PIAC: A Taxonomy for Probing Musical Understanding and Hallucination in ALMs

***Abstract*—Audio-Language Models (ALMs) have shown measurable progress on music understanding benchmarks. Across a growing number of datasets, state-of-the-art models demonstrate increasingly strong performance. Yet despite these improvements, ALMs regularly produce descriptions of music that are fluent and confident, but incorrect. This discrepancy between apparent and actual musical understanding raises the question: how close are we really to true artificial musical intelligence (AMI) from audio?**

This paper focuses on how models perceive, analyze, feel, and contextualize music. In this paper, we argue that current evaluation practices substantially overestimate the musical understanding capabilities of ALMs. We present an overview of recent models and the benchmarks used to evaluate them, identify systematic shortcomings in how musical understanding is assessed, and propose an alternative, four-level taxonomy and evaluation framework, called PIAC (Perceptual, Inferential, Affective, Contextual). PIAC is designed to surface hallucinations and probe deeper musical intelligence. To support reproducible evaluation, we release an accompanying evaluation tool at toward-ami.com.

***Index Terms*—Signal, information processing, aligning human-machine generalization.**

I. Introduction

Audio-Language Models (ALMs) have shown measurable progress on music understanding benchmarks. Recent model families such as NVIDIA's Flamingo line [1], [2], [3] demonstrate substantial gains over earlier generations, with a 15% accuracy improvement on benchmarks like MMAR [4], and similar performance on MMAU [5] to GaMMA [6]. MERT [7] shows striking pitch detection performance on NSynth [8], achieving 94.4% accuracy. These improvements suggest that ALMs are increasingly capable of extracting salient musical cues from audio.

A natural long-term objective for ALMs is to become genuinely knowledgeable about music and to enable new modes of interaction for how we listen to, experience, and create it. Real-world applications include natural-language music search, pedagogical assistants, listener guides, composer support tools—a *secondary pair of ears* for sound designers and musicians alike. More advanced systems could enable richer multimodal interactions, such as generating imagery inspired by music or assisting video editors in selecting musically appropriate visual material. More broadly, ALMs hold the potential to make musical knowledge more accessible and to contextualize, interpret, and describe music in a manner comparable to an expert musician or teacher.

Achieving this vision requires the ability to describe music across multiple levels of abstraction. A competent system must be able to summarize an entire piece while also attending to fine-grained expressive details; explain *why* a passage is effective or emotionally impactful; anticipate plausible harmonic or formal continuations; describe how an interpretation could be improved; and relate musical content to stylistic or historical knowledge. These capabilities presuppose a grounded and structured notion of musical understanding: one that spans perception, analysis, affect, and contextual knowledge.

Consider, for example, the following type of description of a classical passage:

> "Before transitioning back to the second theme in the re-exposition, the leading tone (an A4), played louder than the surrounding notes, strongly augments the tension introduced by the previous diminished chords. At the same time, the music subtly accelerates, contributing to a feeling of despair. This leads up to a climax which bursts out at 02:15, an exclamation typical of the Romantic style but atypical of Prokofiev."

Or advice given to an emerging songwriter:

> You played a four-chord progression (D, A, F, B), and the recurring melodic figure subtly evokes Philip Glass through its rhythmic insistence and looping contour. One way to extend the phrase is to reinterpret the B as a dominant function, briefly resolving to E minor. This would preserve the established tension while also echoing the more turbulent material you played at the beginning of the piece.

Each example combines understanding of:

- **Perceptual content** (e.g., loudness, timing, acceleration),
- **Inferential content** (e.g., leading tones, harmonic function, form),
- **Affective content** (e.g., despair, tension, turbulence),
- **Contextual content** (e.g., stylistic norms, composer-specific tendencies).

Similar mixtures of understanding arise when describing groove and microtiming in jazz; identifying production choices in electronic music; explaining why a film cue heightens suspense; or proposing improvements to a composition for more consistency in the melodic material.

In practice, however, current ALMs often produce the appearance of musical understanding. As documented in [9], hallucinations in multimodal settings can be traced back to over-reliance on unimodal priors learned during training and spurious inter-modality correlations. Quantifying the prevalence of these hallucinations is challenging: the inherent subjectivity

of musical evaluation, the open-ended nature of music-to-text applications, and language being an imprecise probe of musical perception and interpretation make it hard to gauge an ALM's true understanding. The reliance on text further complicates evaluation: unlike humans, who can demonstrate musical understanding through physical reactions like facial expression or movements, we can observe a model's textual output, with little visibility into which auditory cues it relied upon.

