MUSICLISTENBENCH: CAN AUDIO LLMS HEAR THE DIFFERENCE YET?

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio large language models (LLMs) are typically evaluated on song genre, mood, and instrumentation. However, their ability to recognize the same melody across two short clips remains untested. We introduce *MusicListenBench*, a benchmark comprising paired listening tests for melody, harmony, timbre, and tempo. We generate all audio from symbolic music, ensuring precise ground-truth answers. Each test item comes with two variants. A FLIP variant changes the music, so the answer must change. A STAY variant changes only the sound, such as room echo or hiss, so the answer must stay the same. The two variants have opposite answers, so a model that ignores the audio scores 50%, whatever letter it prefers. Seven of eight open audio LLMs score within 4 points of this floor, and three of them give the same letter to 88% to 99% of items. Of four commercial models, both GPT-Audio models are at chance, and the best, Gemini 2.5 Pro, reaches 74.2%, 12.8 points below human listeners (87.0%). Post-training with GRPO on 10k synthetic clean pairs raises Qwen2.5-Omni from 61.2% to 99.0% on held-out clean pairs, 11.5 points above our listeners, and the gain transfers to a task left out of training. Yet on FLIP/STAY the trained model only reaches the level of our listeners (86.0%). It now misses almost no change (1.9%, against 24.7% for Gemini 2.5 Pro), yet it errs on STAY pairs as often as Gemini does (26.1% against 26.9%). On timbre and rhythm it still changes its answer on 20.8% and 16.8% of pairs where only the sound changed, while our listeners made at most one error on each of these tasks. Nine in ten of these errors come from one change, a quiet background hiss. Training taught the model to ignore EQ changes it never saw (1.4% errors), but not hiss or transposition. Models learn to hear a difference much faster than they learn which differences matter. We release (anonymous) data, code[1], and an open leaderboard[2].

1 INTRODUCTION

Audio large language models (audio LLMs) can describe aspects like the genre, mood and instruments of a recording (Yang et al., 2024; Sakshi et al., 2025; Weck et al., 2024). A more basic skill sits under these descriptions: telling whether aspects of two sounds are the same. Do two short clips play the same melody? The same chord? The same instrument? Which one is faster? Such same/different questions are a standard tool of hearing research (Hautus et al., 2021), and tests of human musical ability are built from them (Law & Zentner, 2012; Wallentin et al., 2010). Our human listeners answer them correctly 87.5% of the time. If a model cannot answer them, it is unclear what its descriptions of music are based on.

Recent benchmarks find audio LLMs at or near chance on basic music and sound perception, such as telling whether a melody was shifted in pitch (Carone et al., 2026; 2025; Sun et al., 2026). Two problems remain. First, a same/different score mixes two skills: noticing a change that matters and ignoring a change that does not. Such scores are also inflated by text priors and by a preference for one answer letter (Weck et al., 2024; Zang et al., 2025; López et al., 2025; Zheng et al., 2024). Second, these benchmarks only evaluate models as they are, so they cannot tell whether training fixes the problem.

[1]Code Link
[2]Leaderboard Link

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

We introduce MusicListenBench (Figure 1). It is built on a simple idea. A model that hears the difference must do two things: change its answer when the music changes, and keep its answer when only the sound changes. So every test item comes with two variants. In the FLIP variant we change the queried attribute, for example one note of the melody, and the correct answer switches. In the STAY variant we change only how the music sounds, for example by adding room echo or noise, and the correct answer stays. The two variants ask the same question and have opposite answers. A model that ignores the audio therefore scores 50%, whatever letter it prefers (exactly on the three same/different tasks and on average on rhythm; Section 3.3). STAY pairs also block a shortcut that clean pairs allow. On clean same/different pairs, a model can answer by checking whether the two clips sound alike, without knowing what a melody or a chord is. In a STAY pair the clips sound different even when the answer is 'same'. All audio is rendered from symbolic music, so every answer is known exactly, and the test set uses pitch ranges, a soundfont and sound effects that never appear in training.

We use the benchmark to answer the title question in three steps. (1) *Not yet.* On FLIP/STAY, seven of eight open audio LLMs are within 4 points of the 50% floor, and three of them give the same letter to 88% to 99% of items. Both GPT-Audio models are at chance (49.6% and 52.1%). The best commercial model, Gemini 2.5 Pro (74.2%), is 12.8 points below our listeners (87.0%) and below them on every task (Section 5.1). (2) *Training fixes clean accuracy.* GRPO on our training split raises Qwen2.5-Omni from 61.2% to 99.0% on held-out clean pairs, 11.5 points above our listeners. It also improves a task left out of training (harmony: 59.2% to 96.4%) and barely changes captioning (Sections 5.2 and 5.4). (3) *Models learn to notice change much more than to ignore it.* On FLIP/STAY the same model only reaches the level of our listeners (86.0%). Its miss rate falls from 36.7% to 1.9%, but its false-flip rate only falls from 41.0% to 26.1%, no lower than that of Gemini 2.5 Pro without training (26.9%). The model learns to ignore EQ changes it never saw in training (false flips fall from 33.7% to 1.4%), but not background hiss or transposition. On timbre and rhythm, where our listeners made at most one error each, hiss causes 85 of its 94 false flips (Section 5.3).

Our contributions are:

- MusicListenBench: 10,000 training and 3,000 test items with exact labels, graded difficulty, checks against shortcuts and held-out test conditions. We release them with the generator, a scoring script and a public leaderboard that accepts submissions from anyone (Section 3.5).
- The FLIP/STAY protocol. It separates noticing a change that matters from ignoring one that does not, and a model that ignores the audio scores 50% whatever letter it prefers.
- An evaluation of eight open and four commercial audio LLMs against human listeners.
- A GRPO study on four models. Training closes the gap on clean pairs but not on STAY pairs. We also test transfer to an unseen task and side effects on captioning.

2 Related work

Benchmarks for audio LLMs. AIR-Bench, MMAU and MMAR test audio LLMs on thousands of questions about speech, sound and music (Yang et al., 2024; Sakshi et al., 2025; Ma et al., 2026), and MuChoMusic focuses on music (Weck et al., 2024). They use real audio and questions written by people or LLMs. This makes them broad, but text priors and letter bias can enter the scores (Weck et al., 2024; Rouditchenko et al., 2025). Text-only models reach up to 56.4% on MuChoMusic (Zang et al., 2025), and audio LLM scores change when the order or wording of the options changes (López et al., 2025). Three benchmarks are closest to ours. MUSE tests ten music perception skills, including same/different judgments, on 200 recorded stimuli with 200 human listeners (Carone et al., 2026). Carone et al. (2025) compare MIDI and audio input on syncopation, transposition and chord quality. SonicBench renders controlled sounds for twelve physical attributes, including timbre and tempo, and puts both clips in one file as we do (Sun et al., 2026). All three evaluate models only as they are, without training. We differ in three ways. Every item has a FLIP and a STAY variant, so we can separate noticing a change from ignoring one, and a model that ignores the audio scores 50%. All 3,000 test items have exact labels and graded difficulty. And we provide a training split with held-out test conditions, so the benchmark can measure what post-training changes. Live arenas such as Music Arena (Kim et al., 2026) rank text-to-music systems by human votes, which scores what a model produces rather than what it hears.

Table 1: The four evaluation tasks. *Difference*: how the clips differ when they differ in the queried attribute (for rhythm, how the faster clip differs). *Clip*: length of one clip. Each task has 2,500 training and 250 clean test items.

Task	Question	Difference	Levels	Clip
Melody	Same melody?	one note of a four-note motif moves	100, 200, 400, 700 cents	2.5 s
Harmony	Same chord quality?	the third of the chord moves	100, 200, 300 cents	2.0 s
Timbre	Same instrument?	instrument family (piano, guitar, violin, flute)	one (categorical)	2.5 s
Rhythm	Which clip is faster?	click rate of a hi-hat track	rate ratio 1.10, 1.20, 1.35, 1.50	4.0 s

Paired tests. Same/different and two-choice designs are standard in hearing research (Hautus et al., 2021). The Musical Ear Test asks listeners whether two short melodies are the same when one note may have changed (Wallentin et al., 2010), much like our melody task, and PROMS uses same/different trials for melody, tempo and timbre (Law & Zentner, 2012). In language and vision, minimal pairs (Warstadt et al., 2020), contrast sets (Gardner et al., 2020) and paired image-caption tests (Thrush et al., 2022) check whether a model's answer follows the right change in the input. MMVP and NaturalBench pair images that need different answers and give credit only when all answers in a pair or group are right (Tong et al., 2024; Li et al., 2024). CheckList separates directional tests, where an edit should change the output, from invariance tests, where it should not (Ribeiro et al., 2020). FLIP and STAY are the listening versions of these two test types. Pairing them also cancels the option-letter bias that language models show on multiple-choice questions (Zheng et al., 2024).

