

# Advancing the Foundation Model for Music Understanding

**Yi Jiang<sup>1</sup>, Wei Wang<sup>2\*</sup>, Xianwen Guo<sup>2</sup>, Huiyun Liu<sup>2</sup>, Hanrui Wang<sup>2</sup>, Youri Xu<sup>2</sup>, Haoqi Gu<sup>2</sup>, Zhongqian Xie<sup>2</sup>, Chuanjiang Luo<sup>2</sup>**

<sup>1</sup>Zhejiang University <sup>2</sup>NetEase Cloud Music

## Abstract

The field of Music Information Retrieval (MIR) is fragmented, with specialized models excelling at isolated tasks. In this work, we challenge this paradigm by introducing a unified foundation model named MuFun for holistic music understanding. Our model features a novel architecture that jointly processes instrumental and lyrical content, and is trained on a large-scale dataset covering diverse tasks such as genre classification, music tagging, and question answering. To facilitate robust evaluation, we also propose a new benchmark for multi-faceted music understanding called MuCUE (Music Comprehensive Understanding Evaluation). Experiments show our model significantly outperforms existing audio-large language models across the MuCUE tasks, demonstrating its state-of-the-art effectiveness and generalization ability.

## 1 Introduction

Music, a universal and multifaceted form of human expression, presents a formidable challenge for computational understanding. The field of Music Information Retrieval (MIR) has made significant strides in decoding its complex structures. However, progress has historically been characterized by a paradigm of fragmentation. Models are typically engineered as highly specialized experts, excelling at isolated tasks such as genre classification, beat tracking, or instrument recognition. While effective in their narrow domains, this specialization comes at a cost: a lack of holistic, integrative understanding that mirrors human cognition.

This fragmentation engenders significant limitations. Expert models struggle to generalize across tasks and often fail when confronted with complex, multi-faceted queries that require synergistic reasoning. For instance, answering “Why does this song evoke a sense of melancholy?” necessitates a concurrent analysis of its harmony, tempo, instrumentation, and lyrical content—a capability beyond the scope of single-task models. While recent general-purpose audio-language models possess impressive cross-modal capabilities, they are not intrinsically optimized for the unique structural and semantic nuances of music, often lacking the domain-specific acuity for fine-grained MIR tasks.

To bridge this gap and break the prevailing paradigm, we introduce a unified foundation model for holistic music un-

derstanding. Our model features an architecture capable of concurrently processing both instrumental audio and lyrical content, and is trained on a vast and diverse corpus of data spanning a multitude of MIR tasks. Instead of cultivating a collection of disparate specialists, our objective is to build a single, versatile generalist model that learns a shared, rich representation of music, enabling it to perform a wide array of tasks from a single set of weights.

A parallel challenge in the pursuit of holistic music understanding lies in its evaluation. The absence of a unified, comprehensive benchmark makes it difficult to compare models systematically or to measure true progress towards general musical intelligence. To address this critical need, we propose the Music Comprehensive Understanding Evaluation (MuCUE) benchmark. MuCUE’s core innovation is its standardized format, framing a wide spectrum of tasks—from low-level perception (e.g., pitch and chord recognition) to high-level cognition (e.g., mood and structural analysis)—as multiple-choice questions (MCQs). This approach not only facilitates objective and scalable evaluation but also provides a rigorous tool for probing the emergent reasoning abilities of foundation models, thereby guiding future research.

To summarize, our main contributions are as follows:

- **A Unified Foundation Model for Music:** We introduce a novel, end-to-end trainable foundation model that holistically understands music by jointly processing instrumental audio and lyrical content. Its architecture and a specialized long-context (390s) training regimen enable it to move beyond single-task limitations and achieve state-of-the-art performance on a wide array of MIR tasks.
- **The MuCUE Benchmark:** We propose the Music Comprehensive Understanding Evaluation (MuCUE), a new and extensive benchmark designed to systematically assess music AI capabilities. By framing diverse tasks from low-level perception to high-level cognition as multiple-choice questions, MuCUE provides a standardized and rigorous tool for measuring true progress in the field.

## 2 Related Work

Recent advancements in music-language understanding leverage frozen audio encoders (often MERT(Li et al. 2023)) integrated with large language models (LLMs) via

\* Correspondence to: shakespeare@zju.edu.cn

lightweight adapters to overcome music-text data scarcity. MU-LLaMA(Liu et al. 2023b) (built on LLaMA(Touvron et al. 2023)) pioneered this approach using audio-adapted LLaMA layers and the MusicQA dataset (synthesized from captions and tags), demonstrating strong QA and captioning performance. MusiLingo(Deng et al. 2024) refined this paradigm by aligning frozen MERT embeddings with LLMs like Vicuna(Chiang et al. 2023) through a simple linear projector; its key contribution is the high-quality MusicInstruct (MI) dataset for instruction-tuning, enabling robust open-ended QA and outperforming MU-LLaMA. Similarly, LLARK(Gardner et al. 2023) employs an adapter-based architecture trained on augmented data to excel at instruction-following tasks, including detailed captioning and musical reasoning. Expanding beyond understanding, M2UGen(Liu et al. 2023a) introduces a unified LLaMA 2-based framework combining comprehension (music QA, captioning) with cross-modal generation (text/image/video-to-music, editing), utilizing large synthetic instruction datasets (MusicCaps, MUEdit etc.) and LoRA fine-tuning to achieve SOTA across both understanding and creative tasks.

While these adapter-based models advance music-text alignment, significant limitations in understanding persist. Crucially, reliance on synthetic or limited datasets (e.g., MusicQA, MI derived primarily from MusicCaps(Agostinelli et al. 2023), MagnaTagATune(Law et al. 2009)) restricts scope and depth, potentially perpetuating biases and superficial connections. Scaling QA to complex, subjective musical concepts (emotion, structure, cultural context) remains challenging, as most datasets focus on factual tags or short descriptive captions. Evaluation inadequacies are pronounced: reliance on NLP metrics like BLEU or ROUGE poorly captures musical nuance, subjective meaning, or aesthetic relevance, especially for open-ended QA and long-form captioning.

There are also much breakthroughs of general audio large language models recently(Chu et al. 2024; Abouelenin et al. 2025; Xu et al. 2025; KimiTeam et al. 2025; Huang et al. 2025; Zeng et al. 2024; Fu et al. 2025; Li et al. 2025). For instance, Qwen2-Audio(Chu et al. 2024) establishes a high standard as an audio-language model by integrating a Whisper-large-v3(Radford et al. 2023) encoder with a Qwen-7B(Bai et al. 2023) LLM through a refined three-stage training pipeline (pre-training, SFT, DPO) to achieve state-of-the-art performance. Expanding the scope beyond audio-language pairs, Qwen2.5-Omni(Xu et al. 2025) operates as a truly end-to-end multimodal system processing text, image, audio, and video, distinguished by a "Thinker-Talker" framework for real-time streaming speech generation and a novel positional embedding (TM-RoPE) for synchronizing audio-visual inputs. Concurrently, Kimi-Audio(KimiTeam et al. 2025) proposes a universal audio foundation model, built on a hybrid architecture and pre-trained on over 13 million hours of diverse audio, aiming to unify perception, reasoning, and generation within a single, open-source framework. Although these models are not explicitly trained for Music Information Retrieval (MIR), their state-of-the-art speech recognition capabilities provide a strong foundation for interpreting lyrical music. Conse-

quently, their proficiency in processing lyrical content enables them to effectively perform ancillary MIR tasks such as lyrics transcription and, by extension, music genre classification or mood detection where lyrical themes are indicative.

A significant research gap emerges from this dichotomy, defining two distinct frontiers in music-language modeling. On one hand, adapter-based models are purpose-built for music but are often constrained by their architectural design, such as frozen audio backbones and shallow adapters, limiting their capacity for deep semantic representation. They can be characterized as specialized yet brittle. On the other hand, general-purpose multi-modal systems possess powerful, end-to-end trained architectures but lack the domain-specific optimization required to interpret the complex, non-linguistic syntax of music, such as harmony, structure, and expressive nuance. These models are powerful yet unspecialized for nuanced MIR tasks. Our work is strategically positioned at the confluence of these two paradigms, aiming to synergize the architectural power of modern multi-modal systems with the deep, domain-specific focus essential for MIR. The key differentiators of our approach are summarized in Table 1. Notably, our model benefits from full-parameter tuning across all components and, more critically, is trained on an extended audio context of up to 390 seconds—an order of magnitude greater than that of prior models. This strategic combination of a state-of-the-art foundation, comprehensive fine-tuning, and a novel long-context training regimen is designed to create a model that is both broadly capable and musically astute.

