design as a memory-efficient, lightweight tool, retrieval is capped at the top-5 results per query. We therefore introduce an adjusted completeness metric to account for this architectural constraint.

Because manual annotation of the entire dataset is prohibitive, we introduce a relevance-by-type heuristic to approximate ground truth relevance. This heuristic was developed specifically for this evaluation and is not intended as a general relevance assessment framework. Relevance is determined by the alignment between the semantic requirements expressed in a query and the file extensions associated with the retrieved items. For example, for a query such as "Can you find recordings that combine fast tempo audio with dance-related lyrics?", files with extensions such as *.mp3* or *.wav* are considered relevant for the audio component, while *.txt* files are considered relevant for the lyrics component. Retrieved files that match the predefined datatype requirements are categorized as relevant, whereas files of unrelated types are categorized as irrelevant.

The relevance-by-type heuristic enables automated computation of retrieval metrics without requiring exhaustive manual annotation. However, it evaluates whether the system retrieves the appropriate modalities required to answer a query rather than whether the retrieved files are semantically correct or sufficient to answer the query. Consequently, the resulting accuracy and completeness metrics should be interpreted as measures of modality-level retrieval performance and may overestimate true semantic retrieval quality. To mitigate this limitation, we complement the quantitative evaluation with a human evaluation that explicitly rates the correctness, completeness, clarity, and usefulness of the generated responses.

A steps-saved metric evaluates the reduction in user interaction cost (C_{RAG}), calculated as:

$$C_{\text{RAG}} = N_{\text{folders}} + N_{\text{files}} + N_{\text{formats}} \tag{1}$$

representing the unique entities synthesized by the system per response. To establish a baseline, the manual retrieval cost (C_{Manual}) assumes an equivalent search complexity where a user would need to manually inspect the same set of corresponding folders, files, and file formats to locate and verify that information. The total savings are defined as:

$$\text{Steps Saved} = \sum C_{\text{Manual}} - C_{\text{RAG}} \tag{2}$$

We explicitly note that C_{Manual} serves as an approximate index of relative interaction and search navigation effort rather than a direct measurement of elapsed time. Because the time required for a human to open specialized formats and traverse physical directories differs substantially from automated system synthesis operations, this metric should be interpreted strictly as an indicator of structural navigation complexity abstracted away from the user via natural language automation.

To evaluate the precision of the retrieval component, we define accuracy as the proportion of retrieved items that satisfy the relevance-by-type heuristic:

$$\text{Accuracy} = \frac{R_{\text{retrieved}}}{N_{\text{retrieved}}} \tag{3}$$

where $R_{\text{retrieved}}$ denotes the number of relevant items among the retrieved results and $N_{\text{retrieved}}$ is the total number of items retrieved.

Completeness (recall) traditionally measures the fraction of all relevant items in a dataset that are successfully retrieved. However, because our system is limited to a top-5 retrieval threshold to maintain a low memory footprint, standard recall is naturally capped in large-scale datasets. To provide a fairer assessment of the system's retrieval logic under this constraint, we define the *adjusted completeness* as:

$$C_{\text{adj}} = \frac{R_{\text{retrieved}}}{\min(R_{\text{total}}, 5)} \tag{4}$$

where R_{total} denotes the total number of relevant items for the query in the dataset (under the relevance-by-type heuristic).

The adjustment normalizes the metric relative to the system's maximum possible output, enabling consistent performance evaluation across queries with varying densities of relevant items. For each experiment, these metrics are computed using the structured JSON output generated by the RAG system, which explicitly lists the sources and formats utilized.

3.4.2 Human Evaluation

To evaluate the system's capacity for semantic retrieval and musicological reasoning, we designed a human assessment framework focused on two key functional capabilities: *crossmodal retrieval*, related to the system's ability to navigate the semantic gap between disparate modalities, (e.g., retrieving symbolic MIDI scores or audio recordings from textual musicological concepts), and *dataset summarization*, related to the system's ability to generate coherent and informative overviews of the dataset.