Reinforcement learning for audio and music models. Herremans & Roy (2026) review preference alignment for music generation and its open problems, among them how reliable a preference model can be when musical quality is subjective. Such work optimizes what a model writes; we ask whether a model can hear, the skill any such judge needs first. Group relative policy optimization (GRPO; Shao et al., 2024) has improved audio question answering, for example for Qwen2-Audio on MMAU (Li et al., 2025).

3 MUSICLISTENBENCH

3.1 ITEMS AND TASKS

Each item is one audio file and one question. The file contains clip A, 1.0 s of silence and clip B. Putting both clips in one file, as SonicBench does (Sun et al., 2026), lets us test models that accept only one audio input. After GRPO, one model answers 99.0% of clean items correctly in this format (Section 5.2), so the format itself does not stop models from comparing the clips. The question is one of 8 to 10 hand-written wordings per task. It ends with a fixed instruction that says which letter means which answer, and the model answers with one letter. Table 1 lists the four tasks. Melody, harmony and timbre ask whether the two clips are the same. Rhythm asks which clip is faster (some wordings ask which is slower). Melody, harmony and rhythm have several difficulty levels, and test items are spread evenly across them.

3.2 EXACT LABELS AND NO EASY CUES

We write each item as symbolic music with music21 (Cuthbert & Ariza, 2010) and render it with FluidSynth (FluidSynth Developers, 2000) and General MIDI soundfonts. From each generated item we render a small set of files: the base pair, one edited clip that changes the answer, and two or three clips that change only the sound. Because the audio is generated, every answer is known exactly, and no human labeling is needed. We also remove easy cues. All clips in a task have exactly the same length, so a faster rhythm clip has more clicks, not a shorter file. All clips have the same loudness (-23 LUFS). Finally, we check that simple audio statistics cannot predict the answer. A logistic regression and a one-split decision tree try to predict the answer from duration, level, spectral

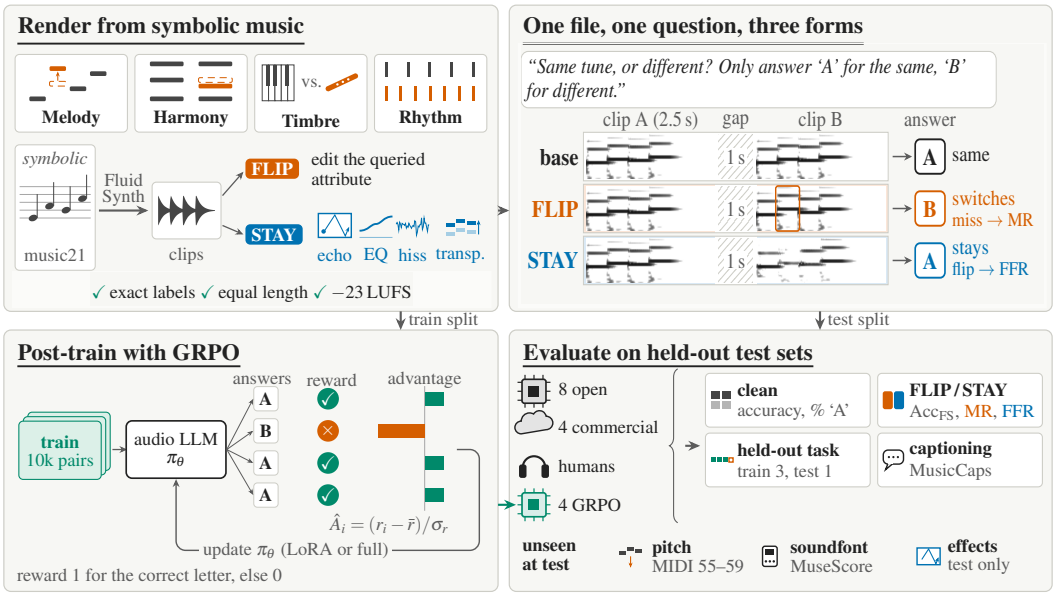


Figure 1: **MusicListenBench overview.** Items are rendered from symbolic music, so every answer is exact. A FLIP version changes the queried attribute and the answer; a STAY version changes only the sound and keeps the answer. We test open and commercial audio LLMs and human listeners under held-out conditions, and post-train open models with GRPO on clean pairs.

centroid and onset count. On every task they reach an AUC of at most 0.55, where 0.5 is chance. For rhythm we leave out onset count, because counting clicks is the task itself. These checks cover cues within the file. A cue in the relation between the two clips, such as how alike they sound, is what the STAY variants test (Section 3.3). Appendix A lists all parameters.

3.3 FLIP AND STAY: TWO WAYS TO FAIL

For each of the 1,000 clean test items we build two variants by replacing clip B.

- FLIP: clip B is replaced by an edited clip that changes the queried attribute, such as one moved note, a moved third or another instrument. For rhythm, the two clips swap places. The correct answer switches.
- STAY: clip B is replaced by a clip that sounds different but keeps the queried attribute. We use four such changes: room echo (reverberation time 0.2, 0.5 or 1.0 s, simulated with pyroomacoustics; Scheibler et al., 2018), a mild EQ change (± 3 dB shelves at 200 Hz and 4 kHz), background hiss at -50 dBFS and, for melody and harmony only, a transposition of up to 200 cents. The correct answer stays.

Transposition tests relative pitch. A melody moved up as a whole is still the same melody (Dowling, 1978), but a melody with one moved note is not. The four STAY changes are spread evenly over the items.

We report two error rates. The miss rate (MR) is the share of FLIP items answered wrongly: the model keeps the old answer although the music changed. The false-flip rate (FFR) is the share of STAY items answered wrongly: the model changes its answer although only the sound changed. Their average gives the FLIP/STAY accuracy, $\mathrm{Acc}_{\mathrm{FS}} = 1 - (\mathrm{MR} + \mathrm{FFR})/2$. This score has a useful property. The two variants of an item have opposite answers. On melody, harmony and timbre, where 'A' always means same, they also have opposite letters, so a model that ignores the audio gets exactly one of them right and scores exactly 50%, whatever letter it prefers. Rhythm questions ask for the faster or the slower clip, and each variant draws its own wording, so in 131 of 250 rhythm items both variants share the correct letter. There, a model that ignores the audio scores 50% only on average: always answering 'A' gives 47.8% on rhythm and 49.45% overall, against 52.8% on the clean test.

The two rates catch different failures. On the three same/different tasks, a model that answers 'different' whenever the two clips sound different rarely misses a FLIP change, but it false-flips on every STAY pair whose answer is 'same'. Each rate alone also moves with letter bias (always answering 'A' gives an MR of 53.0% and an FFR of 48.1%), but their average stays near 50%.

3.4 Splits and held-out conditions

The training split has 10,000 items (2,500 per task). We use it only for post-training. The test split has the 1,000 clean items and the 2,000 FLIP/STAY items. Files made from one generated item never appear in both splits. The test split also differs from training in three ways.

- *Pitch range.* Melody and timbre items start from MIDI notes 60 to 72 in training and 55 to 59 at test. Harmony roots come from MIDI 48 to 60 in training and 62 to 67 at test.
- *Soundfont.* Training audio is rendered with FluidR3_GM (Wen, 2000) or GeneralUser GS (Collins, 2024), and all test audio with MuseScore_General (Collins, 2020).[3]
- *Sound effects.* Room echo, EQ and hiss appear only at test, and GRPO trains on clean pairs only.

In test items the letters are fixed (A means same), and 528 of the 1,000 answers are 'A'. In training items we swap the meaning of the letters for about half of the items, so the letter never predicts the answer. The four tasks also allow a leave-one-task-out test: train on three tasks and test on the fourth.

3.5 Protocol and public leaderboard

MusicListenBench comes with a public leaderboard[4], and anyone can submit a model. The site reports clean accuracy per task, FLIP/STAY accuracy, MR, FFR and the share of 'A' answers, so letter bias is always visible. It lists zero-shot models and models that use our training split separately (Appendix E). Because the test items come from a generator, we can draw a fresh test set with new seeds if the public one becomes stale.