### 3 Model Architecture

The architecture of our model is grounded in the effective and scalable designs of recent multimodal large language models. This paradigm, which couples a powerful pre-trained audio encoder with a large language model backbone, provides a robust foundation for complex reasoning tasks. Our model is designed to accept an interleaved sequence of audio and text inputs and generate a coherent text output. Formally, for any input sequence of the form  $[A_1, T_1, A_2, T_2, \dots, A_n, T_n]$ , where  $A_i$  represents an audio file and  $T_i$  a text segment, each modality is first transformed into a sequence of embedding vectors. These embedding sequences are then concatenated and fed into the language model to produce the final output. The overall architecture, depicted in Figure 1, comprises three core components: a language model backbone, an audio encoder, and a connector module to bridge the two modalities.

#### Language Model Backbone

Our language model backbone is initialized from Qwen3-8B-Base (Yang et al. 2025). We selected this model for its state-of-the-art performance in language understanding and multilingual capabilities. Its strong foundational skills are crucial for interpreting the complex, often abstract, relationships within music and for generating nuanced, descriptive text. By leveraging such a powerful pre-trained LLM, we can focus our efforts on effectively translating musical in-

	MU-LLaMA	MusiLingo	LLARK	M2UGen	Qwen2-Audio	Kimi-Audio	ours
LLM	LLaMA-2 7B	Vicuna-7B	Llama2-7b	Llama2-7b	Qwen-7B	Qwen2.5-7B	Qwen3-8B
audio encoder	MERT	MERT	Jukebox	MERT	whisper	whisper	whisper
audio token frequency(Hz)	fixed embedding	0.17	10	fixed embedding	25	12.5	10
max train duration(s)	29	30	25	10	30(likely)	30(likely)	390
tune modules	adapter	adapter	projection, LLM	adapters, LLM(lora)	unknown	all	all

Table 1: Comparison of some Audio Language Models(ALMs)

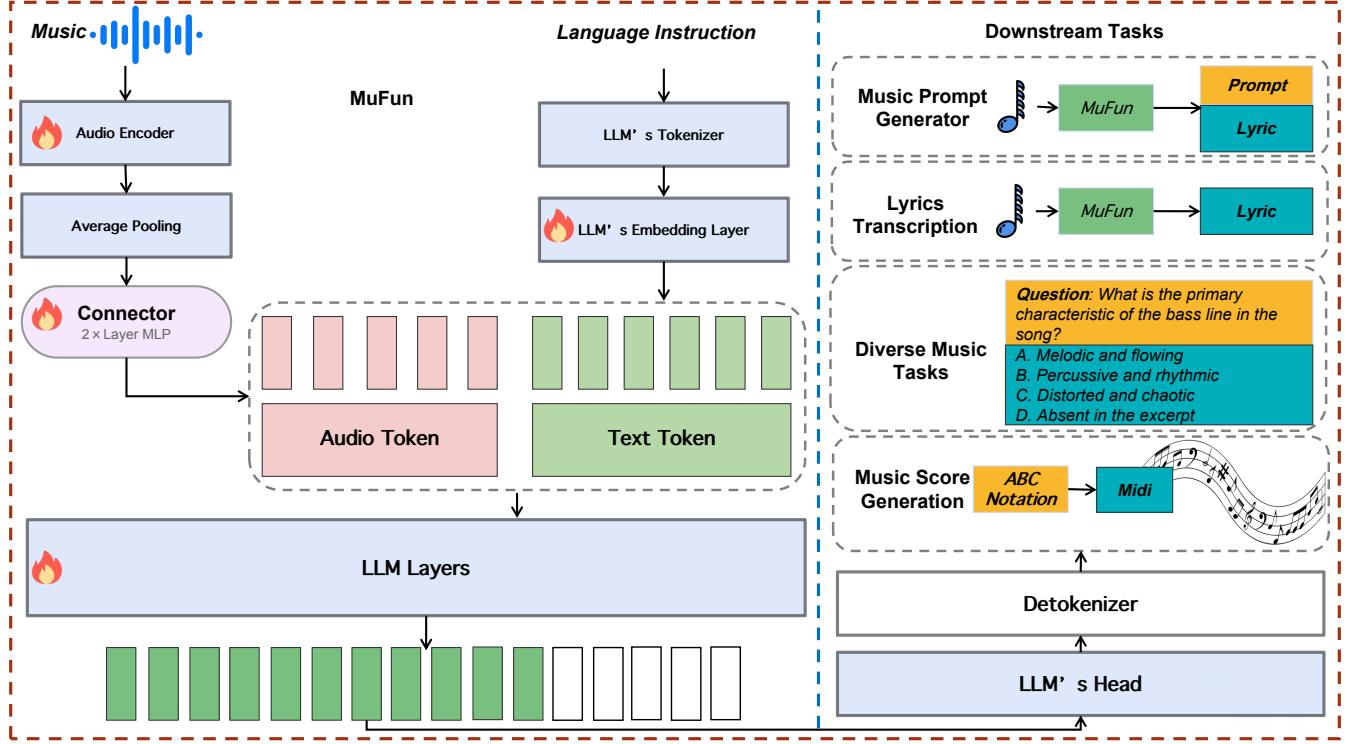


Figure 1: Overview of the MuFun model architecture

formation into a "language" that the LLM can comprehend.

### Audio Encoding and Multi-Layer Feature Fusion

The audio processing module is responsible for converting raw audio waveforms into meaningful feature representations. For this, we initialize the encoder from Whisper-large-v3 (Radford et al. 2023) as our audio backbone. The motivation for this choice is twofold: Whisper is pre-trained on an enormous and diverse dataset of audio, enabling it to learn a highly robust and general-purpose representation of acoustic phenomena that is transferable to music. Furthermore, its transformer-based architecture is well-suited for capturing temporal dependencies.

To create a comprehensive representation of the audio, we do not rely solely on the final output layer of the Whisper encoder. Instead, we adopt a multi-layer feature fusion strategy. Specifically, we extract the hidden states from four distinct layers of the encoder—layers 0, 7, 15, and 32—and concatenate them. This results in a rich feature vector with a

dimension of 5120 ( $1280 \times 4$ ). The rationale behind this approach is that different layers of a deep network capture different levels of abstraction. Early layers (e.g., layer 0) tend to preserve low-level acoustic details like timbre and pitch, while deeper layers (e.g., layer 32) capture more abstract, semantic information like melodic contours and rhythmic patterns. By providing the model with this multi-resolution view, we empower it to access both fine-grained textural details and high-level structural information simultaneously, a critical requirement for holistic music understanding.

### Temporal Downsampling and Long-Context Handling

The Whisper encoder processes a 30-second audio clip into a sequence of 1500 embedding vectors, corresponding to a temporal frequency of 50 Hz. This high density of tokens can be computationally burdensome for the LLM and may not align well with the typical information density of text. To address this, we apply a temporal downsampling step. We use a 1D mean pooling layer with a kernel size and stride of

Table 2: Training Configuration across Different Stages of Model Development

Sub-stages	Pretraining					Finetuning	
	Warmup	Align1	Align2	Context Extending	Short Music	Long Music	
Training steps	400	2500	3500	540	1000	700	
Batch size	384	384	384	144	384	144	
Tune modules	connector	all	all	all	all	all	
Tasks	speech transcription; music score transcription	speech transcription; music score transcription	speech transcription; music score transcription; lyrics transcription; pitch, instrument identification	music score transcription; lyrics transcription	various MIR tasks (short audio or segments)	various MIR tasks (mainly song level)	
Max audio duration (s)	30	30	30	390	30	390	
GPU hours	10	180	244	175	68	180	



Figure 2: Pipeline of the Audio processing Module

5 along the time dimension. This operation reduces the audio token frequency to a more manageable 10 Hz, achieving two goals: 1) it significantly reduces the sequence length, improving computational efficiency, and 2) it smooths the representation, encouraging the model to focus on more salient temporal events.

A key ability of our model is to process long-form, song-level audio. To handle inputs exceeding the 30-second window of the Whisper encoder, we employ a straightforward yet effective chunking strategy. The long audio stream is first segmented into 30-second non-overlapping chunks. Each chunk is processed independently by the audio encoder and pooling layer. The resulting embedding sequences are then concatenated in their original order to form a single, continuous sequence representing the entire audio piece. This mechanism extends the model’s effective receptive field to any audio duration, enabling true song-level analysis.