The evaluation was conducted by five independent annotators with backgrounds in musicology, psychology, or informatics. All annotators are interdisciplinary researchers whose work lies at the intersection of rhythm and music research. Each of the 56 system responses was independently evaluated by all five annotators using the structured questionnaire described in Section 3.4. For each evaluation criterion, annotators assigned a score on a 5-point Likert scale (1 = Poor, 5 = Excellent) and provided written comments justifying their assessments. These comments were used to contextualize the numerical ratings and support the results' interpretation.

The evaluation criteria included four general dimensions (correctness, completeness, clarity, and usefulness) and two task-specific dimensions: crossmodal retrieval and dataset summarization, which were applied only to queries in the corresponding query category (Appendix Table 2). This separation allows us to isolate performance differences between within-modality retrieval and crossmodal reasoning tasks.

For each query and criterion, the final score was computed as the mean rating across the five annotators. Category-level scores were then calculated by averaging the query-level scores within each evaluation category:

$$M = \frac{1}{|Q|} \sum_{q \in Q} s(q) \tag{5}$$

where Q denotes the set of queries in the corresponding category and $s(q)$ represents the mean rating assigned to query q.

To assess the reliability of the aggregated ratings, we computed a two-way random-effects intraclass correlation coefficient (ICC(2,k)) across the five annotators. This measure was selected because the reported evaluation scores represent averages across multiple expert judgments.

4. RESULTS

Across 56 queries, the system achieved a mean modality-level retrieval accuracy of 0.78 and an adjusted completeness of 1.00, demonstrating that 78% of retrieved items matched the exact file format requirements of the query under our relevance-by-type heuristic at the top-5 constraint (Figure 2). Excluding negative retrieval cases ($R_{\text{total}} = 0$), accuracy reached 1.00, proving the system consistently isolates relevant formats when they exist. Lower accuracy scores stem entirely from negative retrieval scenarios where a requested modality was absent. In these instances, the system retrieved alternative files (e.g., *.txt* or *.csv*) to verify and explain the absence; while penalized under the strict heuristic, this reflects correct, informative behavior.

Regarding interaction efficiency, the system averaged 8.96 synthesis steps per query (C_{RAG}). Unlike manual workflows requiring directory navigation and file inspections, the proposed natural-language interface abstracts away this underlying cross-modal complexity transparently, effectively hiding structural navigation friction from the user.

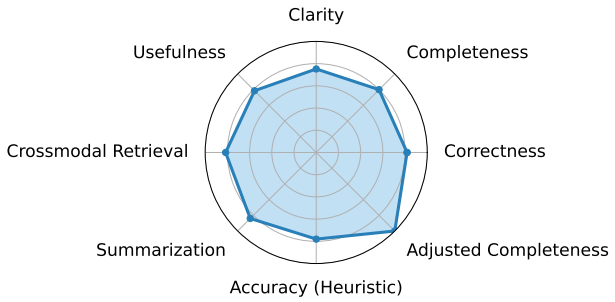


Figure 2. Radar chart evaluating system performance across six qualitative dimensions (Clarity, Usefulness, Crossmodal Retrieval [CMR], Summarization, Completeness, Correctness) and two quantitative dimensions (Accuracy [Heuristic] and Adjusted Completeness).

Human evaluation yielded mean scores of 4.09 for correctness, 4.00 for completeness, 3.76 for clarity, and 3.93 for usefulness. Regarding correctness, the system successfully extracted precise metadata and identified modal presence or absence (e.g., physiological data/video in GFMD, lyrics in KORK) without hallucinations. Deductions resulted from occasional factual errors (e.g., misclassifying Aegina, Ikaria, Chios, and Donousa into the Dodecanese archipelago), reasoning slips (e.g., failing to link filename prefixes to a rehearsal/concert distinction), or genre mislabeling.

For completeness, the model provided detailed, context-aware responses, explaining file roles within the research context. However, this coverage remained largely explicit. In queries like "Which lyrics correspond to recordings with fast tempo?", the system noted the absence of direct alignments but did not infer implicit links via shared metadata. Similarly, while extracting paradata effectively, it left part of the meaning implicit by failing to semantically interpret participant codes (e.g., mapping "AU8" to "Audience").

In terms of clarity, output formatting (bolding, bullet points, headers) ensured readability. However, clarity received the lowest score due to formatting errors in four instances where the model outputted unparsed raw text rather than valid JSON, disrupting data flow. Additionally, one evaluator penalized clarity because certain outputs were overly long or lacked bulleted navigation.