4 Experimental setup

Models. We test eight open audio LLMs: Qwen2-Audio (Chu et al., 2024), Qwen2.5-Omni (Xu et al., 2025), Audio Flamingo 3 (Ghosh et al., 2026), Phi-4-multimodal (Abouelenin et al., 2025), Kimi-Audio (Ding et al., 2025), MiMo-Audio (Xiaomi, 2025), Fun-Audio-Chat (Team et al., 2025) and Step-Audio 2 (Wu et al., 2025). We also test four commercial models through their APIs: GPT-Audio-1.5 (OpenAI, 2026), GPT-Audio-mini (OpenAI, 2026), Gemini 2.5 Pro (Comanici et al., 2025) and Gemini 3.8 Flash (Google DeepMind, 2026). For open models we read the answer by comparing the probabilities of the tokens 'A' and 'B'; reading the generated letter instead gives the same answer on every item. For the API models we use the generated letter. Appendix C lists all checkpoints and API settings.

Human reference. Twelve listeners, from casual listeners to musicians with years of practice, answered 24 items each: 288 answers in total, 72 per task and 96 each from the clean, FLIP and STAY tests. They heard the same audio and questions as the models, over headphones (Appendix B).

GRPO. We post-train four open models with GRPO: Qwen2-Audio, Qwen2.5-Omni, Audio Flamingo 3 and Phi-4-multimodal. We use LoRA (Hu et al., 2021) or full fine-tuning. The reward is 1 for the correct letter and 0 otherwise, plus a small bonus (0.2) for producing a parseable option (group size 8; learning rate 10^{-5} for LoRA and 10^{-6} for full fine-tuning; 8 prompts per step, i.e., 64 sampled rollouts). We report the checkpoint at step 1,000, except for Audio Flamingo 3 with full fine-tuning, where validation accuracy peaks at step 500 before the model over-optimizes the binary reward (rising KL and entropy collapse), so we stop there. Appendix D gives all settings.

[3] MuseScore_General is derived from FluidR3Mono, a mono version of FluidR3, and keeps its samples for instruments without separate credit. For those instruments the test soundfont shares source recordings with FluidR3_GM (Appendix A).

[4] https://musiclistenbench-review.pages.dev/

Table 2: Results without training (%). *Clean*: accuracy on 250 items per task (1,000 in All). *FLIP/STAY*: accuracy ($\mathrm{Acc_{FS}}$), miss rate (MR) and false-flip rate (FFR) on 2,000 items; lower is better for MR and FFR. Bold: best open model per column; underlined: best commercial model. The last two rows are references; the human row has 24 clean answers per task and 96 FLIP and 96 STAY answers in total.

	Clean accuracy					FLIP/STAY		
Model	Mel.	Har.	Tim.	Rhy.	All	$\mathrm{Acc_{FS}}$	MR	FFR
Open models								
Qwen2-Audio	46.0	50.4	43.2	49.2	47.2	50.7	47.7	50.9
Qwen2.5-Omni	**68.8**	**59.2**	**66.8**	50.0	**61.2**	**61.2**	**36.7**	**41.0**
Audio Flamingo 3	51.2	48.8	58.0	52.8	52.7	51.4	49.6	47.6
Phi-4-multimodal	54.4	47.2	56.8	50.8	52.3	49.1	53.5	48.3
Kimi-Audio	57.6	47.2	53.2	52.8	52.7	51.0	52.3	45.7
MiMo-Audio	54.0	54.4	58.0	**59.6**	56.5	52.2	49.4	46.1
Fun-Audio-Chat	50.0	51.6	60.0	49.6	52.8	53.6	46.6	46.1
Step-Audio 2	52.4	50.0	62.8	52.4	54.4	52.7	49.1	45.5
Commercial models (API)								
GPT-Audio-mini	47.6	51.6	54.4	50.8	51.1	49.6	50.6	50.3
GPT-Audio-1.5	45.6	47.6	52.0	44.0	47.3	52.1	45.8	50.1
Gemini 2.5 Pro	80.0	68.8	88.4	79.6	79.2	74.2	24.7	26.9
Gemini 3.8 Flash	53.6	54.4	85.6	72.4	66.5	64.0	33.0	39.1
Human (12)	79.2	87.5	95.8	87.5	87.5	87.0	5.2	20.8
Always 'A'					52.8	49.5	53.0	48.1

Leave-one-task-out. For Qwen2-Audio and Qwen2.5-Omni we train LoRA GRPO for 500 steps on three tasks (7,500 items) and test on the 250 clean items of the fourth task. We use the final checkpoint and do not look at test accuracy to choose it.

Captioning. To check for side effects, Qwen2-Audio and Qwen2.5-Omni caption the 2,735 available clips of the MusicCaps evaluation split (Agostinelli, 2023) before and after GRPO. We compare each caption with the reference using BLEU-4, METEOR, ROUGE-L and BERTScore-F1 (Papineni et al., 2002; Banerjee & Lavie, 2005; Lin, 2004; Zhang et al., 2019), on the same clips for every model (Appendix G).

Chance level. A score is not significantly different from 50% (two-sided, $p > 0.05$, normal approximation) between 43.8% and 56.2% for 250 items, and between 45.6% and 54.4%, 46.9% and 53.1%, and 47.8% and 52.2% for 500, 1,000 and 2,000 items.

5 Results

5.1 Can audio LLMs hear the difference?

Table 2 answers the title question: not yet. On clean pairs, seven of the eight open models score between 47.2% and 56.5%, close to the 52.8% of always answering 'A'. FLIP/STAY removes the benefit of preferring a letter. Five open models land inside the chance range (49.1% to 52.2%), Fun-Audio-Chat and Step-Audio 2 are just above it (53.6% and 52.7%), and only Qwen2.5-Omni is far above it (61.2%). Both GPT-Audio models are inside the chance range (49.6% and 52.1%). Gemini 3.8 Flash reaches 64.0% and Gemini 2.5 Pro 74.2%, the best score of any untrained model, yet still 12.8 points below human listeners (87.0%; $p < 0.001$, Fisher exact test).

Much of this failure is letter bias. Qwen2-Audio answers 'B' for 98.6% of clean items. Phi-4-multimodal answers 'A' for 95.9%, Audio Flamingo 3 for 87.7% and GPT-Audio-mini for 81.1% (Figure 5 and Table 14 in Appendix F). If we average their accuracy over the two answer classes, all four are at chance (49.9%, 49.7%, 50.6% and 49.4%).

Per-task clean scores also carry this bias, so we compare tasks on FLIP/STAY (Table 15 in Appendix F). Timbre is where most models rise above chance: three open and three commercial models are clearly

Table 3: GRPO results (%) on the held-out test sets. *none*: no training. †Step 500; all other runs step 1,000. Letter bias for the same runs is in Figure 5 and Appendix F.

		Clean accuracy					FLIP/STAY		
Model	GRPO	Mel.	Har.	Tim.	Rhy.	All	Acc_{FS}	MR	FFR
Qwen2-Audio	none	46.0	50.4	43.2	49.2	47.2	50.7	47.7	50.9
	LoRA	70.8	60.8	92.4	51.6	68.9	63.8	29.7	42.7
	full	70.4	61.6	80.8	52.0	66.2	61.5	31.7	45.4
Qwen2.5-Omni	none	68.8	59.2	66.8	50.0	61.2	61.2	36.7	41.0
	LoRA	99.6	96.8	99.6	100.0	99.0	86.0	1.9	26.1
	full	100.0	85.2	100.0	99.6	96.2	83.5	4.2	28.9
Audio Flamingo 3	none	51.2	48.8	58.0	52.8	52.7	51.4	49.6	47.6
	LoRA	57.2	74.8	97.6	50.8	70.1	64.4	29.7	41.4
	full†	68.0	80.8	99.6	69.2	79.4	70.4	22.5	36.7
Phi-4-multimodal	none	54.4	47.2	56.8	50.8	52.3	49.1	53.5	48.3
	LoRA	52.0	49.6	56.8	50.8	52.3	49.1	53.4	48.4
	full	55.6	54.0	88.8	50.8	62.3	58.9	42.9	39.3

above 50% on timbre, but only Qwen2.5-Omni and the two Gemini models are on melody. Gemini 3.8 Flash reaches 84.2% on timbre but only 57.2% on melody and 51.2% on harmony, where our listeners score 85.4% and 66.7%. Gemini 2.5 Pro is more even (73.2% on melody, 63.8% on harmony), but it trails our listeners most on the two tasks they find easiest, timbre (82.4%) and rhythm (77.4%, against 97.9% for both). Most of its errors are 'same' answers to clips that differ: it detects about half of the moved notes, whether a note moves by 100 or by 700 cents, and hears piano and guitar as the same instrument in 34 of 36 pairs (Appendix I).