## The Connector Module

The final architectural component is the connector, which serves as a bridge between the audio and language modalities. Its purpose is to project the 5120-dimensional audio embeddings into the 4096-dimensional space of the Qwen3 language model. For this, we use a 2-layer Multilayer Perceptron (MLP). The MLP first expands the input dimension by a factor of two, applies a GELU non-linear activation function, and then projects it down to the target dimension of the LLM. Using a non-linear MLP instead of a simple linear projection affords greater expressive power(Liu et al. 2024), allowing for a more complex and nuanced alignment

between the learned representations of music and language. This trainable ”translator” is vital for harmonizing the two modalities effectively.

## 4 Training

The development of our model’s comprehensive musical understanding is facilitated by a meticulously designed, multi-stage training regimen. This protocol is not a monolithic process but rather a strategic curriculum designed to progressively build capabilities. We employ a curriculum learning approach, systematically advancing from foundational audio-text alignment to sophisticated, long-context musical reasoning. The complexity of the tasks and the length of the audio context are gradually increased, ensuring a stable and efficient learning trajectory. The entire training protocol, summarized in Table 2, is divided into two primary phases: a four-stage pre-training phase to build a robust foundation, and a dual-track fine-tuning phase to specialize the model for diverse MIR applications.

We construct our training data (primarily from public available datasets) in a straight-forward way, unlike many prior works which mostly utilize instruction format. We tend to put as many text labels as possible in a single sample for better efficiency, for which details and some samples are provided in the appendix. Experiments are conducted using NVIDIA A100 40 GB GPUs, at most 16 ones across two nodes.



Figure 3: Handing for Full Song Length Audio

## Pre-training

The objective of the pre-training phase is to establish a strong, fundamental alignment between the audio and language modalities and to imbue the model with core perceptual abilities. This is accomplished over four distinct stages:

**Stage 1: Warmup (Connector Initialization)** The training begins with a brief 400-step warmup stage. In this initial step, we freeze the parameters of both the audio encoder and the LLM, training only the connector module. The rationale is to establish a stable bridge between the two powerful, pre-trained backbones without the risk of destabilizing their well-formed representations with large, chaotic gradients from a randomly initialized component. The tasks are limited to basic speech and music score transcription on 30-second clips, providing a clean, direct signal for the connector to learn the initial modality mapping.

**Stage 2: Initial Full-Parameter Alignment (Align1)** Once the connector is stabilized, we unfreeze all model parameters and begin end-to-end training for 2500 steps. By co-adapting the entire model on the same foundational transcription tasks, we allow the audio encoder and the LLM to gently adjust to one another, deepening the cross-modal alignment beyond the connector. This stage encourages the model to learn a shared representational space more profoundly.

**Stage 3: Capability Enrichment (Align2)** With the model architecture fully aligned, we broaden its musical knowledge base over 3500 steps. The task set is expanded to include more complex and musically salient objectives: lyrics transcription, pitch identification, and instrument identification. This strategic shift moves the model’s learning from simple acoustic-to-text mapping towards genuine musical concept recognition. It is at this stage that the model learns the specific acoustic signatures corresponding to textual concepts like “guitar,” “C4 pitch,” or lyrical content.

**Stage 4: Long-Context Extension** A pivotal and defining stage of our pre-training is the extension to long-form audio. In these 540 steps, the training audio duration is dramatically increased from 30 seconds to a max of 390 seconds. To accommodate the significant memory demands of these long sequences, the batch size is necessarily reduced from 384 to 144. The tasks are focused on music score and lyrics transcription over these extended contexts. The purpose of this stage is to force the model to learn long-range temporal dependencies that are invisible in short clips, such as the verse-chorus structure of a song or the development of a melodic theme. This endows our model with the capacity for true song-level analysis, a critical differentiator from prior work.

## Fine-tuning

Following the comprehensive pre-training phase, the model has developed a robust and general understanding of musical concepts. The fine-tuning phase aims to adapt this generalist foundation into a highly proficient expert for a wide array of MIR applications. This is accomplished through a sequential, two-stage fine-tuning curriculum that further refines the model’s capabilities, first by mastering diverse tasks on short audio segments and then by applying this knowledge to the complexity of full-length musical pieces.

**Stage 1: Short Music Fine-tuning** The first stage of fine-tuning involves training the model for 1000 steps on a diverse mixture of MIR tasks using 30-second audio segments. The objective here is to expose the model to the rich variety of music understanding tasks found in the wild—from genre and mood classification to more technical analyses like key and tempo detection. By focusing on short, information-dense clips, we can efficiently train the model across this wide task distribution, sharpening its ability to perform fine-grained, segment-level analysis. This stage solidifies the model’s grasp of specific musical attributes before it is asked to integrate them over longer time horizons.

**Stage 2: Long Music Fine-tuning** Building directly upon the refined capabilities from the previous stage, the model then undergoes a final 700 steps of fine-tuning, this time using the full 390-second audio context. The purpose of this stage is to transfer the task-specific knowledge learned on short clips to scenarios requiring holistic, song-level reasoning. The model learns to apply its understanding of genre, mood, and structure to entire musical narratives, tracking their evolution and interplay throughout a complete song. This sequential process—first mastering the concepts, then applying them at scale—ensures that the model develops a cohesive and hierarchical understanding, making it adept at both microscopic precision and macroscopic interpretation. This final step is crucial for equipping the model with the comprehensive analytical skills evaluated in our MuCUE benchmark.

## 5 Evaluation

### The MuCUE Benchmark

The evaluation of large audio models’ deep musical intelligence remains a significant challenge, largely due to the lack of a unified and multi-faceted benchmark. To bridge this critical gap, we present MuCUE (Music Comprehensive Understanding Evaluation), a novel benchmark that systematically assesses a wide array of music perception and cognition tasks. Its core innovation lies in framing all evaluation tasks—from low-level pitch and chord recognition to high-level genre, mood, and structural analysis—as multiple-choice questions (MCQs). This standardized format facilitates straightforward, scalable evaluation and is particularly suited for probing the emergent reasoning abilities of generative and foundation models. MuCUE thereby provides the research community with a holistic and rigorous tool to measure the true progress of music AI, identify model weaknesses, and guide future innovations in the field.

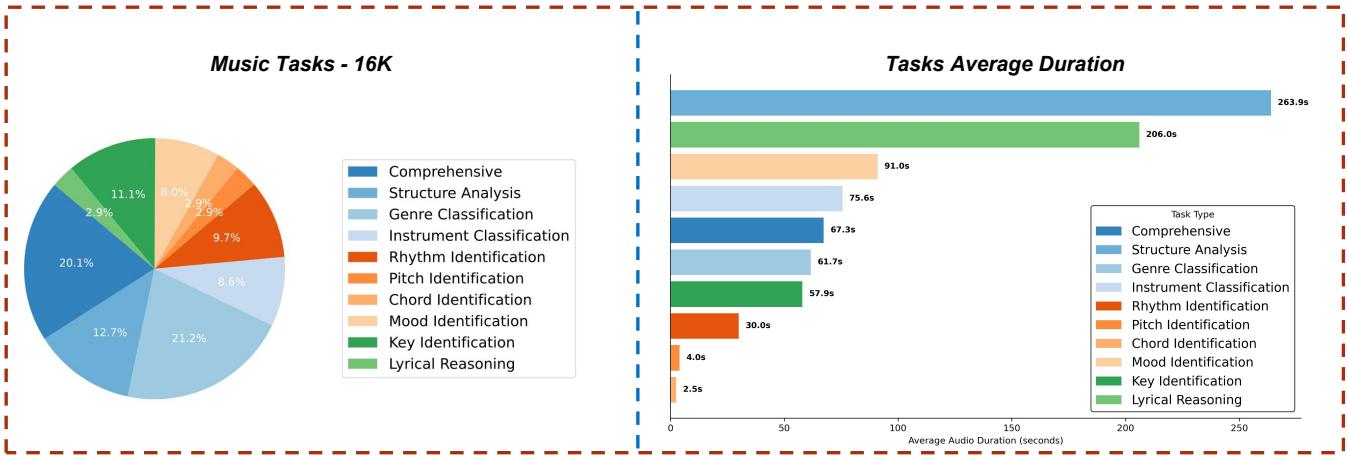


Figure 4: Dataset distribution of MuCUE. MuCUE contains a total of 16,463 samples. The left side shows the distribution of sample quantities across 10 task categories, while the right side displays the distribution of audio durations among the 10 task categories.

The construction process of evaluation data are detailed in appendix. For datasets in large quantity we hold out a portion for evaluation in the first place and leave the rest for training, so there is no possibility of contamination at least for our model.

