Jamendo-MT-QA: A Benchmark for Multi-Track Comparative Music Question Answering

Anonymous ACL submission

Abstract

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 benchmarks do not systematically evaluate reasoning across multiple tracks. Building on the Jamendo-QA dataset, we introduce Jamendo-MT-QA, a dataset and benchmark for *multi-track comparative question answering*. From Creative Commons-licensed tracks on Jamendo, we construct **36,519** comparative QA items over **12,173** track pairs, with each pair yielding three question types: yes/no, short-answer, and sentence-level questions. We describe an LLM-assisted pipeline for generating and filtering comparative questions, and benchmark representative audio–language models using both automatic metrics and LLM-as-a-Judge evaluation.

1 Introduction

Music is often experienced relationally: listeners compare tracks by mood, energy, instrumentation, or stylistic similarity (e.g., "this song feels darker than the previous one"). Despite this, most existing music benchmarks (Liu et al., 2025; Yuan et al., 2023) focus on single-item tagging (Law et al., 2009), captioning (Doh et al., 2023b; Kim et al., 2019; Agostinelli et al., 2023), or classification (Tzanetakis and Cook, 2002; Defferrard et al., 2017, 2018), and do not directly test a model's ability to reason *across* multiple tracks. Recent analyses further suggest that performance on some Music-QA benchmarks may be driven by text-based cues rather than genuine audio perception, highlighting the need for evaluation settings that explicitly probe reasoning over audio inputs (Zang et al., 2025).

In parallel, large language models (LLMs) and audio-language models (Rubenstein et al., 2023; Wu et al., 2019) have achieved strong results on textual QA and multimodal understanding. Yet it remains unclear how well these systems support *comparative* music reasoning, where a model must integrate information from two or more tracks and express the relation in natural language.

Jamendo-QA (Koh et al., 2025b) recently introduced a large-scale dataset of QA pairs and captions aligned with Jamendo music audio, targeting general music understanding. Building on this resource, we propose Jamendo-MT-QA, a dataset for **Multi-Track comparative Question Answering**. Each data point corresponds to a *track pair*, for which we generate three comparative questions (yes/no, short-answer, and sentence-level), collectively forming a *QA group*.

Beyond dataset construction, we empirically investigate the difficulty of multi-track comparative question answering by benchmarking several recent audio–language models. Our evaluation reveals that, despite strong performance on single-track audio understanding, existing models struggle to generate well-grounded sentence-level comparisons across multiple tracks. These findings motivate the need for a dedicated benchmark that explicitly targets comparative music reasoning.

Contributions. Our main contributions are:

- We construct Jamendo-MT-QA on top of Jamendo-QA, comprising **36,519** comparative QA items over **12,173** track pairs, covering genre, mood, instrumentation, and production-related attributes.
- We design an **LLM-assisted pipeline** for generating and filtering three comparative questions per track pair (yes/no, short-answer, sentence-level).
- We provide a **baseline benchmark** of representative multi-audio and caption-based audio-language models under a unified evaluation

protocol, using both automatic metrics and LLM-as-a-Judge evaluation.

2 Related Work

Music-QA and Captioning. Prior datasets such as MUSIC-AVQA (Li et al., 2022), MusicQA (Ouyang et al., 2025), and MusicXQA (Chen et al., 2025) focus on answering questions about music or music-related video, often using symbolic representations or constrained QA formats. Other large-scale resources, including MTG-Jamendo (Bogdanov et al., 2019), JamendoMaxCaps (Roy et al., 2025), and LP-MusicCaps (Doh et al., 2023a), provide tags or captions but not explicit QA pairs. Jamendo-QA (Koh et al., 2025b) dataset addresses Music-QA at scale for single tracks; our work extends this direction to multi-track comparative reasoning.

Multimodal Audio-Language Models and Benchmarks. Audio-language models based on contrastive learning, such as CLAP (Elizalde et al., 2022) and MuLan (Huang et al., 2022), align music or audio with natural language in a shared representation space and have shown strong performance on retrieval and captioning (Koh et al., 2025a). Beyond representation learning, encoder–decoder architectures and audio-conditioned LLMs (Chu et al., 2024; Ghosh et al., 2024; Tang et al., 2024) enable more flexible audio–text generation and understanding. Recent models such as MU-LLaMA (Liu et al., 2023, 2024), Music Flamingo (Ghosh et al., 2025), and ChatMusician (Yuan et al., 2024) further extend this line of work by supporting instruction-following and interactive reasoning grounded in music audio.

In parallel, several benchmarks have been proposed to evaluate audio-language models across diverse tasks and domains. The MAE benchmark (Chen et al., 2024) evaluates multi-audio processing capabilities across speech and general sound domains. AIR-Bench (Yang et al., 2024) introduces an instruction-following benchmark for audio-language models spanning speech, environmental sounds, and music. AudioBench (Wang et al., 2025) further provides a broad, task-diverse evaluation suite for AudioLLMs, covering captioning, question answering, retrieval, and reasoning over general audio. These benchmarks primarily assess general audio understanding or the ability to process multiple audio inputs, typically in single-track or non-comparative settings.

In contrast, Jamendo-MT-QA specifically targets *comparative reasoning* between music tracks, requiring models to integrate perceptual attributes across tracks and generate structured comparative answers.

Comparative and Relational Question Answering. Comparative question answering has been extensively studied in the NLP domain, particularly in multi-hop and relational QA settings (Schnitzler et al., 2024; Li et al., 2024). Representative benchmarks such as HotpotQA (Yang et al., 2018) and WikiMultiHopQA (Ho et al., 2020) require models to aggregate evidence across multiple documents, while DROP (Dua et al., 2019) emphasizes discrete and logical reasoning over textual contexts.

Recent work on audio-language benchmarks has highlighted that high accuracy on existing QA tasks can mask deficiencies in genuine multimodal perception and reasoning. For example, in Music-QA, text-only models can achieve strong results even without access to audio inputs, suggesting dataset biases and shortcut cues rather than true perceptual inference (Zang et al., 2025). Such diagnostic perspectives motivate benchmarks that more explicitly probe relational and reasoning capabilities beyond surface-level pattern matching.

However, existing analyses in the music and audio domain have primarily focused on single-track understanding or perceptual grounding in isolation (Zeng et al., 2021; Wu et al., 2025). They have not systematically examined *comparative* or *relational* reasoning across multiple audio inputs, where models must integrate perceptual attributes, temporal structure, and cross-track relationships. Our work draws inspiration from both multi-hop QA and recent diagnostic benchmark analyses by formulating comparative Music-QA as a multi-track reasoning problem, and by providing a dataset and benchmark that explicitly require integrating information across multiple music tracks.

3 Jamendo-MT QA Dataset

Base Dataset: Jamendo-QA. Jamendo-QA (Koh et al., 2025b) is built from Creative Commons-licensed tracks on the Jamendo platform and provides question–answer pairs and captions aligned with music audio. Each sample links an audio file, rich metadata (genre, tempo, gender, etc.), and automatically generated QA pairs. The dataset spans a wide range of genres and musical attributes. In this work, we treat Jamendo-QA as the foundation