5.2 Can post-training close the gap?

GRPO raises clean accuracy in seven of the eight runs, by 10.0 to 37.8 points (Table 3). Qwen2.5-Omni reaches 99.0% with LoRA and 96.2% with full fine-tuning. Audio Flamingo 3 reaches 79.4% and Qwen2-Audio 68.9%. The test audio uses other pitch ranges and another soundfont than the training audio, so these gains do not come from memorizing training clips. Training also reduces letter bias: the gap between accuracy on items whose answer is 'A' and on items whose answer is 'B' falls from 75.3 to 17.0 points for Audio Flamingo 3 (full) and from 97.2 to 39.6 points for Qwen2-Audio (LoRA; Figure 5 in Appendix F).

Two results stand out. First, only Qwen2.5-Omni (100.0% and 99.6%) and Audio Flamingo 3 with full fine-tuning (69.2%) learn the rhythm task; all other runs stay between 50.8% and 52.0%. Second, Phi-4-multimodal with LoRA learns nothing and moves further toward always answering 'A' (from 95.9% to 99.5% of items). GRPO learns from reward differences between answers sampled for the same item, and a model that almost always samples one letter rarely produces them, which is a likely reason for this failure.

5.3 Models learn to notice change more than to ignore it

Clean accuracy hides what GRPO teaches. Figure 2 shows how the miss rate and the false-flip rate of each model change with training. Training lowers the miss rate much more than the false-flip rate. In six of the seven runs that learned, the miss rate fell 2.2 to 3.2 times more than the false-flip rate ($p < 0.01$ for each run; Table 16). The exception is Phi-4-multimodal with full fine-tuning, where both rates fell by similar amounts (10.6 and 9.0 points). For Qwen2.5-Omni with LoRA, the miss rate falls from 36.7% to 1.9%, but the false-flip rate only falls from 41.0% to 26.1%. The trained model now misses 13 times fewer changes than Gemini 2.5 Pro (1.9% against 24.7%), but its false-flip rate is no lower (26.1% against 26.9%). Figure 4 in Appendix F places every model and our listeners on these two axes.

Human listeners also make more false flips than misses (20.8% against 5.2%), but most of their false flips come from one task: at least 11 of their 20 are on harmony items (Table 9). On timbre and

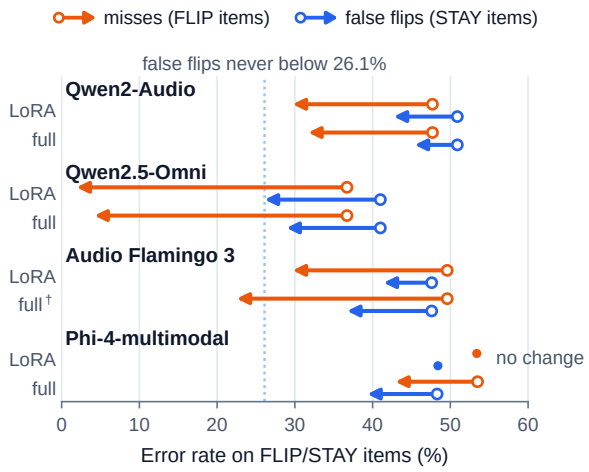


Figure 2: Misses and false flips on FLIP/STAY before and after GRPO, one row per run. Each arrow starts at the error rate without training (open circle) and ends at the rate after GRPO. Orange: miss rate (FLIP items); blue: false-flip rate (STAY items). Orange arrows are 2.2 to 3.2 times longer than blue ones in six of the seven runs that learned, and no run gets below 26.1% false flips (dotted line). Phi-4-multimodal with LoRA does not move. †Step 500. Values are in Table 3.

rhythm, listeners made at most one error each in 48 FLIP/STAY answers. The trained Qwen2.5-Omni misses fewer changes than our listeners (1.9% against 5.2%), yet it changes its answer on 20.8% of timbre and 16.8% of rhythm STAY pairs. Over these two tasks, the model false-flips on 94 of 500 STAY pairs. Across the runs that learned, clean accuracy is 3.4 to 13.0 points higher than FLIP/STAY accuracy, so clean accuracy alone overstates how well the models listen.

False flips by sound change. Each STAY pair applies one of four sound changes, and GRPO trains on none of them (Table 19 in Appendix H). Qwen2.5-Omni with LoRA learns to ignore EQ, where its false-flip rate falls from 33.7% to 1.4%, and partly room echo (37.2% to 16.6%). It does not learn to ignore hiss (49.0% before training, 52.4% after) or transposition (48.4% and 45.2%). On timbre and rhythm, the gap to our listeners comes from hiss alone: hiss causes 85 of the 94 false flips, and on the other 334 STAY pairs of these tasks (EQ or room echo) the model errs on 9 (2.7%). On rhythm, EQ and room echo change its answer on 1 of 167 items, and hiss on 41 of 83. The model also detects a single moved note in 99.6% of melody FLIP pairs, yet errs on 40.3% of melody pairs that are only transposed (46.8% before training). It hears a change of pitch but has not learned relative pitch. Full fine-tuning, Qwen2-Audio and Audio Flamingo 3 show the same pattern (Figure 6 in Appendix H).

This ranking of the changes holds before training too: without GRPO, Qwen2.5-Omni errs least on EQ and most on hiss and transposition, and so does Gemini 3.8 Flash. Gemini 2.5 Pro, which we did not train, also errs most on transposition and hiss (46.8% and 32.5%, against 22.4% for EQ and 17.2% for room echo), and on rhythm hiss raises its error rate to 42.2% (EQ 16.7%, room echo 13.3%). So the weakness to hiss is not caused by our training. GRPO sharpens the ranking but does not change it. We see a simple reason. The reward only checks the letter on clean pairs, so nothing in training asks the model to ignore a change in sound. Any invariance it gains is a side effect, and this side effect covers EQ and part of room echo, but not hiss or transposition.

5.4 Transfer to an unseen task, and side effects

Transfer to an unseen task. Training on three tasks improves the fourth, with a clear pattern (Figure 3). Harmony improves in both models. Qwen2-Audio goes from 50.4% to 71.6%, higher than the 60.8% it reaches when harmony is in training. Qwen2.5-Omni goes from 59.2% to 96.4%, close to its 96.8% with harmony in training. Timbre improves for Qwen2.5-Omni (66.8% to 96.8%), and melody gains 8 to 10 points. For Qwen2-Audio, timbre only moves from 43.2% to 50.8%, which is still chance. Rhythm does not improve (1.2 and 1.6 points), although Qwen2.5-Omni reaches 100.0% on rhythm when rhythm is in training. Rhythm is the only task about timing and the only one that is not a same/different question. So the models learn part of a general skill for comparing two sounds, either a real comparison of pitch and timbre or only a check of whether the clips sound alike; both help any same/different task, and neither says which clip is faster.

Side effects on captioning. GRPO on synthetic pairs keeps captioning close to the base model (Table 17 in Appendix G). On 2,735 MusicCaps clips, BERTScore-F1 changes by -1.28 to $+1.95$

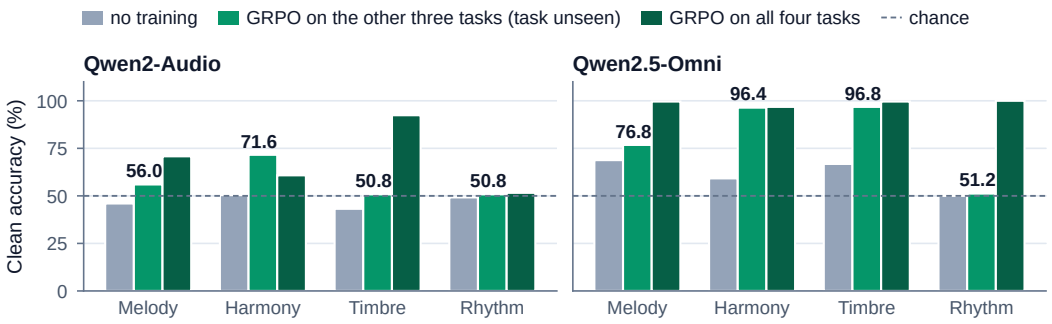


Figure 3: Transfer to a task left out of training (clean accuracy, 250 items per bar). Grey: no training. Green, with values: LoRA GRPO for 500 steps on the other three tasks, so the task itself is never seen. Dark green: LoRA GRPO for 1,000 steps on all four tasks. Without being trained on them, harmony improves in both models and timbre in Qwen2.5-Omni. Rhythm does not, although Qwen2.5-Omni reaches 100.0% when rhythm is in training.

points and BLEU-4 by -0.22 to $+1.73$. Mean caption length changes by less than one word, and no run produces degenerate captions (paired tests in Appendix G).