To compensate, most benchmarks rely heavily on multiple-choice questions, which we show to consistently yield higher scores than open-ended alternatives probing the same skills. While convenient and scalable, this format inflates apparent capability and does not always reflect how we would want to interact with ALMs in real-world use cases. Current benchmarks oftentimes remain at the surface without probing for a profound analysis or interpretation, an observation that could be attributed to the complexity and subjectivity of music evaluation.

In addition, success on current benchmarks does not imply that a model can be trusted in open-ended musical interpretation. These issues are compounded by a deeper challenge inherent to multimodality. Hallucination in ALMs arises from ambiguity and weak grounding between audio and language, yielding convincing but unsubstantiated interpretations.

In this paper, we (1) survey existing benchmarks and evaluation practices and (2) discuss how evaluation choices, particularly answer format and question design, systematically shape reported performance. We (3) introduce a listener-centered taxonomy and evaluation framework aimed at disentangling perceptual, inferential, affective, and contextual understanding. Finally, we (4) release a living public website, available at `toward-ami.com`, which maintains an updated inventory of music understanding benchmarks and provides an LLM-as-a-judge evaluation toolkit for rapid, systematic assessment of audio-language models. To the best of our knowledge, `toward-ami` is the first public evaluation platform dedicated specifically to musical understanding. Our broader goal is to provide the community with tools and perspectives that move us closer to true artificial musical intelligence.

II. Related Work

There exist several benchmarks designed to probe music understanding, or more broadly, audio understanding. Some benchmarks target highly specific skills. A canonical example is NSynth [8], which focuses on pitch and instrument identification. Such benchmarks offer narrow coverage but high focus, making them well suited for isolating individual perceptual competencies. PitchBench [10] complements NSynth-style pitch identification by providing a systematic evaluation suite designed to probe multiple layers of both absolute and relative pitch perception.

A number of benchmarks aim for broader skill coverage across music and audio understanding. MuChoMusic [11], for example, consists of 1,187 questions, evaluating music understanding cross a wide range of tasks. MMAU-Pro [12] comprises approximately 5,305 questions spanning speech, sound, and music. It organizes skills into two main categories: perceptual skills (e.g. harmony, pitch, melody, rhythmic patterns, tempo, artist, timbre, instrument, and melodic structure recognition) and reasoning skills (e.g. emotion interpretation, form analysis, genre recognition). MMAR [4], by contrast, contains 206 music-focused questions. BASS covers a distinct set of capabilities, including artist collaboration, musicological analysis, structural segmentation, and structured lyric transcription.

Other benchmarks are constructed by aggregating or repurposing existing datasets. Examples include MARBLE [13] and CMI-Bench [14], which combine multiple sources to increase task diversity. Their tasks span key detection, music captioning, emotion tagging, pitch estimation, genre classification, and lyrics transcription, among others.

Across benchmarks, several orthogonal axes can be identified. One major distinction lies in answer format, with benchmarks relying on multiple-choice questions (MCQ), open-ended questions (OEQ), or a combination of both. Dataset sizes range from a few hundred to several thousand questions. Benchmarks also vary in scope: some are music-specific, while others include music as one modality among many, alongside speech or sound. Difficulty and depth of musical understanding likewise differ substantially. An updated overview of existing benchmarks is maintained on our accompanying website.

Benchmarks also differ in how they conceptualize musical understanding. MMAR, for instance, organizes questions into four layers: cultural, signal, semantic, and perception. However, the boundaries between these layers are sometimes unclear: music theory or statistical concepts are classified under the perception layer, while content-level analysis appears under the semantic layer. MMAU-Pro distinguishes between perceptual and reasoning skills, whereas MuChoMusic separates knowledge-based and reasoning-based questions.

While there exists some overlap, these categorizations are not consistent nor grounded in a clear model of how human listeners engage with music, and do not provide a principled account of how different types of musical skills should be evaluated.

III. PIAC Framework

Our aim is to design a framework that enables rigorous evaluation of *artificial musical intelligence from audio*, specifically musical understanding, in Audio-Language Models (ALMs), aligned with how human listeners engage with rendered music as well as *real-world use cases* in which ALMs are expected to operate.

Specifically, the design criteria are the following.

First, it should distinguish *how accurate a model is* from *how accurate a model can be*. Musical questions differ fundamentally in their level of ambiguity: some admit a single correct answer, while others permit multiple reasonable interpretations. By explicitly categorizing questions according to their ambiguity, the evaluation framework becomes aware of the limits of determinacy and can assess models accordingly.