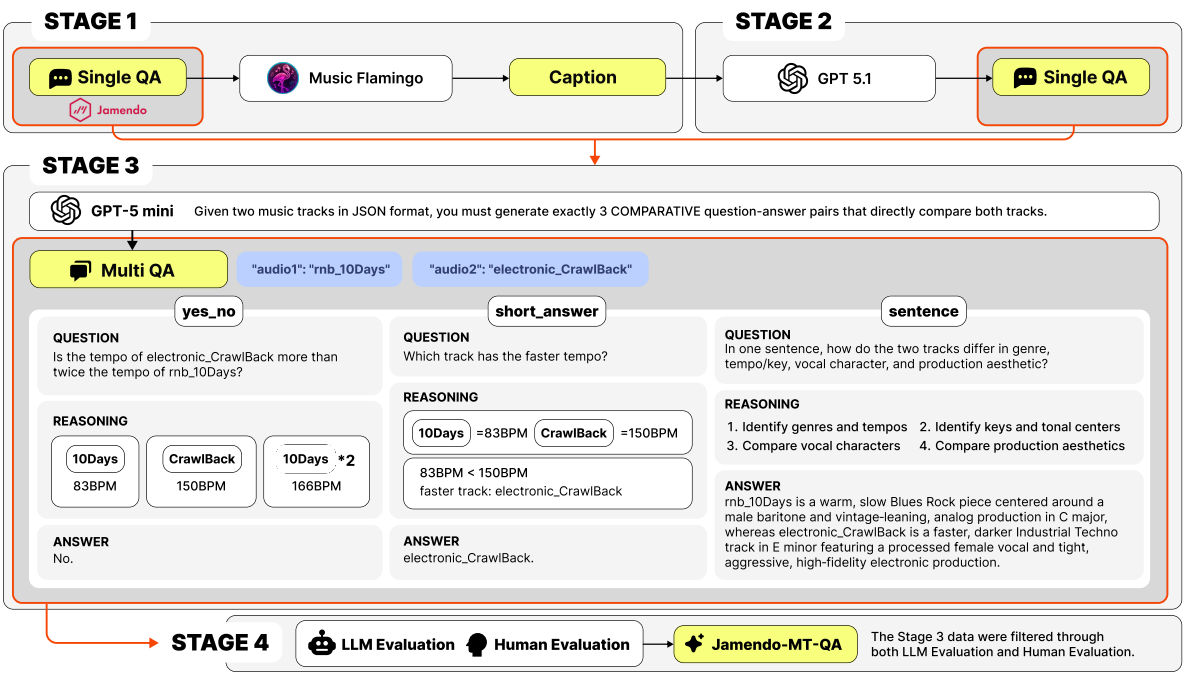

Figure 1: Overview of the Jamendo-MT-QA construction pipeline. Starting from the Jamendo-QA dataset, we first generate rich single-track captions using Music Flamingo, expand them into single-track QA pairs using GPT-5.1, and construct comparative QA groups consisting of three question types (yes/no, short-answer, sentence-level) for each track pair via an LLM-assisted generation pipeline. Finally, we apply human evaluation and LLM-as-a-Judge assessment for quality control and filtering, resulting in the final Jamendo-MT-QA dataset.

for our multi-track comparative QA dataset.

3.1 Stage 1: High-quality music captioning

For each track, structured natural language description using Music Flamingo (Ghosh et al., 2025). Given a raw audio, Music Flamingo produces detailed captions covering genre, tempo, key, instrumentation, production style, vocal characteristics, and lyrical themes. These captions form a dense semantic representation of each track that serves as the basis for downstream QA generation. To ensure audio–text consistency, we further conduct a human evaluation in which annotators with access to the underlying audio assess the alignment between the generated captions and the musical content. The detailed evaluation method and results are provided in Appendix A.1

3.2 Stage 2: Single-track QA expansion

We then expand each caption into multiple high-quality single-track QA pairs using GPT-5.1. The model generates diverse factual and descriptive questions about the track (e.g., genre, tempo, key, instrumentation, mood), producing an enriched single-track QA resource.

3.3 Stage 3: Multi-track Comparative QA Generation

To construct comparative QA pairs, we combine two track descriptions (caption, metadata, and single-track QA) into a structured JSON input and prompt the model to act as a *music comparison expert*. In this stage, we use **GPT-5 mini** as the primary generator for dataset construction. For each track pair, the model produces exactly three comparative questions corresponding to yes/no, short-answer, and sentence-level formats. All questions are required to explicitly reference both tracks. The full output schema is provided in Appendix B.

We experimented with several alternative LLMs for Stage 3 generation, including Qwen3-32B (Yang et al., 2025), InternLM3-8B (Cai et al., 2024), Gemma 3 12B (Gemma Team et al., 2025), and Claude 4.5 Haiku (Anthropic, 2025). Based on qualitative comparison, we select GPT-5 mini as the generator used for all dataset construction. GPT-5 mini consistently produces step-by-step reasoning, which improves transparency and supports the use of the *reasoning* field as a high-quality supervisory signal. Figure 2 illustrates a qualitative example of the generated comparative QA group

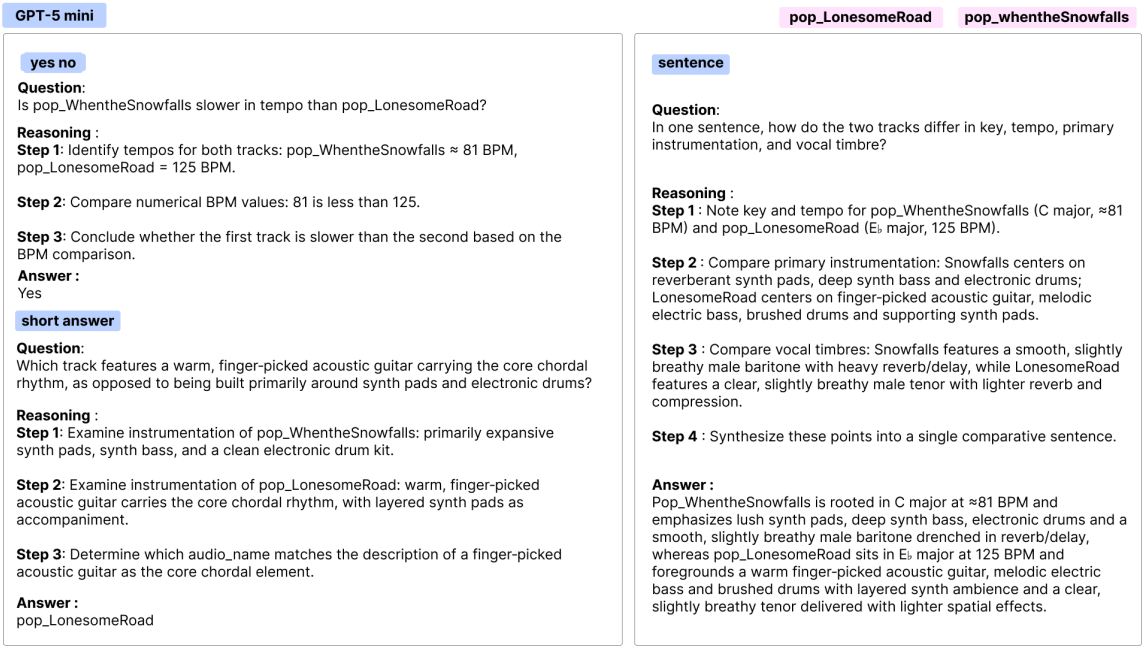

Figure 2: Qualitative example of a Stage 3 comparative QA group generated by **GPT-5 mini** for the same track pair. The generator produces three comparative question types (yes/no, short-answer, sentence-level) and provides explicitly step-by-step reasoning, which improves interpretability and makes the *reasoning* field more suitable as a supervision signal for downstream training.

for a single track pair. A broader qualitative comparison against other candidate generators is deferred to Appendix G.

3.4 Stage 4: Evaluation and Quality-Based Filtering

After generating the full set of 39,291 comparative QA items in Stage 3, we evaluate and refine the dataset through a final quality-control stage. This stage consists of: (1) a two-phase evaluation (human evaluation and LLM-as-a-Judge), and (2) quality-based filtering to produce the final released dataset.

Human Evaluation. To obtain a human reference for evaluation reliability, we randomly sample 300 QA items from the GPT-5-mini generated dataset. Four annotators independently rate each item using four criteria: Correctness, Comparative Validity, Reasoning Quality, and Difficulty.

Table 1 reports per-annotator scores together with aggregate statistics, including the human mean and mean absolute deviation (MAD), for four evaluation criteria. While individual item-level agreement is difficult to quantify due to score compression on the 1–5 Likert scale, we observe consistent *mean-level trends* across annotators for the three semantic criteria (Correctness, Comparative Validity, and Reasoning Quality).

Table 1 also includes scores from the GPT-5 Mini LLM judge on the same subset. Across all semantic criteria, LLM-based scores closely match human mean ratings, indicating that the LLM judge can serve as a scalable proxy for semantic quality assessment.

LLM-as-a-Judge Evaluation. Based on the observed alignment between human and LLM judgments on the 300-item subset, we employ GPT-5-mini as an LLM-as-a-Judge to evaluate the full dataset at scale. Rather than treating the LLM as a replacement for human evaluation, we use it as a *consistent semantic scorer* whose behavior is anchored by human reference statistics. This approach enables scalable quality control while preserving comparability with human judgments at the level of aggregated trends. As shown in Table 1, the LLM judge reproduces the relative ordering and mean scores of the semantic criteria, even though individual item-level agreement may vary.