Regarding usefulness, the system effectively translates technical paradata into musicological insights (e.g., relating note density to dance rhythm), enabling effortless exploration of over 100,000 files. It successfully reduced technical friction by clarifying sensor paradata (e.g., MoveSense physiological streams). Usefulness was primarily bounded by the aforementioned geographic misclassifications, which could mislead domain-specific analysis.

Task-specific dimensions performed strongly. Cross-modal retrieval achieved a mean score of 4.08, bridging fluid semantic concepts and technical features. The system successfully evaluated rhythmic onset density in MIDI or spectral centroids in audio to answer text queries about dance, linked linguistic concepts (e.g., "bride") to symbolic score structures (e.g., "wedding Syrtos" PDFs), and connected physiological CSV data to MP3 audio features. Its primary limitation was a conservative reliance on explicit links over multi-step contextual inferences. Dataset summarization scored 4.20, showing robust capacity to synthesize large volumes of multi-modal data into coherent musicological narratives (e.g., oral transmission, regional variation, performance contexts) and to handle cross-lingual extraction from Greek gracefully.

To validate score reliability, a two-way random-effects intraclass correlation coefficient matrix was computed (Table 1). The overall average measures consensus across the four baseline criteria demonstrated moderate reliability ($\text{ICC}(2,k) = 0.6058$), balancing out the expected subjectivity of individual open-text grading ($\text{ICC}(2,1) = 0.2351$).

Correctness achieved high consensus across both datasets ($\text{ICC}(2,k) = 0.7167$). The highest alignment emerged in GFMD cross-modal retrieval ($\text{ICC}(2,k) = 0.8397$), where judges strongly agreed on the system's negative retrieval validation. Conversely, low agreement on GFMD Clarity ($\text{ICC}(2,k) = 0.1518$) exposed divergent evaluator tolerances toward raw syntax blocks: some heavily penalized the formatting breach (scoring a 1), while others prioritized underlying text accuracy. A severe consensus failure occurred in GFMD summarization ($\text{ICC}(2,k) = -7.5000$), where a limited target pool ($N = 3$) combined with varying evaluator preferences regarding optimal sum-

Table 1. Inter-Rater Reliability Across Evaluation Metrics via Average Intraclass Correlation Coefficients (ICC(2,k))

Evaluation Metric	Combined Datasets	Greek Folk Music (GFMD)	KORK Dataset
Overall (4 Common Criteria)	**0.60**	**0.53**	**0.61**
Correctness	0.72	0.68	0.71
Completeness	0.65	0.63	0.63
Clarity	0.38	0.15	0.51
Usefulness	0.67	0.67	0.58
Cross-Modal Score	0.78	0.84	0.49
Summarization Score	0.00[a]	-7.50[b]	0.53

[a] Zero variance between certain evaluator subsets resulted in an anchored baseline score of 0.00.
[b] Negative ICC values indicate that the systematic variance between evaluators outpaced the variance between small-sample target outputs ($N = 3$).

mary length caused inter-rater variance to outpace target variance. Conversely, the KORK dataset exhibited stable consensus across Clarity (0.5145) and Summarization (0.5294) due to the absence of unparsed JSON errors.

5. CONCLUSIONS AND DISCUSSION

Our experiments demonstrate the feasibility of a lightweight hierarchical RAG system for interacting with multimodal music datasets. The proposed framework enables natural language exploration of heterogeneous musical resources while reducing the technical barriers associated with navigating complex repositories. The system achieved strong human-evaluated performance in crossmodal reasoning (mean score = 4.08), indicating its ability to establish meaningful semantic relationships across textual descriptions, symbolic scores, and audio features without relying on explicit database mappings. By supporting complex natural language queries, the framework provides an intuitive interface for multidisciplinary music research.

At the same time, the evaluation revealed several limitations, including occasional geographic misclassifications, conservative reasoning behavior, and JSON parsing errors associated with smaller locally deployed LLMs. Although larger cloud-based models would likely improve output robustness and factual accuracy, they introduce trade-offs in terms of privacy, copyright, and operational costs. Our design therefore prioritizes a fully local, privacy-preserving, and computationally lightweight deployment.