Importantly, the different levels of ambiguity allow us to probe *depth of understanding*: whether a model's response is grounded in the audio and supported by reasoning, or whether it merely produces fluent but unsubstantiated musical discourse.

Second, it reflects the layered manner in which humans experience music, mapping directly to the four modes of listener engagement introduced in this paper: hearing (perceptual), analyzing (inferential), feeling (affective), and contextualizing. Furthermore, this framework is designed to support real-world human-model interactions across pedagogical, artistic, and streaming use cases.

To this end, we propose a listener-centered taxonomy organized around four modes of musical engagement, each corresponding to a distinct degree of ambiguity: **Perceptual**, **Inferential**, **Affective**, and **Contextual** content. For each category we discuss (1) the nature of the information, (2) its epistemic status, in particular the degree to which consensus is expected, (3) representative examples, (4) an appropriate evaluation methodology, and (5) an illustrative question.

*A. **Perceptual**: information directly measurable from the audio signal, admitting a single ground truth*

a) Information: Perceptual content comprises anything that is measurable from the audio itself. This category requires neither prior knowledge nor active reasoning.

b) Ambiguity: The defining characteristic is the absence of reasonable disagreement, given a predefined answer format: a given note corresponds to A4 or it does not; a note onset occurs at a specific time or it does not.

c) Coverage: This category includes *objective* signal-level attributes or audio features, such as pitch, timing, duration, loudness, instrumentation, and lyrics. Musical descriptors that don't fall into this category include meter, as it is not physically present in the signal and may be interpreted differently by different listeners (such as 4/4 or 2/2), lacking a clear ground truth.

d) Evaluation: Evaluation within this category can proceed via exact semantic match or within a predefined tolerance. A constrained answer often required to increase comparability between prediction and reference.

e) Example question: Q: "What note is played by the violin at 0:31:23? Answer in scientific pitch notation." *A:* "A4."

*B. **Inferential**: information derived from the audio through expert listening and analytical reasoning, for which expert consensus is expected, although limited disagreement may remain*

a) Information: Inferential content consists of musical properties that can be derived from the audio through expert listening and analytical reasoning.

b) Ambiguity: The claims are *intersubjective*: trained listeners tend to converge on an answer, but some degree of disagreement remains possible.

c) Coverage: This category includes harmonic function, formal segmentation, phrase structure, genre attribution, voice leading, and other aspects of musical organization.

d) Evaluation: Evaluation at this level should account for justified expert-annotated alternatives. Answers must be explicitly or implicitly substantiable in perceptual content. A claim about a theme's boundaries, for instance, should be traceable to observable features such as recurring melodic material or harmonic closure, even though identifying the theme itself requires some degree of interpretation.

e) Example question: Q: "[In a situation where a B♭ is played as a grace note to an A, above an A7 chord] What is the name of this chord?" *A:* "A7," or "A9."

*C. **Affective**: information related to the subjective experience of the listener, for which no consensus is expected or desirable*

a) Information: Affective content captures how music is experienced by a listener and relates to expression and emotional response.

b) Ambiguity: By definition, these claims are *subjective* and should not be resolved by consensus, as there is no single right answer.

c) Coverage: This category includes perceived mood, character, tension, intimacy, energy, aesthetic quality, and personal response.

d) Evaluation: Evaluation should accommodate multiple correct answers by focusing on core criteria: the plausibility of the affective description (whether it is musically reasonable), internal consistency (ensuring different parts of the response cohere), and grounding (verifying that affective claims are supported by perceptual or inferential references). For example, describing a passage as "tender" gains substance if it is supported by references to soft dynamics, legato articulation, and sustained harmonic resolution.

e) Example question: Q: "What is the most dramatic spot in the audio?" *A:* "Any spot that can be conceived as the most dramatic spot in the audio by an engaged human listener."

*D. **Contextual**: factual knowledge about the music external to the audio signal, admitting a single ground truth*

a) Information: Contextual content consists of factual information that is associated with the music through historical or physical context and admits a single ground truth.

b) Ambiguity: Contextual claims are not ambiguous in principle: they admit a single ground truth. When contextual information is unknown or unavailable, the correct response is to acknowledge that uncertainty.

c) Coverage: This includes the composer, performer, name, date of recording, or the audio's reception and influence. In contrast, detecting genre from audio is inferential and ambiguous.

d) Evaluation: Evaluation should proceed via exact semantic match.

e) Example question: Q: "Who is the composer of the piece I just played?" *A:* "Mozart."

