BASS 𝄞: BENCHMARKING AUDIO LMS FOR MUSICAL STRUCTURE AND SEMANTIC REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Music understanding is a complex task that often requires reasoning over both structural and semantic elements of audio. We introduce BASS, designed to evaluate music understanding and reasoning in audio language models across four broad categories: structural segmentation, lyric transcription, musicological analysis, and artist collaboration. BASS comprises 2658 questions spanning 12 tasks, 1993 unique songs and covering over 138 hours of music from a wide range of genres and tracks, crafted to assess musicological knowledge and reasoning in real-world scenarios. We evaluate 14 open-source and frontier multimodal LMs, finding that even state-of-the-art models struggle on higher-level reasoning tasks such as structural segmentation and artist collaboration, while performing best on lyric transcription. Our analysis reveals that current models leverage linguistic priors effectively but remain limited in reasoning over musical structure, vocal, and musicological attributes. BASS provides an evaluation framework with widespread applications in music recommendation and search and has the potential to guide the development of audio LMs.

1 INTRODUCTION

Audio LMs have advanced rapidly in their ability to understand and reason about speech, environmental sounds, and music (Goel et al., 2025; Ghosh et al., 2025b; Gemini Team et al., 2023; KimiTeam et al., 2025; Xu et al., 2025; Xie et al., 2025; Wijngaard et al., 2025). Much of this progress has been shaped by benchmarks that define the tasks models are ultimately trained to optimize (Huang et al., 2025). In music, however, existing benchmarks remain limited. Many treat music as one audio domain among many (Huang et al., 2025; Ahia et al., 2025; Sakshi et al., 2025; Kumar et al., 2025; Ma et al., 2025b), while benchmarks focused on music often rely on short excerpts or caption-derived tasks (Weck et al., 2024; Agostinelli et al., 2023; Manco et al., 2023). These design choices leave an important evaluation gap. Many musically meaningful reasoning tasks require access to complete recordings rather than short excerpts. Identifying repeated sections, localizing transitions, attributing multiple vocalists across time, or reasoning about higher-level musical attributes often requires access to the full temporal context of a recording. It is unclear if improvements on short-context or constrained multiple-choice benchmarks translate to robust music understanding in the form users naturally engage with. For example, MuChoMusic (Weck et al., 2024) constructs multiple-choice questions from captioning datasets and uses predominantly short clips, limiting evaluation of long-range musical structure. Related efforts such as CMI-Bench (Ma et al., 2025a) and MusicQA (Liu et al., 2023) broaden task coverage, but focus primarily on classical music information retrieval, captioning, or general question answering rather than long-form reasoning over complete recordings.

We introduce BASS, a benchmark for evaluating music understanding and reasoning in audio LMs over complete musical recordings. As illustrated in Figure 1, BASS comprises 12 tasks spanning four categories. **Structural segmentation** (§2) requires identifying and localizing musical sections within a recording. **Structural lyric transcription** (§2) evaluates transcription under structural constraints and at varying levels of granularity. **Musicological analysis** (§2) probes reasoning over abstract musical attributes. **Artist collaboration** (§2) requires identifying, attributing, and temporally localizing multiple vocalists within a song. The tasks are designed around practical music-analysis queries that naturally arise in musicology, production, metadata creation, and retrieval, and whose solutions require full-track temporal contexts.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

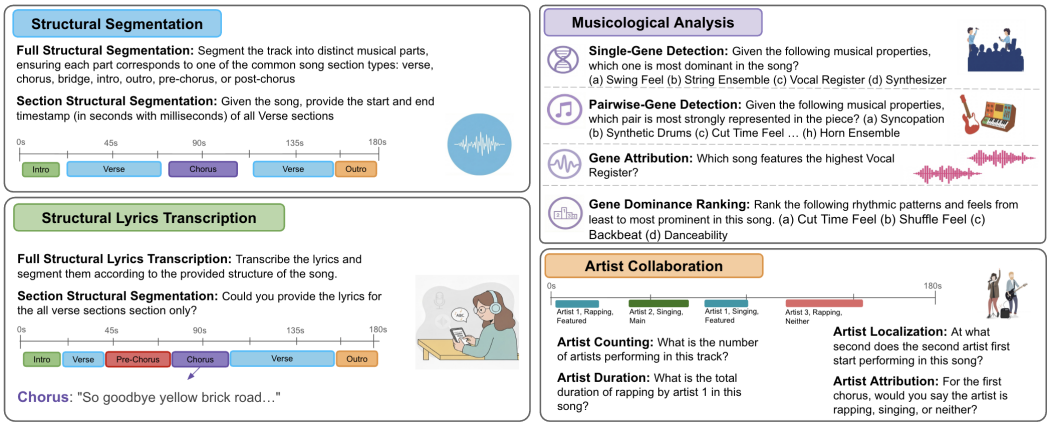

Figure 1: Overview of the tasks included in BASS, designed to evaluate audio LMs on musical understanding requiring long-term structure or hierarchical reasoning. The benchmark contains twelve tasks across four categories: structural segmentation, structural lyrics transcription, musicological analysis, and artist collaboration.

Using BASS, we evaluate 14 audio LMs. Results show that even frontier systems such as Gemini 2.5 Pro and strong open-source models such as Qwen3-Omni struggle substantially on several tasks. Models perform relatively well on lyrics transcription, where linguistic priors can often be exploited, but remain substantially weaker on tasks requiring temporal structure, multi-artist attribution, and higher-level musical reasoning. These findings suggest that current audio LMs remain limited in long-context musical understanding. BASS provides a challenging testbed for measuring this gap and for guiding future research on long-form music understanding and reasoning in audio LMs.

2 BASS BENCHMARK

BASS evaluates long-form music understanding and reasoning in contemporary music, focusing on tasks that require models to integrate information across time, musical sections, performers, and musical attributes. We organize the benchmark around four questions about a complete recording: *where* are its structural sections **(structural segmentation)**, *what* is sung in each section **(structural lyrics transcription)**, *how* does it sound **(musicological analysis)**, and *who* performs, when, and in what vocal style **(artist collaboration)**. It comprises 2,658 questions across 12 tasks over 138 hours of music spanning eight genres and nine languages (Table 5). Songs average 4.16 minutes (49 s to 11.7 min), and every question is asked over the full recording.

We derive ground truth from expert-annotated datasets, but we do not take questions directly from them. Instead, we use the audio and annotations to formulate tasks that mirror how users naturally ask music understanding questions in real-world applications. These tasks require answering natural-language questions rather than predicting the original labels, an approach also taken by prior audio benchmarks (Weck et al., 2024; Kumar et al., 2025; Huang et al., 2025). All questions were written by the authors with guidance from co-authors with music expertise, using paraphrased templates for tasks that share a fixed question form and individually written questions otherwise. Answers were derived from the annotations using the rules described below. The music-expert co-authors then reviewed questions to check that they were unambiguous, answerable from the audio alone, and correctly derived from its source annotation; questions failing any check were revised or removed. Recordings were sourced from YouTube via the source datasets' URLs and authors manually verified against the annotated song by listening to them. As copyright prevents redistributing audio, we release YouTube links with MusicBrainz[1] recording IDs so tracks remain recoverable if links disappear, following standard practice (Nieto et al., 2019; Oramas et al., 2025).

`Structural Segmentation:` Popular songs are organized into repeated, contrasting sections. Structural segmentation, which requires identifying these sections and locating them in time is a long-

[1] https://musicbrainz.org/

standing MIR task (Paulus et al., 2010; Ehmann et al., 2011) that supports applications ranging from music summarization to personalised recommendation and music generation; streaming platforms like Spotify already leverage structural information for playlisting and highlight extraction (Montecchio et al., 2020; Music Tomorrow, 2025; Choi et al., 2016). We build this category from the from the Harmonix Set (Nieto et al., 2019), which provides expert structural annotations for popular music and consolidate its 61 section labels into five core types (Intro, Verse, Chorus, Bridge, Outro) so that models are evaluated on widely shared structural concepts. These static annotations are then transformed into natural language question-answer pairs directly accessible to any audio LM. Within this category, BASS includes two tasks: **Full Structural Segmentation** (FSS) requires the models to segment the entire song into its major sections and return start–end timestamps for each. **Section Structural Segmentation** (SSS) targets a single section type per question (e.g., all Chorus occurrences) or a specific instance (e.g., the second Verse), requiring models to localize repeated structural elements and reason about their ordering within the song.

`Structural Lyric Transcription:` Lyrics are one of the main ways listeners engage with popular music: streaming services serve lyrics at scale while users listen. Lyrics transcription supports applications such as synchronized lyric display and lyric-based search; yet existing benchmarks treat it as a flat, structure-agnostic task. Structure matters because lyrics are organized by section: choruses repeat, verses carry new content, and lyric resources present lyrics section by section. BASS extends standard lyric transcription (Wang et al., 2022; Mesaros & Virtanen, 2008) by requiring models to transcribe lyrics within each structural section, which means they must segment the song and assign the recognized lyrics to the correct section. Reference lyrics and section labels come from Genius,[2] a platform whose transcriptions and section labels are collaboratively edited by hundreds of contributors refining transcriptions and segmentations of popular songs, ensuring higher reliability than single-annotator sources. To evaluate music beyond English, the category also includes songs in Chinese, Korean, Vietnamese, Zulu, Nigerian Pidgin, Spanish, Portuguese, and French, which span diverse phonologies and song structures. Within this category, BASS includes two tasks:

Full Structural Lyrics Transcription (FSLT): requires models to transcribe and segment lyrics into each section in a song. Gold-standard references cover common sections (Intro, Verse, Chorus, Bridge, Outro) as well as frequently occurring sections such as Pre-Chorus and Post-Chorus. To reduce ambiguity, each question lists all relevant sections in the song and asks the model to generate the corresponding lyrics for each.

Section Structural Lyrics Transcription (SSLT): requires models to transcribe lyrics for a specific section rather than the full song (e.g., Bridge). It includes two question types: *all section-level*, where the model is given a section name and must transcribe every occurrence of that section; and *single section-level*, where a specific occurrence is specified (e.g., the second Verse) and the model must transcribe only that instance.

`Musicological Analysis:` Listeners and musicians describe music through attributes such as rhythm, instrumentation, timbre, and vocal style. These attributes form the foundation of music classification and recommendation (Castelluccio, 2006; Oramas et al., 2025) among other applications and BASS evaluates these attributes over full songs against expert ratings. on musicological analysis: the task of identifying and reasoning about specific, measurable attributes of songs. Following the Music Genome Project (Castelluccio, 2006), we call these attributes *genes*. We derive this category from MGPHot (Oramas et al., 2025), in which professional musicologists rated 58 genes, grouped into seven categories (rhythm, instrumentation, sonority, harmony, compositional focus, vocals, and lyrics), on a 0–1 scale for 21,320 songs that appeared on the Billboard Hot 100 between 1958 and 2022, where 0 indicates absence and 1 high dominance of the attribute. BASS does not ask models to predict these ratings. Instead, we use them to construct new natural-language questions that require comparisons over full songs: which gene dominates a song, which of several songs expresses a gene most strongly, and how genes rank against each other. The category contains four tasks:

Single-Gene Detection (SGD) requires identifying the most dominant gene in a given recording from four options, given each gene's name and description. Questions are constructed from songs where the gene is strongly expressed (≥ 0.6) in the MGPHot dataset. Distractors are chosen from attributes

[2] https://genius.com/

with values positively correlated with the ground truth but negatively correlated with each other to avoid similar options within the same gene category (e.g., "Synthesizer" and "Synth Timbre").