The evaluation also highlighted an architectural limitation imposed by the retrieval stage. Annotator feedback indicated that count-based and threshold-based queries occasionally suffered from reduced precision because the model reasons only over the retrieved context rather than the complete dataset. Preliminary experiments with larger retrieval sets produced qualitatively similar results, although hardware limitations prevented a full human evaluation. Systematic analysis of retrieval-depth sensitivity therefore remains an important direction for future work.

Finally, the inter-rater reliability analysis highlights the subjective nature of evaluating open-ended, crossmodal responses. While the reliability of individual annotators was relatively low ($\text{ICC}(2,1) = 0.2351$), the aggregated ratings across five annotators achieved moderate agreement ($\text{ICC}(2,k) = 0.6058$), supporting the use of expert consensus rather than individual judgments. Lower agreement for clarity and summarization further suggests that future evaluations would benefit from more detailed, criterion-specific scoring rubrics.

Future work will investigate replacing the general-purpose embedding model (*all-MiniLM-L6-v2*) with music-specific multimodal retrieval models such as MuLan, CLaMP, or TTMR++, which are explicitly designed to capture musical semantics. Beyond musicology and music psychology, the proposed local RAG framework also has potential for supporting conversational exploration of institutional music archives and other privacy-sensitive multimodal collections.

Acknowledgments

This project is supported by the Research Council of Norway through projects 262762 (RITMO) and 357438 (MishMash).

6. REFERENCES

[1] C. C. Liem, M. Müller, D. Eck, G. Tzanetakis, and A. Hanjalic, "The need for music information retrieval with user-centered and multimodal strategies," in *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, ser. MIRUM '11. New York, NY, USA: Association for Computing Machinery, Nov. 2011, pp. 1–6. [Online]. Available: https://doi.org/10.1145/2072529.2072531

[2] A.-M. Christodoulou, O. Lartillot, and A. R. Jensenius, "Multimodal music datasets? Challenges and future goals in music processing," *International Journal of Multimedia Information Retrieval*, vol. 13, no. 3, p. 37, Aug. 2024. [Online]. Available: https://doi.org/10.1007/s13735-024-00344-6

[3] F. Simonetta, S. Ntalampiras, and F. Avanzini, "Multimodal music information processing and retrieval: survey and future challenges," in *2019 International Workshop on Multilayer Music Representation and Processing (MMRP)*, Jan. 2019, pp. 10–18.

[4] P. Knees, M. Schedl, T. Pohle, and G. Widmer, "An innovative three-dimensional user interface for exploring music collections enriched," in *Proceedings of the 14th ACM international conference on Multimedia*, ser. MM '06. New York, NY, USA: Association for Computing

Machinery, Oct. 2006, pp. 17–24. [Online]. Available: https://dl.acm.org/doi/10.1145/1180639.1180652

[5] R. Khulusi, J. Kusnick, C. Meinecke, C. Gillmann, J. Focht, and S. Jänicke, "A Survey on Visualizations for Musical Data," *Computer Graphics Forum*, vol. 39, no. 6, pp. 82–110, 2020.

[6] F. Kleerdorfer, H. Ulf, and B. Krenn, "Making Large Music Collections Accessible using Enhanced Metadata and Lightweight Visualizations," in *Third International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution*, 2007. [Online]. Available: https://sites.google.com/view/2nd-mmfm-workshop/challenge

[7] D. Bretherton, D. A. Smith, m. schraefel, R. Polfreman, M. Everist, L. J. Brooks, and J. Lambert, "Integrating musicology's heterogeneous data sources for better exploration," in *10th International Society for Music Information Retrieval Conference*. International Society for Music Information Retrieval, Oct. 2009, pp. 27–32, num Pages: 6. [Online]. Available: https://eprints.soton.ac.uk/68804/

[8] T. Langer, "Music Information Retrieval & Visualization," *Trends in Information Visualization*, pp. 15–22, 2010.

