Novelty Assessment Report

Paper: Music Flamingo: Scaling Music Understanding in Audio Language Models
PDF URL: https://openreview.net/pdf?id=RS7T9S16Bl
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

We introduce Music Flamingo, a novel large audio–language model, designed to advance music (including song) understanding in foundational audio models. While audio–language research has progressed rapidly, music remains challenging due to its dynamic, layered, and information-dense nature. Progress has been further limited by the difficulty of scaling open audio understanding models, primarily because of the scarcity of high-quality music data and annotations. As a result, prior models are restricted to producing short, high-level captions, answering only surface-level questions, and showing limited generalization across diverse musical cultures. To address these challenges, we curate MF-Skills, a large-scale dataset labeled through a multi-stage pipeline that yields rich captions and question–answer pairs covering harmony, structure, timbre, lyrics, and cultural context. We fine-tune an enhanced Audio Flamingo 3 backbone on MF-Skills and further strengthen multiple skills relevant to music understanding. To improve the model's reasoning abilities, we introduce a post-training recipe: we first cold-start with MF-Think, a novel chain-of-thought dataset grounded in music theory, followed by GRPO-based reinforcement learning with custom rewards. Music Flamingo achieves state-of-the-art results across 10+ benchmarks for music understanding and reasoning, establishing itself as a generalist and musically intelligent audio–language model. Beyond strong empirical results, Music Flamingo sets a new standard for advanced music understanding by demonstrating how models can move from surface-level recognition toward layered, human-like perception of songs. We believe this work provides both a benchmark and a foundation for the community to build the next generation of models that engage with music as richly and meaningfully as humans do. Demo: https://musicflamingo.github.io

Core Task Landscape

This paper addresses: **Music Understanding in Audio Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Audio-Language Model Architectures and Training**
- **Music-Specific Understanding and Reasoning**
- **Music Generation and Editing**
- **General Audio Understanding and Multi-Domain Systems**
- **Evaluation and Benchmarking**

Complete Taxonomy Tree

- Music Understanding in Audio Language Models Survey Taxonomy
- Audio-Language Model Architectures and Training
  - Discrete Token-Based Audio Language Models (4 papers)
  - [1] Audiolm: a language modeling approach to audio generation (ZalÃ¡n Borsos, 2023) View paper
  - [4] Continuous Audio Language Models (Rouard, 2025) View paper
  - [28] Generating Stereophonic Music with Single-Stage Language Models (Xingda Li, 2024) View paper
  - [43] An Independence-promoting Loss for Music Generation with Language Models (Lemercier, 2024) View paper
  - Multi-Modal Audio Encoders and Fusion (7 papers)
  - [2] Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models (Chu Yun-fei, 2023) View paper
  - [5] Audio flamingo 3: Advancing audio intelligence with fully open large audio language models (Goel, 2025) View paper
  - [7] SALMONN: Towards Generic Hearing Abilities for Large Language Models (Tang, 2023) View paper
  - [8] Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities (Kong, 2024) View paper
  - [15] Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities (Ghosh, 2025) View paper
  - [21] Midashenglm: Efficient audio understanding with general audio captions (Dinkel, 2025) View paper
  - [42] Extending Audio Context for Long-Form Understanding in Large Audio-Language Models (Taveekitworachai, 2025) View paper
  - Joint Audio-Text Embedding Models (3 papers)
  - [32] Mulan: A joint embedding of music audio and natural language (Qing-qing Huang, 2022) View paper
  - [36] Semitone-Aware Fourier Encoding: A Music-Structured Approach to Audio-Text Alignment (C Du, 2025) View paper
  - [45] Collap: Contrastive long-form language-audio pretraining with musical temporal structure augmentation (Junda Wu, 2025) View paper
- Music-Specific Understanding and Reasoning
  - Music-Centric Audio-Language Models ★ (4 papers)
  - [0] Music Flamingo: Scaling Music Understanding in Audio Language Models (Anon et al., 2026) View paper
  - [3] Mumu-llama: Multi-modal music understanding and generation via large language models (Liu, 2024) View paper

- [12] Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities (Sreyan Ghosh, 2024) View paper
- [16] Musilingo: Bridging music and text with pre-trained language models for music captioning and query response (Zihao Deng, 2024) View paper
- Music Perception and Knowledge Evaluation (3 papers)
- [13] Evaluation of pretrained language models on music understanding (Bittner, 2024) View paper
- [24] Evaluating multimodal large language models on core music perception tasks (Brandon James Carone, 2025) View paper
- [27] MuChoMusic: Evaluating Music Understanding in Multimodal Audio-Language Models (Benno Weck, 2024) View paper
- Lyrics and Multi-Modal Music Analysis (4 papers)
- [35] Converting Vocal Performances into Sheet Music Leveraging Large Language Models (Jinjing Jiang, 2024) View paper
- [40] Interpreting song lyrics with an audio-informed pre-trained language model (Yixiao Zhang, 2022) View paper
- [46] Multimodal music mood classification using audio and lyrics (Cyril Laurier, 2008) View paper
- [47] Joint sentiment analysis of lyrics and audio in music (Kruspe, 2024) View paper