Pairwise-Gene Detection (PGD) presents eight genes, each with its name and description, and asks which two are most prominent in a song. This task is more challenging than SGD. The correct pair consists of genes that are usually negatively correlated across the dataset (e.g., Acoustic Guitar and Synthetic Sonority) but both strongly expressed in this song, so the answer cannot be inferred from which genes typically co-occur. Each distractor is positively correlated with at least one correct gene and negatively correlated with all other distractors.

Gene Attribution (GA) moves beyond single-song analysis to evaluating musicological attributes across four recordings. The model is given a gene and its description and must select which of four concatenated songs expresses this attribute most prominently. The correct song must have an attribute value greater than 0.8, while distractor songs must have values below 0.5 to ensure sufficient contrast. Evaluation is done by measuring accuracy over the four-way choice task.

Gene Dominance Ranking (GDR); the model is given four genes and must rank them from least to most prominent. To avoid ambiguity, the attribute values are designed to be uniformly spaced. Rankings must be exact to be considered correct; no credit is given for partially correct outputs.

`Artist Collaboration:` Collaboration is a defining feature of popular music, with songs frequently featuring multiple vocalists whose contributions vary across sections and delivery styles (Duinker, 2021; Komaniecki, 2017; Duguay, 2022). Understanding these supports applications such as music metadata generation, artist-based recommendation, and vocal style search. BASS evaluates models on artist roles (main or featured), vocal delivery types (rap or singing), and their temporal distribution across formal sections of a song. We build this category from the Collaborative Song Dataset (CoSoD; Duguay et al., 2023), which annotates 331 multi-artist songs with timestamped formal sections, each artist's role (main or featured), and their vocal delivery (rap or singing). Rather than asking models to reproduce these annotations, we use them to derive new natural-language questions that require counting, timing, and comparing artists' contributions across the song. Artists are always referred to by order of appearance (e.g., "artist 2") rather than by name, so questions cannot be answered from prior knowledge of a song's credits. The category contains four tasks:

Artist Counting asks how many artists perform in a song. Questions cover the whole song, including both main and featured artists, or restrict the count to featured artists only, to artists with a given vocal delivery (e.g., those who rap), to one or more sections (e.g., the chorus), or to a given time interval. Some questions have zero as the correct answer, to test whether models count only what is present: for example, asking how many artists perform in the Bridge of a song that has no Bridge, or between 300 and 330 seconds in a song that ends at 240 seconds. The model is expected to output the exact number of artists, and performance is evaluated using exact-match accuracy between the model's output and the ground-truth count.

Artist Duration asks for a total performance time in seconds. Questions target a given artist (e.g., how long does artist 2 perform in total), a given vocal delivery regardless of who performs it (e.g., how many seconds of the song are rapped), a given artist using a given delivery (e.g., how long does artist 1 rap for), or all occurrences of a section type (e.g., what is the the combined length of all Verses in the song). Each answer requires quantitative reasoning across multiple, non-contiguous segments of the song. Evaluation is done using exact-match accuracy, where the model's output must match the ground-truth duration within a tolerance of ± 3 seconds.

Artist Localization asks for time in seconds at which a given artist first begins performing in the song. The artist is identified by their order of appearance rather than by name. Evaluation uses exact-match accuracy within a ± 3 second window of the ground-truth timestamp.

Artist Attribution asks categorical questions about artist vocal contributions at a given point in a song and the questions take different forms. The first compares two artists' vocal delivery, (e.g., whether the featured artist in the first Verse and the main artist in the second Verse use the same delivery). The second asks whether an artist raps, sings, or neither within a given section (e.g., the first Chorus), where *neither* indicates that the artist does not perform there. The third asks for the same judgment within a given time interval rather than a named section, so the model must locate the interval itself. The fourth asks whether an artist is the main performer, a featured performer, or

neither in a given section. Answering requires first identifying which voice belongs to the referenced artist, then characterizing that voice's delivery or role within the specified part of the song.

3 Experimental Setup

Models We benchmark a wide range of audio LMs, selected to span diverse architectures and model scales, including both reasoning and non-reasoning models. Among open-source models, we include Qwen3-Omni (Xu et al., 2025), Audio Flamingo 3 (Goel et al., 2025), Music Flamingo (Ghosh et al., 2025a), Step-Audio-R1 (Tian et al., 2025), R1-AQA (Li et al., 2025), Audio Reasoner (Xie et al., 2025), Phi 4 Multimodal Instruct (Microsoft et al., 2025), Voxtral (Liu et al., 2025), SALMONN-7B (Tang et al., 2024), Mellow (Deshmukh et al., 2025), Kimi-Audio (KimiTeam et al., 2025), and Gemma 3 Gemma Team et al. (2025). The closed-source models include Gemini 2.5 Pro, Gemini 2.5 Flash (with thinking), and Gemini 2.5 Flash (without thinking) (Gemini Team et al., 2023).[3]

Evaluation Strategy In all our experiments, audio LMs receive a text prompt (instruction) and an audio file as input and generate text as output. We use greedy decoding for all models; other generation settings follow each model's developer default. All audio files are provided at 16kHz mono, consistent with the input format of the majority of evaluated models. For each task, the authors handcraft 10 natural language prompts that are paraphrases of each other, meaning they all ask the same underlying question in slightly different ways, providing sufficient samples to estimate sensitivity to prompt variation without prohibitive computational cost. Each audio sample is evaluated using a single paraphrase, with paraphrases distributed evenly across samples within each task to ensure balanced coverage. To ensure models truly process the audio and avoid relying on shortcuts, we evaluate all tasks using free-form generation, except for Musicological Analysis questions, in which we use a multiple-choice format as explained in §2. Multiple-choice performance is determined using majority voting over four runs to account for sensitivity to option ordering.

Metrics For structural lyric transcription, we use word error rate (WER). For questions with multiple sections, reference and predicted sections are matched, the final WER is the averaged across sections. For structural segmentation, performance is measured using intersection over union (IoU), computed using mir_eval (Raffel et al., 2014), where an overlap counts as an intersection only if the predicted section labels match the reference. For other tasks, we use exact match accuracy (EMA). For questions requiring timestamps or durations, predictions within a 3-second offset are considered correct consistent with the MIREX evaluation standard for structural segmentation [4]. Aggregating the metrics across tasks poses two challenges. First, metrics differ in their direction of optimality: WER is minimized, while EMA is maximized. Furthermore, WER is unbounded.[5] Second, naively averaging EMAs across tasks is invalid due to differing random chance accuracies. To address these challenges, we invert WER as $\text{IWER} = \frac{1}{1+\text{WER}}$, bounding the score between 0 and 1 where higher is better. We normalize EMA as $\text{EMA}_N = \frac{\text{EMA}-R}{1-R}$, where R is random guessing accuracy: reciprocal of the number of options for multiple-choice tasks (e.g., $R = 0.25$ for four-choice and $R = 0.125$ for eight-choice tasks), and $R = 0$ for free-form tasks requiring exact numerical outputs such as counts, durations, and timestamps. EMA_N has an upper bound of 1.0 and can be negative if a model performs worse than random. All results are reported in percentages (%).

4 Results and Discussion

Benchmarking Results In Table 1, we present results containing normalized scores across all tasks in BASS for all models. The raw scores are provided in Appendix Table 7. On average, performance across all models remains below 30%; the best is Gemini-2.5 Pro at 26.54% and Qwen3-Omni following closely at 23.61% and Step-Audio-R1 at 22.24% . All of these models are trained with hybrid thinking modes, and these evaluations were conducted with thinking enabled. Overall, models

[3] We do not evaluate OpenAI models due to cost limitations; in fact, processing audio with these models is roughly 16 times more expensive than using Gemini 2.5 Pro.

[4] https://music-ir.org/mirex/wiki/2025:Music_Structure_Analysis

[5] WER usually falls between 0 and 1 for speech, but many models exceeded 1 on our long, noisy musical recordings, mainly because of excessive insertions, making WER unbounded.

perform best at lyric transcription, while artist collaboration tasks prove to be the most challenging. The performance gap suggests that current audio LMs are stronger on tasks where they can leverage linguistic priors, but are limited in higher level quantitative and temporal reasoning over musical structure, vocal style, and multi-artist profiles. Below, we present a detailed discussion of our results, framed around the research questions that guided our study. Specifically, we examine how audio LMs understand and reason about musical structure and lyrics, how they handle artist roles and vocal styles, and how their performance varies across different genres. While the majority of our analyses were focused on English music data, we extend our findings on multilingual music in this section.

Table 1: Normalized performance of evaluated models on BASS across musicological analysis (Single-Gene Detection—SGD, Pairwise-Gene Detection—PGD, Gene Dominance Ranking—GDR, and Gene Attribution-GA), structural segmentation, artist collaboration, and lyrics transcription tasks. Each score upper bound is 100% and bold values highlight the highest value per task. Values are reported as the point estimate ± the half-width of its 95% confidence interval (see Appendix D); the Δ_{Multi} and Average columns are reported without intervals. The normalization procedure is described in §3. The raw scores are provided in Table 7 in the Appendix.