Quality-Based Filtering. For the final release, we retain only QA groups for which all three QA items receive perfect scores (5/5/5) across the three semantic criteria (Correctness, Comparative Validity, and Reasoning Quality) under LLM-based evaluation. This conservative filtering strategy prioritizes semantic reliability over coverage and is

Metric	R1	R2	R3	R4	Human Mean	GPT-5 Mini Mean	MAD
Correctness	4.93	4.66	4.64	5.00	**4.79**	4.87	0.32
Comparative Validity	4.93	4.75	4.83	4.99	**4.83**	4.61	0.44
Reasoning Quality	4.87	4.71	4.67	5.00	**4.78**	4.37	0.56
Difficulty	1.89	2.58	3.12	2.66	**2.25**	2.17	0.57

Table 1: Human evaluation results on 300 sampled QA items using four annotators (R1–R4), together with LLM-as-a-Judge scores on the same subset. We report human mean scores and mean absolute deviation (MAD) to characterize aggregate-level agreement between human and LLM judgments.

intended to support high-precision benchmarking and supervised training.

Importantly, this step does not aim to simplify the task itself. As shown in Section 3.5, the filtered dataset preserves the original diversity of genres, topics, and cross-genre pairings. The resulting dataset contains **12,173** track pairs (**92.9%** of the original 13,097 pairs), forming the final released benchmark.

3.5 Dataset Statistics

Our final dataset contains **12,173** track pairs, each associated with three comparative questions, resulting in a total of **36,519** QA items. The dataset exhibits a strong dominance of cross-genre comparisons, encouraging fine-grained comparative reasoning beyond genre identity.

Human difficulty ratings collected during Stage 4 indicate that sentence-level questions are substantially more challenging than yes/no and short-answer formats, reflecting their requirement for multi-attribute comparison and coherent natural-language justification.

A detailed statistical analysis, including genre distributions, genre-pair frequencies, question-topic coverage, and question-type difficulty, is provided in Appendix E and Figure 5.

4 Baseline Benchmark

To assess the difficulty of Jamendo-MT-QA and provide reference points for future work, we benchmark several representative audio–language models on a sampled subset of the dataset. Our goal is not to exhaustively optimize model performance, but to characterize how existing models perform on multi-track comparative question answering under a unified evaluation protocol.

4.1 Baselines

Figure 3 summarizes our two baseline setups. Some models can directly process multiple audio inputs (*Multi*), while others operate on single-audio inputs and are evaluated via an intermediate caption representation (*Cap*). This design helps disentangle limitations of multi-audio perception from those of comparative reasoning over high-level semantics.

We evaluate the following baseline models: (1) **GPT-4o Audio** and **GPT-4o-mini Audio**(OpenAI et al., 2024), which accept multiple audio inputs and perform end-to-end audio–text reasoning; (2) **Qwen2-Audio** (Chu et al., 2024) and **Qwen3-Omni** (Xu et al., 2025), open-source multimodal models with multi audio input support; and (3) **Music Flamingo**, a strong audio–language model, evaluated in a caption-based comparative setting by comparing generated captions. Due to cost constraints, GPT-4o Audio and GPT-4o-mini Audio are evaluated on a randomly sampled subset of 2,010 track pairs, while other baselines follow the same evaluation protocol when applicable.

Caption-based Single-Audio Baselines. Not all audio–language models natively support multiple audio inputs. To enable a broader and more diagnostic comparison, we include caption-based single-audio baselines, where each track is first independently converted into a natural language description and comparative reasoning is performed purely in the text domain by an LLM.

This setup allows us to disentangle the contribution of direct multi-audio perception from that of semantic abstraction and comparative reasoning. By evaluating caption-based and multi-audio models under a unified protocol, we assess whether performance gains primarily stem from access to multiple audio inputs or from higher-level reasoning over structured representations.

4.2 Evaluation Metrics

We evaluate three types of questions separately. For *yes/no* questions, we report accuracy. For *short-answer* questions, accuracy is computed by ex-

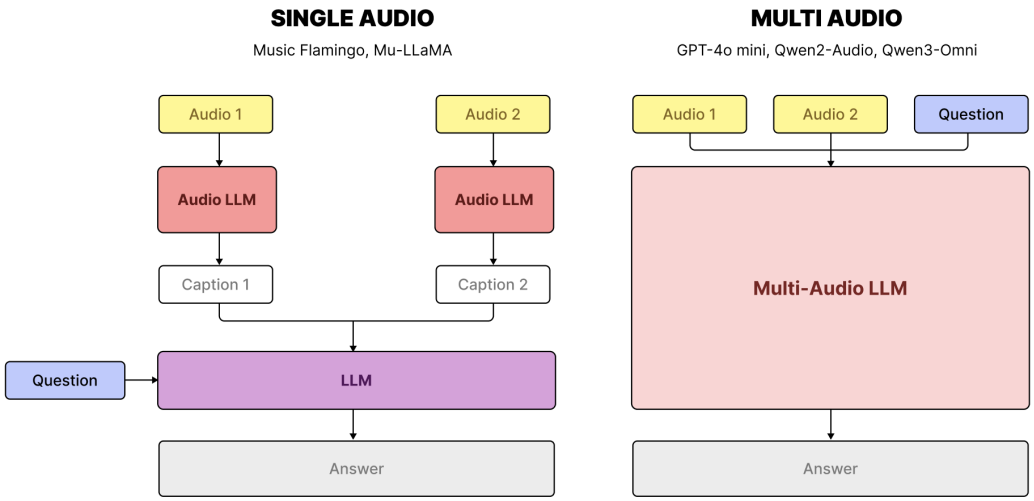

Figure 3: Baseline evaluation setups for Jamendo-MT-QA. **Left: caption-based single-audio baselines** (e.g., Music Flamingo, MU-LLaMA), where each track is independently captioned and an LLM performs comparative reasoning over the two captions plus the question. **Right: multi-audio baselines** (e.g., GPT-4o(-mini) Audio, Qwen2-Audio, Qwen3-Omni), which take two audio inputs and the question to produce an answer end-to-end.

act match between the predicted track identifier and the ground-truth answer. For *sentence-level* questions, we report BLEU (Papineni et al., 2002) and ROUGE-1,2,L (Lin, 2004) as surface-form similarity metrics based on n-gram overlap, and BERTScore (Zhang* et al., 2020) as a semantic similarity metric.

Since such metrics are known to be imperfect for open-ended comparative generation, we additionally employ an LLM-as-a-Judge protocol (Zheng et al., 2023) that assigns a 1–5 score to sentence-level answers based on correctness and comparative soundness. Full details of the LLM-as-a-Judge prompt and scoring rubric used for baseline evaluation are provided in Appendix D.

4.3 Results

We evaluate baseline models on Jamendo-MT-QA under two evaluation settings, summarized in Tables 2 and 3. To additionally assess models that support end-to-end multi-audio reasoning, we report results on a smaller subset of **2,010** track pairs that can be processed by all models, including GPT-based models.

Across models and evaluation settings, BLEU and ROUGE scores remain low, reflecting the difficulty of generating precise surface-form matches for open-ended comparative explanations. In contrast, BERTScore and LLM-as-a-Judge scores are substantially higher, indicating that models often produce semantically correct and comparative answers despite limited lexical overlap. This gap highlights the limitations of n-gram overlap metrics for evaluating sentence-level questions.

A notable result is that the caption-based **Music Flamingo** baseline achieves consistently strong performance across both evaluation settings. As shown in Tables 2 and 3, Music Flamingo attains the highest sentence-level BLEU, ROUGE, and LLM-as-a-Judge scores among caption-based models, and remains competitive with multi-audio systems on the subset of 2,010 track pairs. Despite operating on single-audio inputs, Music Flamingo benefits from high-quality captions that capture rich musical attributes, which support comparative reasoning when combined with an LLM.

Multi-audio models show mixed performance on yes/no and short-answer questions. In particular, **GPT-4o Audio** and **GPT-4o-mini Audio** achieve strong accuracy on the subset of 2,010 track pairs, while open-source multi-audio models exhibit more variable results. However, gains on sentence-level questions remain modest across all multi-audio models, suggesting that direct access to multiple audio inputs alone does not guarantee strong comparative explanation quality.