[9] L. Pugin, "The Challenge of Data in Digital Musicology," *Frontiers in Digital Humanities*, vol. 2, Aug. 2015. [Online]. Available: https://www.frontiersin.org/journals/digital-humanities/articles/10.3389/fdigh.2015.00004/full

[10] G. G. Hendrix, E. D. Sacerdoti, D. Sagalowicz, and J. Slocum, "Developing a natural language interface to complex data," *ACM Trans. Database Syst.*, vol. 3, no. 2, pp. 105–147, Jun. 1978. [Online]. Available: https://dl.acm.org/doi/10.1145/320251.320253

[11] I. Androutsopoulos, G. D. Ritchie, and P. Thanisch, "Natural language interfaces to databases – an introduction," *Natural Language Engineering*, vol. 1, no. 1, pp. 29–81, Mar. 1995. [Online]. Available: https://www.cambridge.org/core/journals/natural-language-engineering/article/abs/natural-language-interfaces-to-databases-an-introduction/21C30448C70DD4988E6DA0D54205FB56

[12] S. Liu, A. S. Hussain, C. Sun, and Y. Shan, "Music Understanding LLaMA: Advancing Text-to-Music Generation with Question Answering and Captioning," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 286–290. [Online]. Available: https://ieeexplore.ieee.org/document/10447027/

[13] G. Li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu, "Learning to Answer Questions in Dynamic Audio-Visual Scenarios," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 19 086–19 096. [Online]. Available: https://ieeexplore.ieee.org/document/9879157/

[14] A.-M. Christodoulou, K. Glette, O. Lartillot, and A. R. Jensenius, "MusiQAl: A Dataset for Music Question-Answering through Audio-Video Fusion," *Transactions of the International Society for Music Information Retrieval*, vol. 8, no. 1, Jul. 2025. [Online]. Available: https://transactions.ismir.net/articles/10.5334/tismir.222

[15] A. Swartz, "MusicBrainz: a semantic Web service," *IEEE Intelligent Systems*, vol. 17, no. 1, pp. 76–77, Jan. 2002. [Online]. Available: https://ieeexplore.ieee.org/document/988466/

[16] Y. Raimond and M. Sandler, "Evaluation of the Music Ontology Framework," in *The Semantic Web: Research and Applications*, E. Simperl, P. Cimiano, A. Polleres, O. Corcho, and V. Presutti, Eds. Berlin, Heidelberg: Springer, 2012, pp. 255–269.

[17] O. Mayor, Q. Llimona, M. Marchini, P. Papiotis, and E. Maestre, "repoVizz: a framework for remote storage, browsing, annotation, and exchange of multimodal data," in *Proceedings of the 21st ACM international conference on Multimedia*. Barcelona Spain: ACM, Oct. 2013, pp. 415–416. [Online]. Available: https://dl.acm.org/doi/10.1145/2502081.2502247

[18] B. Weck, I. Manco, E. Benetos, E. Quinton, G. Fazekas, and D. Bogdanov, "MuChoMusic: Evaluating Music Understanding in Multimodal Audio-Language Models," Aug. 2024. [Online]. Available: http://arxiv.org/abs/2408.01337

[19] Y. Zang, S. O'Brien, T. Berg-Kirkpatrick, J. McAuley, and Z. Novack, "Are you really listening? Boosting Perceptual Awareness in Music-QA Benchmarks," Jun. 2025, arXiv:2504.00369 [cs]. [Online]. Available: http://arxiv.org/abs/2504.00369

[20] S. Krishna, K. Krishna, A. Mohananey, S. Schwarcz, A. Stambler, S. Upadhyay, and M. Faruqui, "Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generation," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 4745–4759. [Online]. Available: https://aclanthology.org/2025.naacl-long.243/

[21] S. Sharma, D. S. Yoon, F. Dernoncourt, D. Sultania, K. Bagga, M. Zhang, T. Bui, and V. Kotte, "Retrieval Augmented Generation for Domain-specific Question Answering," May 2024.

[22] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive

NLP Tasks," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[23] R. Zhao, H. Chen, W. Wang, F. Jiao, X. L. Do, C. Qin, B. Ding, X. Guo, M. Li, X. Li, and S. Joty, "Retrieving Multimodal Information for Augmented Generation: A Survey," in *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 4736–4756.