- Music Generation and Editing
  - Text-to-Music Generation (4 papers)
  - [10] MusicLM: Generating Music From Text (Weck, 2023) View paper
  - [14] Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies (Ke Chen, 2024) View paper
  - [34] Diffusion based Text-to-Music Generation with Global and Local Text based Conditioning (Zhang, 2025) View paper
  - [37] Tango 2: Aligning Diffusion-based Text-to-Audio Generations through Direct Preference Optimization (Navonil Majumder, 2024) View paper
  - Music Editing and Instruction Following (2 papers)
  - [31] Instruct-MusicGen: Unlocking Text-to-Music Editing for Music Language Models via Instruction Tuning (Zhang YiXiao, 2024) View paper
  - [50] Arrange, Inpaint, and Refine: Steerable Long-term Music Audio Generation and Editing via Content-based Controls (Lin Li-wei, 2024) View paper
  - Multi-Modal Music Generation Systems (3 papers)
  - [17] MUGen: Multi-modal Music Understanding and Generation with the Power of Large Language Models (S Liu, 2023) View paper
  - [18] Uniaudio: An audio foundation model toward universal audio generation (Yang, 2023) View paper
  - [49] Uniaudio: Towards universal audio generation with large language models (D Yang, 2024) View paper
- General Audio Understanding and Multi-Domain Systems
  - Multi-Domain Audio-Language Models (6 papers)
  - [6] Audiogpt: Understanding and generating speech, music, sound, and talking head (Rongjie Huang, 2024) View paper
  - [22] Audio-llm: Activating the capabilities of large language models to comprehend audio data (Dong-Ting Li, 2024) View paper
  - [33] Enhancing audio comprehension in large language models: Integrating audio knowledge (Daniel Ogof, 2024) View paper
  - [41] Make Some Noise: Towards LLM audio reasoning and generation using sound tokens (Mehta, 2025) View paper
  - [44] Musicagent: An ai agent for music understanding and generation with large language models (Dingyao Yu, 2023) View paper
  - [48] Wavjourney: Compositional audio creation with large language models (Xubo Liu, 2025) View paper
  - Video-Audio-Language Integration (3 papers)
  - [20] video-SALMONN: Speech-Enhanced Audio-Visual Large Language Models (Sun Guang-zhi, 2024) View paper
  - [29] On the audio hallucinations in large audio-video language models (Nishimura, 2024) View paper
  - [38] VoxDialogue: Can Spoken Dialogue Systems Understand Information Beyond Words? (X Cheng, 2025) View paper
  - Domain-Specific Audio Applications (2 papers)
  - [19] Music Genre Classification using Large Language Models (Mohamed El Amine Meguenani, 2024) View paper
  - [25] Naturelm-audio: an audio-language foundation model for bioacoustics (Robinson, 2024) View paper
- Evaluation and Benchmarking
  - Comprehensive Audio Understanding Benchmarks (4 papers)
  - [9] AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension (Qian Yang, 2024) View paper
  - [11] Mmau: A massive multi-task audio understanding and reasoning benchmark (Sakshi, 2024) View paper
  - [30] MMAR: A Challenging Benchmark for Deep Reasoning in Speech, Audio, Music, and Their Mix (Ma, 2025) View paper
  - [39] Audio-language models for audio-centric tasks: A survey (Su Yi, 2025) View paper
  - Music-Specific Evaluation and Datasets (2 papers)
  - [23] PAGURI: a user experience study of creative interaction with text-to-music models (Francesca Ronchini, 2025) View paper
  - [26] Audio dialogues: Dialogues dataset for audio and music understanding (arushi goel, 2024) View paper

Narrative

Core task: music understanding in audio language models. The field has evolved around several complementary branches. Audio-Language Model Architectures and Training explores foundational designs that bridge audio encoders with large language models, exemplified by systems like Qwen-Audio[2] and SALMONN[7]. Music-Specific Understanding and Reasoning focuses on models tailored to musical content, addressing tasks such as genre classification, mood analysis, and music-centric question answering. Music Generation and Editing encompasses text-to-music synthesis (e.g., MusicLM[10], MusicLDM[14]) and instruction-based editing workflows. General Audio Understanding and Multi-Domain Systems develop versatile architectures that handle speech, environmental sounds, and music within unified frameworks, while Evaluation and Benchmarking provides datasets and metrics (e.g., AIR-Bench[9], MMAU[11]) to assess model capabilities across diverse audio tasks.