E. Examples

Consider a short rubato passage spanning four bars:

- Perceptual: "The tempo decelerates from 120 to 90 BPM over four seconds."
- Inferential: "The ritardando delays the cadential resolution."
- Affective: "The music creates a sense of suspension and anticipation before the release."
- Contextual: "Such rubato is a commonly used rhetorical device in Romantic performance practice."

All four statements refer to the same musical material but operate at a different level. Assessing a model's response requires determining whether higher-level claims are appropriately grounded in lower-level ones.

> "The pianist plays at 66 BPM, slowly, with conviction, in a church."

- Perceptual: "66 BPM" is measurable, and therefore perceptual.
- Inferential: "Slowly" is inferential, as it maps a measurable quantity onto a relative musical category.
- Affective: "With conviction" is affective, reflecting the listener's experience.
- Contextual: "In a church" is contextual, as it refers to an external fact about the recording situation.

F. Taxonomy Structure and Advantages

1) Evaluative flexibility: A core advantage of the proposed taxonomy is its flexibility: it allows identical or highly similar queries to fall into different categories depending on the listener's mode of engagement and the underlying nature of the ground truth. For example, consider an inquiry regarding a track's recording location. If the answer relies on documented historical fact (e.g., a well-known studio session at Abbey Road), the question belongs to the **contextual** domain and should therefore be evaluated against world knowledge. Conversely, if the location must be deduced purely from acoustic cues (e.g., estimating room size from speech characteristics or reverberation profile), the task becomes **inferential**. In this latter scenario, evaluation shifts toward expert consensus. This multi-layered breakdown ensures that the framework can dynamically accommodate the subtle contextual framing of any musical query.

2) Intercategorical relationships: The four categories are inherently interrelated: perception provides the raw material for inference; inference supports affective judgment; and contextual knowledge situates both within broader musical traditions. Tracing these interconnections offers a mechanism for diagnosing model hallucinations. For instance, when a model produces a convincing inferential or affective description, the taxonomy allows us to test its foundational grounding: did it correctly perceive the underlying audio signal, or is it merely relying on generic textual priors? If an ALM characterizes a passage as "turbulent," the framework assesses whether it can identify the specific perceptual features driving that judgment. Because hallucination frequently originates from weak or incorrect perceptual grounding masked by plausible language, analyzing these intercategorical relationships isolates where a model's understanding fails.

3) Hierarchical layering: The taxonomy enables a stratified approach to evaluation. A query such as "how does the tempo change?" only touches superficially on musical understanding. In contrast, our framework may probe layered understanding over (1) measurable changes in the audio signal (Perceptual), (2) the structural function those changes serve in the composition (Inferential), (3) the psychological effect they elicit in the listener (Affective), and (4) the historical or stylistic traditions from which they arise (Contextual).

4) Complementarity with skill-based benchmarks: Our taxonomy complements existing skill-based benchmarks by foregrounding the epistemic status of musical claims. We argue that the structural foundations of musical understanding (with the exception of speech) ultimately reduce to *tone*, *time*, or their interaction, where tone encompasses pitch, loudness, and timbre, and time situates sound within a temporal structure. By mapping these foundational acoustic dimensions across our four epistemic categories, the taxonomy provides a framework for evaluating model performance across specific musical topics, listener engagement modes, and varying degrees of ambiguity.

IV. Experiments

We present a concrete implementation of PIAC as an evaluation instrument and use it to address three research questions:

- **Q1**: How do evaluation choices—particularly question formulation and answer format (multiple-choice versus open-ended)—systematically affect reported performance on music understanding benchmarks?
- **Q2**: Where and how do Audio Language Models hallucinate in music understanding tasks, and which categories of musical understanding are most vulnerable?
- **Q3**: How can a taxonomy-driven evaluation be used to probe deeper musical intelligence beyond benchmark surface accuracy?