		Musicological Analysis				Structural Segmentation		Artist Collaboration					Lyrics Transcription				Average
Model	**Size**	SGD	PGD	GDR	GA	Full	Section	Counting	Δ_{Multi}	Duration	Localization	Attribution	Full	Δ_{Multi}	Section	Δ_{Multi}	
Omni Models																	
Gemini-2.5 Pro	-	**64.98** ±6.7	10.40 ±5.6	17.80 ±11.1	31.90 ±12.8	**32.34** ±2.2	**22.98** ±3.3	25.77 ±5.5	14.95	8.25 ±2.9	3.12 ±3.9	28.84 ±10.2	**42.63** ±15.9	−38.98	29.45 ±9.1	4.89	**26.54**
Qwen3-Omni	30B	47.86 ±7.6	1.27 ±3.9	11.51 ±11.1	31.87 ±12.8	28.88 ±2.4	22.07 ±3.3	18.54 ±5.4	17.04	**9.78** ±3.2	9.38 ±7.1	**40.32** ±9.9	25.53 ±9.2	−22.88	**36.25** ±13.3	9.53	23.61
Gemini-2.5-Flash	-	43.48 ±7.7	4.06 ±3.2	**20.90** ±14.3	23.50 ±12.8	24.47 ±2.2	15.15 ±2.6	20.40 ±5.6	9.08	7.66 ±2.9	3.12 ±3.9	31.44 ±9.8	5.13 ±8.1	−2.86	5.89 ±8.5	−1.35	17.10
Large Audio Reasoning Models																	
Step-Audio-R1	33B	43.88 ±7.8	8.89 ±4.5	14.62 ±14.3	**31.91** ±12.8	32.19 ±2.4	17.73 ±2.7	19.34 ±5.3	-	6.42 ±2.6	**7.81** ±6.3	23.74 ±10.2	21.86 ±3.6	−17.07	35.89 ±8.3	−0.40	22.24
R1-AQA	8.2B	12.04 ±7.2	−2.98 ±1.7	−4.35 ±0.0	7.87 ±11.3	5.91 ±0.8	4.16 ±1.3	**35.03** ±5.8	−6.70	0.92 ±1.1	0.00 ±0.0	20.51 ±10.4	40.05 ±0.7	−22.35	36.35 ±7.9	9.27	12.96
Audio-Reasoner	8.4B	11.64 ±7.0	5.37 ±3.8	1.98 ±6.3	7.80 ±11.3	6.63 ±1.6	2.90 ±1.9	18.47 ±4.7	1.46	3.06 ±1.9	0.00 ±0.0	−59.03 ±3.9	2.46 ±2.8	−0.47	0.00 ±3.0	3.24	0.00
Mellow	167M	−32.94 ±0.0	−3.70 ±0.0	−4.35 ±0.0	−33.33 ±0.0	0.00	0.00	0.00 ±1.6	9.92	2.14 ±1.6	0.00 ±0.0	−62.50 ±2.7	53.50 ±0.1	−45.99	0.55 ±5.8	23.92	−6.72
Large Audio Language Models																	
Kimi Audio	7B	16.82 ±7.9	0.55 ±2.5	−4.35 ±0.0	−33.33 ±0.0	13.06 ±1.8	2.64 ±0.9	19.12 ±5.1	8.23	0.92 ±1.1	0.00 ±0.0	20.96 ±10.4	5.17 ±3.7	−2.52	1.80 ±1.6	1.04	3.61
Audio Flamingo 3	8B	25.60 ±9.0	8.01 ±5.6	−4.34 ±0.0	7.87 ±12.1	7.03 ±1.6	0.01 ±0.0	21.92 ±5.3	-	1.84 ±1.4	0.00 ±0.0	21.13 ±10.0	17.59 ±3.8	-	26.57 ±5.1	-	11.10
Gemma 3	5B	13.63 ±7.8	3.30 ±3.8	−1.19 ±4.7	9.22 ±12.1	8.92 ±2.4	4.35 ±1.3	12.25 ±4.0	10.51	0.30 ±0.5	1.56 ±2.4	2.47 ±10.5	40.05 ±1.7	−26.89	36.35 ±6.0	4.99	10.93
Voxtral-Small	24B	16.82 ±7.8	5.42 ±4.9	5.15 ±9.5	−10.67 ±10.0	32.16 ±2.6	8.84 ±2.7	9.09 ±4.1	18.95	2.76 ±1.7	0.00 ±0.0	3.42 ±9.7	26.45 ±3.5	-	6.08 ±4.8	-	8.79
SALMONN-7B	7B	0.50 ±5.6	**37.26** ±0.0	−4.03 ±0.0	4.93 ±11.4	3.18 ±1.3	0.12 ±0.1	27.32 ±5.4	0.87	0.61 ±0.8	0.00 ±0.0	−1.48 ±9.3	11.61 ±2.3	1.88	4.63 ±2.2	−0.14	7.05
Phi4-mm-instruct	5.5B	−4.96 ±7.1	1.27 ±1.7	5.15 ±7.9	−4.93 ±10.7	2.71 ±0.5	0.16 ±0.1	23.59 ±5.6	0.36	3.66 ±2.0	0.00 ±0.0	38.51 ±9.8	4.99 ±0.7	−4.04	22.45 ±0.4	−21.97	7.71
Music Flamingo	8B	32.34 ±8.1	4.07 ±4.4	−4.34 ±0.0	0.67 ±11.4	13.07 ±1.9	2.59 ±0.9	33.54 ±5.9	5.68	4.27 ±2.2	4.69 ±4.7	31.88 ±9.7	15.72 ±4.2	−10.12	11.47 ±14.6	6.06	12.50

How well do audio LMs understand and reason about musical structure? On structural tasks like structural segmentation and lyrics transcription, we observe that models often perform better when the task requires temporal localization across full songs rather than isolated sections. We suspect that reasoning sequentially over earlier sections allows models to use contextual cues from preceding segments to better identify later sections. However, even when models perform temporal localization across full songs (see Figure 8 in the Appendix), they frequently get confused between acoustically similar sections, such as verses and choruses. We speculate that this confusion is due to audio LMs not being explicitly trained to understand musical structure across long contexts. Our section-wise analysis (Figure 2) is consistent with this: in structural segmentation, models are generally more capable of localizing content in sections that occur earlier in the song, such as intros and first verses, than later sections such as bridges and outros. For lyrics transcription, performance is instead lowest on intros (Figure 2, right). Breaking down WER by error type Figure 12 in the Appendix) shows that intros and outros contain substantially more insertion and repetition errors.

Does providing an ASR transcript improve audio LMs' structural lyric transcription? General-purpose speech recognition models such as Whisper(Radford et al., 2023) show strong generalization to automatic lyric transcription (Radford et al., 2023; Wang et al., 2023; Zhuo et al., 2023; Ou et al., 2022; Cífka et al., 2024) and leading lyric transcription systems combine Whisper with a language model that refines its output. We test a similar cascade: we transcribe each song with Whisper and provide the transcript with the audio, to the audio LM performing structural transcription. A transcript could help the model locate section boundaries and align lyrics accurately. Whisper alone achieves a WER of 64.97% on flat, unsegmented transcription. Nevertheless, our results in Table 2 shows that adding the transcript *reduces* performance compared to audio alone for every

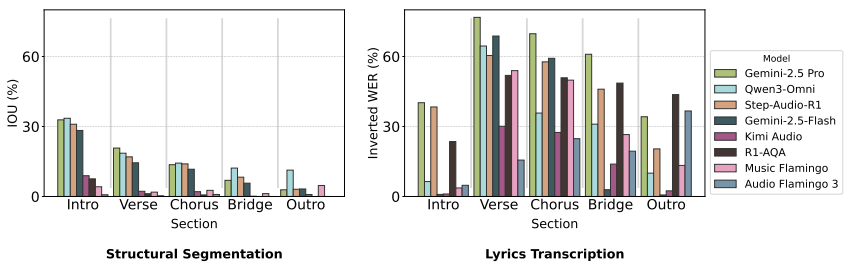

Figure 2: Section-wise performance for Structural Segmentation and Lyric Transcription (higher is better). Segmentation performance declines over time, while transcription performance (IWER) is lowest for the *Intro* compared to *Verse* and *Chorus* sections.

model tested. Two factors likely contribute to this. First, Whisper's transcripts contain errors, which are passed on as incorrect lexical content that models copy rather than correct. Second, the transcript is a single unsegmented block of text: it provides words but not section boundaries, so the model must still align the text to the song's structure, and the transcript may distract from the acoustic cues needed to do so. Together with Whisper's comparatively low WER, this suggests that the bottleneck in structural lyric transcription is not recognizing words but assigning them to the right sections.

Can audio LMs disentangle artist roles and vocal styles over time in collaborative music? Collaborative music presents a complex challenge for audio LMs, with models showing lower performance compared to other tasks. Performance is poor for fine-grained localization and quantifying the duration of artists' contributions within a song. This aligns with prior work on general purpose audio benchmarks that study temporal localization (Ahia et al., 2025; Kumar et al., 2025; Ma et al., 2025b). We see that models are better at identifying rapping than singing (Figure 9). This might be due to rapping more closely resembling spoken language, which models are typically trained on.

Table 2: Cascaded ASR (Whisper + LM) hurts transcription vs. audio-only. IWER reported; higher is better.

Setup	Qwen3 Omni	Gem. 2.5	Phi MM	Gem. 3n
Full Lyric Trans. (FLT) ↑				
Cascaded	6.14	13.25	**3.02**	10.90
Audio only	**11.38**	**42.09**	0.90	**14.45**
Section Lyric Trans. (SLT) ↑				
Cascaded	23.77	29.69	0.72	5.51
Audio only	**26.54**	**46.11**	0.30	**34.11**

Artist collaboration tasks share similarities with conventional speaker diarization, which estimates how many distinct voices are present and when each one speaks. To test whether a dedicated speaker diarization system helps, we run pyannote (Bredin, 2023) on each song and provide its output (speaker turns with timestamps) alongside the audio to Gemini-2.5 Pro. The cascade improves artist localization (11.48% vs. 3.12%) but leaves counting and duration largely unchanged, and it substantially lowers attribution accuracy (18.09% vs. 56.66%). Diarization supplies speaker-change timing, which helps locate when an artist enters, but not vocal delivery or artist roles, and its errors in separating overlapping or layered vocals appear to mislead the model on these categorical questions.

How well can audio LMs reason about musicological attributes? Music understanding involves reasoning over a variety of musicological attributes present in the audio, from vocals and lyrics to rhythm etc. To assess how well audio LMs capture these properties, we analyze model performance across different attributes in Figure 10. Results vary, Gemini-2.5 Pro performs strongest overall, especially on Vocals, Lyrics, and Composition, while Music Flamingo excels mainly in Instrumentation but lags in Harmony. All models evidently struggle with Sonority and Rhythm, suggesting that capturing subtle timbral and temporal nuances remains a key challenge for current audio LMs, consistent with our earlier observations of temporal reasoning difficulties in audio LMs.

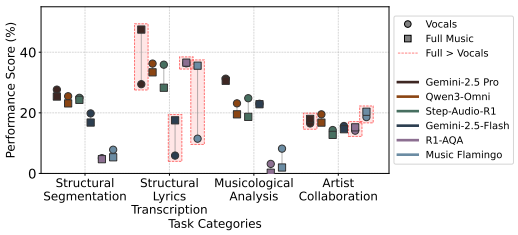

Figure 3: Source separation degrades performance. Mixed audio *(squares)* consistently outperforms vocal stems *(circles)*, especially in structural lyric transcription.

Do audio LMs benefit from source separation for structural and semantic music reasoning? Music recordings usually contain a mix of vocals, instruments, and background sounds, differing significantly from conventional speech. Previous work has investigated the effects of source separation—isolating individual stems; vocals, bass, drums, and other instruments from a mixed recording—on downstream music understanding tasks (de Berardinis et al., 2020; Xu et al., 2014; Rosner & Kostek, 2018; Agrawal & Karamchandani, 2023; Syed et al., 2025), and consistently reported performance gains when models operate only on vocal stems. We investigate if same benefits hold for audio LMs across the tasks in our benchmark. We use the BS-RoFormer model (Lu et al., 2023) to replace the mixed audio in our benchmark with vocal stems and reevaluate all models using the same configuration. As shown in Figure 3, vocal stems degrade performance across tasks to varying degrees. Overall, models perform worse on most tasks, indicating that mixed audio contains richer information and that source separation does not reliably improve performance for reasoning about musical structure and semantics.

Which musical genres challenge audio LMs the most? BASS contains recordings across musical genres as displayed in Figure 5. For this analysis, we focus on musicological task and analyze average model performance to understand the most challenging genres for audio LMs. As shown in Figure 10, models perform best on *Country* and *Rock* but relatively worse on *Pop*.