Recent work reveals a tension between general-purpose audio-language models and music-centric specialization. General systems like Audio Flamingo[8] and its successors (Audio Flamingo 2[15], Audio Flamingo 3[5]) aim for broad audio understanding, whereas music-focused models such as Mumu-llama[3], GAMA[12], and Musilingo[16] prioritize deep musical reasoning and domain-specific knowledge. Music Flamingo[0] sits within this music-centric cluster, emphasizing structured music understanding alongside these specialized approaches. Compared to Mumu-llama[3], which targets comprehensive music question answering, and Musilingo[16], which integrates symbolic music representations, Music Flamingo[0] explores how to adapt flamingo-style architectures specifically for musical content. This specialization trend highlights an open question: whether future progress will favor unified multi-domain models or continue to benefit from domain-tailored designs that capture music's unique structural and perceptual properties.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Mumu-llama: Multi-modal music understanding and generation via large language models

Authors: Liu, Shansong, Shansong Liu, Wu Qilong, Atin Sakkeer Hussain, et al. (12 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Research on large language models has advanced significantly across text, speech, images, and videos. However, multi-modal music understanding and generation remain underexplored due to the lack of well-annotated datasets. To address this, we introduce a dataset with 167.69 hours of multi-modal data, including text, images, videos, and music annotations. Based on this dataset, we propose MuMu-LLaMA, a model that leverages pre-trained encoders for music, images, and videos. For music generation, ...

Relationship Analysis

Both papers belong to the Music-Centric Audio-Language Models category, focusing on developing specialized models trained on music data with music-aware architectures for deep musical understanding. They overlap in addressing multi-modal music understanding, music captioning, and question-answering tasks using large language models with specialized music encoders (MERT). However, Music Flamingo emphasizes scaling music understanding through extensive data curation (MF-Skills with 4M+ samples), chain-of-thought reasoning (MF-Think), and GRPO-based reinforcement learning for theory-grounded analysis, while MuMu-LLaMA focuses on integrating both music understanding and generation capabilities across multiple modalities (text, image, video-to-music) using a unified framework with music editing functionalities.

---

2. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities

Authors: Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, et al. (9 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â music understanding extensively. We do not do this as we do not train GAMA on diverse and large-scale music â for comprehensive music understanding if trained on large-scale music â

Relationship Analysis

Both papers belong to the Music-Centric Audio-Language Models category, focusing on building specialized models for music understanding through audio-language integration. They overlap in addressing music captioning, question-answering, and reasoning tasks by fine-tuning large audio-language models on music-specific datasets with enhanced architectures. However, Music Flamingo emphasizes scaling through a multi-stage data pipeline (MF-Skills) with theory-grounded chain-of-thought reasoning (MF-Think) and GRPO-based reinforcement learning, while GAMA focuses on integrating diverse audio representations (Audio Q-Former, multi-layer aggregator) and instruction-tuning with CompA-R for complex reasoning across general audio, not exclusively music.

---

3. Musilingo: Bridging music and text with pre-trained language models for music captioning and query response

Authors: Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, et al. (8 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â of the recent triumphs of large language models (LLMs), entails integrating â musical tasks. Considering these insights, we introduce a novel music language model designed for music â

Relationship Analysis

Both papers belong to the Music-Centric Audio-Language Models category, focusing on models trained specifically on music data with music-aware architectures for deep musical understanding. They overlap in addressing music captioning and question-answering tasks using large audio-language models with frozen encoders (MERT) and language models. However, Music Flamingo emphasizes scaling to full-length multicultural songs with layered, theory-aware captions and introduces chain-of-thought reasoning via MF-Think and GRPO-based reinforcement learning, while MusiLingo focuses on a simpler linear projection adapter design and instruction tuning with the MusicInstruct dataset derived from existing MusicCaps captions.

Contributions Analysis

Overall novelty summary. Music Flamingo proposes a large audio-language model specialized for music understanding, addressing harmony, structure, timbre, lyrics, and cultural context through curated datasets and a post-training recipe involving chain-of-thought reasoning. The paper resides in the 'Music-Centric Audio-Language Models' leaf, which contains four papers total, including Mumu-llama, GAMA, and Musilingo. This leaf represents a focused research direction within the broader 'Music-Specific Understanding and Reasoning' branch, indicating a moderately populated area where specialized music models are actively being developed alongside general audio-language systems.