6 Discussion and limitations

Clean accuracy is not enough. On our clean test, always answering 'A' scores 52.8% and our best trained model beats our listeners by 11.5 points; on FLIP/STAY the first scores 49.45% and the second only matches them. Listening benchmarks should report paired variants next to clean accuracy. This matters beyond benchmarking: audio LLMs are already used to label, filter and rank music (Herremans & Roy, 2026), and a judge that changes its verdict when a quiet hiss is added is scoring the recording rather than the music.

Toward invariance. Section 5.3 suggests training on STAY-type pairs, or rewarding the same answer across a STAY pair and a changed answer across a FLIP pair; hiss and transposition need it most. Our training split already holds sound-only versions that GRPO did not use (codec, soundfont swap, transposition and, for timbre, a new melody; Table 5), and room echo, EQ and hiss appear only at test, so they can show whether such invariance carries over to new changes.

Limitations. The clips are short, synthetic and rendered from soundfonts, so a high score does not show good listening on real, mixed recordings. Each test item is evaluated once, in its original clip order and letter meaning, and on rhythm the 50% floor holds only on average (Section 3.3); swapped orders and letters and one wording per item would be cleaner. For some instruments, the test soundfont shares source recordings with a training soundfont. API models can change behind the same name.

7 Conclusion

With exact labels and paired FLIP/STAY variants, MusicListenBench shows that most open audio LLMs cannot hear the difference between two short musical clips, and that the best commercial model, Gemini 2.5 Pro, stays 12.8 points below human listeners. A little GRPO lifts the best model above our listeners on clean pairs, but on FLIP/STAY it at best matches them: it learns to notice change and to ignore a mild EQ change, but not a quiet hiss or a transposition. A reward that only checks the letter on clean pairs is already maximal for a model that reacts to any change at all, so invariance has to be asked for. Asking for it is cheap, since the two clips of a STAY pair come from the same generated item, and the reward only has to require the same answer across that pair. We consider the title question answered when a model matches our listeners' 87.0% on FLIP/STAY without ever seeing a STAY-type change in training. Teaching models which differences to ignore is the open problem.

LLM USAGE STATEMENT

We used an LLM assistant (Claude Sonnet) to paraphrase sentences and carry out grammatical checks. All research ideas, experiments and conclusions are the authors', who take full responsibility for the content.

ETHICS STATEMENT

Human listeners. Twelve adults took part in the listening test. They were told what the task was, that participation was voluntary and that they could stop at any time, and they gave consent before starting. We recorded only the answers and a coarse self-report of musical experience. No names, contact details, audio recordings or other identifying information were collected, so the released answers cannot be traced to a participant. Listeners set their own comfortable playback level before the test.

Audio and third-party material. All benchmark audio is synthesized from generated symbolic music with freely available soundfonts, so it contains no recordings of people and no copyrighted musical works. We follow the licence terms of each soundfont (Wen, 2000; Collins, 2024; 2020) and state them with the release. MusicCaps clips are used only to measure captioning side effects. We do not redistribute them; we release captions and scores, and the clips stay with their original source.

REPRODUCIBILITY STATEMENT

The leaderboard is available at `https://musiclistenbench-review.pages.dev`. We release the generator, all audio and item files, the scoring script, and the evaluation code at `https://anonymous.4open.science/r/MusicListenBench-F38B/`. Both are strictly anonymous.

Appendix A gives all generation parameters, Appendix C all model versions, and Appendix D all training settings.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Andrea Agostinelli. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Brandon James Carone, Iran R Roman, and Pablo Ripollés. Evaluating multimodal large language models on core music perception tasks. *arXiv preprint arXiv:2510.22455*, 2025.

Brandon James Carone, Iran R Roman, and Pablo Ripollés. The muse benchmark: Probing music perception and auditory relational reasoning in audio llms. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16097–16101. IEEE, 2026.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

S. Christian Collins. MuseScore_General: the default MuseScore SoundFont. `https://musescore.org/en/node/317991`, 2020. Derived from FluidR3Mono, MIT license; Accessed: September 20, 2026.

S. Christian Collins. GeneralUser GS: a General MIDI SoundFont bank. https://github.com/mrbumpy409/GeneralUser-GS, 2024. Accessed: September 20, 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Michael Scott Cuthbert and Christopher Ariza. music21: A toolkit for computer-aided musicology and symbolic music data. 2010.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

W Jay Dowling. Scale and contour: Two components of a theory of memory for melodies. *Psychological review*, 85(4):341, 1978.

R EBU-Recommendation. Loudness normalisation and permitted maximum level of audio signals. *Eur. Broadcast. Union*, 2011.

FluidSynth Developers. FluidSynth: a real-time SoundFont synthesizer. https://www.fluidsynth.org, 2000. Accessed: September 20, 2026.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Advances in Neural Information Processing Systems*, 38:41819–41886, 2026.

Google DeepMind. Gemini 3.8 Flash model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-8-Flash-Model-Card.pdf, 2026. Accessed 26 September 2026.

Michael J Hautus, Neil A Macmillan, and C Douglas Creelman. *Detection theory: A user's guide*. Routledge, 2021.

Dorien Herremans and Abhinaba Roy. Aligning generative music ai with human preferences: methods and challenges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 39699–39706, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yonghyun Kim, Wayne Chi, Anastasios Angelopoulos, Wei-Lin Chiang, Koichi Saito, Shinji Watanabe, Yuki Mitsufuji, and Chris Donahue. Music arena: Live evaluation for text-to-music. *Advances in Neural Information Processing Systems*, 38, 2026.

Lily NC Law and Marcel Zentner. Assessing musical abilities objectively: Construction and validation of the profile of music perception skills. *PloS one*, 7(12):e52508, 2012.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean De Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *Advances in Neural Information Processing Systems*, 37:17044–17068, 2024.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*, 2025.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Fernando López, Santosh Kesiraju, and Jordi Luque. Robustness assessment of large audio language models in multiple-choice evaluation. *arXiv preprint arXiv:2510.04584*, 2025.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *Advances in Neural Information Processing Systems*, 38, 2026.

OpenAI. GPT-Audio-1.5 Model Documentation. `https://developers.openai.com/api/docs/models/gpt-audio-1.5`, 2026. Accessed: September 20, 2026.

OpenRouter Team. Openrouter: A unified api for ai models. `https://openrouter.ai`, 2026. Accessed: 2026-09-26.

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

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4902–4912, 2020.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. Omni-r1: Do you really need audio to fine-tune your audio llm? In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7. IEEE, 2025.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025.

Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 351–355. IEEE, 2018.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yirong Sun, Yanjun Chen, Xin Qiu, Gang Zhang, Hongyu Chen, Daokuan Wu, Chengming Li, Min Yang, Dawei Zhu, Wei Zhang, et al. Sonicbench: Dissecting the physical perception bottleneck in large audio language models. *arXiv preprint arXiv:2601.11039*, 2026.

Tongyi Fun Team, Qian Chen, Luyao Cheng, Chong Deng, Xiangang Li, Jiaqing Liu, Chao-Hong Tan, Wen Wang, Junhao Xu, Jieping Ye, et al. Fun-audio-chat technical report. *arXiv preprint arXiv:2512.20156*, 2025.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238. IEEE, 2022.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024.

Jean-Marc Valin, Koen Vos, and Timothy Terriberry. Definition of the opus audio codec. Technical report, 2012.