## Main Results and Analysis

The comprehensive evaluation results on the MuCUE benchmark are presented in Table 3. We evaluate some open-weighted models(Chu et al. 2024; Xu et al. 2025; KimiTeam et al. 2025) as well as a proprietary model. Upon tests, some other audio language models like MU-LLaMA(Liu et al. 2023b) have difficulty in doing this kind of MCQs, it becomes unfair to compare in this way, so we exclude them here. The results unequivocally demonstrate the superior performance of our proposed model. Achieving an average score of 65.7, our model establishes a new state-of-the-art, outperforming the next-best model, Qwen2.5-Omni, by a significant margin of over 15 points in average accuracy. This substantial improvement across a diverse set of 26 tasks underscores the efficacy of our unified architecture and targeted training strategy.

Our model’s strength is particularly evident in low-level perceptual tasks that require fine-grained audio analysis. For instance, it achieves remarkable scores of 77.2 on pitch identification (`nsyn_pitch`), 58.8 on chord recognition (`guitarset`), and a commanding 91.2 on instrument classification (`ins_cls`). We attribute this success to our architectural design, specifically the fusion of features from multiple encoder layers (0, 7, 15, and 32), which provides the model with a rich, multi-resolution representation of the audio signal. Furthermore, the inclusion of similar tasks during the pre-training alignment phase directly cultivated these foundational capabilities, enabling the model to excel where others falter.

Beyond granular perception, our model excels at high-level cognitive tasks that depend on understanding long-

range temporal dependencies. Its exceptional performance on music structure analysis, such as segment boundary detection (`salami_segd` at 64.8), and lyrical reasoning (`lyr` at 90.8), highlights this capability. This proficiency is a direct outcome of our novel context-extending training stage, where the model was explicitly trained on audio contexts of up to 390 seconds. This stage enabled the model to learn the hierarchical and narrative structures inherent in complete musical pieces, a feat unattainable by models limited to short 30-second clips.

A nuanced analysis also reveals areas for further investigation. While dominant in most tasks, our model’s performance on overall structural summary (`salami_overall` at 48.7) is surpassed by Gemini. This may suggest that tasks requiring high-level abstract summarization are more heavily influenced by the raw reasoning power of the underlying LLM, where proprietary, larger-scale models may hold an advantage. Conversely, on the popular GTZAN genre classification task, our model (81.3) is competitive but behind Qwen2.5-Omni (88.6). This could be attributed to differences in pre-training data composition, as GTZAN is a ubiquitous benchmark that may feature more prominently in the training of other general-purpose models. These results highlight the intricate interplay between model architecture, training data, and task-specific requirements, providing valuable insights for future work.

## Downstream Applications

Our trained model is more like a base model, which can be further finetuned to adapt to more specific downstream tasks. For instance, we further train an instruct version and try out reinforcement learning like GRPO((Shao et al. 2024)) afterwards. We also make a prompt generator for a music generation model ACE-Step(Guo 2025) which takes in a song and output some prompts and lyrics for generation. More examples are shown in appendix. This demonstrates the model’s strong generalization ability and flexibility.

Table 3: Evaluation Results on various music understanding tasks. We group the datasets by their task category for better readability. The highest score in each row is highlighted in **bold**.

Task	Dataset	Gemini-2.0-flash	Qwen2.5-Omni	Kimi-Audio	Qwen2-Audio	ours
Key Identification	gs_key_30s(Knees et al. 2015)	33.6	23.8	26.0	18.2	<b>50.4</b>
	gtzan_key(Tzanetakis and Cook 2002)	33.7	28.7	28.3	22.0	<b>34.1</b>
Pitch Identification	nsyn.pitch(Engel et al. 2017)	30.8	36.8	31.8	31.2	<b>77.2</b>
Chord Identification	guitarset(Xi et al. 2018)	25.2	13.2	27.2	19.2	<b>58.8</b>
Rhythm Identification	ballroom_tempo(Gouyon et al. 2006)	<b>31.7</b>	28.9	24.4	31.1	29.4
	gtzan_tempo(Tzanetakis and Cook 2002)	<b>41.3</b>	32.4	22.9	27.1	40.7
Instrument Classification	ins_cls(Abdulvahap 2024)	26.0	66.8	79.4	39.8	<b>91.2</b>
	nsyn_ins(Engel et al. 2017)	32.4	40.6	44.4	22.4	<b>74.0</b>
	mtg_ins(Bogdanov et al. 2019)	19.8	55.8	51.2	24.0	<b>68.6</b>
Genre Classification	gtzan(Tzanetakis and Cook 2002)	72.2	<b>88.6</b>	77.8	83.9	81.3
	fma-small(Defferrard et al. 2017)	63.4	66.2	55.8	65.6	<b>72.4</b>
	fma-medium(Defferrard et al. 2017)	62.8	78.0	59.8	77.0	<b>85.2</b>
	mtg_genre(Bogdanov et al. 2019)	57.2	61.6	55.8	46.4	<b>81.4</b>
Mood Identification	ballroom_genres(Gouyon et al. 2006)	<b>57.0</b>	45.8	44.0	35.2	52.4
	mtg_mood(Bogdanov et al. 2019)	38.2	43.4	39.4	29.2	<b>52.8</b>
Structure Analysis	md4q(Panda, Malheiro, and Paiva 2018)	<b>71.9</b>	47.6	61.3	57.8	65.9
	salami_segd(Smith et al. 2011)	40.6	18.6	27.2	19.4	<b>64.8</b>
	salami_pred(Smith et al. 2011)	37.6	32.2	34.6	31.2	<b>64.8</b>
	salami_cnt(Smith et al. 2011)	<b>49.8</b>	36.8	37.8	30.2	43.2
Lyrical Reasoning	salami_overall(Smith et al. 2011)	<b>62.1</b>	55.8	45.3	42.6	48.7
	lyr(internal testset for lyrical reasoning)	88.2	87.4	87.0	60.0	<b>90.8</b>
Comprehensive	mmau-music(Sakshi et al. 2024)	<b>67.1</b>	63.8	66.2	57.8	66.5
	tat(Law et al. 2009)	61.2	59.4	54.0	61.4	<b>80.6</b>
	mucho(Weck et al. 2024)	69.6	66.5	<b>69.7</b>	66.7	63.9
	mcaps(Agostinelli et al. 2023)	62.2	65.6	68.0	74.0	<b>80.0</b>
	mqa(internal testset for music QA)	58.0	76.0	79.0	60.8	<b>88.4</b>
Average		49.8	50.8	49.9	43.6	<b>65.7</b>

variants	instruct	combined finetuning	last hidden
Avgerror on MuCUE(compared with base)	+0.078	-0.143	-1.250

Table 4: Ablation Study Results

## Ablation Study

To validate our key design choices, we conducted several ablation experiments (see Table 4). First, we assess our multi-layer feature fusion by training a variant using only the last hidden layer of the audio encoder with the same data. This resulted in a performance degradation, confirming that a rich, multi-resolution audio representation is critical for our model’s success. Next, we test our sequential finetuning curriculum by combining the short and long-context data into a single stage. This led to a small performance drop, suggesting our staged approach provides a more effective learning path. Finally, adding a post-hoc instruction-tuning stage yielded a marginal gain, indicating that our core training regimen already aligns the model effectively for question-answering tasks within the music domain. Collectively, these results underscore the importance of our proposed feature fusion and sequential training strategies.

## 6 Conclusion and Future Work

In this work, we introduced a unified foundation model called MuFun that significantly advances the state-of-the-art in holistic music understanding. By leveraging a multi-layer feature fusion architecture and a novel, long-context (390s) training curriculum, our model successfully overcomes the task fragmentation common in Music Information Retrieval. Its superior performance is demonstrated on MuCUE, a comprehensive new benchmark we developed to systematically evaluate a wide spectrum of musical abilities via a standardized multiple-choice question format. Our results validate that a single, strategically trained model can achieve both fine-grained perceptual accuracy and high-level cognitive reasoning, setting a new standard for the field.

While our model demonstrates powerful capabilities, future work can address its current limitations. Our immediate goals are to enhance data efficiency through semi-supervised and self-supervised learning to reduce reliance on large annotated corpora. We also plan to extend the model from a pure understanding system into a unified framework for both music analysis and generation. Further research will focus on developing evaluation methods that capture the subjective and creative aspects of music, and on exploring cross-modal applications that connect music with other domains like video and dance, paving the way for more sophisticated computational creativity.