The taxonomy reveals neighboring work in 'Multi-Modal Audio Encoders and Fusion' (seven papers) and 'Multi-Domain Audio-Language Models' (six papers), which explore general audio understanding without music-specific specialization. The 'Music Perception and Knowledge Evaluation' leaf (three papers) addresses music theory tasks like chord recognition, while 'Lyrics and Multi-Modal Music Analysis' (four papers) integrates textual and audio modalities. Music Flamingo bridges these areas by combining multi-modal fusion techniques with music-centric training, distinguishing itself from general audio models like Audio Flamingo and from purely symbolic or theory-focused approaches.

Among 29 candidates examined, the model architecture contribution shows one refutable candidate out of ten examined, suggesting some overlap with prior audio-language model designs. The dataset contributions (MF-Skills and MF-Think) encountered no refutable candidates across ten examined papers, indicating potential novelty in curating music-specific annotations and chain-of-thought data. The post-training recipe contribution identified one refutable candidate among nine examined, reflecting partial overlap with existing reinforcement learning or reasoning enhancement methods. These statistics reflect a limited semantic search scope, not an exhaustive literature review.

Given the search scope of 29 candidates, the work appears to offer incremental architectural refinement alongside potentially novel dataset curation for music understanding. The taxonomy context shows Music Flamingo occupies a moderately active research niche, with sibling papers pursuing similar music-centric goals. The analysis does not cover exhaustive prior work in music information retrieval or symbolic music processing, which may contain additional relevant comparisons beyond the top-K semantic matches examined here.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Music Flamingo model for advanced music understanding

Description: The authors present Music Flamingo, a large audio-language model that moves beyond surface-level music recognition toward layered, human-like perception of songs. The model is designed to handle the dynamic, layered, and information-dense nature of music through enhanced training strategies and reasoning capabilities.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Sparks of large audio models: A survey and outlook

URL: View paper

Brief Assessment

Large Audio Survey[74] provides a broad overview of large audio models across speech, music, and other audio domains, but does not present a specific model architecture for music understanding comparable to Music Flamingo. The survey discusses existing models like MusicLM and AudioLDM but does not claim to introduce a novel music understanding model with the layered perception capabilities described in the original paper.

2. CLAP Learning Audio Concepts from Natural Language Supervision

URL: View paper

Brief Assessment

CLAP[73] focuses on learning audio concepts from natural language supervision using contrastive learning across general audio domains (sound events, speech, music). It does not present a large audio-language model specifically designed for advanced, layered music understanding with reasoning capabilities as described in the original contribution.

3. Continuous Audio Language Models

URL: View paper

Brief Assessment

Continuous Audio[4] focuses on continuous audio generation through consistency modeling and VAE frames, not on music understanding or perception tasks that Music Flamingo addresses.

4. A survey of foundation models for music understanding

URL: View paper

Brief Assessment

Foundation Models Survey[72] is a survey paper that reviews existing large-scale music foundation models rather than proposing a novel model. It discusses models like Qwen-Audio, LTU, SALMONN, and others, but does not claim to introduce a new audio-language model for music understanding that would refute Music Flamingo's novelty.

5. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities

URL: View paper

Brief Assessment

GAMA[12] focuses on general audio understanding (non-speech sounds and non-verbal speech) with complex reasoning capabilities, not specifically on music understanding. While both are large audio-language models, GAMA[12] addresses broader audio domains rather than the specialized music perception and layered understanding that Music Flamingo targets.

6. SALMONN: Towards Generic Hearing Abilities for Large Language Models

URL: View paper

Brief Assessment

SALMONN[7] is a general audio-language model covering speech, audio events, and music, whereas the original paper focuses specifically on deep, layered music understanding with theory-aware reasoning. SALMONN[7] does not demonstrate the specialized music captioning, harmonic analysis, or musician-level reasoning that the original contribution claims as novel.

7. Pam: Prompting audio-language models for audio quality assessment

URL: View paper

Brief Assessment

PAM[71] focuses on audio quality assessment across various tasks (text-to-audio, text-to-music, text-to-speech, noise suppression) using audio-language models to compute quality metrics, not on building large audio-language models for music understanding and perception.

8. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities

URL: View paper

Brief Assessment

Audio Flamingo 2[15] focuses on general audio-language modeling across non-speech sounds and music with long-audio understanding capabilities, whereas the original paper presents Music Flamingo as a specialized model specifically designed for layered, human-like music perception with theory-grounded reasoning. The candidate does not demonstrate prior work that challenges the novelty of Music Flamingo's music-specific architecture and training approach.