Across all models, performance is consistently higher on yes/no and short-answer questions than on sentence-level questions. This indicates that identifying the correct comparative direction is often easier than generating well-grounded natural

Model	Type	Yes/No	Short	Sentence-level						
		Acc.	Acc.	BLEU	R-1	R-2	R-L	BERT-F1	GPT	Claude
Music Flamingo	Cap	**77.4%**	**89.7%**	**4.00**	**27.5**	**4.6**	**24.7**	**0.8786**	**3.24**	**3.87**
Qwen2-Audio	Cap	37.4%	39.1%	1.88	19.3	1.8	17.4	0.8489	1.49	1.53
MU-LLaMA	Cap	20.6%	55.3%	2.39	23.8	3.2	16.5	0.8572	2.36	2.01
Qwen2-Audio	Multi	50.9%	80.2%	2.09	21.0	3.4	14.6	0.8472	1.37	1.62
Qwen3-Omni	Multi	**62.9%**	**80.3%**	**3.58**	**29.3**	**6.6**	**20.4**	**0.8632**	**3.11**	**3.48**

Table 2: Results on the full benchmark of 12,173 track pairs.

Model	Type	Yes/No	Short	Sentence-level						
		Acc.	Acc.	BLEU	R-1	R-2	R-L	BERT-F1	GPT	Claude
Music Flamingo	Cap	**82.1%**	**88.1%**	**2.06**	**25.3**	**5.9**	**19.6**	**0.875**	**3.25**	**3.97**
Qwen2-Audio	Cap	26.9%	55.5%	0.58	15.3	1.6	12.1	0.845	1.37	1.60
MU-LLaMA	Cap	23.6%	50.5%	0.68	16.1	1.9	13.3	0.853	1.92	2.13
Qwen2-Audio	Multi	34.4%	77.7%	0.57	14.3	1.4	11.3	0.845	1.63	1.75
Qwen3-Omni	Multi	59.7%	75.5%	**1.22**	19.9	**3.9**	16.0	0.860	2.57	3.25
GPT-4o-mini-Audio	Multi	**77.3%**	73.2%	0.87	17.8	2.7	14.7	0.859	2.92	3.14
GPT-4o-Audio	Multi	69.6%	**84.4%**	1.04	**20.0**	3.6	**16.7**	**0.870**	**3.17**	**3.57**

Table 3: Results on the subset of 2,010 track pairs.

language explanations that articulate perceptual differences across tracks. Sentence-level comparative reasoning therefore remains the primary bottleneck for current audio–language models.

Overall, these results show that performance on Jamendo-MT-QA is not determined solely by multi-audio perception. Instead, the quality of intermediate semantic representations and a model's ability to reason over them play a critical role, particularly for sentence-level comparative explanations. This finding is consistent with prior work in multimodal learning, which shows that strong abstraction and language alignment can partially compensate for limited perceptual access in complex reasoning tasks (Elizalde et al., 2022; Huang et al., 2022; Ghosh et al., 2025; Liu et al., 2023).

Key Observations. We draw three main conclusions from the baseline benchmark.

First, models consistently perform better on yes/no and short-answer questions than on sentence-level questions, indicating that identifying comparative direction is easier than generating explicit comparative explanations.

Second, caption-based single-audio models are not uniformly inferior to multi-audio models, highlighting the importance of representation quality and reasoning capability in comparative Music-QA.

Third, the large discrepancy between surface-form automatic metrics (BLEU/ROUGE) and LLM-as-a-Judge scores underscores the inadequacy of n-gram overlap metrics for evaluating comparative music explanations, motivating the use of semantically grounded evaluation protocols.

Model	Comp.Col.	Attr.Hal.	Gran.Mis.
Music Flamingo	35.3%	64.0%	0.7%
MU-LLaMA	37.3%	60.7%	2.0%
GPT-4o-Audio	56.7%	35.3%	8.0%
Qwen3-Omni	23.4%	75.9%	0.7%

Table 4: Error type distribution across models on sentence-level questions (LLM judge score < 3). Classified by GPT-4o-mini.

4.4 Error Analysis

To better understand the failure modes of current audio–language models, we conduct an error analysis on incorrectly answered *sentence-level* questions, combining qualitative categorization with a quantitative breakdown of dominant error types. Table 4 and Figure 4 summarize the distribution of dominant error types across models, computed over incorrectly answered sentence-level questions.

Error Categories. We define the following error types: **(A) Comparative Collapse**, where the model avoids explicit comparison and produces a generic summary of both tracks; **(B) Attribute Hallucination**, where the model introduces musical attributes not supported by the audio or captions; **(C) Granularity Mismatch**, where the compari-

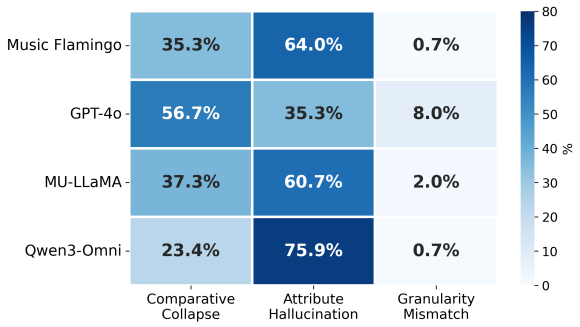

Figure 4: Heatmap visualization of error type distributions across models. Values indicate the percentage of each error category among incorrectly answered sentence-level questions (LLM judge score < 3).

son is too coarse or too fine-grained relative to the question intent.

Model-Specific Failure Patterns. We observe clear qualitative differences across models (Table 4; Figure 4). **Music Flamingo** primarily fails through *attribute hallucination*, suggesting that while the comparative structure is often preserved, factual grounding of fine-grained musical attributes remains challenging in caption-based settings. **MU-LLaMA** shows a similar pattern, but with a slightly higher rate of *comparative collapse*, indicating weaker comparative commitment even when high-level content is partially captured. In contrast, **GPT-4o-Audio** exhibits the highest rate of *comparative collapse*, frequently producing fluent but non-committal summaries that avoid explicit relational judgments. Finally, **Qwen3-Omni** is dominated by *attribute hallucination*, often introducing unsupported instrument or production details despite making explicit comparisons. Representative examples for each model illustrating these failure modes are provided in Appendix F.

Overall, these results highlight that Jamendo-MT-QA does not merely measure accuracy, but also exposes *qualitatively distinct failure modes* in multi-track music reasoning, underscoring its value as a diagnostic benchmark.

5 Conclusion

We introduced Jamendo-MT-QA, a multi-track comparative question answering dataset built on top of Jamendo-QA, targeting Music-QA systems that must reason about relationships between tracks rather than isolated audio clips. Our LLM-assisted pipeline generates three comparative questions per track pair, covering yes/no, short-answer, and sentence-level reasoning.

A key aspect of our dataset construction is the integration of human evaluation and LLM-as-a-Judge evaluation. By validating strong alignment between human judgments and LLM-based scores on a representative subset, we enable scalable quality control while maintaining consistency with human semantic evaluation. This approach allows us to filter and curate a large dataset without sacrificing annotation reliability.

In addition, we benchmark representative audio–language models on Jamendo-MT-QA. While current models achieve reasonable performance on yes/no and short-answer comparisons, sentence-level comparative generation remains challenging, especially under surface-form automatic metrics.

We believe Jamendo-MT-QA provides a valuable resource for future research on comparative music understanding, multi-track reasoning, and audio-language modeling.

Limitations

Our benchmark relies on LLM-generated comparative QA, which may inherit generator-specific biases despite human and judge-based filtering. Automatic metrics (BLEU/ROUGE) are limited for sentence-level comparative answers, and LLM-as-a-Judge introduces model-dependent scoring noise. Finally, some baselines require caption-based evaluation due to lack of multi-audio support, which may underestimate the potential of end-to-end audio reasoning models.

Ethical Considerations

Our dataset is derived from Creative Commons-licensed Jamendo audio, and we respect the original licensing conditions by releasing only annotations and metadata. While question generation is automated, we monitor and filter for potentially sensitive or biased content. Nevertheless, the dataset may reflect existing biases in genre labels, artist demographics, and descriptive language. We encourage practitioners to consider these factors when training or evaluating models on this resource.

Human Subjects. The human annotation study was conducted for research quality control on synthetic QA data and did not involve the collection of personal or sensitive information. Annotators provided informed consent, and the study was determined to be exempt from formal IRB review under institutional guidelines.

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. Musiclm: Generating music from text. *Preprint*, arXiv:2301.11325.

Anthropic. 2025. Claude 4.5 and claude 4.5 sonnet. https://www.anthropic.com/news/claude-sonnet-4-5.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The mtg-jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, United States.