## References

- Abdulvahap. 2024. Music Instrument Sounds for Classification.
- Abouelenin, A.; Ashfaq, A.; Atkinson, A.; Awadalla, H.; Bach, N.; Bao, J.; Benhaim, A.; Cai, M.; Chaudhary, V.; Chen, C.; et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.
- Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.
- Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 4211–4215.
- Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; Hui, B.; Ji, L.; Li, M.; Lin, J.; Lin, R.; Liu, D.; Liu, G.; Lu, C.; Lu, K.; Ma, J.; Men, R.; Ren, X.; Ren, X.; Tan, C.; Tan, S.; Tu, J.; Wang, P.; Wang, S.; Wang, W.; Wu, S.; Xu, B.; Xu, J.; Yang, A.; Yang, H.; Yang, J.; Yang, S.; Yao, Y.; Yu, B.; Yuan, H.; Yuan, Z.; Zhang, J.; Zhang, X.; Zhang, Y.; Zhang, Z.; Zhou, C.; Zhou, J.; Zhou, X.; and Zhu, T. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*.
- Bogdanov, D.; Won, M.; Tovstogan, P.; Porter, A.; and Serra, X. 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.
- Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%\* ChatGPT Quality.
- Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; Zhou, C.; and Zhou, J. 2024. Qwen2-Audio Technical Report. *arXiv preprint arXiv:2407.10759*.
- Defferrard, M.; Benzi, K.; Vandergheynst, P.; and Bresson, X. 2017. FMA: A Dataset for Music Analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*.
- Deng, Z.; Ma, Y.; Liu, Y.; Guo, R.; Zhang, G.; Chen, W.; Huang, W.; and Benetos, E. 2024. MusiLingo: Bridging Music and Text with Pre-trained Language Models for Music Captioning and Query Response. In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*. Association for Computational Linguistics.
- Engel, J.; Resnick, C.; Roberts, A.; Dieleman, S.; Norouzi, M.; Eck, D.; and Simonyan, K. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. In *International conference on machine learning*, 1068–1077. PMLR.
- Fu, C.; Lin, H.; Wang, X.; Zhang, Y.-F.; Shen, Y.; Liu, X.; Li, Y.; Long, Z.; Gao, H.; Li, K.; et al. 2025. VITA-1.5: Towards GPT-4o Level Real-Time Vision and Speech Interaction. *arXiv preprint arXiv:2501.01957*.
- Galvez, D.; Diamos, G.; Ciro, J.; Cerón, J. F.; Achorn, K.; Gopi, A.; Kanter, D.; Lam, M.; Mazumder, M.; and Reddi, V. J. 2021. The People’s Speech: A Large-Scale Diverse English Speech Recognition Dataset for Commercial Usage. *CoRR*, abs/2111.09344.
- Gardner, J.; Durand, S.; Stoller, D.; and Bittner, R. M. 2023. Llark: A multimodal instruction-following language model for music. *arXiv preprint arXiv:2310.07160*.
- Gouyon, F.; Klapuri, A.; Dixon, S.; Alonso, M.; Tzanetakis, G.; Uhle, C.; and Cano, P. 2006. An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio Speech and Language Processing*, 14(5): 1832–1844. ISSN 1558-7916; br/; Contribution: organisation=sgn,FACT1=1.
- Guo, J. G. W. Z. S. W. S. X. J. 2025. ACE-Step: A Step Towards Music Generation Foundation Model. <https://github.com/ace-step/ACE-Step>. GitHub repository.
- Huang, A.; Wu, B.; Wang, B.; Yan, C.; Hu, C.; Feng, C.; Tian, F.; Shen, F.; Li, J.; Chen, M.; Liu, P.; Miao, R.; You, W.; Chen, X.; Yang, X.; Huang, Y.; Zhang, Y.; Gong, Z.; Zhang, Z.; Li, B.; Wan, C.; Hu, H.; Ming, R.; Yuan, S.; Zhang, X.; Zhou, Y.; Li, B.; Ma, B.; An, K.; Ji, W.; Li, W.; Wen, X.; Ma, Y.; Liang, Y.; Mou, Y.; Ahmadi, B.; Wang, B.; Li, B.; Miao, C.; Xu, C.; Feng, C.; Wang, C.; Shi, D.; Sun, D.; Hu, D.; Sai, D.; Liu, E.; Huang, G.; Yan, G.; Wang, H.; Jia, H.; Zhang, H.; Gong, J.; Wu, J.; Liu, J.; Sun, J.; Zhen, J.; Feng, J.; Wu, J.; Wu, J.; Yang, J.; Wang, J.; Zhang, J.; Lin, J.; Li, K.; Xia, L.; Zhou, L.; Gu, L.; Chen, M.; Wu, M.; Li, M.; Li, M.; Liang, M.; Wang, N.; Hao, N.; Wu, Q.; Tan, Q.; Pang, S.; Yang, S.; Gao, S.; Liu, S.; Liu, S.; Cao, T.; Wang, T.; Deng, W.; He, W.; Sun, W.; Han, X.; Deng, X.; Liu, X.; Zhao, X.; Wei, Y.; Yu, Y.; Cao, Y.; Li, Y.; Ma, Y.; Xu, Y.; Shi, Y.; Wang, Y.; Zhong, Y.; Luo, Y.; Lu, Y.; Yin, Y.; Yan, Y.; Yang, Y.; Xie, Z.; Ge, Z.; Sun, Z.; Huang, Z.; Chang, Z.; Yang, Z.; Zhang, Z.; Jiao, B.; Jiang, D.; Shum, H.-Y.; Chen, J.; Li, J.; Zhou, S.; Zhang, X.; Zhang, X.; and Zhu, Y. 2025. Step-Audio: Unified Understanding and Generation in Intelligent Speech Interaction. *arXiv:2502.11946*.
- KimiTeam; Ding, D.; Ju, Z.; Leng, Y.; Liu, S.; Liu, T.; Shang, Z.; Shen, K.; Song, W.; Tan, X.; Tang, H.; Wang, Z.; Wei, C.; Xin, Y.; Xu, X.; Yu, J.; Zhang, Y.; Zhou, X.; Charles, Y.; Chen, J.; Chen, Y.; Du, Y.; He, W.; Hu, Z.; Lai, G.; Li, Q.; Liu, Y.; Sun, W.; Wang, J.; Wang, Y.; Wu, Y.; Wu, Y.; Yang, D.; Yang, H.; Yang, Y.; Yang, Z.; Yin, A.; Yuan, R.; Zhang, Y.; and Zhou, Z. 2025. Kimi-Audio Technical Report. *arXiv:2504.18425*.
- Knees, P.; Faraldo, Á.; Herrera, P.; Vogl, R.; Böck, S.; Hörschläger, F.; and Goff, M. L. 2015. Two Data Sets for Tempo Estimation and Key Detection in Electronic Dance Music Annotated from User Corrections. In *International Society for Music Information Retrieval Conference*.
- Law, E.; West, K.; Mandel, M. I.; Bay, M.; and Downie, J. S. 2009. Evaluation of algorithms using games: The case of music tagging. In *ISMIR*, 387–392.