9. Can synthetic audio from generative foundation models assist audio recognition and speech modeling?

URL: View paper

Brief Assessment

Synthetic Audio[76] focuses on using synthetic audio from generative models (AudioGen, AudioLDM2, MusicGen) for audio recognition and speech modeling tasks, not on building large audio-language models for music understanding like Music Flamingo.

10. Llark: A multimodal foundation model for music

URL: View paper

Prior Art Analysis

LLARK[75] demonstrates that a multimodal foundation model for music understanding was developed prior to Music Flamingo. Both models address the same core challenge: moving beyond surface-level music recognition toward deeper, layered understanding of songs. LLARK[75] presents a multimodal architecture combining a pretrained generative audio encoder with a pretrained language model for music understanding tasks including captioning, reasoning, and classification. The candidate paper explicitly states it addresses music understanding through instruction-tuning and metadata augmentation, producing detailed captions that capture harmony, structure, timbre, and cultural context—the same multi-layered approach claimed as novel by Music Flamingo.

Evidence

Evidence 1 - **Rationale**: Both papers claim to introduce novel models specifically designed to address the unique challenges of music understanding, acknowledging music's complex, layered nature as a core motivation. - **Original**: we introducemusic flamingo, a novel large audio-language model, designed to advance music (including song) understanding in foundational audio models. while audio-language research has progressed rapidly, music remains challenging due to its dynamic, layered, and information-dense nature. - **Candidate**: Music has a unique and complex structure which is challenging for both expert humans and existing ai systems to

understand, and presents unique challenges relative to other forms of audio. we present ll ark , an instruction-tuned multimodal model for music understanding.

Evidence 2 - **Rationale**: Both papers claim their respective models achieve unprecedented levels of music understanding across multiple tasks, moving beyond surface-level recognition to deeper comprehension. - **Original**: music flamingo achieves state-ofthe-art results across 10+ benchmarks for music understanding and reasoning, establishing itself as a generalist and musically intelligent audio-language model. beyond strong empirical results, music flamingo sets a new standard for advanced music understanding by dem... - **Candidate**: our evaluations demonstrate ll ark 's music understanding, captioning, and reasoning capabilities at a level of quality unseen so far from a single model. our study points to several directions for future work.

Contribution 2: MF-Skills and MF-Think datasets for music understanding

Description: The authors introduce two large-scale datasets: MF-Skills contains over 4 million samples with detailed multi-aspect captions and question-answer pairs covering full-length, multi-cultural songs; MF-Think provides 300,000 chain-of-thought examples grounded in music theory to enable deliberate reasoning about music.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Improving BERT for symbolic music understanding using token denoising and pianoroll prediction

URL: View paper

Brief Assessment

BERT Symbolic[52] focuses on symbolic music representation (MIDI tokens) and pre-training objectives for note-level understanding, not on creating large-scale datasets with multi-aspect captions, question-answer pairs, or chain-of-thought reasoning for audio-language models covering full-length songs with lyrics and cultural context.

---

2. MusiXQA: Advancing Visual Music Understanding in Multimodal Large Language Models

URL: View paper

Brief Assessment

MusiXQA[51] focuses on visual music sheet understanding through synthetic sheet images and visual QA tasks, whereas the original paper addresses audio-based music understanding with detailed multi-aspect captions covering harmony, structure, timbre, and lyrics from actual audio recordings. These are fundamentally different modalities and tasks.

---

3. Multimodal music datasets? Challenges and future goals in music processing

URL: View paper

Brief Assessment

Multimodal Datasets[56] focuses on defining multimodal music datasets and categorizing existing datasets for various music processing tasks, but does not present new large-scale datasets with detailed multi-aspect captions and chain-of-thought examples like MF-Skills and MF-Think.

---

4. Towards Unified Music Emotion Recognition across Dimensional and Categorical Models

URL: View paper

Brief Assessment

Unified Emotion[59] focuses on music emotion recognition datasets with categorical and dimensional labels (e.g., MTG-Jamendo, DEAM, PMEmo), not comprehensive music understanding datasets covering harmony, structure, timbre, lyrics, and cultural context as described in the original paper's MF-Skills and MF-Think contributions.

---

5. ERLD-HC: Entropy-Regularized Latent Diffusion for Harmony-Constrained Symbolic Music Generation

URL: View paper

Brief Assessment

ERLD-HC[60] focuses on symbolic music generation (MIDI) with harmony constraints using diffusion models and CRF, not on music understanding datasets or audio-language models.

---

6. Songcreator: Lyrics-based universal song generation

URL: View paper

Brief Assessment

SongCreator[57] focuses on lyrics-based song generation (vocals + accompaniment), not on music understanding datasets. The candidate does not present datasets for music understanding covering harmony, structure, timbre, and lyrics as the original paper does.