A. Experimental Protocol

We follow a five-step evaluation protocol.

a) Step 1: Taxonomic categorization.: Each benchmark question is assigned to one PIAC category based on the type of information queried and the degree of listener consensus expected.

b) Step 2: Category-specific evaluation criteria.: Evaluation criteria are adapted per category. Perceptual and contextual questions are evaluated against a single ground truth (or predefined tolerances), inferential questions against expert-consensus alternatives, and affective questions using plausibility, internal consistency, and grounding with respect to the audio.

c) Step 3: Question refinement.: Benchmark questions are reformulated to probe deeper understanding by converting multiple-choice questions into structured open-ended prompts

Model	PIAC	MCQ (Apparent)	OEQ (Actual)	Hallucination	Gap
Audio Flamingo Next	Perceptual	50.0%	20.6%	76.5%	+29.4
	Inferential	45.5%	40.2%	59.8%	+5.3
	Affective	83.3%	50.0%	41.7%	+33.3
	Contextual	60.7%	32.1%	71.4%	+28.6
Gemini	Perceptual	56%	24%	74%	+32
	Inferential	56%	43%	57%	+13
	Affective	75%	67%	17%	+8
	Contextual	71%	21%	79%	+50

TABLE I
APPARENT VERSUS ACTUAL MUSICAL UNDERSTANDING ACROSS PIAC CATEGORIES FOR AUDIO FLAMINGO NEXT AND GEMINI. MCQ REFLECTS MULTIPLE-CHOICE ACCURACY, OEQ REFLECTS OPEN-ENDED CORRECTNESS, AND THE GAP DENOTES MCQ MINUS OEQ. HALLUCINATION RATES INDICATE THE PROPORTION OF OPEN-ENDED RESPONSES CONTAINING UNGROUNDED OR FABRICATED CLAIMS.

and by separating perceptual evidence from higher-level interpretation.

d) Step 4: Evaluation.: Models are evaluated on original multiple-choice questions (MCQs), their open-ended counterparts (OEQs), and decomposed follow-up questions using the appropriate category-specific criteria.

e) Step 5: Analysis.: We analyze discrepancies across formats and categories to identify cases where models appear competent under standard benchmarks but fail under more diagnostic probing.

V. RESULTS

We report results on the music subset of the MMAR benchmark using *Audio Flamingo Next* and *Gemini 3 Flash*.

A. Q1: Effect of Evaluation Format

Table 1 compares MCQ accuracy and OEQ performance for Audio Flamingo Next across PIAC categories. While the model achieves an overall MCQ accuracy of 50.49%, open-ended evaluation reveals a substantially lower mean score (0.38), with large category-dependent gaps.

Perceptual and contextual questions show the largest discrepancies (gaps of +29.4 and +28.6 points). Inferential questions exhibit a smaller gap, while affective questions achieve the highest OEQ scores, reflecting their lower reliance on precise perceptual grounding.

B. Q2: Hallucination Patterns

Open-ended evaluation exposes systematic hallucination patterns that are largely invisible under MCQ scoring.

a) Perceptual.: The model frequently fabricates precise signal-level values such as timestamps, repetition counts, or intensity ranges (e.g., seconds, number of occurrences), revealing failures in counting and temporal localization.

b) Inferential.: The most common failure mode consists of confident but incorrect musical analysis, including hallucinated keys, modes, chord progressions, or instrument roles, even when the corresponding MCQ is answered correctly.

c) Contextual.: The model substitutes plausible but incorrect world knowledge—such as wrong decades, regions, or stylistic origins—indicating reliance on high-frequency priors rather than evidence from the audio.

d) Affective.: Errors take the form of fluent but weakly grounded interpretations, where expressive language masks missing perceptual or contextual support.

In one representative example, the model correctly identifies the stylistic period of an excerpt and provides coherent justification, yet incorrectly describes the instrumentation, overlooking clearly audible string parts. High-level reasoning is thus built on false perceptual premises.

C. Q3: Probing Deeper Musical Intelligence

Many benchmark questions operate at a granularity insufficient to diagnose musical understanding. For example, asking whether tempo increases or decreases captures only a coarse trend and does not test whether a model perceives timing structure in a musically meaningful way.

A taxonomy-aligned evaluation instead decomposes such questions into perceptual probes (e.g., beat positions or local tempo estimates) and inferential probes (e.g., phrasing or expressive intent). Applying this decomposition reveals cases where models answer the benchmark question correctly but fail on the associated sub-questions, exposing gaps in underlying competence.

D. Apparent vs. Actual Performance

The same pattern holds for Gemini 3 Flash. While the model achieves 59.2% MCQ accuracy overall, OEQ performance drops to 38.3%, a gap of more than 20 points. Contextual questions degrade most sharply: although correct options are often selected, open-ended responses frequently fabricate composers, titles, or historical details.

Across skills, the largest discrepancies occur in loudness and mood, timbre and lyrics, and fine-grained harmony, where hallucination rates approach 90%. These results confirm that MCQ accuracy substantially overestimates musical understanding.