- Li, T.; Liu, J.; Zhang, T.; Fang, Y.; Pan, D.; Wang, M.; Liang, Z.; Li, Z.; Lin, M.; Dong, G.; et al. 2025. Baichuan-audio: A unified framework for end-to-end speech interaction. *arXiv preprint arXiv:2502.17239*.
- Li, Y.; Yuan, R.; Zhang, G.; Ma, Y.; Chen, X.; Yin, H.; Lin, C.; Ragni, A.; Benetos, E.; Gyenge, N.; Dannenberg, R.; Liu, R.; Chen, W.; Xia, G.; Shi, Y.; Huang, W.; Guo, Y.; and Fu, J. 2023. MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training. *arXiv:2306.00107*.
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26296–26306.
- Liu, S.; Hussain, A. S.; Sun, C.; and Shan, Y. 2023a. M2UGen: Multi-modal Music Understanding and Generation with the Power of Large Language Models. *arXiv preprint arXiv:2311.11255*.
- Liu, S.; Hussain, A. S.; Sun, C.; and Shan, Y. 2023b. Music Understanding LLaMA: Advancing Text-to-Music Generation with Question Answering and Captioning. *arXiv preprint arXiv:2308.11276*.
- McKee, D.; Salamon, J.; Sivic, J.; and Russell, B. 2023. Language-Guided Music Recommendation for Video via Prompt Analogies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Mitton, M. 2025. bread-midi-dataset (Revision 95c2155).
- Panda, R.; Malheiro, R.; and Paiva, R. P. 2018. Musical texture and expressivity features for music emotion recognition. In *19th International Society for Music Information Retrieval Conference (ISMIR 2018)*, 383–391.
- Pegoraro Santana, I. A.; Pinhelli, F.; Donini, J.; Catharin, L.; Mangolin, R. B.; da Costa, Y. M. e. G.; Delisandra Feltrim, V.; and Domingues, M. A. 2020. Music4All: A New Music Database and Its Applications. In *2020 International Conference on Systems, Signals and Image Processing (IWS-SIP)*, 399–404.
- Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.
- Sakshi, S.; Tyagi, U.; Kumar, S.; Seth, A.; Selvakumar, R.; Nieto, O.; Duraiswami, R.; Ghosh, S.; and Manocha, D. 2024. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. *arXiv:2410.19168*.
- Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Smith, J. B. L.; Burgoyne, J. A.; Fujinaga, I.; De Roure, D.; and Downie, J. S. 2011. Design and creation of a large-scale database of structural annotations. In *ISMIR*, volume 11, 555–560. Miami, FL.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Tzanetakis, G.; and Cook, P. 2002. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5): 293–302.
- von Werra, L.; Belkada, Y.; Tunstall, L.; Beeching, E.; Thrush, T.; Lambert, N.; Huang, S.; Rasul, K.; and Galloüédec, Q. 2020. TRL: Transformer Reinforcement Learning. <https://github.com/huggingface/trl>.
- Weck, B.; Manco, I.; Benetos, E.; Quinton, E.; Fazekas, G.; and Bogdanov, D. 2024. MuChoMusic: Evaluating Music Understanding in Multimodal Audio-Language Models. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*.
- Xi, Q.; Bittner, R. M.; Pauwels, J.; Ye, X.; and Bello, J. P. 2018. GuitarSet: A Dataset for Guitar Transcription. In *International Society for Music Information Retrieval Conference*.
- Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.
- Yang, 2020. zhvoice.
- Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Zeng, A.; Du, Z.; Liu, M.; Wang, K.; Jiang, S.; Zhao, L.; Dong, Y.; and Tang, J. 2024. GLM-4-Voice: Towards Intelligent and Human-Like End-to-End Spoken Chatbot. *arXiv:2412.02612*.
- Zhou, B.; Hu, Y.; Weng, X.; Jia, J.; Luo, J.; Liu, X.; Wu, J.; and Huang, L. 2024. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.

## A Model Output Examples

Here we show some outputs of fintunes of our base model, refer to Figs. 5 to 8.

## B Training Details

Our main training code is adapted from TinyLLaVA Factory(Zhou et al. 2024) to support audio input. As for reinforcement learning, we modify the HuggingFace TRL library(von Werra et al. 2020).

For distributed training, we use the DeepSpeed ZeRO Stage 3 optimization strategy. In each sub-stage, the model is initialized from previous one and trained with a learning rate of 2e-5 and cosine scheduler. To manage memory and enhance performance, the training utilizes bfloat16 mixed-precision, Flash Attention 2, and gradient checkpointing.

The training data is constructed in a straight-forward way, unlike many prior works which mostly utilize instruction format. We tend to put as many text labels as possible in a single sample for better efficiency. Some open music datasets used in training are listed in Table5, examples shown in Figs. 9 and 10. We also provide the specific data recipes for each training stage, see Tables 6 to 11.

## C Evaluation Details

### Data Construction

We collect 10+ open datasets as well as some in-house data. Since these datasets have audio and text label pairs, multiple-choice questions can be easily generated for evaluation. The correct choice can be inferred from the text label pair, while the incorrect choices can be randomly selected from the text label in other pairs. For example, the fma-small dataset has 8 genre labels, we can randomly select 3 incorrect choices from the other 7 genres. The exception is the MusicCaps, for which we directly prompt gpt-4o to generate the multiple-choice questions based on the corresponding text descriptions. For MMAU(Sakshi et al. 2024), the music part of test-mini v05.15.25 version is used.

### Experiments

The input text prompt is 'Choose the correct option for the question based on the audio.', followed by the question and choices. For open-weighted model and our model, we all use BF16 precision for inference and temperature zero during sampling. The gemini model is accessed via Vertex AI Platform with default request parameters.

## D Downstream Applications

### Instruction-tuning

The instruction-tuning data for full song understanding are scarce, so we prompt various LLMs to generate some QA pairs based on some text information related to a song. This dataset has around 30k samples with all song-level audios, on which the base model is further trained. Afterwards, we also try methods of reinforcement learning like GRPO(Shao et al. 2024)) on it.

### Prompt Generator

We make a prompt generator for a music generation model ACE-Step(Guo 2025) which takes in a song and output some prompts and lyrics for generation, by training on around 20k text song pairs. On a held-out test set of 100 songs, our finetuned model achieves a similarity score of 0.98 (using MERT(Li et al. 2023) model’s embedding) compared with original generated song.

### Music Score Transcription

The model is trained on more instrumental piece ABC code pairs to have better performance on music score transcription task. ABC code can be used to synthesized back the music. On a held-out test set of size 256, our finetuned model achieves a similarity score of 0.92 (using MERT’s embedding) compared with original piece.

### Lyrics Transcription with Timestamps

We also try to train a model that could output both lyrics and timestamps given a song.

Dataset	Count	Audio Length	Labels
Giantsteps-key(Knees et al. 2015)	604	120s	Key
Free Music Archive(Defferrard et al. 2017)	100k	30s	Similar to website API
MagnaTagATune(Law et al. 2009)	25k	29s	188 simple binary tags
MTG-Jamendo(Bogdanov et al. 2019)	55k	song	195 tags (genre, instrument, mood/theme)
MusicCaps(Agostinelli et al. 2023)	5.5k	10s	Expert-written text descriptions and some tags
YouTube8M-MusicTextClips(McKee et al. 2023)	4k	10s	Text descriptions
bread-midi-dataset(Mitton 2025)	851k	various	midi
Music Instrument Sounds(Abdulvahap 2024)	42.3k	3s	Instrument
NSynth(Engel et al. 2017)	306k	4s	Note, instrument
GuitarSet(Xi et al. 2018)	360	30s	Pitch, beat, tempo, chord
Music4All(Pegoraro Santana et al. 2020)	109k	30s	tags, lyrics
SALAMI(Smith et al. 2011)	1447	song	structure

Table 5: open music datasets used in training

dataset	nums
CommonVoice(Ardila et al. 2020)	100k
bread-midi-dataset(Mitton 2025)	38k

Table 6: warmup

dataset	nums
CommonVoice(Ardila et al. 2020)	1014k
bread-midi-dataset(Mitton 2025)	115k

Table 7: align-1

dataset	nums
lyrics seg(internal trainset)	4k
MagnaTagATune(Law et al. 2009)	71k
NSynth(Engel et al. 2017)	88k
MusicCaps(Agostinelli et al. 2023)	10k
Music Instrument Sounds(Abdulvahap 2024)	52k
GuitarSet(Xi et al. 2018)	4k
Giantsteps-key(Knees et al. 2015)	3k
FMA-medium(Defferrard et al. 2017)	24k
FMA-small(Defferrard et al. 2017)	7k
Music4All(Pegoraro Santana et al. 2020)	109k

Table 10: finetuning-short

dataset	nums
CommonVoice(Ardila et al. 2020)	100k
lyrics seg(internal trainset)	400k
NSynth(Engel et al. 2017)	200k
Music Instrument Sounds(Abdulvahap 2024)	42k
People's Speech (Galvez et al. 2021)	310k
zhvoice (Yang 2020)	300k

Table 8: align-2

dataset	nums
lyrics full(internal trainset)	6k
finetuning-short	10k
MTG-Jamendo(Bogdanov et al. 2019)	59k
music general QA (internal trainset)	10k
lyrical reasoning (internal trainset)	9k
SALAMI(Smith et al. 2011)	5k

Table 11: finetuning-long

dataset	nums
bread-midi-dataset(Mitton 2025)	10k
lyrics full(internal trainset)	100k