Zheng Cai and 1 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Jian Chen, Wenye Ma, Penghang Liu, Wei Wang, Tengwei Song, Ming Li, Chenguang Wang, Jiayu Qin, Ruiyi Zhang, and Changyou Chen. 2025. Musixqa: Advancing visual music understanding in multimodal large language models. *Preprint*, arXiv:2506.23009.

Yiming Chen, Xianghu Yue, Xiaoxue Gao, Chen Zhang, Luis Fernando D'Haro, Robby T. Tan, and Haizhou Li. 2024. Beyond single-audio: Advancing multi-audio processing in audio large language models. *Preprint*, arXiv:2409.18680.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2017. FMA: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*.

Michaël Defferrard, Sharada P. Mohanty, Sean F. Carroll, and Marcel Salathé. 2018. Learning to recognize musical genre from audio. In *The 2018 Web Conference Companion*. ACM Press.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023a. Lp-musiccaps: Llm-based pseudo music captioning. *Preprint*, arXiv:2307.16372.

SeungHeon Doh, Minz Won, Keunwoo Choi, and Juhan Nam. 2023b. Toward universal text-to-music retrieval. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2022. Clap: Learning audio concepts from natural language supervision. *Preprint*, arXiv:2206.04769.

Gemma Team and 1 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sang gil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Music flamingo: Scaling music understanding in audio language models. *Preprint*, arXiv:2511.10289.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *Preprint*, arXiv:2406.11768.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. 2022. Mulan: A joint embedding of music audio and natural language. *Preprint*, arXiv:2208.12415.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.

Junyoung Koh, Soo Yong Kim, Gyu Hyeong Choi, and Yongwon Choi. 2025a. Aiba: Attention-based instrument band alignment for text-to-audio diffusion. *Preprint*, arXiv:2509.20891.

Junyoung Koh, Soo Yong Kim, Yongwon Choi, and Gyu Hyeong Choi. 2025b. Jamendo-qa: A large-scale music question answering dataset. *Preprint*, arXiv:2509.15662.

Edith Law, Kris West, Michael I. Mandel, Mert Bay, and J. S. Downie. 2009. Evaluation of algorithms using games: The case of music tagging. In *International Society for Music Information Retrieval Conference*.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ruosen Li, Zimu Wang, Son Quoc Tran, Lei Xia, and Xinya Du. 2024. MEQA: A benchmark for multi-hop event-centric question answering with explanations. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Cheng Liu, Hui Wang, Jinghua Zhao, Shiwan Zhao, Hui Bu, Xin Xu, Jiaming Zhou, Haoqin Sun, and Yong Qin. 2025. Musiceval: A generative music dataset with expert ratings for automatic text-to-music evaluation. *Preprint*, arXiv:2501.10811.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2023. Music Understanding LLaMA: Advancing Text-to-Music Generation with Question Answering and Captioning. *arXiv preprint arXiv:2308.11276*.

Shansong Liu, Atin Sakkeer Hussain, Qilong Wu, Chenshuo Sun, and Ying Shan. 2024. Mumu-llama: Multimodal music understanding and generation via large language models. *arXiv preprint arXiv:2412.06660*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Zhihao Ouyang, Ju-Chiang Wang, Daiyu Zhang, Bin Chen, Shangjie Li, and Quan Lin. 2025. Mqad: A large-scale question answering dataset for training music large language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Abhinaba Roy, Renhang Liu, Tongyu Lu, and Dorien Herremans. 2025. Jamendomaxcaps: A large scale music-caption dataset with imputed metadata. *arXiv:2502.07461*.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, and 11 others. 2023. Audiopalm: A large language model that can speak and listen. *Preprint*, arXiv:2306.12925.

Julian Schnitzler, Xanh Ho, Jiahao Huang, Florian Boudin, Saku Sugawara, and Akiko Aizawa. 2024. Morehopqa: More than multi-hop reasoning. *Preprint*, arXiv:2406.13397.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. *Preprint*, arXiv:2310.13289.

G. Tzanetakis and P. Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Mengyue Wu, Heinrich Dinkel, and Kai Yu. 2019. Audio caption: Listen and tell. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 830–834.

Shangda Wu, Guo Zhancheng, Ruibin Yuan, Junyan Jiang, SeungHeon Doh, Gus Xia, Juhan Nam, Xiaobing Li, Feng Yu, and Maosong Sun. 2025. CLaMP 3: Universal music information retrieval across unaligned modalities and unseen languages. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2605–2625, Vienna, Austria. Association for Computational Linguistics.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings*

of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, Liumeng Xue, Ziyang Ma, Qin Liu, Tianyu Zheng, Yizhi Li, Yinghao Ma, Yiming Liang, Xiaowei Chi, Ruibo Liu, and 13 others. 2024. ChatMusician: Understanding and generating music intrinsically with LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6252–6271, Bangkok, Thailand. Association for Computational Linguistics.

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, and 1 others. 2023. Marble: Music audio representation benchmark for universal evaluation. *Advances in Neural Information Processing Systems*, 36:39626–39647.

Yongyi Zang, Sean O'Brien, Taylor Berg-Kirkpatrick, Julian J. McAuley, and Zachary Novack. 2025. Are you really listening? boosting perceptual awareness in music-qa benchmarks. *CoRR*, abs/2504.00369.

Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. 2021. MusicBERT: Symbolic music understanding with large-scale pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 791–800, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Human and LLM Evaluation Criteria

This appendix provides the full descriptions of the five evaluation criteria used in the human evaluation and the four criteria used in the LLM-as-a-Judge evaluation.

A.1 Human Evaluation Criteria

We conduct human evaluation in two separate stages. In **Stage 1**, we verify the *audio–text alignment* of the enriched captions generated by Music Flamingo, where annotators have access to the underlying audio. In **Stage 4**, we use a multi-criteria rubric to evaluate and filter comparative QA items based on the provided text descriptions of Track A and Track B.

Annotator	Mean	Std
Annotator A	4.43	0.65
Annotator B	4.47	0.62
Annotator C	4.60	0.71
Annotator D	4.85	0.48
Average	**4.59**	–

Table 5: Human evaluation results for audio–text alignment of enriched captions in Stage 1, reported on a 1–5 scale.

Stage 1: Audio-Text Alignment of Enriched Captions (Music Flamingo) To ensure that the enriched captions faithfully describe the actual music content, four human annotators listen to the audio and rate the caption-level alignment on a **1–5 scale**: *"Does the enriched caption accurately align with the musical content of the audio (e.g., instrumentation, vocals, tempo, mood, and production cues)?"* Scores range from **1 (complete mismatch)** to **5 (perfect alignment)**, with intermediate values (2–4) reflecting partial correctness and varying degrees of coverage and specificity. Given the consistently high alignment scores and low variance across annotators, we found that this evaluation size was sufficient to validate caption quality for downstream QA generation.

Table 5 summarizes the results of the human evaluation for audio–text alignment of enriched captions. Across 1200 individual ratings, the captions achieve a high overall average score of **4.59**, indicating strong consistency between the generated textual descriptions and the underlying music audio. The equal-weight average across annotators (4.59) further confirms that this alignment is consistent and not dominated by a single evaluator. These results suggest that the enriched captions provide a reliable semantic foundation for subsequent QA generation stages.

Stage 4: Comparative QA Evaluation and Filtering Criteria For the evaluation and filtering in Stage 4, annotators rate each model answer (or candidate QA item, depending on the filtering stage) on a **1–5 scale** according to the following dimensions, grounded in the provided textual descriptions of Track A and Track B (i.e., captions/metadata,

without requiring direct audio access):

(1) Correctness. *"Is the answer factually correct based on the provided descriptions of Track A and Track B?"* Scores range from **1 (completely incorrect)** to **5 (completely correct)**, with 2–4 indicating partially correct answers (e.g., correct direction but missing key evidence).

(2) Comparative Validity. *"Does the answer make a valid comparative statement between Track A and Track B?"* Scores range from **1 (no comparison or invalid comparison)** to **5 (excellent comparative statement)**, with 2–4 capturing weak, ambiguous, or only partially comparative responses.

(3) Reasoning Quality. *"Is the reasoning coherent, logically grounded in the provided descriptions, and properly supporting the answer?"* Scores range from **1 (incoherent/illogical)** to **5 (excellent, well-supported reasoning)**, with intermediate values reflecting partially grounded reasoning or missing justification.

(4) Difficulty. *"How difficult is the question to answer correctly?"* Scores range from **1 (very easy)** to **5 (very difficult)**, where higher difficulty indicates that answering requires subtle distinctions, multi-attribute comparison, or non-trivial inference from the descriptions.