Does step-by-step reasoning improve music understanding performance? Many tasks in BASS require compositional skills and may benefit from step-by-step reasoning. While many audio LMs are not trained with hybrid reasoning modes (i.e., capable of switching between a deeper, step-by-step reasoning mode and a faster, direct response mode), we investigate how such reasoning capabilities impact performance across all tasks in our benchmark. We perform this analysis on Qwen3-Omni.[6] In our experiments, we use vLLM's reasoning control method[7] to switch Qwen3-Omni between "thinking" and "non-thinking modes", performing all experiments across three independent seeds.

Our results in Table 3 indicate that enabling "thinking" significantly improved model performance on 6 out of 12 tasks in BASS. We observed the largest performance gains on Artist Attribution (+24.53), Gene Attribution (+17.97), and Section Lyrics Transcription (+13.33), all of which are statistically significant ($p < 0.001$) after Holm-Bonferroni correction. Tasks like Full Lyrical Transcription and (-12.77) and Gene Dominance Ranking (-2.11) show small negative performance deltas where "thinking" degrades performance.

Table 3: Performance comparison of Qwen3-Omni with and without "thinking". Tasks where enabling "thinking" led to statistically significant improvements ($p < 0.001$ after Holm-Bonferroni correction across three seeds) are highlighted in green.

Task	No Think	Think	Δ
Artist Collaboration			
Artist Counting	25.04	24.59	−0.46
Artist Duration	4.59	8.97	+4.38
Artist Attribution	11.10	35.63	+24.53
Artist Localization	3.64	7.29	+3.65
Musicological Analysis			
SGD	45.47	46.27	+0.80
PGD	3.30	4.91	+1.61
GDR	17.79	15.68	−2.11
Gene Attribution	13.00	30.97	+17.97
Structural Segmentation			
Full Seg.	24.05	27.19	+3.14
Section Seg.	15.06	19.84	+4.77
Structural Lyric Trans.			
Full Trans.	28.08	15.31	−12.77
Section Trans.	27.86	41.19	+13.33

Do Audio LMs struggle to connect memorized metadata to musical audio? BASS contains musical recordings across genres, particularly popular songs. Given that many recent LMs are trained on large-scale web data, which may include overlapping musical content or metadata, it is important to assess the extent to which model performance reflects memorization rather than audio-based reasoning. To investigate this, we provide models only with the artist name and song title along with the question, omitting the audio. We then compare performance in this text-only condition to performance when the musical audio recording is provided. This analysis is performed on Qwen3-Omni, Gemini, and Gemini Pro, focusing specifically on the artist collaboration, musicological analysis, and structural lyrics transcription tasks. We excluded questions that require timestamp localization, since we do not provide any audio for these evaluations.

[6] Closed-source models like Gemini Flash also have "Thinking" and non-"Thinking" variants, but it is unclear if these use the same underlying model.

[7] https://docs.vllm.ai/en/latest/features/reasoning_outputs/#server-level-default-chat-template-kwargs

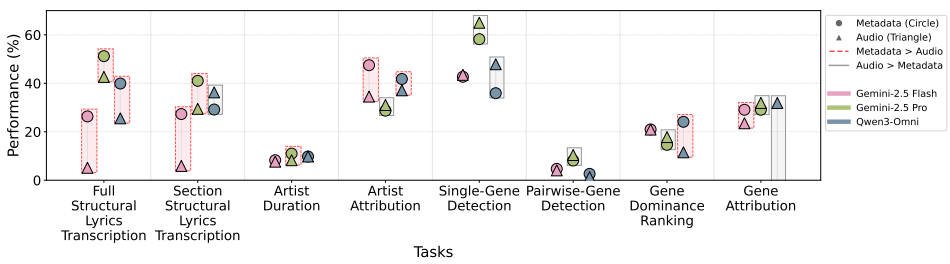

Figure 4: Performance comparison when audio is replaced with metadata (artist name and song title). With only metadata, models achieve higher relative performance on lyrics transcription than on other tasks, suggesting a stronger reliance on memorized artist–song associations in the training data.

Our results in Figure 4 show that providing only the metadata (artist name and song title) surprisingly boosts performance on lyrics transcription tasks compared to providing the audio alone, with Gemini Flash having the largest improvement. While this improvement indicates that models can leverage memorized artist–song associations in the training data, it also highlights the challenges they face when text priors are absent, leaving substantial room for improvement. On other tasks, performance either drops or remains similar to the audio condition. This aligns with our expectations, since these tasks rely more on information present in the audio than on memorized metadata. It is still a little surprising that, even though the LM component of the audio LM might have seen the lyrics during training, the audio LMs struggles to connect this prior knowledge to the musical audio at inference.

How well do audio LMs understand on multilingual music? As audio LMs are deployed globally, their ability to understand non-English music is increasingly important. On FSLT, we observe a clear degradation in performance on multilingual data, suggesting that multimodal LMs struggle to localize and transcribe long-form audio outside of English. On SSLT and Artist Counting, however, performance *improves* on multilingual data. We hypothesize that LMs cannot lean as heavily on their linguistic priors in non-English settings, forcing them to rely on the audio signal itself to segment songs or localize artists, which may yield more accurate predictions. Across the eight non-English languages evaluated, models performed best on Spanish, Portuguese, French, and Vietnamese (Table 4). Spanish, Portuguese, and French are all high-resource languages that share the Latin alphabet with English, allowing the LMs to draw more readily on their linguistic priors than in lower-resource or non-Latin-script languages.

Table 4: Average task performance on the multilingual subsets: Chinese (ZH), Korean (KO), Portuguese (PT), Spanish (ES), French (FR), Nigerian Pidgin (PCM), Vietnamese (VI), Zulu (ZU).

Model	ZH	KO	PT	ES	FR	PCM	VI	ZU
Gemini-2.5 Pro	13.6	56.9	32.1	34.7	28.7	25.0	40.7	23.3
Qwen3-Omni	24.6	29.9	43.3	28.6	28.6	29.3	34.8	29.1
Gemini-2.5-Flash	5.2	19.1	17.1	22.6	17.9	14.7	23.8	16.0
Music Flamingo	21.0	19.2	38.3	29.9	35.1	28.4	37.3	21.9
R1-AQA	28.3	28.0	32.6	31.0	36.7	28.1	44.4	24.0
Gemma 3n	27.7	23.5	30.4	25.9	21.4	18.8	20.8	31.3
SALMONN-7B	8.0	15.5	21.3	19.6	27.0	16.5	31.4	14.1
Kimi Audio	31.3	23.2	41.7	19.7	23.5	21.5	26.7	40.0
Phi4-mm-instruct	6.9	8.4	10.8	34.8	17.2	11.3	32.5	15.0
Mellow	10.5	10.9	16.2	10.7	21.4	16.2	7.5	9.2

5 Conclusion

In this paper, we introduce BASS, designed to evaluate music understanding and reasoning in AI systems. BASS covers four categories—*Structural Segmentation*, *Lyric Transcription*, *Musicological Analysis*, and *Artist Collaboration* —across 12 tasks requiring both structural and semantic reasoning. We evaluate 14 open-source and frontier audio LMs, including models trained with hybrid thinking, and find that even the best-performing model achieves only 26% average performance. Our analysis shows that state-of-the-art models struggle on higher-level reasoning tasks such as structural segmentation and artist collaboration, while performing relatively better on lyric transcription. BASS highlights model strengths and gaps to guide future development of audio LMs with strong long-form music understanding and reasoning capabilities.

6 LIMITATIONS

Scope of source annotations. BASS is built on existing music expert-annotated datasets and inherits their coverage. Genre representation reflects these sources, with Pop and Hip Hop the most common, and vocal delivery is annotated as rapping or singing, so other vocal styles are not yet covered. As expert annotations for further genres, languages, and vocal styles become available, our well documented construction pipeline can be applied directly to extend BASS.

Audio availability. Due to licensing restrictions, BASS distributes YouTube links rather than audio files. This follows common practice in music information retrieval, where benchmarks built on commercial music distribute references to recordings rather than the recordings themselves (Nieto et al., 2019; Oramas et al., 2025); since advancing music understanding requires evaluation on real-world music, most of which is copyrighted, this is a constraint of the domain rather than of our benchmark specifically. As links may change over time, we also provide MusicBrainz[8] recording IDs, which identify each track independently of any hosting platform, and we recommend that users report the number of recordings retrieved alongside their results.

Evaluation Repeating the full evaluation over multiple runs, across all models and experimental conditions, would cost several thousand US dollars in API and compute, a substantial expense in an academic setting; we therefore evaluate each model in a single run. Our confidence intervals capture uncertainty from the choice of songs, and our thinking-mode comparison (Table 3) was run across three seeds. Results for proprietary models may also vary over time, as providers can update models served under the same identifier.

AI USE STATEMENT

We used AI tools to aid in implementing experiments. We have not used generative AI tools for generating synthetic datasets, and helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, assisting with translation, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, interpretting results are not applicable to this work. Additionally, we used generative AI tools for aiding in polishing writing and finding relevant literature. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

BASS is built from publicly available, expert-annotated music datasets, with audio obtained from YouTube. Because the recordings are copyrighted, we do not distribute audio files. Instead, we release YouTube links together with MusicBrainz[9] recording IDs, which identify each track independently of any hosting platform, allowing users to access the audio legally while respecting copyright. The benchmark contains works by a diverse range of artists, and we encourage responsible and respectful use. To promote transparency and reproducibility, we will release the benchmark publicly, along with our evaluation code, metrics, and data construction pipeline, so that the research community can build on our work.

REPRODUCIBILITY STATEMENT

We describe the data sources and annotation procedure for every task in Section 2, and the models, prompting strategy, and metrics in Section 3. Dataset statistics and the exact prompts and response-format instructions for all tasks are given in the Appendix, and the computational resources used are listed in Appendix C.

[8] https://musicbrainz.org/
[9] https://musicbrainz.org/

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. Musiclm: Generating music from text, 2023. URL https://arxiv.org/abs/2301.11325.

Vanshaj Agrawal and Sunil Karamchandani. Audio source separation as applied to vocals-accompaniment extraction. *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, 5:100254, 2023. ISSN 2772-6711. doi: https://doi.org/10.1016/j.prime.2023.100254. URL https://www.sciencedirect.com/science/article/pii/S2772671123001493.

Orevaoghene Ahia, Martijn Bartelds, Kabir Ahuja, Hila Gonen, Valentin Hofmann, Siddhant Arora, Shuyue Stella Li, Vishal Puttagunta, Mofetoluwa Adeyemi, Charishma Buchireddy, Ben Walls, Noah Bennett, Shinji Watanabe, Noah A. Smith, Yulia Tsvetkov, and Sachin Kumar. Blab: Brutally long audio bench, 2025. URL https://arxiv.org/abs/2505.03054.

Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*, 2023.

Michael Castelluccio. The music genome project. *Strategic Finance*, Dec 2006.