Mikkel Wallentin, Andreas Højlund Nielsen, Morten Friis-Olivarius, Christian Vuust, and Peter Vuust. The musical ear test, a new reliable test for measuring musical competence. *Learning and Individual Differences*, 20(3):188–196, 2010.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. Muchomusic: Evaluating music understanding in multimodal audio-language models. *arXiv preprint arXiv:2408.01337*, 2024.

Frank Wen. FluidR3_GM: a General MIDI SoundFont. `https://musical-artifacts.com/artifacts/738`, 2000. MIT license; Accessed: September 20, 2026.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025.

LCT Xiaomi. Mimo-audio: Audio language models are few-shot learners. *CoRR, abs/2512.23808*, 2025.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, 2024.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, volume 2024, pp. 19426–19454, 2024.

A Benchmark construction

Generation. Table 4 gives the generator settings. All settings, including the split rules, were fixed in configuration files before any model was trained.

Table 4: Generator settings. MIDI note 60 is middle C. Training and test pitch ranges do not overlap.

Task	Setting
Melody	Four-note motif, 0.4 s per note, in a 2.5 s clip. When the melodies differ, one note moves by 100, 200, 400 or 700 cents. Starting pitch from MIDI 60 to 72 (train) or 55 to 59 (test).
Harmony	Triad in a 2.0 s clip. When the chord qualities differ, the third moves by 100, 200 or 300 cents. Major thirds move down and minor thirds move up, so a 100-cent move turns major into minor and minor into major. The moved third stays at least 50 cents away from the root and the fifth. Root from MIDI 48 to 60 (train) or 62 to 67 (test).
Timbre	Short melody in a 2.5 s clip on piano, guitar, violin or flute. Two different instruments always come from different families; close pairs such as acoustic and electric piano are excluded. Starting pitch as for melody.
Rhythm	Hi-hat click track in a 4.0 s clip, 50 ms clicks. The click rates of the two clips differ by a ratio of 1.10, 1.20, 1.35 or 1.50. The time of the first click is drawn independently of the answer.

Audio versions of each item. From each generated item we render a base pair (clips A and B), one edited clip B that changes the answer, and two or three clips B that only change the sound. Table 5 lists the sound-only changes. Each keeps the queried attribute fixed, so the correct answer stays the same.

Table 5: Sound-only changes by split and task.

Change	Setting	Split	Tasks
Room echo	pyroomacoustics simulation, reverberation time 0.2, 0.5 or 1.0 s	test	all
EQ	shelving filters, ± 3 dB at 200 Hz and 4 kHz	test	all
Hiss	added noise at -50 dBFS	test	all
Codec	Opus at 48 kbps (Valin et al., 2012) or MP3 at 128 kbps	train	all
Soundfont swap	same instrument, other soundfont	train	melody, harmony, rhythm
New melody, key	melody and key drawn again, same instrument	train	timbre
Transposition	0 to 200 cents	train, test	melody, harmony

Length, loudness and format. After every change we decode (for the codecs), trim and check that the number of samples equals the clip length of the task. A test file (clip A, 1.0 s of silence, clip B) is therefore 6.0 s long for melody and timbre, 5.0 s for harmony and 9.0 s for rhythm, with the same number of samples for every file in a task. Every clip is normalized to -23 LUFS (EBU-Recommendation, 2011) and stored as 16-bit PCM WAV at 44.1 kHz, mono.

FLIP and STAY items. These items use no new audio; they recombine test renders. Each clean test item gives one FLIP and one STAY item. Every STAY item keeps the original clip A and takes clip B from a sound-only change. For melody, harmony and timbre, the edit changes clip B only, and clip A of the edited render is byte-identical to the original clip A. The rhythm edit swaps the two clips, so rhythm FLIP items use the swapped pair. The correct answer of every FLIP item differs from that of its base item (1,000 of 1,000). For melody, harmony and timbre the correct letter changes too. For rhythm the faster clip changes, but each version draws its own wording (faster or slower), so the correct letter changes in 130 of the 250 rhythm FLIP items (Section 3.3). The sound-only changes rotate over items (Table 6).

Questions. Each task has 8 to 10 hand-written wordings. Rhythm has 10 wordings, some asking for the faster clip and some for the slower one, and the correct letter follows the wording. Every question ends with a fixed instruction that says what each letter means, for example "Only answer

Table 6: Composition of the FLIP/STAY test (2,000 items).

		STAY change				
Task	FLIP	EQ	Hiss	Room	Transposition	Total
Melody	250	63	63	62	62	500
Harmony	250	63	63	62	62	500
Timbre	250	84	83	83	n/a	500
Rhythm	250	84	83	83	n/a	500
Total	1,000	294	292	290	124	2,000

'A' for the same, 'B' for different . . . ". One wording is drawn per test file with a fixed seed, so the base, FLIP and STAY versions of an item usually have different wordings. Test items keep a fixed letter meaning. Training items swap the meaning for about half of the items, balanced so that the letter carries no information about the answer. Tables 7 and 8 list every wording and the exact instruction under both letter meanings.

Table 7: All 34 question wordings, verbatim. Each prompt is *wording* + *fixed instruction* (Table 8) followed by the audio placeholder `<audio>`, to which the item's concatenated clip is attached. Gold letters are balanced (train: 49.6% A; test: 52.8% A).

Task	Question wording	Correct letter is the
Melody	Two melodic clips are joined by a brief silence. Do they play the same melody, or a different one?	–
	You will hear two short melodies separated by a short pause. Is the melody the same in both, or different?	–
	Listen to two melodic phrases with a brief gap between them. Same tune, or different?	–
	A pair of melodies is presented back to back with a short silence in between. Are they the same melody?	–
	Two musical phrases follow one another after a brief pause. Is the second phrase the same melody as the first, or different?	–
	You are given two clips of melodic material with a short silent gap. Do they share the same melody?	–
	Listen carefully to the two melodic clips, divided by a brief silence. Same melody, or not?	–
	Two short tunes are played in sequence with a pause between them. Is it the same melody twice, or two different melodies?	–
Harmony	Two chords are played in sequence with a brief silence between them. Is the chord the same in both, or different?	–
	You will hear two chords separated by a short pause. Same chord, or different?	–
	Listen to two chords joined by a brief gap of silence. Do they sound like the same kind of chord, or different?	–
	A pair of chords is presented back to back with a short silence in between. Is the harmony the same, or different?	–
	Two short chords follow each other after a brief pause. Same chord type, or different?	–
	You are given two chord clips with a short silent gap. Do they share the same chord type?	–

continued on next page

Table 7 (continued)

Task	Question wording	Correct letter is the
	Listen to the two chords, divided by a brief silence. Same harmonic quality, or not?	–
	Two chords are played one after the other with a pause between them. Is it the same chord twice, or two different qualities?	–
Timbre	Two clips are joined by a brief silence. Are they played by the same instrument, or different instruments?	–
	You will hear two short clips separated by a short pause. Same instrument in both, or different?	–
	Listen to two clips with a brief gap between them. Is it the same instrument, or a different one?	–
	A pair of clips is presented back to back with a short silence in between. Are they performed on the same instrument?	–
	Two clips follow one another after a brief pause. Is the second clip the same instrument as the first, or different?	–
	You are given two clips with a short silent gap. Do they share the same instrument?	–
	Listen carefully to the two clips, divided by a brief silence. Same instrument, or not?	–
	Two clips are played in sequence with a pause between them. Is it the same instrument twice, or two different instruments?	–
Rhythm	Two audio segments are given, with a brief silence in between. Which segment moves faster?	faster clip
	Two audio segments are given, with a brief silence in between. Which segment moves slower?	slower clip
	You are presented with two sound clips, separated by a short silent interval. Which clip carries more speed?	faster clip
	You are presented with two sound clips, separated by a short silent interval. Which part proceeds at a slower pace?	slower clip
	Listen to a pair of audio segments with a brief silent space separating them. Which audio feels slower?	slower clip
	You are given two clips with a small silent gap in between. Which segment moves faster?	faster clip
	Two auditory recordings are given, separated by a small pause. Which part has a faster rhythm?	faster clip
	You are provided with two audio clips separated by a short silent gap. Which part proceeds at a faster pace?	faster clip
	Two sound recordings are presented, separated by a brief silent segment. Which clip runs more leisurely?	slower clip
	You have two auditory segments, separated by a short pause. Which segment plays more rapidly?	faster clip

Statistics. The benchmark contains 11,825 generated items: 3,025 each for melody, harmony and timbre and 2,750 for rhythm. Melody, harmony and timbre include extra catch trials with two identical clips, which we do not use in this paper. The items used here are 10,000 training items, 1,000 clean test items and 2,000 FLIP/STAY items (250 of each kind per task in the test split). Of the 1,000 clean test items, 528 have the answer 'A' and 472 the answer 'B'.