A.2 LLM-as-a-Judge Criteria

The automatic evaluation follows the same rubric as the human evaluation, with the exception of the Audio-Text Alignment, which cannot be reliably assessed by text-only LLM judges without access to the underlying audio.

Accordingly, LLM evaluators score each answer using the following four dimensions: **Correctness**, **Comparative Validity**, **Reasoning Quality**, and **Difficulty**. Each criterion is rated on a **1–5 scale**, where intermediate scores (2–4) represent partial correctness, weak or ambiguous comparisons, or varying degrees of reasoning quality and difficulty.

B Output Schema for Comparative QA Generation

The following listing shows the JSON output format used during Stage 3 comparative QA generation. For each track pair, the model generates three comparative question–answer pairs: yes/no, short-answer, and sentence-level.

```
{
  "audio1": "{audio1}",
  "audio2": "{audio2}",
  "qa_pairs": [
    {
      "type": "yes_no",
      "question": "...",
      "reasoning": "...",
      "answer": "yes or no"
    },
    {
      "type": "short_answer",
      "question": "...",
      "reasoning": "...",
      "answer": "{audio1} or {audio2}"
    },
    {
      "type": "sentence",
      "question": "...",
      "reasoning": "...",
      "answer": "complete comparative sentence"
    }
  ]
}
```

C Prompt Template for Multi-Track QA Generation

Listing 1 shows the prompt template used for multi-track comparative QA generation in Stage 3. The prompt explicitly enforces comparative constraints and a fixed three-question structure (yes/no, short-answer, and sentence-level) for each audio pair.

Listing 1: Prompt template for multi-track comparative QA generation.

```
You are a music comparison expert. Given two
    music tracks in JSON format,
you must generate exactly 3 COMPARATIVE question-
    answer pairs that directly
compare both tracks.

IMPORTANT: Every question MUST compare both
    tracks.
Do NOT ask about a single track alone.

Music Track 1:
{music1_str}

Music Track 2:
{music2_str}

Requirements:
1. YES/NO question - Must compare an attribute
    between both tracks
2. short-answer question - Answer should be the
    audio name
3. SENTENCE question - Detailed comparison in a
    complete sentence
```

D LLM-as-a-Judge Prompt and Scoring Rubric

D.1 Judge Prompt for Sentence-level Evaluation

For sentence-level comparative questions, we employ an LLM-as-a-Judge protocol to evaluate se-

mantic correctness and comparative soundness. The following prompt is used to score model predictions on a 0–5 scale.

```
You are an expert evaluator for music QA.
Compare the prediction with the ground truth
    answer.

## Question
{question}

## Ground Truth Answer
{ground_truth}

## Model Prediction
{prediction}

## Scoring Criteria (0-5 scale)
- 0: Completely wrong or irrelevant
- 1: Mostly wrong with very minor correct
    elements
- 2: Partially correct but missing key points
- 3: Correct on main points but missing details
    or has minor errors
- 4: Mostly correct with only minor omissions
- 5: Fully correct, captures all key information

Evaluate semantic similarity and correctness.
The prediction does not need to use exact
    wording.

Focus on:
- genre comparison
- tempo or energy
- vocal characteristics
- production style
- mood

Respond with ONLY a JSON object:
{"score": <0-5>, "explanation": "<brief reason>"}
```

D.2 Scoring Rubric

For sentence-level comparative questions, we evaluate model outputs using a graded semantic scoring scheme that captures both correctness and comparative soundness beyond surface-form overlap. Table 6 summarizes the 0–5 rubric used by LLM judges.

D.3 Evaluation by Question Type

Different question types are evaluated using task-appropriate strategies:

E Statistical Analysis of the Dataset

This section provides a statistical analysis of the Jamendo Multi-Track comparative QA dataset, focusing on genre–genre pair composition, track-level genre distribution, same-versus-different genre comparisons, and question topic diversity.

Score	Description
0	Completely wrong or irrelevant
1	Mostly wrong with very minor correct elements
2	Partially correct but missing key points
3	Correct on main points but missing details or minor errors
4	Mostly correct with only minor omissions
5	Fully correct, capturing all key comparative information

Table 6: Scoring rubric used by LLM judges for sentence-level comparative answers.

Question Type	Evaluation Method	Score Range
Yes/No	Rule-based matching	0 or 1
Short-answer	Rule-based matching (Track A/B)	0 or 1
Sentence-level	LLM-as-a-Judge	0–5
Open-ended	LLM-as-a-Judge	0–5

Table 7: Evaluation methods used for different question types.

E.1 Basic Statistics

Table 8 summarizes the fundamental properties of the dataset. Each audio pair is associated with exactly three comparative QA items, resulting in a total of 36,519 questions.

Statistic	Value
Total music pairs	12,173
Total QA pairs	36,519
Unique genre pairs	789

Table 8: Basic statistics of the multi-track comparative QA dataset.

E.2 Track-Level Genre Distribution

Table 9 presents the most frequent genres across all tracks participating in pairwise comparisons. The distribution is dominated by *rock*, *indie*, and *pop*, which is consistent with the characteristics of Creative Commons music collections such as Jamendo.

E.3 Genre Pair Combinations

Table 10 lists the most frequent genre–genre combinations among the 12,173 track pairs. While same-genre comparisons such as *rock–rock* and *indie–indie* appear regularly, most pairs span different genres, encouraging cross-style comparative reasoning.

Figure 5 provides a complementary visualization of the genre structure in Jamendo-MT-QA. The inner ring illustrates the marginal genre distribution of individual tracks, confirming that the dataset spans a broad range of popular music styles, including rock, indie, pop, folk, electronic, rap, and

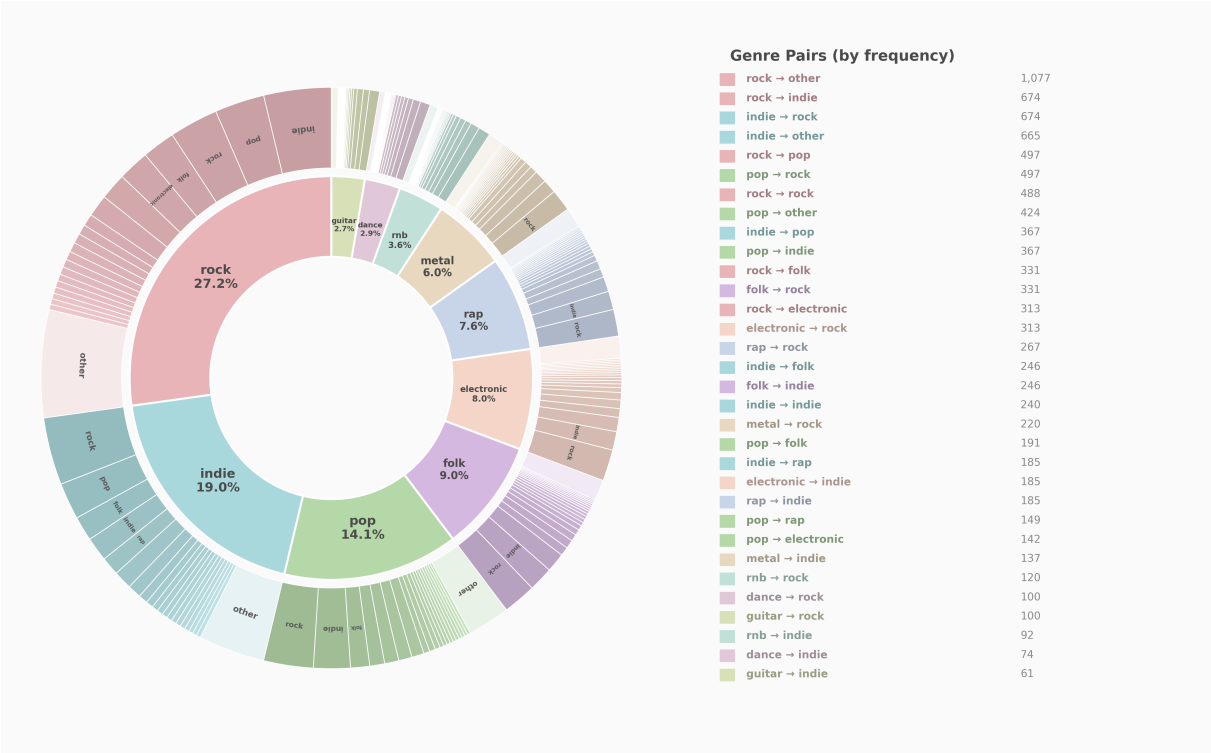

Figure 5: Genre composition and genre-pair distribution in Jamendo-MT-QA. The inner ring shows the marginal distribution of individual genres across all tracks, while the outer ring visualizes the frequency of genre–genre pairings used for comparative question generation. For clarity, only the top-10 most frequent genres are shown in the visualization; therefore, percentages may differ from table-based statistics computed over all tracks.