Table 9: context-extending

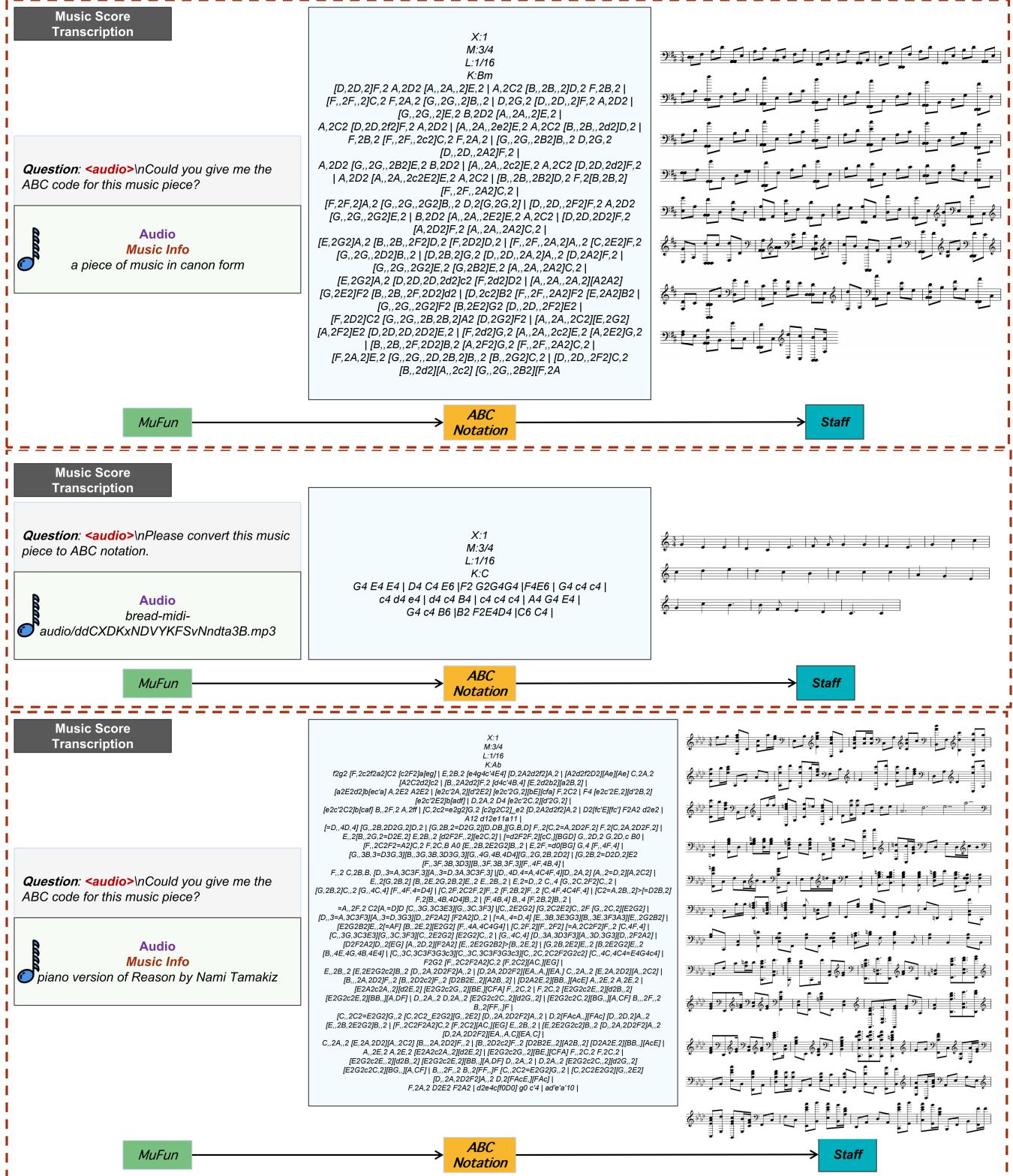


Figure 5: Music Score Transcription Demo

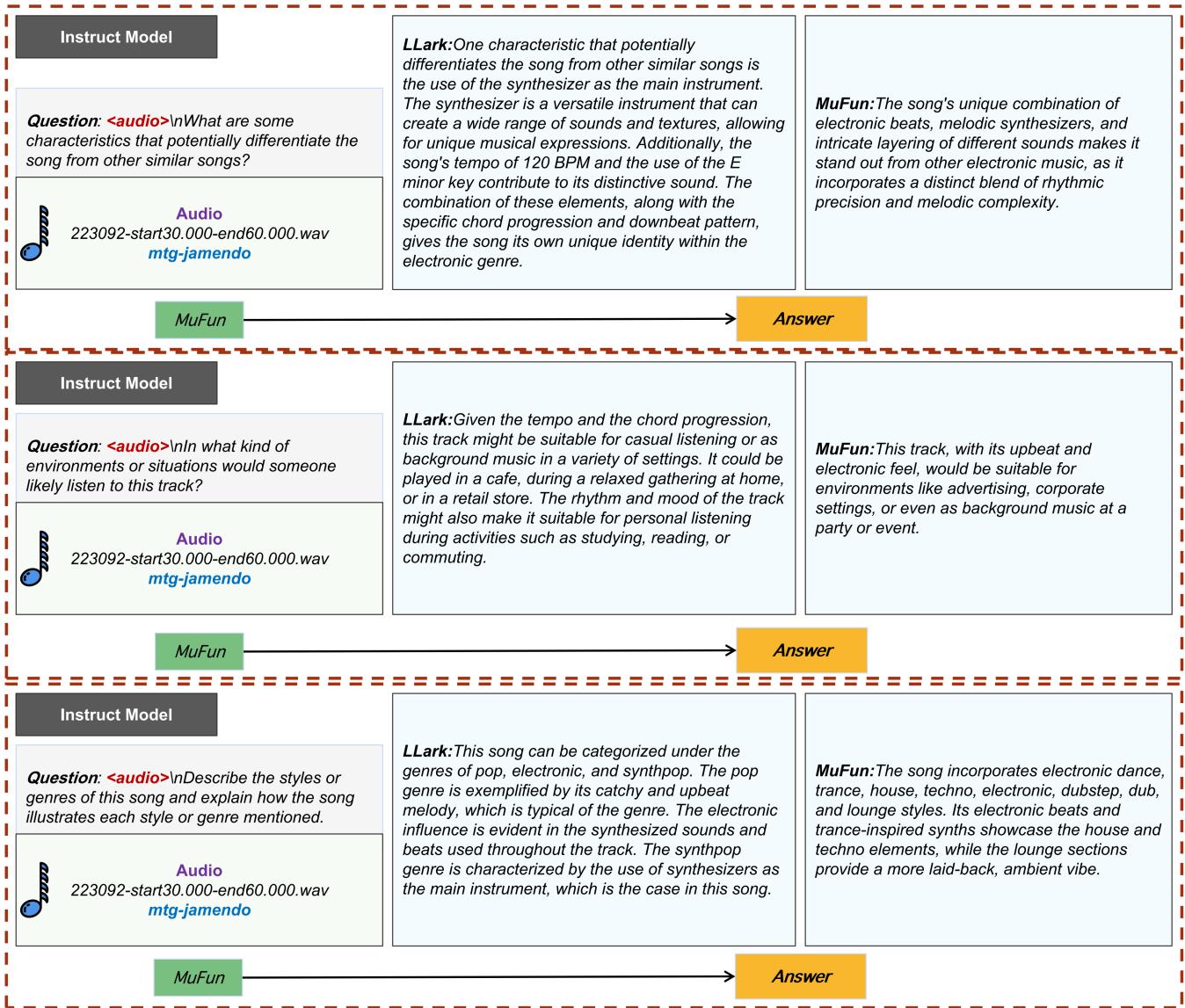


Figure 6: Instruct Model Demo

**Lyrics Transcription**

**Question:** <audio>\n这首歌的歌词 (带时间戳)

**Audio**  
**Music Info**  
a piece of Chinese pop music

[00:06.61]该怎么去形容你最贴切  
 [00:12.01]我拿什么当你总比较才算特别  
 [00:17.42]对你的感觉强烈 却又不太了解  
 [00:23.90]只凭直觉  
 [00:26.12]你像窝在被子里的舒服  
 [00:31.82]却又像风捉摸不住  
 [00:36.73]享受爱是散发的香水味  
 [00:42.56]是爱不释手的 红色高跟鞋  
 [00:54.78]oh ~  
 [01:04.67]你像窝在被子里的舒服  
 [01:10.03]却又像风捉摸不住  
 [01:15.08]享受爱是散发的香水味  
 [01:21.01]是爱不释手的  
 [01:25.62]我爱你有种左灯右行的冲突  
 [01:31.71]疯狂却太礼貌没有退路  
 [01:36.81]你能否让我停止这种追逐  
 [01:43.25]只贪恋上最后唯一的理由  
 [01:47.66]红色高跟鞋

MuFun → lyrics transcription

```

graph LR
    MuFun[MuFun] --> transcription[lyrics transcription]
  