Table 8: Fixed instruction appended verbatim to every wording (before the audio placeholder), under both letter meanings. The swapped order applies only to same/different tasks, to 51.4% of their training items; the test split always uses the original order.

Applies to	Instruction (verbatim)
Melody/Harmony/Timbre, original order	"Only answer 'A' for the same, 'B' for different. Do not add any explanation, punctuation, or extra text."
Melody/Harmony/Timbre, swapped order	"Only answer 'B' for the same, 'A' for different. Do not add any explanation, punctuation, or extra text."
Rhythm, both splits	"Only answer 'A' for the first, 'B' for the second. Do not add any explanation, punctuation, or extra text."

Soundfont overlap. MuseScore_General is derived from FluidR3Mono, a mono version of FluidR3. Its documentation states that instruments without separate credit still use FluidR3Mono samples, and its changelog lists new piano samples [5]. For instruments that keep FluidR3Mono samples, the test soundfont changes the file and its instrument settings, but the source recordings overlap with FluidR3_GM, one of the two training soundfonts. The soundfont change is therefore weaker for those instruments than for piano.

B Human reference study

12 people took part with various levels of musical expertise. Each answered 24 test items, so we have 288 answers. They heard the same audio files and questions as the models. We removed three test runs and one empty submission. Accuracy was 87.5% on clean items (84 of 96; 95% interval 79.4 to 92.7), 94.8% on FLIP items (91 of 96; 88.4 to 97.8) and 79.2% on STAY items (76 of 96; 70.0 to 86.1). By task, over all items, it was 97.2% for timbre (70 of 72), 94.4% for rhythm (68 of 72), 83.3% for melody (60 of 72) and 73.6% for harmony (53 of 72). Per person, accuracy ranged from 66.7% to 95.8%. The intervals are Wilson intervals that treat all answers as independent; each person gave 24 answers, so the true uncertainty is somewhat larger.

Table 9: Human answers by task. Clean: 24 answers per task. FLIP/STAY: 48 answers per task, both variants together. Over all tasks there are 5 misses and 20 false flips, so at least 11 of the false flips are on harmony and at most one each on timbre and rhythm.

	Clean		FLIP/STAY	
Task	Correct	%	Correct	%
Melody	19 of 24	79.2	41 of 48	85.4
Harmony	21 of 24	87.5	32 of 48	66.7
Timbre	23 of 24	95.8	47 of 48	97.9
Rhythm	21 of 24	87.5	47 of 48	97.9
All	84 of 96	87.5	167 of 192	87.0

C Models and inference

Open models. All open models are tested without training and with their default chat templates. Each item is one audio file and one text question. We compare the log-probabilities of the tokens 'A' and 'B' at the first answer position and take the larger. Greedy decoding gives the same letter on every item. Table 10 lists the exact checkpoint (Hugging Face name and revision), parameter count, and audio sampling rate for each of the eight models.

Commercial models. GPT-Audio-1.5 is called through the OpenAI API with the model `gpt-audio-1.5`. Gemini 3.8 Flash is called through the Gemini API with the model `gemini-3.8-flash` (high thinking). Both are called at temperature 0. Each trial's two clips are

[5] https://github.com/CarlGao4/Muse-Sounds

Table 10: Open-model checkpoints. Revisions are the local Hugging Face cache commits actually used for evaluation; parameter counts are computed from the checkpoint safetensors headers.

Model	Hugging Face checkpoint (revision)	Params	Audio sr
Qwen2-Audio	`Qwen/Qwen2-Audio-7B-Instruct` (`0a09522`)	8.4B	16 kHz
Qwen2.5-Omni	`Qwen/Qwen2.5-Omni-7B` (`ae9e169`)	10.7B[†]	16 kHz
Audio Flamingo 3	`nvidia/audio-flamingo-3-hf` (`7d4bae6`)	16.5B	16 kHz[‡]
Phi-4-MM	`microsoft/Phi-4-multimodal-instruct` (`93f923e`)	5.6B	16 kHz
Kimi-Audio	`moonshotai/Kimi-Audio-7B-Instruct` (`9a82a84`)	∼7.9B[§]	16 kHz
MiMo-Audio	`XiaomiMiMo/MiMo-Audio-7B-Instruct` (`c359441`)	8.0B[¶]	24 kHz
Fun-Audio-Chat	`FunAudioLLM/Fun-Audio-Chat-8B` (`7bf72dc`)	9.5B	16 kHz
Step-Audio 2 mini	`stepfun-ai/Step-Audio-2-mini` (`e36fdd5`)	8.3B[‖]	16 kHz

[†]Full checkpoint (thinker 8.9B + talker 1.4B + token2wav 0.5B); evaluation instantiates the thinker only. [‡]From the model card; the cached snapshot contains no processor config to confirm. [§]Model-card figure; local count not completed. [¶]Excludes the separate MiMo-Audio-Tokenizer (1.3B, rev `5df9914`). [‖]Excludes the token2wav vocoder weights.

concatenated into a single mono 16-bit, 44.1 kHz WAV and sent inline (base64) with the question. Neither API exposes option log-probabilities for these models, so both are scored by their generated letter. For Gemini we constrain decoding to the two option letters, so every response is a valid letter; no invalid answers occurred (0 in 3,000 calls). Chat Completions offers no such constraint for audio models, so GPT-Audio-1.5 returns a completion of at most 16 tokens, from which we read the first standalone option letter (word-bounded, case-insensitive). Responses containing no standalone letter (10.3% clean, 6.9% control) are recorded as a prediction of the first option; under the balanced answer key this fallback scores at chance, and all such items are flagged in the released logs. GPT-Audio-mini and Gemini 2.5 Pro are called through the OpenRouter API (OpenRouter Team, 2026) with the model IDs `openai/gpt-audio-mini` and `google/gemini-2.5-pro`, on the same WAV files and prompts. We read the letter from the generated text. All 3,000 responses of each of these two models were exactly 'A' or 'B', so no fallback was needed. All API runs were conducted in September 2026.

D GRPO DETAILS

GRPO uses only the clean base pairs of the training split; none of the sound-only versions in Table 5 is used for training. We report the checkpoint at step 1,000 for all runs except Audio Flamingo 3 with full fine-tuning, which uses step 500. Leave-one-task-out runs use LoRA for 500 steps on 7,500 items and the final checkpoint.

Table 11: GRPO training details. Batch size is given as prompts per GPU per step (B) and rollouts per prompt (group size G); all runs use one gradient update per step. Wall-clock times are elapsed times measured on a shared node (*cf.* Footnote 11). Q2A = Qwen2-Audio-7B-Instruct, Omni = Qwen2.5-Omni-7B, AF3 = Audio-Flamingo-3, Phi-4 = Phi-4-Multimodal.

	LoRA	Full fine-tuning
Group size G (rollouts/prompt)	8	8
Sampling temperature / top-p	1.0 / 1.0	1.0 / 1.0
Batch size B (prompts/GPU/step)	4	4
Effective batch (rollouts/step)	64 (8×8)[a]	64 (8×8)
Learning rate (schedule)	1×10^{-5}, constant	1×10^{-6}, constant
Optimizer	AdamW	Adafactor
Gradient clipping	1.0	1.0
KL coefficient (vs. frozen reference)	0.10	0.10
LoRA rank / α / dropout	16 / 32 / 0	—
Trainable parameters	LM-decoder linear projections only[b]	LM decoder (+ lm_head)
Max. answer length	8 new tokens	8 new tokens
Training steps	1000 (all models)	1000 (Q2A, Omni, Phi-4); 500 (AF3)
Hardware	2 GPUs per run[c]	2× NVIDIA A100 (80 GB)
Wall-clock time (h)	Q2A ≈2.3, Omni ≈5.9, AF3 ≈3.9, Phi-4 ≈0.8	Q2A ≈2.7, Omni ≈2.8, AF3 ≈2.7, Phi-4 ≈25.5[d]

Reward (per rollout): $r = 1.0 \cdot \mathbb{1}[\text{correct}] + 0.2 \cdot \mathbb{1}[\text{parses to A/B}] + 0.1 \cdot s_{\text{task}} \cdot (\pm 1)$, where the last term is an anti-collapse nudge ($+$ if the rollout predicts the currently under-represented letter for that task, $-$ otherwise; s_{task} is an EMA with decay 0.95 of the per-task predicted-letter skew). Zero-advantage groups are resampled up to 4 extra times before being dropped.