Genre	Count	Ratio
rock	4,890	20.09%
indie	3,424	14.06%
pop	2,527	10.38%
folk	1,611	6.62%
electronic	1,446	5.94%
rap	1,358	5.58%
metal	1,072	4.40%
rnb	655	2.69%
dance	515	2.12%
guitar	479	1.97%

Table 9: Top genres across all tracks in the dataset.

metal.

The outer ring depicts the distribution of genre–genre pairs used for comparative question construction. Consistent with the statistics reported in Table 10, the majority of comparisons involve cross-genre pairs rather than same-genre pairs. This design choice encourages models to rely on fine-grained musical attributes (e.g., instrumentation, production, tempo, or vocal characteristics) rather than genre identity alone.

Overall, this visualization highlights the structural diversity of the dataset and supports the claim that Jamendo-MT-QA systematically targets *cross-genre comparative reasoning* rather than trivial within-genre comparisons.

Genre 1	Genre 2	Count
indie	rock	674
pop	rock	497
rock	rock	488
indie	pop	367
folk	rock	331
electronic	rock	313
rap	rock	267
folk	indie	246
indie	indie	240
metal	rock	220

Table 10: Most frequent genre–genre pairs in the dataset.

E.4 Same vs. Different Genre Pairs

Table 11 quantifies the proportion of same-genre and different-genre track pairs. A large majority of pairs (over 91%) involve different genres, ensuring that models must rely on fine-grained musical attributes rather than genre identity alone.

Category	Count	Ratio
Same-genre pairs	1,065	8.75%
Different-genre pairs	11,108	91.25%

Table 11: Distribution of same-genre and different-genre track pairs.

E.5 Question Topic Distribution

Table 12 summarizes the semantic topics addressed by the comparative questions. The dataset emphasizes perceptually salient musical attributes such as vocals, instrumentation, production, and tempo, while also covering higher-level concepts including key, mood, genre, and structural properties.

Topic	Count	Ratio
vocal	15,509	42.47%
tempo	13,509	36.99%
instrument	13,384	36.65%
production	12,301	33.68%
key	11,086	30.36%
mood	4,646	12.72%
genre	3,203	8.77%
structure	594	1.63%

Table 12: Distribution of question topics across the dataset.

Overall, the analysis confirms that the dataset exhibits substantial diversity in both genre combinations and question semantics. The dominance of cross-genre pairs (91.25%) and the wide coverage of musical attributes highlight the dataset's suitability for evaluating multi-track comparative reasoning in music understanding.

E.6 Question-Type Difficulty

To provide additional context on the relative hardness of different QA formats, we report human-rated difficulty statistics for each question type. Difficulty is rated on a 1–5 Likert scale during Stage 4 evaluation, where higher scores indicate that answering requires more subtle cross-track distinctions, multi-attribute comparison, or non-trivial inference from the provided descriptions. As shown in Table 13, sentence-level questions are substantially more difficult than yes/no and short-answer questions, supporting our claim that open-ended comparative explanation is the primary bottleneck for current models.

Question Type	Count	Avg. Difficulty
Yes/No	12,173	2.5
Short-Answer	12,173	2.5
Sentence-Level	12,173	2.9
Total	36,519	2.6

Table 13: Distribution of question types and their average difficulty scores from Stage 4 human evaluation. Difficulty is reported on a 1–5 scale, where higher values indicate more challenging comparative reasoning.

F Quantitative Error Distribution

This section provides implementation details for the quantitative error analysis reported in Section 4.4.

Scope. We analyze *incorrectly answered* sentence-level questions, defined as items whose LLM-as-a-Judge score is below 3. For each model, percentages are computed by normalizing over the number of incorrect sentence-level items for that model, so the error-type proportions sum to 100%.

Primary Error Labeling. Each incorrect item is assigned a *primary* error type among the four categories in Section 4.4. We use GPT-4o-mini to classify the dominant failure mode based on the question, reference answer, and the model's prediction. When multiple issues co-occur (e.g., collapse *and* hallucination), the classifier is instructed to select the most salient error that explains the final answer's failure.

F.1 Representative Error Examples

Table 15 provides representative failure cases for each model, illustrating typical breakdown patterns across the four error categories.

G Qualitative Comparison of Stage 3 Generators

In Stage 3, we experimented with multiple LLM generators for comparative QA construction before selecting **GPT-5 mini** as the primary generator. To illustrate qualitative differences, Figures 6, 7, 8, and 9 show the three-question QA group (yes/no, short-answer, sentence-level) generated for the *same* audio pair across four alternative models. Red highlights in these figures mark recurring issues that motivated our decision to not use these models for final dataset construction.

Qwen3-32B. We observed that the short-answer question sometimes yields uncommon or subjective word choices (e.g., stylistic adjectives rather than a strict selection), which can break the intended answer format. In addition, sentence-level outputs

Model	Temp.	Top-p	Top-k	Max Tokens
MU-LLaMA	0.6	0.8	–	512
Music Flamingo	0.0 (greedy)	–	–	512
Qwen3-Omni	0.01	0.1	1	256
GPT-4o-mini	0.0 (greedy)	–	–	512

Table 14: Decoding parameters used for baseline inference and LLM-as-a-Judge evaluation.

often mirror the reasoning text with minimal reformulation, reducing diversity between *reasoning* and *answer*.

InternLM3-8B. The model frequently produces reasoning that closely repeats the question prompt, and we occasionally observed incorrect selections in the short-answer type, indicating weaker grounding and higher error rates for comparative decisions.

Gemma-3-12B. While overall quality is competitive, the reasoning is sometimes less detailed or less step-wise than desired, which reduces interpretability and usefulness for downstream training or analysis.

Claude-4.5-Haiku. This model generally produces strong comparative QA. However, **GPT-5 mini** more consistently provides step-by-step, grounded reasoning, which we found preferable for reliable dataset construction and for using the reasoning field as a high-quality supervisory signal.

H Baseline Inference Settings

H.1 Generation Parameters

Table 14 summarizes decoding parameters used for baseline inference. Unless otherwise noted, we use the model's default tokenizer and stop conditions.

H.2 Prompt Templates

We evaluate two baseline setups (Figure 3): (i) *caption-based* baselines, where each track is first converted into a textual description and an LLM answers comparative questions using the two captions; and (ii) *multi-audio* baselines, where models directly process two audio inputs along with the question. To reduce prompt sensitivity, we use a single standardized template per setup, with only minimal model-specific formatting (e.g., special tokens for audio inputs).

H.2.1 Caption-based comparison prompt.

```
You are a music expert. Based on the following
    descriptions of two music tracks, answer the
     question.

=== Track A: {audio1_name} ===
{caption1}

=== Track B: {audio2_name} ===
{caption2}

Question: {question}

Answer:
```

H.2.2 Multi-audio comparison prompt.

```
[Audio 1] [Audio 2]
You are given two music tracks.
Track A: {audio1_name}
Track B: {audio2_name}

Question: {question}

Please compare the two tracks and answer the
    question concisely.
```

Caption prompts. For caption-based models, we prompt each model to produce a detailed single-track description. For Music Flamingo, we request structured coverage (genre, tempo, key, instrumentation, vocals, production, mood). For MU-LLaMA, we use a concise captioning instruction ("Describe this music in detail.") and then apply the standardized caption-based comparison prompt above.

Qwen3-32B

yes no

Question:
Do both tracks use melancholic moods and atmospheric production elements?

Reasoning :
Both tracks are described as melancholic and use atmospheric production elements (e.g., synth pads and reverb) to enhance their emotional and sonic characteristics.

Answer :
yes

short answer

Question :
Which track has a more **subdued** tempo?

Reasoning :
Track 2 has a slower tempo of ≈81 BPM, compared to Track 1's 125 BPM, making it more subdued in rhythmic intensity.

Answer :
pop_WhentheSnowfalls

pop_LonesomeRoad **pop_whentheSnowfalls**

sentence

Question:
How do the instrumentation styles of pop_LonesomeRoad and pop_WhentheSnowfalls differ in their approaches to creating mood?