Keunwoo Choi, George Fazekas, and Mark Sandler. Towards playlist generation algorithms using rnns trained on within-track transitions, 2016. URL https://arxiv.org/abs/1606.02096.

Ondřej Cífka, Hendrik Schreiber, Luke Miner, and Fabian-Robert Stöter. Lyrics transcription for humans: A readability-aware benchmark, 2024. to appear; preprint arXiv:2408.06370.

Jacopo de Berardinis, Angelo Cangelosi, and Eduardo Coutinho. The multiple voices of musical emotions: Source separation for improving music emotion recognition models and their interpretability. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, pp. 310–317, Montréal, Canada, 2020.

Soham Deshmukh, Satvik Dixit, Rita Singh, and Bhiksha Raj. Mellow: a small audio language model for reasoning, 2025. URL https://arxiv.org/abs/2503.08540.

Michèle Duguay, Kate Mancey, and Johanna Devaney. Collaborative song dataset (cosod): An annotated dataset of multi-artist collaborations in popular music. In *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR 2023)*, pp. 71–79, 2023. doi: 10.5281/zenodo.10265225. URL https://doi.org/10.5281/zenodo.10265225.

Michèle Duguay. Analyzing vocal placement in recorded virtual space. *Music Theory Online*, 28(4): 1–?, 2022.

Ben Duinker. Song form and the mainstreaming of hip-hop music. 107:93–135, 01 2021. doi: 10.52214/cm.v107i.7177. URL http://dx.doi.org/10.52214/cm.v107i.7177.

Andreas F Ehmann, Mert Bay, J Stephen Downie, Ichiro Fujinaga, and David De Roure. Music structure segmentation algorithm evaluation: Expanding on mirex 2010 analyses and datasets. In *12th International Society for Music Information Retrieval Conference, ISMIR 2011*, pp. 561–566, 2011.

Gemini Team, Rohan Anil, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, and .... Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi,

Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/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. Music flamingo: Scaling music understanding in audio language models, 2025a. URL https://arxiv.org/abs/2511.10289.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=xWu5qpDK6U.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv*, 2025.

Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T. Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, William Chen, Chih-Kai Yang, Xuanjun Chen, Chi-Yuan Hsiao, Puyuan Peng, Shih-Heng Wang, Chun-Yi Kuan, Ke-Han Lu, Kai-Wei Chang, Fabian Ritter-Gutierrez, Kuan-Po Huang, Siddhant Arora, You-Kuan Lin, Ming To Chuang, Eunjung Yeo, Kalvin Chang, Chung-Ming Chien, Kwanghee Choi, Cheng-Hsiu Hsieh, Yi-Cheng Lin, Chee-En Yu, I-Hsiang Chiu, Heitor R. Guimarães, Jionghao Han, Tzu-Quan Lin, Tzu-Yuan Lin, Homu Chang, Ting-Wu Chang, Chun Wei Chen, Shou-Jen Chen, Yu-Hua Chen, Hsi-Chun Cheng, Kunal Dhawan, Jia-Lin Fang, Shi-Xin Fang, Kuan-Yu Fang, Chi-An Fu, Hsien-Fu Hsiao, Ching Yu Hsu, Shao-Syuan Huang, Lee Chen Wei, Hsi-Che Lin, Hsuan-Hao Lin, Hsuan-Ting Lin, Jian-Ren Lin, Ting-Chun Liu, Li-Chun Lu, Tsung-Min Pai, Ankita Pasad, Shih-Yun Shan Kuan, Suwon Shon, Yuxun Tang, Yun-Shao Tsai, Jui-Chiang Wei, Tzu-Chieh Wei, Chengxi Wu, Dien-Ruei Wu, Chao-Han Huck Yang, Chieh-Chi Yang, Jia Qi Yip, Shao-Xiang Yuan, Vahid Noroozi, Zhehuai Chen, Haibin Wu, Karen Livescu, David Harwath, Shinji Watanabe, and Hung-yi Lee. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities

of spoken language models with 180 tasks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Robert Komaniecki. Analyzing collaborative flow in rap music. *Music theory online*, 23(3), 2017. ISSN 1067-3040. doi: 10.30535/mto.23.4.8.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, et al. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992*, 2025.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering, 2025. URL https://arxiv.org/abs/2503.11197.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, Amélie Martin, Anmol Agarwal, Antoine Roux, Arthur Darcet, Arthur Mensch, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Chris Bamford, Christian Wallenwein, Christophe Renaudin, Clémence Lanfranchi, Darius Dabert, Devendra Singh Chaplot, Devon Mizelle, Diego de las Casas, Elliot Chane-Sane, Emilien Fugier, Emma Bou Hanna, Gabrielle Berrada, Gauthier Delerce, Gauthier Guinet, Georgii Novikov, Guillaume Martin, Himanshu Jaju, Jan Ludziejewski, Jason Rute, Jean-Hadrien Chabran, Jessica Chudnovsky, Joachim Studnia, Joep Barmentlo, Jonas Amar, Josselin Somerville Roberts, Julien Denize, Karan Saxena, Karmesh Yadav, Kartik Khandelwal, Kush Jain, Lélio Renard Lavaud, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Marie Pellat, Mathilde Guillaumin, Mathis Felardos, Matthieu Dinot, Maxime Darrin, Maximilian Augustin, Mickaël Seznec, Neha Gupta, Nikhil Raghuraman, Olivier Duchenne, Patricia Wang, Patryk Saffer, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Rémi Delacourt, Romain Sauvestre, Roman Soletskyi, Sagar Vaze, Sandeep Subramanian, Saurabh Garg, Shashwat Dalal, Siddharth Gandhi, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Thibault Schueller, Thibaut Lavril, Thomas Robert, Thomas Wang, Timothée Lacroix, Tom Bewley, Valeriia Nemychnikova, Victor Paltz, Virgile Richard, Wen-Ding Li, William Marshall, Xuanyu Zhang, Yihan Wan, and Yunhao Tang. Voxtral, 2025. URL https://arxiv.org/abs/2507.13264.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. Music understanding llama: Advancing text-to-music generation with question answering and captioning, 2023. URL https://arxiv.org/abs/2308.11276.

Wei-Tsung Lu, Ju-Chiang Wang, Qiuqiang Kong, and Yun-Ning Hung. Music source separation with band-split rope transformer, 2023. URL https://arxiv.org/abs/2309.02612.

Yinghao Ma, Siyou Li, Juntao Yu, Emmanouil Benetos, and Akira Maezawa. Cmi-bench: A comprehensive benchmark for evaluating music instruction following, 2025a. URL https://arxiv.org/abs/2506.12285.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, tianrui wang, Yuping Wang, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, EngSiong Chng, and Xie Chen. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *The Thirty-ninth Annual*

Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2025b. URL https://openreview.net/forum?id=fgmrBJemlQ.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, Elio Quinton, György Fazekas, and Juhan Nam. The song describer dataset: a corpus of audio captions for music-and-language evaluation. In *Machine Learning for Audio Workshop at NeurIPS 2023*, 2023.

Annamaria Mesaros and Tuomas Virtanen. Automatic recognition of lyrics in singing. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx)*, pp. 1–4, 2008.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Nicola Montecchio, Pierre Roy, and François Pachet. The skipping behavior of users of music streaming services and its relation to musical structure. *PLOS ONE*, 15(9):e0239418, September 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0239418. URL http://dx.doi.org/10.1371/journal.pone.0239418.

Music Tomorrow. Inside Spotify's recommendation system: A complete guide (2025 update), 2025. URL https://music-tomorrow.com/blog/how-spotify-recommendation-system-works-complete-guide. Accessed: 2026-01-04. Explains Spotify's use of temporal structural units for content-based recommendations.

Oriol Nieto, Matthew C. McCallum, Matthew E. P. Davies, Andrew Robertson, Adam M. Stark, and Eran Egozy. The harmonix set: Beats, downbeats, and functional segment annotations of western popular music. In Arthur Flexer, Geoffroy Peeters, Julián Urbano, and Anja Volk (eds.), *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, pp. 565–572, 2019. URL http://archives.ismir.net/ismir2019/paper/000068.pdf.

Sebastian Oramas, Fabien Gouyon, Stuart Hogan, Chris Landau, and Anahid Ehmann. Mgphot: A dataset of musicological annotations for popular music (1958–2022). *Transactions of the International Society for Music Information Retrieval*, 8(1):108–120, 2025.

Longshen Ou, Xiangming Gu, and Ye Wang. Transfer learning of wav2vec 2.0 for automatic lyric transcription. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, pp. 891–899, Bengaluru, India, 2022.

Jouni Paulus, Meinard Müller, and Anssi Klapuri. State of the art report: Audio-based music structure analysis. In J. Stephen Downie and Remco C. Veltkamp (eds.), *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010, Utrecht, Netherlands, August 9-13, 2010*, pp. 625–636. International Society for Music Information Retrieval, 2010. URL http://ismir2010.ismir.net/proceedings/ismir2010-107.pdf.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis. Mir_eval: A transparent implementation of common MIR metrics. In Hsin-Min

Wang, Yi-Hsuan Yang, and Jin Ha Lee (eds.), *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, pp. 367–372, 2014. URL http://www.terasoft.com.tw/conf/ismir2014/proceedings/T066_320_Paper.pdf.

Aldona Rosner and Bozena Kostek. Automatic music genre classification based on musical instrument track separation. *J. Intell. Inf. Syst.*, 50(2):363–384, April 2018. ISSN 0925-9902. doi: 10.1007/s10844-017-0464-5. URL https://doi.org/10.1007/s10844-017-0464-5.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TeVAZXr3yv.

Jaza Syed, Ivan Meresman Higgs, Ondřej Cífka, and Mark Sandler. Exploiting music source separation for automatic lyrics transcription with whisper, 2025. URL https://arxiv.org/abs/2506.15514.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models, 2024. URL https://arxiv.org/abs/2310.13289.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, Chengyuan Yao, Hexin Liu, Eng Siong Chng, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, and Gang Yu. Step-audio-r1 technical report, 2025. URL https://arxiv.org/abs/2511.15848.

Ju-Chiang Wang, Yun-Ning Hung, and Jordan BL Smith. To catch a chorus, verse, intro, or anything else: Analyzing a song with structural functions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 416–420. IEEE, 2022.

Jun-You Wang, Chon-In Leong, Yu-Chen Lin, Li Su, and Jyh-Shing Roger Jang. Adapting pretrained speech model for mandarin lyrics transcription and alignment. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. Muchomusic: Evaluating music understanding in multimodal audio-language models. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. Audsemthinker: Enhancing audio-language models through reasoning over semantics of sound, 2025. URL https://arxiv.org/abs/2505.14142.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models, 2025. URL https://arxiv.org/abs/2503.02318.

Jieping Xu, Xirong Li, Yun Hao, and Gang Yang. Source separation improves music emotion recognition. In *Proceedings of the International Conference on Multimedia Retrieval*, pp. 423–426. ACM, 2014.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni: A single multimodal model for text, image, audio, and video. *arXiv preprint arXiv:2509.17765*, 2025.