VI. DISCUSSION

These experiments position PIAC as both a diagnostic framework and an evaluation methodology. By making explicit what type of musical information a question targets and what epistemic status the answer should have, the taxonomy enables clearer interpretation of benchmark results and more reliable claims about model capabilities.

VII. CONCLUSION

We introduced a listener-centered taxonomy of musical understanding and demonstrated its use in restructuring the evaluation of Audio Language Models. Aligning question formulation, evaluation criteria, and epistemic expectations exposes hallucination, clarifies apparent versus actual competence, and enables open-ended evaluation grounded in how humans listen to and understand music.

ACKNOWLEDGEMENT

The authors acknowledge the use of artificial intelligence (ChatGPT and Claude Sonnet) to polish the writing in this text, as well as Claude Opus for refining the code for the experiments.

REFERENCES

[1] S. Ghosh et al., *Audio Flamingo Next: Next-Generation Open Audio-Language Models for Speech, Sound, and Music*, 2026. DOI: 10.48550/ARXIV.2604.10905 Accessed: May 2, 2026.

[2] S. Ghosh et al., *Music Flamingo: Scaling Music Understanding in Audio Language Models*, Nov. 2025. DOI: 10.48550/arXiv.2511.10289 arXiv: 2511.10289 [eess]. Accessed: Jan. 29, 2026.

[3] A. Goel et al., *Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models*, Jul. 2025. DOI: 10.48550/arXiv.2507.08128 arXiv: 2507.08128 [cs]. Accessed: Feb. 4, 2026.

[4] Z. Ma et al., *MMAR: A Challenging Benchmark for Deep Reasoning in Speech, Audio, Music, and Their Mix*, May 2025. DOI: 10.48550/arXiv.2505.13032 arXiv: 2505.13032 [cs]. Accessed: Mar. 10, 2026.

[5] S. Sakshi et al., *MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark*, Oct. 2024. DOI: 10.48550/arXiv.2410.19168 arXiv: 2410.19168 [eess]. Accessed: Mar. 10, 2026.

[6] Z. You, Z. Yu, M. Liu, B. Zhu, Y. Wan, and Z. Wu, *GaMMA: Towards Joint Global-Temporal Music Understanding in Large Multimodal Models*, May 2026. DOI: 10.48550/arXiv.2605.00371 arXiv: 2605.00371 [cs.SD]. Accessed: May 30, 2026.

[7] Y. Li et al., *MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training*, Dec. 2024. DOI: 10.48550/arXiv.2306.00107 arXiv: 2306.00107 [cs.SD]. Accessed: May 19, 2026.

[8] J. Engel et al., *Neural audio synthesis of musical notes with wavenet autoencoders*, 2017. eprint: arXiv:1704.01279.

[9] S. Leng et al., "The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio," *arXiv*, 2024. [Online]. Available: https://arxiv.org/abs/2410.12787

[10] M. L. Dujardin, S.-Z. Yu, C. C. Thomas-Smith, D. M. Chan, and K. Nguyen, *PitchBench: Measuring Pitch Hearing in Audio-Language Models*, May 2026. DOI: 10.48550/arXiv.2605.26176 arXiv: 2605.26176 [cs.SD]. Accessed: Jul. 1, 2026.

[11] B. Weck, I. Manco, E. Benetos, E. Quinton, G. Fazekas, and D. Bogdanov, *MuChoMusic: Evaluating Music Understanding in Multimodal Audio-Language Models*, Aug. 2024. DOI: 10.48550/arXiv.2408.01337 arXiv: 2408.01337 [cs]. Accessed: Mar. 10, 2026.

[12] S. Kumar et al., *MMAU-Pro: A Challenging and Comprehensive Benchmark for Holistic Evaluation of Audio General Intelligence*, Aug. 2025. DOI: 10.48550/arXiv.2508.13992 arXiv: 2508.13992 [eess]. Accessed: Mar. 10, 2026.

[13] R. Yuan et al., *MARBLE: Music Audio Representation Benchmark for Universal Evaluation*, Nov. 2023. DOI: 10.48550/arXiv.2306.10548 arXiv: 2306.10548 [cs]. Accessed: May 3, 2026.

[14] Y. Ma, S. Li, J. Yu, E. Benetos, and A. Maezawa, *CMI-Bench: A Comprehensive Benchmark for Evaluating Music Instruction Following*, Jun. 2025. DOI: 10.48550/arXiv.2506.12285 arXiv: 2506.12285 [eess]. Accessed: Mar. 10, 2026.