---

7. Are we there yet? a brief survey of music emotion prediction datasets, models and outstanding challenges

URL: View paper

Brief Assessment

Emotion Prediction[54] focuses on emotion recognition datasets and models for affective music analysis, not on comprehensive music understanding datasets covering harmony, structure, timbre, and lyrics with chain-of-thought reasoning as proposed in the original paper.

---

8. MGPHot: A Dataset of Musicological Annotations for Popular Music (1958â□□2022)

URL: View paper

Brief Assessment

MGPHot[55] focuses on musicological annotations (rhythm, harmony, instrumentation, etc.) for Billboard Hot 100 songs, not on multi-aspect captions with question-answer pairs or chain-of-thought reasoning for training audio-language models.

---

9. Foundation models for music: A survey

URL: View paper

Brief Assessment

Foundation Models Survey[53] is a broad survey paper covering foundation models for music. It does not present specific datasets that would refute the novelty of MF-Skills and MF-Think, which are specialized datasets with detailed multi-aspect captions and chain-of-thought examples for music understanding.

---

10. The interconnections of music structure, harmony, melody, rhythm, and predictivity

URL: View paper

Brief Assessment

Music Interconnections[58] focuses on extracting hierarchical structure from MIDI data and analyzing interactions between structure, harmony, melody, and rhythm. It does not present large-scale datasets with captions and question-answer pairs for training audio-language models.

Contribution 3: Post-training recipe with chain-of-thought and reinforcement learning

Description: The authors propose a novel post-training approach that combines supervised fine-tuning on chain-of-thought examples from MF-Think with GRPO-based reinforcement learning using custom-designed rewards. This enables the model to perform explicit step-by-step musical reasoning rather than simple pattern matching.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Robust chain of thoughts preference optimization

URL: View paper

Brief Assessment

Robust CoT[68] focuses on offline preference optimization using chain-of-thought for self-improvement in language models, not on music understanding or audio-language models. The technical approach and application domain are fundamentally different.

---

2. Incentivizing Consistent, Effective and Scalable Reasoning Capability in Audio LLMs via Reasoning Process Rewards

URL: View paper

Brief Assessment

Reasoning Process Rewards[67] focuses on audio LLMs and addresses test-time inverse scaling through process-oriented rewards, while the original paper targets music understanding with a different reward structure (structured thinking reward for captions). The domains and specific reward mechanisms differ substantially.

---

3. SoundMind: RL-Incentivized Logic Reasoning for Audio-Language Models

URL: View paper

Brief Assessment

SoundMind[62] focuses on audio logical reasoning tasks using rule-based RL rewards for audio-language models, not general music understanding with GRPO-based reinforcement learning as in the original paper.

---

4. Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning

URL: View paper

Brief Assessment

EchoInk-R1[66] applies GRPO to audio-visual question answering, while the original paper focuses on music understanding with custom rewards for music-specific reasoning (harmony, lyrics, structure). The candidate does not demonstrate prior work in music-domain chain-of-thought reasoning with reinforcement learning.

---

5. Audio-cot: Exploring chain-of-thought reasoning in large audio language model

URL: View paper

Brief Assessment

Audio-CoT[64] focuses on applying chain-of-thought reasoning methods to large audio-language models across sound, speech, and music domains, but does not describe a post-training recipe combining supervised fine-tuning on chain-of-thought examples with GRPO-based reinforcement learning using custom rewards specifically for music understanding tasks.

---

6. Thinking in cocktail party: Chain-of-Thought and reinforcement learning for target speaker automatic speech recognition

URL: View paper

Brief Assessment

Cocktail Party[70] applies chain-of-thought and reinforcement learning to target speaker automatic speech recognition in cocktail party scenarios, not to music understanding or reasoning tasks. The technical domain and application context differ fundamentally from the original paper's music-focused contributions.

---

7. Motion-R1: Chain-of-Thought Reasoning and Reinforcement Learning for Human Motion Generation

URL: View paper

Brief Assessment

Motion-R1[61] applies chain-of-thought reasoning and reinforcement learning to human motion generation, not music understanding. The technical domains and modalities (motion sequences vs. audio-language) are fundamentally different, preventing meaningful novelty comparison.

---

8. Adaptive Divergence Regularized Policy Optimization for Fine-tuning Generative Models

URL: View paper

Brief Assessment

Adaptive Divergence[63] focuses on adaptive divergence regularization for generative models (text-to-image, LLMs) and does not address music-specific chain-of-thought reasoning or music understanding tasks that are central to the original paper's contribution.