Reasoning :
pop_LonesomeRoad emphasizes acoustic elements and intimate vocals to create a wistful, personal mood, while pop_WhentheSnowfalls relies on lush synth textures, electronic drum kits, and ethereal reverb to create a dreamy and atmospheric mood.

Answer :
pop_LonesomeRoad uses acoustic guitar and intimate vocals for a personal, wistful mood, while pop_WhentheSnowfalls uses electronic synths and reverberant textures for a dreamy, atmospheric mood.

Figure 6: Qualitative example of Stage 3 comparative QA generation using **Qwen3-32B**. The short-answer questions occasionally contain uncommon or subjective word choices, and sentence-level answers often closely mirror the reasoning text. Red highlights mark cases where reasoning and answers are nearly identical, reducing diversity and usefulness for training.

InternLM3 - 8B

yes no

Question:
Do both tracks utilize atmospheric synth pads to create a spacious backdrop?

Reasoning :
Comparing the use of atmospheric elements in both tracks.

Answer :
yes

short answer

Question :
What is the difference in the dominant vocal range between the lead vocalists of both tracks?

Reasoning :
Comparing the vocal characteristics of the two tracks.

Answer :
tenor vs baritone

pop_LonesomeRoad **pop_whentheSnowfalls**

sentence

Question:
How do the central lyrical themes of both tracks differ in their emotional expression?

Reasoning :
Comparing the emotional and lyrical content of the two tracks.

Answer :
While 'Lonesome Road' focuses on heartbreak and longing, 'When the Snow Falls' emphasizes yearning and hope during a snowy season.

Figure 7: Qualitative example of Stage 3 comparative QA generation using **InternLM3-8B**. The model frequently produces reasoning that closely resembles the question itself, and we observe occasional incorrect answers in the short-answer type. These issues indicate weaker grounding and reliability for comparative QA generation. Red highlights denote incorrect or weakly supported outputs.

Gemma 3-12B

pop_LonesomeRoad **pop_whentheSnowfalls**

yes no

Question:
Does both tracks utilize reverb and delay to enhance their atmosphere?

Reasoning :
Both descriptions explicitly mention the use of reverb and delay to create a spacious or dreamy atmosphere.

Answer :
yes

short answer

Question :
What is the tempo difference between the two tracks?

Reasoning :
The tracks have significantly different tempos - 125 BPM versus approximately 81 BPM.

Answer :
44

sentence

Question:
Considering the lyrical themes, which track focuses more on the aftermath of a relationship and which focuses on longing for someone during a specific season?

Reasoning :
LonesomeRoad's lyrics directly address heartbreak and moving on, while WhentheSnowfalls centers around longing for a loved one during winter.

Answer :
LonesomeRoad focuses on the aftermath of a relationship, while WhentheSnowfalls centers on longing during a specific season.

Figure 8: Qualitative example of Stage 3 comparative QA generation using **Gemma 3 12B**. The model generally produces reasonable questions and answers; however, the reasoning is sometimes less detailed or less step-wise, which reduces interpretability for dataset construction and downstream supervision. Red highlights indicate parts where the reasoning lacks sufficient specificity.

Claude Haiku 4.5

pop_LonesomeRoad **pop_whentheSnowfalls**

yes no

Question:
Is the tempo of 'When the Snow Falls' faster than the tempo of 'Lonesome Road'?

Reasoning :
When the Snow Falls' has a tempo of approximately 81 BPM, while 'Lonesome Road' moves at 125 BPM. Since 81 BPM is slower than 125 BPM, the answer is no.

Answer :
No

short answer

Question :
Which track features an acoustic guitar as the primary instrument carrying the core chordal rhythm?

Reasoning :
'When the Snow Falls' is built around expansive synth pads and synth bass with electronic drums, with no mention of acoustic guitar. 'Lonesome Road' explicitly features a warm, finger-picked acoustic guitar that carries the core chordal rhythm. Therefore, the answer is 'Lonesome Road'.

Answer :
pop_LonesomeRoad

sentence

Question:
How do the lyrical themes and emotional focal points differ between these two tracks?

Reasoning :
When the Snow Falls' focuses on longing for a loved one during winter with themes of anticipation and reunion (seasonal romance and distance metaphor). 'Lonesome Road' explores heartbreak and moving on from a past relationship with themes of persistent longing and difficulty letting go. Both are melancholic but address different emotional narratives - one about unrequited or distant love, the other about loss and struggling to move forward.

Answer :
When the Snow Falls' centers on yearning and anticipation for reunion with a distant loved one using winter imagery, while 'Lonesome Road' explores the pain of heartbreak and the difficulty of moving on from a past relationship despite others' advice.

Figure 9: Qualitative example of Stage 3 comparative QA generation using **Claude 4.5 Haiku**. The model typically produces strong comparative questions and fluent sentence-level answers. However, compared to GPT-5 mini, the reasoning is less consistently structured in explicit step-by-step form, which motivated our preference for GPT-5 mini in final dataset construction.

Model	Error Type	Question (Sentence-level)	Reference	Model Output
GPT-4o-Audio	**(A)** Comp.Col.	In one sentence, how do the vocal style and emotional tone differ between Track A and Track B?	Track A features a restrained, airy vocal delivery with a more melancholic tone, while Track B has a fuller vocal presence and a warmer, more uplifting feel.	Track A and Track B both present expressive vocals and emotional atmospheres. Track A leans into a softer, intimate delivery, while Track B maintains a similarly emotive style but with a slightly different texture overall.
GPT-4o-Audio	**(B)** Attr.Hal.	In one sentence, compare the rhythmic groove and instrumentation between Track A and Track B.	Track A has a steady mid-tempo groove with prominent electric guitars and live drums, whereas Track B uses a lighter rhythmic feel with more electronic percussion and synth layers.	Track A features a hip-hop inspired beat with heavy bass and sampled percussion, while Track B relies on acoustic strumming and orchestral strings, creating a more cinematic rhythm section.
Music Flamingo	**(B)** Attr.Hal.	In one sentence, how do the primary instruments differ between Track A and Track B?	Track A is driven by electric guitars and live drums, whereas Track B emphasizes synth textures and programmed percussion.	Track A is led by piano and orchestral strings, while Track B centers on brass sections and saxophone riffs.
Music Flamingo	**(A)** Comp.Col.	In one sentence, compare the mood and atmosphere of Track A and Track B.	Track A feels darker and more introspective, while Track B is brighter and more upbeat.	Both tracks create strong atmospheres and convey distinctive moods through their instrumentation and arrangement.
MU-LLaMA	**(A)** Comp.Col.	How do the key, tempo, instrumentation, and vocal delivery of each track create their distinct emotional atmospheres?	Track A uses contrasting musical elements to produce a dramatic, cinematic feel, while Track B creates an uplifting trance atmosphere.	Both Track A and Track B share an upbeat and energetic quality, but they create distinct emotional atmospheres through their instrumentation and vocal delivery.
MU-LLaMA	**(B)** Attr.Hal.	In one sentence, how do the two tracks differ in genre, tempo, vocal timbre, and production approach?	Track A is an introspective, guitar-driven Indie/Alternative Rock song at about 100 BPM with a melancholic male tenor.	Track A has an indie genre with a weird and tribal feel, likely slower in tempo, while Track B is an upbeat electronic fusion of traditional Indian and Western styles.
Qwen3-Omni	**(A)** Comp.Col.	In one sentence, how do the two tracks compare in terms of overall mood, instrumentation, and lyrical themes?	Track A is an upbeat, polished indie-folk pop track with lyrics about adventure, while Track B is a reflective acoustic piece about personal loss.	While both tracks feature a prominent ukulele and have an uplifting, positive mood, they differ significantly in theme and instrumentation.
Qwen3-Omni	**(B)** Attr.Hal.	Compare how tempo, harmonic complexity, production style, and instrumentation differ between the two tracks.	Track A sits around 120 BPM with shifting feels and distorted guitars, while Track B is a slower, melodic rock track at 85 BPM.	Track A is a slow, atmospheric piece with ambient textures, while Track B is a fast-paced electronic track with aggressive synths.

Table 15: Representative error examples for sentence-level comparative questions. Each row shows one incorrectly answered item (LLM-judge score < 3), together with its assigned error type. Comp.Col.: Comparative Collapse; Attr.Hal.: Attribute Hallucination.