Le Zhuo, Ruibin Yuan, Jiahao Pan, Yinghao Ma, Yizhi Li, Ge Zhang, Si Liu, Roger Dannenberg, Jie Fu, Chenghua Lin, Emmanouil Benetos, Wenhu Chen, Wei Xue, and Yike Guo. Lyricwhiz: Robust multilingual zero-shot lyrics transcription by whispering to chatgpt. In *Proceedings of the 24th International Society for Music Information Retrieval Conference (ISMIR)*, Milan, Italy, 2023.

A Dataset Details

A.1 Dataset Statistics

We provide a comprehensive breakdown of data statistics across tasks in BASS in Table 5. In Figure 5, we present distribution of musical genres across the dataset, which is led by Pop (30.3%) and Hip Hop (25.6%), followed by a diverse range of categories including Electronic, Rock, and R&B.

Category	# Q	# S	Avg.	Min.	Max.
Structural Segmentation	298	298	3.37	1.70	10.78
Structural Lyric Trans.	884	884	3.78	1.60	11.72
Musicological Analysis	609	484	5.57	0.82	19.44
Artist Collaboration	867	327	3.83	1.34	6.97
Total	2658	1993	4.16	0.82	19.44

Table 5: Data statistics of BASS by task category. # Q is the number of questions and # S the number of distinct songs; Avg./Min./Max. are song length in minutes, computed over the distinct songs in each category. Because a song may be used by several tasks within a category, # S is smaller than the sum of the per-task song counts (see Table 5); for the same reason the corpus-level song count and mean length are not the sum or weighted mean of the rows above.

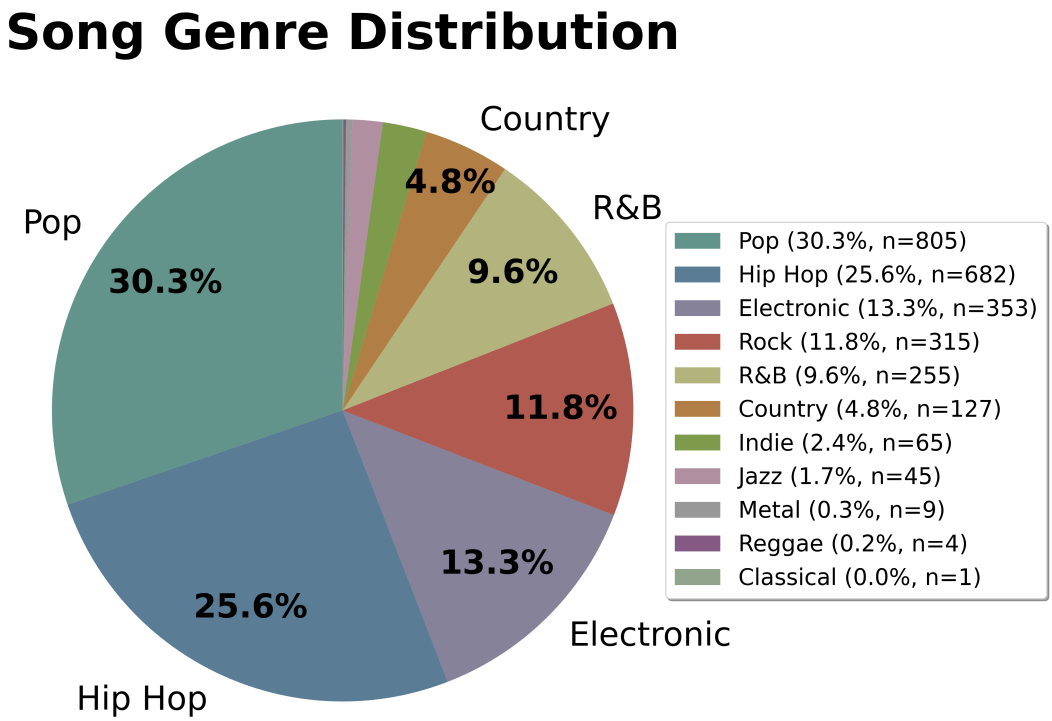

Figure 5: Distribution of musical genres across songs in BASS

A.2 Prompts for all tasks

In Table 6, we present the specific prompts and response formatting instructions for all tasks in the benchmark.

B TASK-SPECIFIC RESULTS

In §4 in the main paper, we present an overview of our key findings and discuss their implications for audio LMs' reasoning across musical structure, lyrics, artist roles, and genres. Here, we present finegrained results for each task.

B.1 STRUCTURAL SEGMENTATION

For structural segmentation tasks, our results show that models generally perform better when localizing all sections within a song (Full Segmentation) compared to localizing a single, isolated section (Sectional Segmentation). Across all evaluated models, the average absolute normalized score difference between these two tasks is 7.63%. In Figure 6, we present the average model performance per section across the top performing models. We observe that models are more capable of segmenting sections that occur earlier in the song, such as the intro and verses (1st verse). There is a gradual decline in performance as the songs progress toward the outro section.

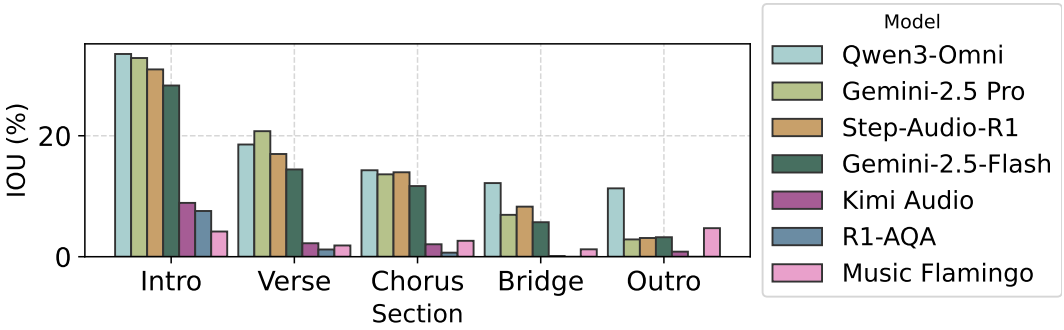

Figure 6: Section-wise performance for structural segmentation. Performance is highest for earlier sections and gradually decreases toward the end of the recording.

Figure 15 shows a confusion matrix across all models. We observe that the verse and chorus are most commonly mistaken for each other. We do not expect acoustic information alone to distinguish verses from choruses across songs; cues such as repeated melody and lyrics in choruses should be more informative. We speculate this confusion is due to audio LMs not being explicitly trained to understand musical structure across long range dependencies.

B.2 STRUCTURAL LYRICS TRANSCRIPTION

For this task, we observe trends broadly consistent with structural segmentation: models perform substantially better on full than on section-level transcription. Across all evaluated models, the average normalized performance difference between full-song and section-level transcription is 4.35% (measured via IWER §3).

While global performance patterns resemble those seen in structural segmentation, we note a distinct gap in the *Intro* sections. We hypothesize that this arises because *Intros* and *Outros* often lack lyrics; as a result, models are more likely to produce insertion errors, possibly hallucinating words or repeating phrases from other sections.

To investigate this, we analyze fine-grained WER and error types. We find that models produce more deletions in *Verses* and more insertions in *Intros* and *Outros* (Appendix, Figure 11), contributing to lower transcription performance in these sections. We further characterize this behavior by quantifying repeated phrases in the predicted transcription, using

$$\frac{\sum_{\phi \in \text{repeats}} \max(0, c_\phi - r_\phi) \cdot |\phi|}{|\text{ref_words}| + k},$$

where ϕ is a repeated phrase, c_ϕ is its count in the prediction, r_ϕ is its count in the reference, $|\phi|$ is the phrase length in words, $|\text{ref_words}|$ is the total number of words in the reference, and k is a smoothing constant to prevent inflated scores for very short references. As shown in Figure 12,

repetition errors are indeed most prevalent in *Intro* and *Outro* sections on both full and section-level transcriptions, supporting our hypothesis.

B.3 Musicological Analysis

For this task, we observe the highest performance on Single-Gene Detection where models are prompted to select a single musicological attribute that is strongly dominant in the song (e.g., presence of a rhythmic groove) out of four options. Performance drops substantially in the Pair-Gene Detection task, where models must select a pair of attributes from eight options, making the task more challenging; even Gemini achieves only around 10% accuracy. Performance is similarly low for Gene Dominance Ranking, which requires ordering attributes by dominance. However, results improve for Gene Attribution, where models operate at a higher level by selecting a single song with a dominant attribute from a set of four songs.

B.4 Artist Collaboration

Across all task categories, this category appears to be the most challenging. While overall performance is low, models consistently perform better on artist counting and artist attribution than others. These tasks share an underlying focus on recognizing vocal style, artist roles, and temporal localization. Previous benchmarks have reported audio LMs lacking temporal localization skills (Ahia et al., 2025; Kumar et al., 2025; Ma et al., 2025b), so this result is not surprising. In Figure 7, we analyze tasks that require identifying artist roles within a song at varying levels of granularity. Models are generally more accurate at identifying main artists than featured artists, and perform worst on cases where artists are neither main nor featured.

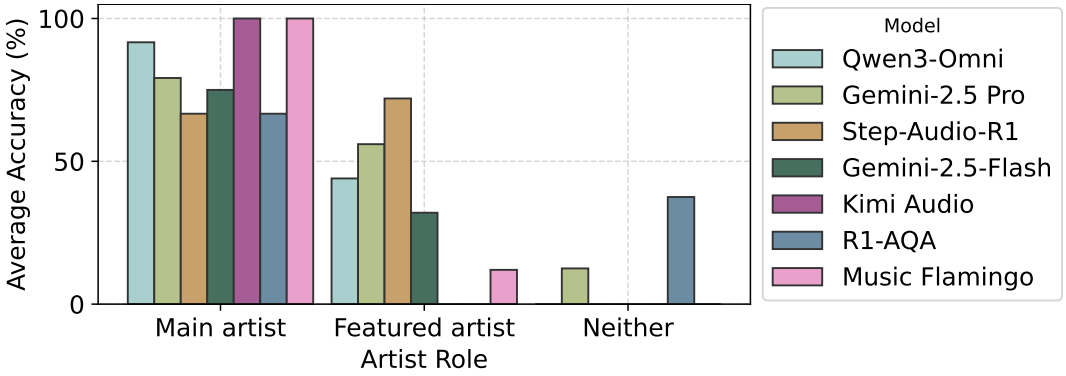

Figure 7: Model accuracy across different artist roles. Accuracy is highest for main artists, lower for featured artists, and lowest for artists in neither category.

For vocal style delivery, we analyze performance differences between identifying singing and rapping (Figure 9). Most models achieve higher accuracy on questions related to rapping, with the widest gap for Gemini Flash.

Finally, we quantified prediction errors across artist counting and duration-related tasks (see Appendix Figure 13 and 14). Gemini Pro achieves near-zero error, while Audio-Flamingo, Audio-Reasoner, and Phi-mm Instruct tend to underestimate durations, and Gemini Flash and R1 AQA overestimate. Similar trends appear in artist counting, with most models underestimates artist counts on the full songs. Extreme outliers (e.g., Gemma, Qwen) are omitted from the plots due to hallucinated errors exceeding 10^7 seconds that would dominate the scale of the plot.