---

9. Audio-thinker: Guiding audio language model when and how to think via reinforcement learning

URL: View paper

Prior Art Analysis

Audio-Thinker[69] demonstrates that similar post-training approaches combining chain-of-thought reasoning with reinforcement learning for audio-language models were explored prior to the original paper's submission. The candidate paper explicitly describes using reinforcement learning (specifically GRPO) to enhance reasoning capabilities in large audio-language models, incorporating chain-of-thought style reasoning processes. Both papers address the same fundamental challenge: enabling audio-language models to perform explicit step-by-step reasoning through a combination of supervised fine-tuning on reasoning examples followed by reinforcement learning with custom rewards.

Evidence

Evidence 1 - **Rationale**: Both papers propose reinforcement learning frameworks for audio-language models that enhance reasoning through custom reward functions. The original paper's claim to novelty in combining chain-of-thought with GRPO-based RL is challenged by Audio-Thinker[69]'s similar approach. - **Original**: we propose a post-training stage specifically designed to enhance reasoning. for this stage, we further introducemf-think, a dataset of 300k chain-of-thought examples grounded in music theory, which we use for cold-start reasoning training. finally, we apply grpo-based reinforcement learning shao et... - **Candidate**: we propose audiothinker, a reinforcement learning framework designed to enhance the reasoning capabilities of lalms, with a focus on improving adaptability, consistency, and effectiveness. our approach introduces an adaptive think accuracy reward, enabling the model to adjust its reasoning strategie...

Evidence 2 - **Rationale**: Both papers use structured prompting with explicit thinking tags to train models on chain-of-thought reasoning before applying reinforcement learning, demonstrating similar methodological approaches to the post-training recipe. - **Original**: supervised fine-tuning with mf-think.to equip the model with advanced reasoning capabilities, we first perform sft of the music foundation model on our curated mf-think dataset. During this stage, we append the prompt: output the thinking process in and final answer in </ an... - **Candidate**: we prompt the model to first assess whether a query requires reasoning, and then either generate a reasoning process if needed or provide a direct answer otherwise. Details of the prompt are provided in appendix a.1. reasoning model

Evidence 3 - **Rationale**: Audio-Thinker[69] cites multiple prior works (r1-aqa, sari) that already applied GRPO and chain-of-thought training to audio-language models, indicating the approach was established in the field before the original paper's contribution. - **Original**: grpo for music reasoning and understanding.building on the advancements in the grpo algorithm, we adhere to the standard grpo algorithm to train our model. grpo obviates the need for an additional value function and uses the average reward of multiple sampled outputs for the same question to estimat... - **Candidate**: r1-aqa li et al. (2025a) utilizes the grpo algorithm to fine-tune the qwen2-audio model for audio question-answering tasks, enhancing reasoning accuracy with less data through reward-driven optimization. concurrently, sari wen et al. (2025) fine-tunes qwen2.5-omni xu et al. (2025) using reinforcemen...

Appendix: Text Similarity Detection

Textual similarity detection checked 32 papers and found 4 similarity segment(s) across 2 paper(s).

The following **2 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. CLAP Learning Audio Concepts from Natural Language Supervision

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

2. Multimodal music datasets? Challenges and future goals in music processing

Detected in: Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Music Flamingo: Scaling Music Understanding in Audio Language Models View paper
- [1] Audiolm: a language modeling approach to audio generation View paper
- [2] Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models View paper
- [3] Mumu-llama: Multi-modal music understanding and generation via large language models View paper
- [4] Continuous Audio Language Models View paper
- [5] Audio flamingo 3: Advancing audio intelligence with fully open large audio language models View paper
- [6] Audiogpt: Understanding and generating speech, music, sound, and talking head View paper
- [7] SALMONN: Towards Generic Hearing Abilities for Large Language Models View paper
- [8] Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities View paper
- [9] AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension View paper
- [10] MusicLM: Generating Music From Text View paper
- [11] Mmau: A massive multi-task audio understanding and reasoning benchmark View paper
- [12] Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities View paper
- [13] Evaluation of pretrained language models on music understanding View paper
- [14] Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies View paper
- [15] Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities View paper
- [16] Musilingo: Bridging music and text with pre-trained language models for music captioning and query response View paper
- [17] MUGen: Multi-modal Music Understanding and Generation with the Power of Large Language Models View paper
- [18] Uniaudio: An audio foundation model toward universal audio generation View paper
- [19] Music Genre Classification using Large Language Models View paper
- [20] video-SALMONN: Speech-Enhanced Audio-Visual Large Language Models View paper
- [21] Midashenglm: Efficient audio understanding with general audio captions View paper
- [22] Audio-llm: Activating the capabilities of large language models to comprehend audio data View paper
- [23] PAGURI: a user experience study of creative interaction with text-to-music models View paper
- [24] Evaluating multimodal large language models on core music perception tasks View paper
- [25] Naturelm-audio: an audio-language foundation model for bioacoustics View paper
- [26] Audio dialogues: Dialogues dataset for audio and music understanding View paper
- [27] MuChoMusic: Evaluating Music Understanding in Multimodal Audio-Language Models View paper
- [28] Generating Stereophonic Music with Single-Stage Language Models View paper
- [29] On the audio hallucinations in large audio-video language models View paper
- [30] MMAR: A Challenging Benchmark for Deep Reasoning in Speech, Audio, Music, and Their Mix View paper
- [31] Instruct-MusicGen: Unlocking Text-to-Music Editing for Music Language Models via Instruction Tuning View paper
- [32] Mulan: A joint embedding of music audio and natural language View paper
- [33] Enhancing audio comprehension in large language models: Integrating audio knowledge View paper
- [34] Diffusion based Text-to-Music Generation with Global and Local Text based Conditioning View paper
- [35] Converting Vocal Performances into Sheet Music Leveraging Large Language Models View paper
- [36] Semitone-Aware Fourier Encoding: A Music-Structured Approach to Audio-Text Alignment View paper