```

Figure 7: Lyrics Transcription Demo

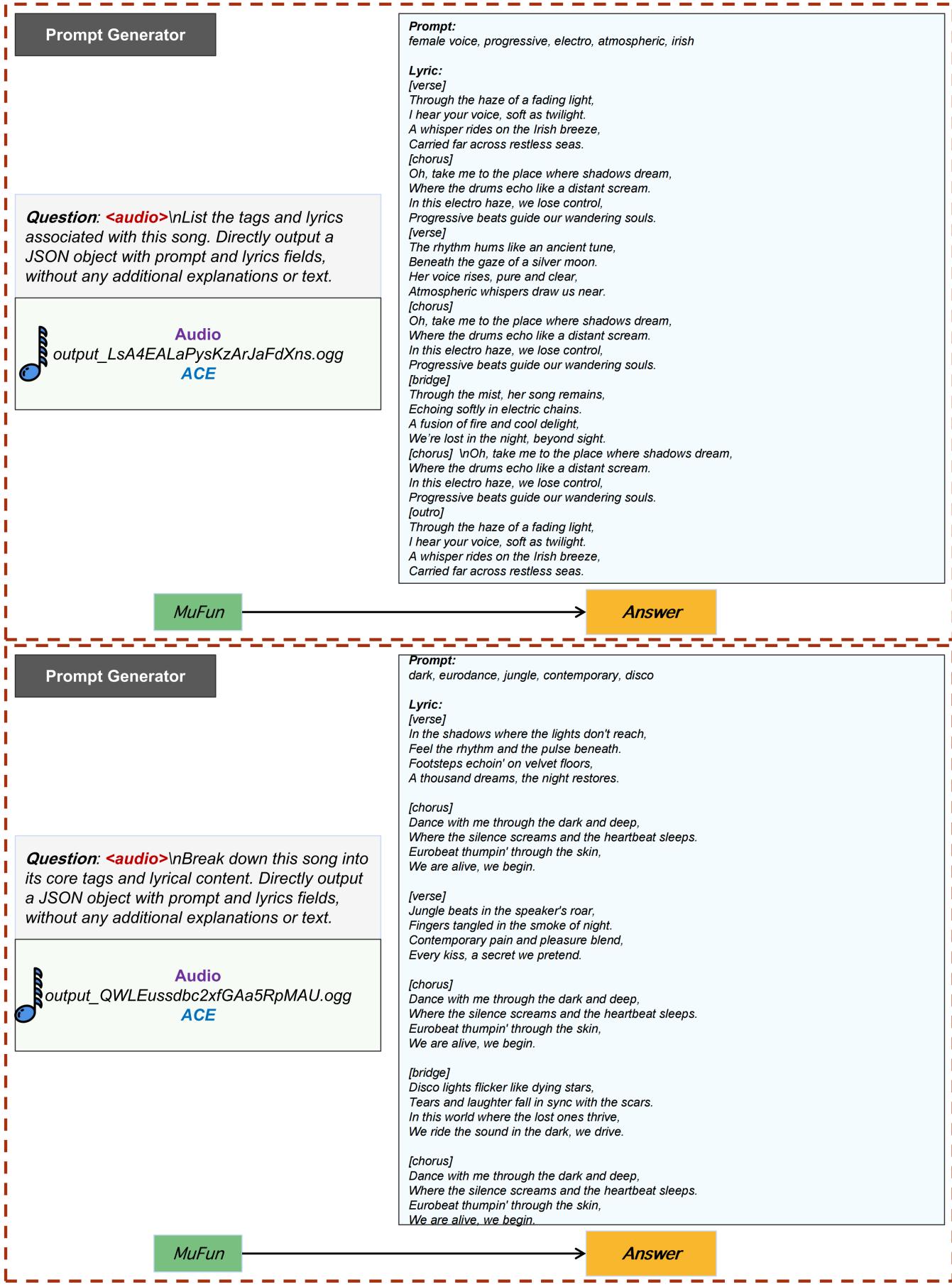


Figure 8: Prompt Generator Demo

<p><b>Question:</b> &lt;audio&gt;\nCan you give me the pitch designation for the note just played?</p> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 10px;">  <b>reed_acoustic_008-068-050.wav</b>  <b>NSynth</b> </div>	<p>Ab4</p>
<p><b>Question:</b> &lt;audio&gt;\nCould you analyze this music segment and list the instruments?</p> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 10px;">  <b>Ukulele/195.wav</b>  <b>Music Instrument Sounds</b> </div>	<p>ukulele</p>
<p><b>Question:</b> &lt;audio&gt;\nTag this short piece of music.</p> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 10px;">  <b>memories_of_tomorrow-waiting_for_dawn-09-not_like_this-30-59.mp3</b>  <b>MagnaTagATune</b> </div>	<p>female, fast, woman, dance, female vocals, techno, beat</p>
<p><b>Question:</b> &lt;audio&gt;\nCan you transcribe the lyrics of the song I'm about to play?</p> <div style="border: 1px solid #ccc; padding: 5px; margin-top: 10px;">  <b>song_en/3986241.0.mp3(Internal Dataset)</b>  <b>lyrics full</b> </div>	<p>When I was young I'd listen to the radio Waiting for my favorite songs When they played I'd sing along It made me smile Those were such happy times and not so long ago how I wondered where they'd gone But they're back again just like a long lost friend all the songs I loved so well Every shalalala every wo'wo Still shines Every shing-a-ling-a-ling That they're starting to sing So fine When they get to the part Where he's breaking her heart It can really make me cry Just like before It's yesterday once more (shoobie do lang lang) Looking back on how it was in years gone by And the good times that I had Makes today seem rather sad So much has changed It was songs of love that I would sing to then and I'd memorize each word Those old melodies still sound so good to me As they melt the years away Every shalalala every wo'wo still shines Every shing-a-ling-a-ling That they're starting to sing So fine All my best memories come back clearly to me Some can even make me cry Just like before It's yesterday once more (shoobie do lang lang) Every shalalala every wo'wo still shines Every shing-a-ling-a-ling That they're starting to sing So fine Every shalalala every wo'wo still shines.</p>

Figure 9: Training Data Overview

<p><b>Question:</b> &lt;audio&gt;\nListen to this audio and give me the text version.</p> <p> common_voice_en_22322977.mp3 <b>CommonVoice</b></p>	<p>For this reason, they are sometimes called squeegee merchants.</p>
<p><b>Question:</b> &lt;audio&gt;\nI need the ABC notation for this tune. Can you assist?</p> <p> nZKStvsWn4tYjcKED2JnGX.mp3 <b>bread-midi-dataset</b></p>	<p>X:1 M:3/4 L:1/16 K:Bm D2DD D2DD D2DD   D2DD G8&lt; B8   D2DD   D2DD D2DD D2DD   G4 B8  </p>
<p><b>Question:</b> &lt;audio&gt;\nPlease help me find the lyrics of this music clip.</p> <p> 2153998118.mp3(Internal Dataset) <b>lyrics seg</b></p>	<p>Sure, the lyrics of this music clip are: Angel in the front tryna guide my steps.</p>
<p><b>Question:</b> &lt;audio&gt;\nCan you tell me the genre of this musical piece?(choose the genre from: Rock, Instrumental, Hip-Hop, Experimental, Electronic, International, Folk, Pop)</p> <p> fma_small/071/071231.mp3 <b>FMA</b></p>	<p>Hip-Hop</p>
<p><b>Question:</b> &lt;audio&gt;\nGive information about this 30s song clip.(such as the song name, artist, album name, release year, language, pitch class and modality, BPM, genres, danceability, energy, valence, popularity and other tags)\n</p> <p> audios/FbboTtnNfY9BjS2K.mp3 <b>Music4All</b></p>	<pre>song_name: Live to Rise artist: Soundgarden album_name: Echo Of Miles: Scattered Tracks Across The Path release_year: 2014 lang: en pitch_class: E modality: minor BPM: 84 genres: rock,grunge,alternative rock danceability: 0.48 energy: 0.68 valence: 0.27 popularity: 44 other_tags: rock,grunge,alternative rock</pre>
<p><b>Question:</b> &lt;audio&gt;\nGenerate some tags(genre, instrument, mood/theme) for this song.</p> <p> 75/1420475.mp3 <b>MTG-Jamendo</b></p>	<pre>genre: disco, chillout, dance, funk instrument: percussion, bass, synthesizer, keyboard, drums mood/theme: summer, happy</pre>

Figure 10: Training Data Overview