C Computational Resources

For each models smaller than 8B parameters, a single A40 or L40 GPU was used to run the experiments. For Voxtral-Small-24B and Qwen3-Omni, 4 A40 or L40 GPUs were used to run the experiments. For Gemini 2.5 Pro and Gemini 2.5 Flash, we queried from the models through GCP.

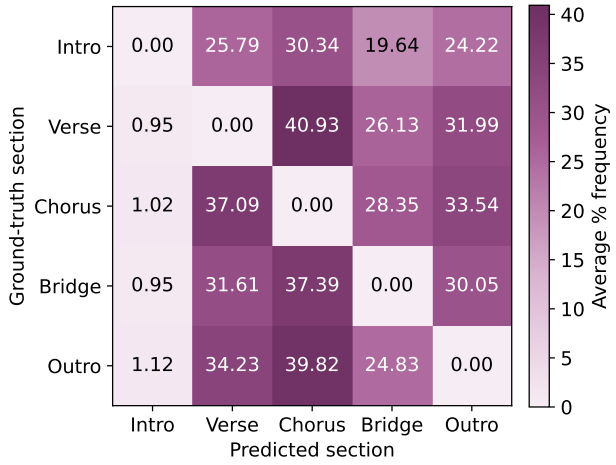

Figure 8: Average section prediction errors across models for structural segmentation, highlighting frequent confusion between choruses and verse sections.

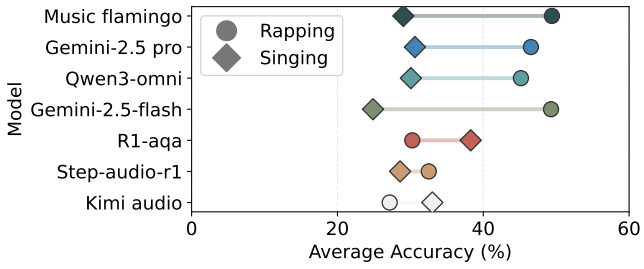

Figure 9: Model accuracy for singing versus rapping vocal delivery. Most models identify rapping more accurately than singing, likely due to rapping's prosodic similarity to speech.

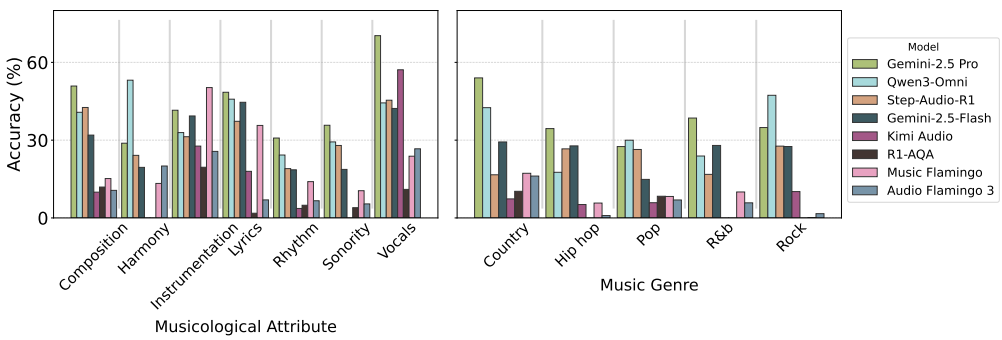

Figure 10: Model performance across musicological attributes and genres in the musicological analysis task. **(Left)** Average performance across specific attributes; while models excel in vocals and lyrics, attributes like rhythm, instrumentation, and sonority remain challenging even for frontier models. **(Right)** Performance breakdown by genre, showing that models generally perform best on Country and Rock while experiencing a noticeable dip on Hip Hop and Pop.

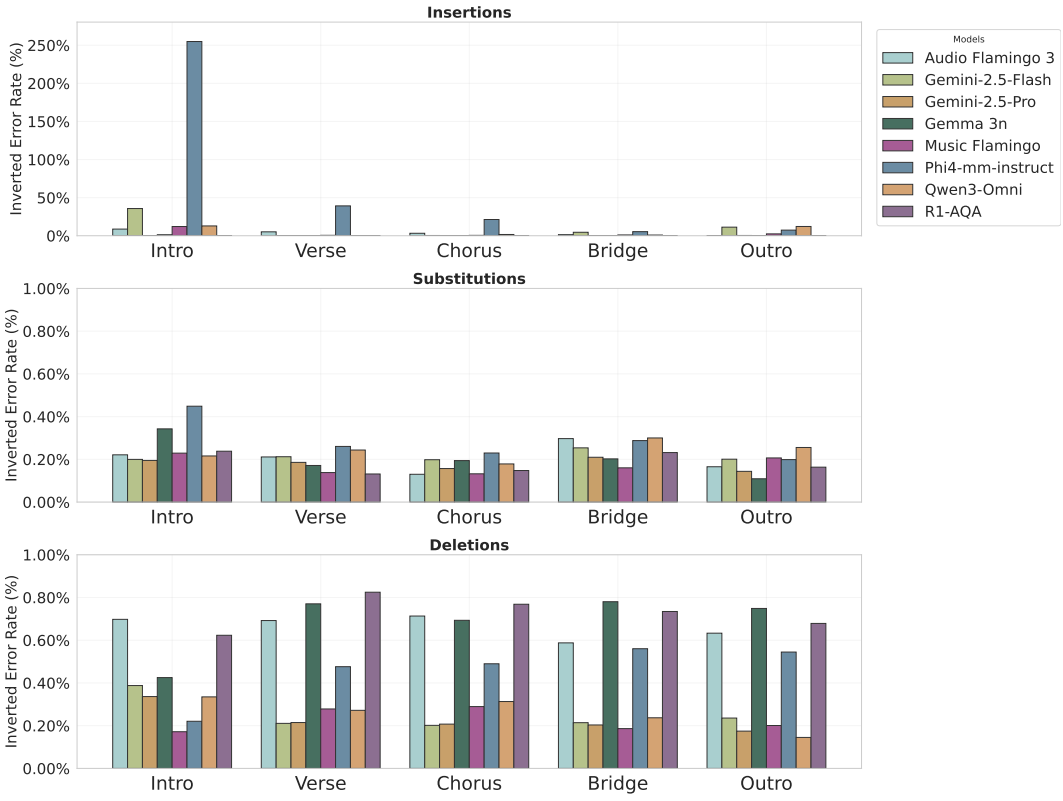

Figure 11: Analysis of inverted error rates by song section. Breakdown of Inverted **insertion**, **substitution**, and **deletion** error rates across sections. Note the significantly higher scale for **insertion** errors **(up to 250%)** compared to **substitutions** and **deletions** (under 1.00%), particularly in the *Intro* section.

D Confidence Intervals

To quantify the uncertainty in our reported scores that arises from the finite set of songs in the benchmark, we compute 95% confidence intervals using a bootstrap over songs.

Because BASS contains multiple questions per song, questions from the same song are not independent: model errors on them are likely correlated. Bootstrapping at the question level would

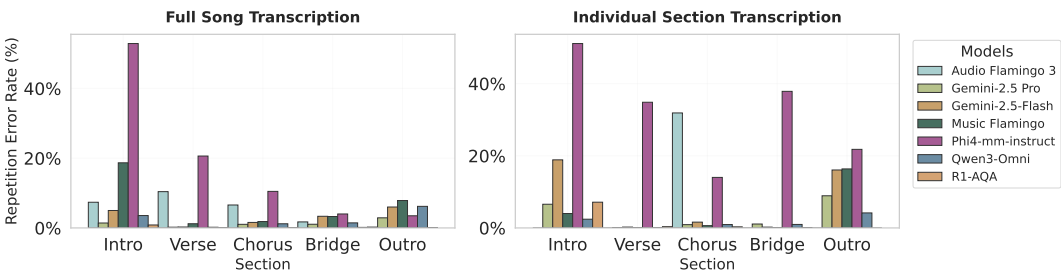

Figure 12: Repetition error rates by section for Full and Section transcription. We observe higher repetition errors (phrase hallucinations) in the *Intro* across both tasks, with Section transcription consistently exhibiting higher repetition rates than Full transcription.

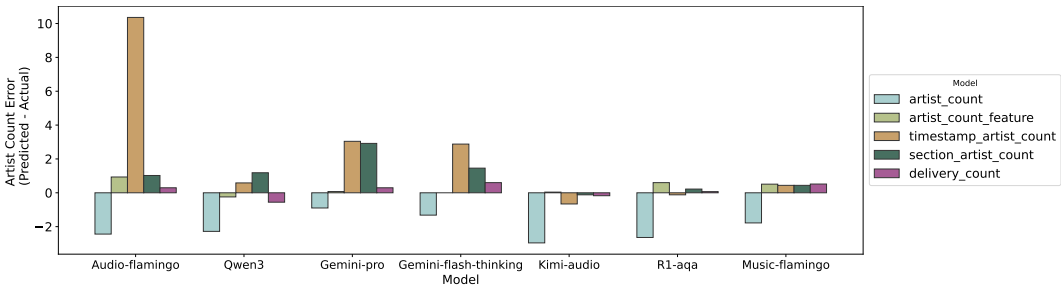

Figure 13: Model performance across artist counting subtasks. The y – axis shows error (Predicted – Actual), indicating overestimation (positive) or underestimation (negative) of artist and delivery counts.

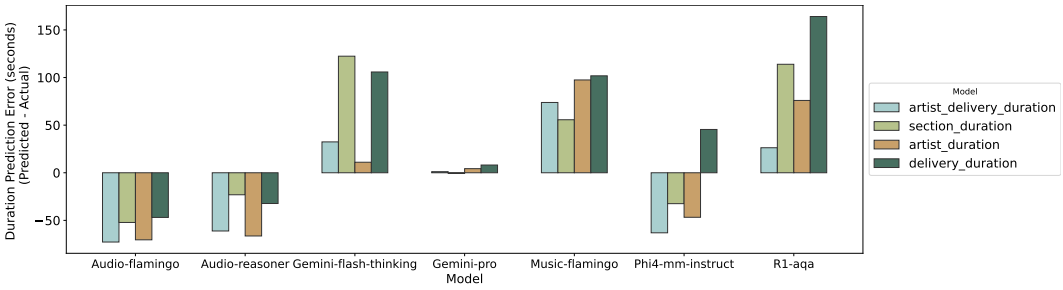

Figure 14: Temporal prediction errors (seconds) across artist duration tasks. The y – axis shows error (Predicted – Actual), indicating overestimation (positive) or underestimation (negative) of duration.