- [37] Tango 2: Aligning Diffusion-based Text-to-Audio Generations through Direct Preference Optimization View paper
- [38] VoxDialogue: Can Spoken Dialogue Systems Understand Information Beyond Words? View paper
- [39] Audio-language models for audio-centric tasks: A survey View paper
- [40] Interpreting song lyrics with an audio-informed pre-trained language model View paper
- [41] Make Some Noise: Towards LLM audio reasoning and generation using sound tokens View paper
- [42] Extending Audio Context for Long-Form Understanding in Large Audio-Language Models View paper
- [43] An Independence-promoting Loss for Music Generation with Language Models View paper
- [44] Musicagent: An ai agent for music understanding and generation with large language models View paper
- [45] Collap: Contrastive long-form language-audio pretraining with musical temporal structure augmentation View paper
- [46] Multimodal music mood classification using audio and lyrics View paper
- [47] Joint sentiment analysis of lyrics and audio in music View paper
- [48] Wavjourney: Compositional audio creation with large language models View paper
- [49] Uniaudio: Towards universal audio generation with large language models View paper
- [50] Arrange, Inpaint, and Refine: Steerable Long-term Music Audio Generation and Editing via Content-based Controls View paper
- [51] MusiXQA: Advancing Visual Music Understanding in Multimodal Large Language Models View paper
- [52] Improving BERT for symbolic music understanding using token denoising and pianoroll prediction View paper
- [53] Foundation models for music: A survey View paper
- [54] Are we there yet? a brief survey of music emotion prediction datasets, models and outstanding challenges View paper
- [55] MGPHot: A Dataset of Musicological Annotations for Popular Music (1958â□□2022) View paper
- [56] Multimodal music datasets? Challenges and future goals in music processing View paper
- [57] Songcreator: Lyrics-based universal song generation View paper
- [58] The interconnections of music structure, harmony, melody, rhythm, and predictivity View paper
- [59] Towards Unified Music Emotion Recognition across Dimensional and Categorical Models View paper
- [60] ERLD-HC: Entropy-Regularized Latent Diffusion for Harmony-Constrained Symbolic Music Generation View paper
- [61] Motion-R1: Chain-of-Thought Reasoning and Reinforcement Learning for Human Motion Generation View paper
- [62] SoundMind: RL-Incentivized Logic Reasoning for Audio-Language Models View paper
- [63] Adaptive Divergence Regularized Policy Optimization for Fine-tuning Generative Models View paper
- [64] Audio-cot: Exploring chain-of-thought reasoning in large audio language model View paper
- [65] Perception, reason, think, and plan: A survey on large multimodal reasoning models View paper
- [66] Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning View paper
- [67] Incentivizing Consistent, Effective and Scalable Reasoning Capability in Audio LLMs via Reasoning Process Rewards View paper
- [68] Robust chain of thoughts preference optimization View paper
- [69] Audio-thinker: Guiding audio language model when and how to think via reinforcement learning View paper
- [70] Thinking in cocktail party: Chain-of-Thought and reinforcement learning for target speaker automatic speech recognition View paper
- [71] Pam: Prompting audio-language models for audio quality assessment View paper
- [72] A survey of foundation models for music understanding View paper
- [73] CLAP Learning Audio Concepts from Natural Language Supervision View paper
- [74] Sparks of large audio models: A survey and outlook View paper
- [75] Llark: A multimodal foundation model for music View paper
- [76] Can synthetic audio from generative foundation models assist audio recognition and speech modeling? View paper