[24] B. Wang, L. Zhuo, Z. Wang, C. Bao, W. Chengjing, X. Nie, J. Dai, J. Han, Y. Liao, and S. Liu, "Multimodal Music Generation with Explicit Bridges and Retrieval Augmentation," Dec. 2024, arXiv:2412.09428 [cs]. [Online]. Available: http://arxiv.org/abs/2412.09428

[25] M. Huang, P. Zhang, and B. Zhang, "MotionRAG-Diff: A Retrieval-Augmented Diffusion Framework for Long-Term Music-to-Dance Generation," Jun. 2025, arXiv:2506.02661 [cs]. [Online]. Available: http://arxiv.org/abs/2506.02661

[26] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992.

[27] D. Kim and C. Team, "Chroma-the ai-native open-source embedding database," 2023. [Online]. Available: https://docs.trychroma.com/docs/overview/introduction

[28] "janhq/jan," Jun. 2026, original-date: 2023-08-17T02:17:10Z. [Online]. Available: https://github.com/janhq/jan

[29] "MusicLab: Abels KORK - RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion," Jun. 2024. [Online]. Available: https://www.uio.no/ritmo/english/projects/musiclab/2024/abels-kork/index.html

[30] A.-M. Christodoulou and O. Lartillot, "A Multimodal Dataset of Greek Folk Music," in *Proceedings of the 12th International Conference on Digital Libraries for Musicology*. Seoul Republic of Korea: ACM, 2025, pp. 19 – 27. [Online]. Available: https://dl.acm.org/doi/epdf/10.1145/3748336.3748339

[31] M. Fenniak, M. Stamy, M. Thoma, and M. Peveler, "PyPDF2: A pure-python PDF library capable of splitting, merging, cropping, and transforming PDF files," 2022. [Online]. Available: ttps://pypi.org/project/PyPDF2

[32] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," Austin, Texas, 2015, pp. 18–24. [Online]. Available: https://conference.scipy.org/proceedings/scipy2015/brian_mcfee.html

[33] C. Raffel, "Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching," Ph.D. dissertation, Columbia University, 2016. [Online]. Available: https://www.proquest.com/openview/4dc9b57b251e7806ff9041a948b7c320/1?pq-origsite=gscholar&cbl=18750

[34] B. Laczkó and A. R. Jensenius, "Reflections on the Development of the Musical Gestures Toolbox for Python," in *Proceedings of the Nordic Sound and Music Computing Conference*. Aalborg University Copenhagen, 2021, accepted: 2021-11-25T18:33:58Z. [Online]. Available: https://www.duo.uio.no/handle/10852/89331

[35] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023, arXiv:2302.13971 [cs]. [Online]. Available: http://arxiv.org/abs/2302.13971

7. APPENDIX

Table 2. Question categories used to assess the system's capabilities in dataset exploration.

Category	Example Queries
Dataset Structure: Data, Metadata, Paradata	What types of files exist in this dataset?
	Which folders contain audio recordings?
	How many MIDI files are present?
	What metadata is available for each recording?
	What information exists about how the data was collected or processed?
Crossmodal Retrieval	Which MIDI files correspond to audio recordings longer than 30 seconds?
	Which lyrics correspond to recordings with a fast tempo?
	Retrieve scores associated with recordings mentioning dance in lyrics.
	Which physiological recordings correspond to specific musical segments?
	Can you find video recordings corresponding to x audio excerpt?
Content-Based Retrieval (Single Modality)	Which audio recordings are longer than one minute?
	Which MIDI files contain more than one instrument?
	Which files reference traditional Greek dance forms?
	Which participants have ECG or respiration data available?
	Which files include audience's and which performer's physiological data?
Semantic and Cultural Queries	Which pieces are examples of balos dance?
	Which recordings come from the region x?
	Which pieces mention love in their lyrics?
Dataset Understanding and Summarization	Can you summarize the contents of the dataset x?
	Can you describe the types of musical material included?
	What kinds of musical traditions are represented in dataset x?
Creative and Exploratory Queries	Can you show recordings with similar rhythmic patterns?
	Can you find pieces with similar harmonic characteristics?
	Can you suggest recordings that could inspire a new composition?
	Can you compare performance styles between x and y?
	Can you find rare or unusual recordings?
Complex Reasoning	Can you identify recordings of balos across the dataset and describe differences in instrumentation, performance context, and lyrical content?
	Can you find recordings that combine fast tempo audio with dance-related lyrics?