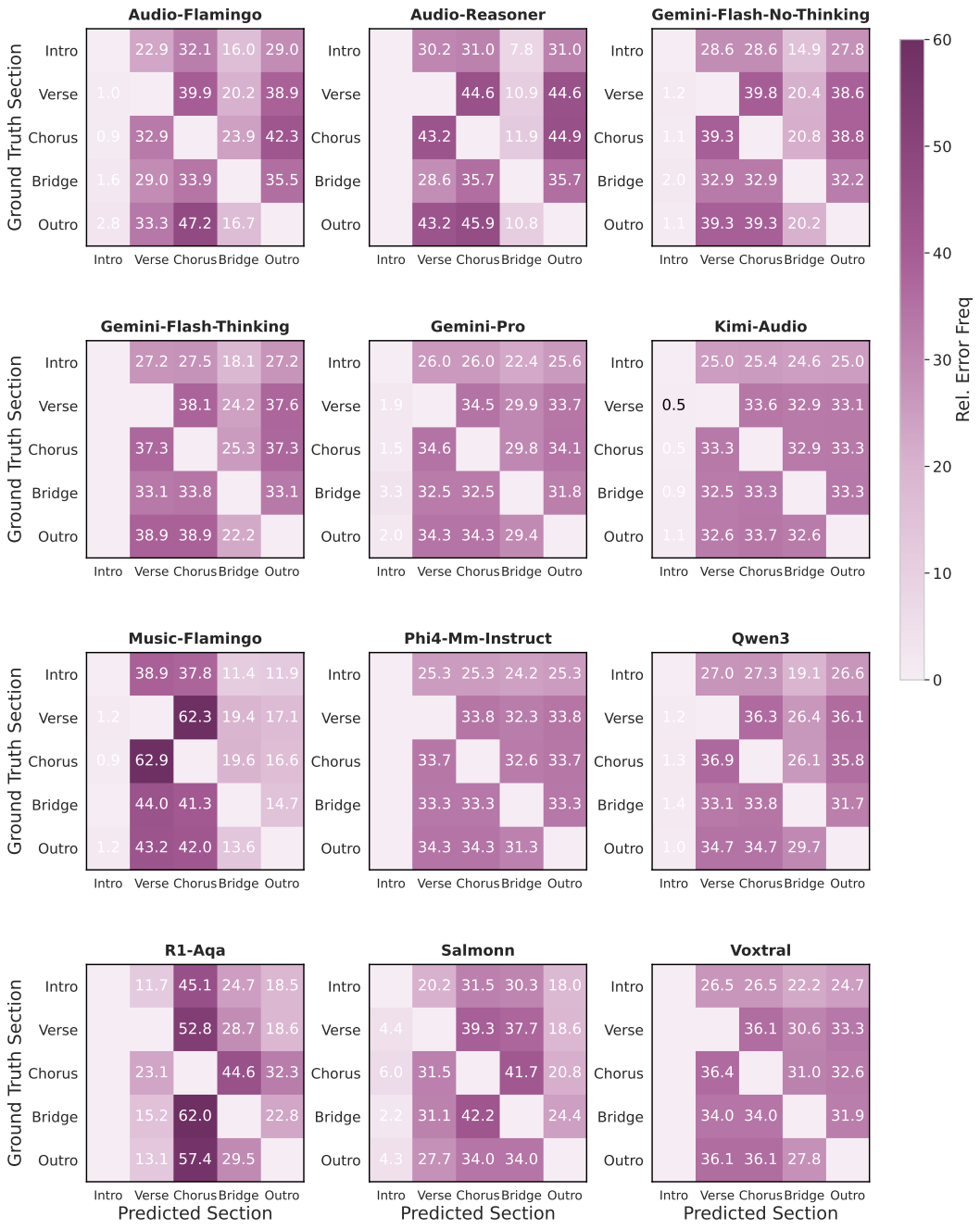

Figure 15: Section prediction errors across models for structural segmentation, highlighting frequent confusion between choruses and verse sections.

therefore treat correlated observations as independent and understate uncertainty. We instead use a song-level clustered bootstrap, stratified by genre. In each bootstrap iteration, we resample songs with replacement within each genre, keeping the number of songs per genre fixed, retain all questions associated with each sampled song, and recompute [the task-level and genre-level] scores. We repeat this procedure 10,000 times and report the 2.5th and 97.5th percentiles of the resulting score distribution as the 95% confidence interval.

These intervals capture uncertainty due to the sampling of songs, but not variance arising from stochastic decoding across repeated runs, which our single-run evaluation cannot measure. They should therefore be interpreted as a lower bound on total uncertainty.

Table 6: Tasks, example questions, and specific formatting instructions for all categories.

Category	Task	Subtasks	Example Question & Expected Output Format
Structural Segmentation	Full Structural Segmentation	–	Divide the track into sections corresponding to any of these categories: Verse, Chorus, Bridge, Intro, Outro, Pre-Chorus, Post-Chorus. For each section, provide the following: section title, start and end timestamps of the entire section. **Format:** `[ {"section": "<title>", "start": <ts>, "end": <ts> }, ... ]`
	Section Structural Segmentation	–	Given the song, provide the start and end timestamp (in seconds with milliseconds) of all Verse sections. **Format:** `[ {"start": <ts>, "end": <ts> }, ... ]`
Lyrics Transcription	Full Structural Lyrics Transcription	–	You are provided with the structural segmentation of a song. Please transcribe the lyrics for each section. **Format:** `[ {"section": "<section1>", "lyrics": "..." }, ... ]`
	Section Structural Lyrics Transcription	–	Could you provide the lyrics for the all verse sections section only? The transcription of each occurrence of the given section should be treated as a separate item in a list. **Format:** `[ {"section": "section_name", "lyrics": "..." } ]`
Musicological Analysis	Single-Gene Detection	–	Given the following musical properties, which one is most dominant in the song? Answer Choices: Compound Meter, Percussion, Vocal Grittiness, Synthesizer. **Format:** Respond with only the name of the musicological feature.
	Pairwise-Gene Detection	–	Which pair of features is most strongly present in the track? **Format:** Respond with only a list of two names. Example: `['<property1>', '<property2>']`
	Gene Attribution	–	Which song features the highest Vocal Register? (Property measures pitch range of lead singer). **Format:** Respond with only the number (1, 2, 3, or 4) corresponding to the song.
	Gene Dominance Ranking	–	Rank the following rhythmic patterns and feels from least to most prominent. **Format:** Respond with only a list of the names ranked. Example: `['<prop1>', '<prop2>']`
Artist Collaboration	Artist Counting	Standard / Featured	What is the number of artists performing? / How many performers are featured artists?
		Vocal Delivery	In this song, what number of artists are performing rap sections?
		Section / Temporal	How many artists are performing in the Post-chorus? / From 211 to 249 seconds? **Format (All Counting):** Respond with only the number.
	Artist Duration	Target / Vocal	What is the total duration of performance by artist 2? / How many seconds are devoted to rapping?
		Artist Delivery / Section	Total duration of rrapped parts by artist 1? / How long do Bridge sections last? **Format (All Duration):** Respond with only the number in seconds with milliseconds.
	Artist Localization	–	At what second does artist 2 first start performing in this song? **Format:** Respond with only the timestamp number in seconds with milliseconds.
	Artist Attribution	Vocal Comparison	Do the featured artist in the first Verse and the main artist in the second Verse use the same type of vocal delivery? **Format:** Respond with only: `same` or `different`.
		Section Style / Role	For the first Chorus, is the artist rapping, singing, or neither? Is it the main or featured artist? **Format:** `Rapping`, `Singing`, `Neither` OR `Main artist`, `Featured Artist`, `Neither`.

Table 7: Raw Performance scores **(before normalization)** of evaluated models on BASS across musicological analysis (Single-Gene Detection—SGD, Pairwise-Gene Detection—PGD, Gene Dominance Ranking—GDR, and Song-Level Gene Detection—Song-GD), structural segmentation, artist collaboration, and lyrics transcription tasks. The metric for Musicological Analysis and Artist Collaboration task is Exact Match accuracy (%). The metric for Structural Segmentation is IOU (%). The metric for Lyrics Transcription is WER (not in percentages).

		Musicological Analysis				Structural Segmentation		Artist Collaboration				Lyrics Transcription	
Model	**Size**	SGD	PGD	GDR	GA	Full	Section	Cnt.	Dur.	Loc.	Attr.	Full	Section
Gemini-2.5 Pro	-	73.73	10.88	21.20	48.90	32.34	22.98	25.77	8.25	3.12	56.66	1.35	2.40
Qwen3-Omni	30B	60.90	6.16	15.20	48.90	28.88	22.07	18.54	9.78	9.38	63.66	2.92	1.76
Step-Audio-R1	-	57.91	10.27	18.18	48.94	32.19	17.73	19.34	6.42	7.81	53.13	3.57	3.61
Gemini-2.5-Flash	-	57.61	6.82	24.20	42.60	24.47	15.15	20.40	7.66	3.12	56.81	18.51	15.97
Music Flamingo	8B	49.25	4.76	0.00	25.50	13.07	2.59	33.54	4.27	4.69	56.97	5.36	7.72
R1-AQA	8.2B	34.03	0.00	0.00	30.90	5.91	4.16	35.03	0.92	0.00	50.84	1.50	1.75
Gemma 3	5B	35.22	4.76	3.03	31.91	8.92	4.35	12.25	0.30	1.56	40.80	1.37	1.70
Audio Flamingo 3	-	44.20	8.18	0.00	30.90	7.03	0.01	21.92	1.84	0.00	49.80	4.68	2.76
Voxtral-Small-24B	24B	37.61	8.84	9.12	17.00	32.16	8.84	9.09	2.76	0.00	38.52	2.78	15.46
SALMONN-7B	7B	25.37	1.36	0.30	28.70	3.18	0.12	27.32	0.61	0.00	34.86	7.61	20.59
Kimi Audio	7B	37.61	4.76	0.00	0.00	13.06	2.64	19.12	0.92	0.00	52.07	18.33	54.56
Phi4-mm-instruct	5.5B	33.73	1.36	9.12	21.30	2.71	0.16	23.59	3.67	0.00	38.51	79.29	364.54
Mellow	167M	0.30	0.00	0.00	0.00	0.00	0.00	0.00	2.14	0.00	0.00	19.02	3.46
Audio-Reasoner	-	33.73	3.40	6.06	30.85	6.63	2.91	18.47	3.06	0.00	2.31	-	-

Table 8: Normalized performance of evaluated models on multilingual data across Full Structural Lyric Transcription, Section Structural Lyric Transcription, Standard Artist Counting, and Section Artist Counting tasks. Each score is upper bounded by 100$ and bold values highlight the highest value per task.

	Artist Collaboration		Lyrics Transcription	
Model	Standard Count	Section Count	Full Transcr.	Section Transcr.
Gemini-2.5 Pro	**41.05**	40.39	3.65	34.34
Step-Audio-R1	-	-	-	-
Qwen3-Omni	26.32	44.83	2.65	**45.78**
Gemini-2.5-Flash	28.42	30.54	2.27	4.54
Music Flamingo	32.63	45.81	5.60	17.53
R1-AQA	24.47	32.18	**17.70**	45.62
Audio Flamingo 3	-	-	-	-
Gemma 3	3.16	42.36	13.16	41.34
Voxtral-Small	6.32	**49.75**	-	-
SALMONN-7B	26.32	30.05	13.49	4.49
Phi4-mm-instruct	26.60	21.29	0.95	0.48
Audio-Reasoner	10.64	29.21	1.99	3.24
Kimi Audio	10.64	44.06	2.65	2.84
Mellow	2.11	17.73	7.51	24.47