[a]Qwen2.5-Omni LoRA ran on 6 GPUs: 24 prompts × 8 = 192 rollouts/step. [b]Attention and MLP projections of the language-model decoder ($q/k/v/o$/gate/up/down; fused qkv/o/gate_up/down for Phi-4); audio encoders and projectors frozen. [c]Single 8-GPU node: 2× A100 (80 GB) + 6× A40 (46 GB); the exact A100/A40 assignment per LoRA run was not recorded. [d]Elapsed time between first and last checkpoint writes on a node shared with other jobs; the Phi-4 full-FT run overlapped several concurrent jobs and its elapsed time is substantially inflated relative to exclusive-GPU compute (cf. ≈2.7–2.8 h for the other full-FT runs).

E Leaderboard submission format

A submission is one JSON Lines file with one record per test item, holding the item ID and the answer letter ('A' or 'B'). It also carries the model name and version, how the answer was read (letter probabilities or generated text), whether the model used our training split in any form, the date of the run and an optional link to code. Missing or invalid answers count as wrong. The scoring script is the one used for Tables 2 and 3. Training on test items is not allowed.

F Full results

Figure 4 places every model on the two error rates of FLIP/STAY, and Figure 5 shows the letter bias of the four trained models. The tables below give the per-task numbers behind Tables 2 and 3.

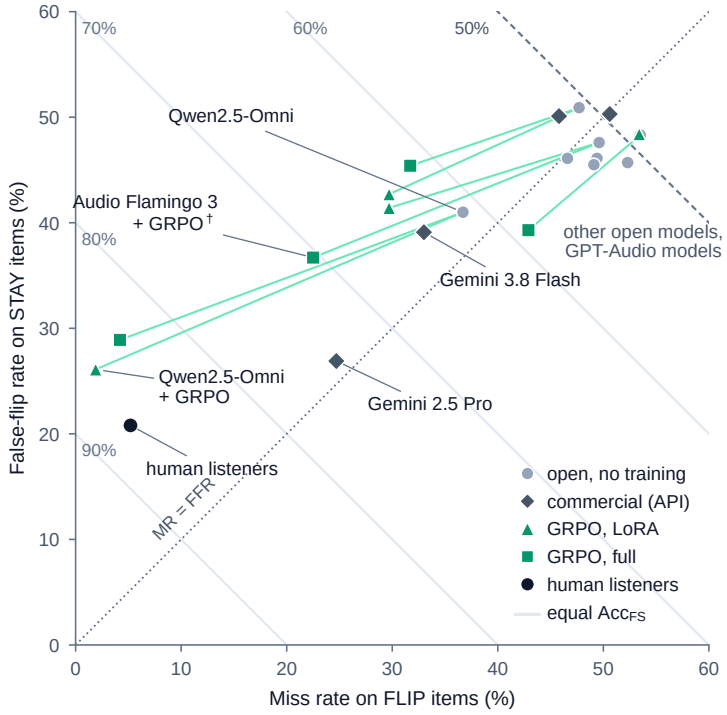


Figure 4: Every model on the two FLIP/STAY error rates, before and after GRPO. The accuracy $\mathrm{Acc}_{\mathrm{FS}}$ is constant along each grey diagonal; a model that ignores the audio lies on the dashed diagonal (50%). Grey circles: open models without training. Dark diamonds: commercial models. Green: GRPO runs (triangles LoRA, squares full fine-tuning), joined to their untrained model by thin lines. Black: our listeners. Without training, the models that listen (Gemini 2.5 Pro, Gemini 3.8 Flash, Qwen2.5-Omni) lie near the dotted line of equal rates. GRPO moves models to the left, toward fewer misses, far more than down: the best run reaches 1.9% misses but keeps 26.1% false flips, the level of Gemini 2.5 Pro (26.9%). †Step 500.

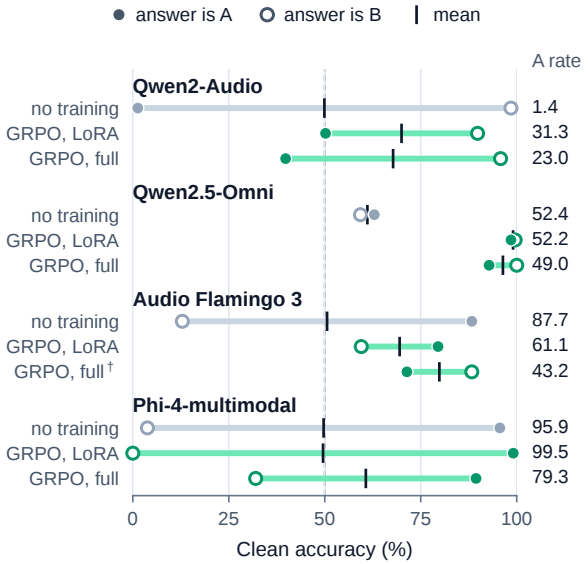


Figure 5: Letter bias on the clean test. Filled dot: accuracy on items whose correct answer is 'A'. Open dot: accuracy on items whose answer is 'B'. Tick: their mean, the class-balanced accuracy. Grey rows: no training; green rows: after GRPO. Right column: share of 'A' answers (52.8% of correct answers are 'A'). †Step 500. Without training, three of the four models sit near the two ends of the scale, with a mean near 50%. GRPO moves the dots closer together for Qwen2-Audio, Audio Flamingo 3 and Phi-4-multimodal with full fine-tuning, and lifts both dots of Qwen2.5-Omni above 92%. Phi-4-multimodal with LoRA moves further toward always answering 'A'.

Table 12: Per-task miss rate and false-flip rate (%) without training. Commercial rows: the two models analyzed item by item in Appendix I.

	MR				FFR			
Model	Mel.	Har.	Tim.	Rhy.	Mel.	Har.	Tim.	Rhy.
Qwen2-Audio	46.0	50.4	46.0	48.4	54.0	49.6	52.8	47.2
Qwen2.5-Omni	38.0	36.0	24.8	48.0	42.0	45.6	33.6	42.8
Audio Flamingo 3	50.4	49.6	54.4	44.0	46.4	52.4	41.2	50.4
Phi-4-multimodal	54.0	51.6	56.8	51.6	46.8	50.4	43.2	52.8
Kimi-Audio	53.2	49.2	56.0	50.8	41.2	52.4	42.8	46.4
MiMo-Audio	54.0	47.2	54.0	42.4	46.0	50.0	41.6	46.8
Fun-Audio-Chat	50.0	41.2	43.2	52.0	50.4	48.4	36.4	49.2
Step-Audio 2	55.2	49.2	42.0	50.0	49.2	46.4	38.0	48.4
GPT-Audio-mini	50.4	47.6	52.8	51.6	53.6	51.2	43.6	52.8
Gemini 2.5 Pro	23.2	35.2	19.2	21.2	30.4	37.2	16.0	24.0

Table 13: Per-task miss rate and false-flip rate (%) after GRPO. †Step 500.

		MR				FFR			
Model	GRPO	Mel.	Har.	Tim.	Rhy.	Mel.	Har.	Tim.	Rhy.
Qwen2-Audio	LoRA	26.0	40.4	4.4	48.0	45.2	46.8	34.0	44.8
	full	29.6	41.6	6.8	48.8	46.8	45.2	40.8	48.8
Qwen2.5-Omni	LoRA	0.4	4.0	3.2	0.0	32.4	34.4	20.8	16.8
	full	0.0	15.2	1.2	0.4	34.0	41.2	22.8	17.6
Audio Flamingo 3	LoRA	40.4	24.4	2.4	51.6	50.8	41.2	20.4	53.2
	full†	34.4	20.0	2.4	33.2	48.4	38.8	20.4	39.2
Phi-4-multimodal	LoRA	55.6	49.6	56.8	51.6	47.2	50.4	43.2	52.8
	full	48.8	45.6	25.6	51.6	48.4	46.0	10.0	52.8

Table 14: Letter bias on the clean test (%). A rate: share of items answered 'A' (52.8% of correct answers are 'A'). Acc|A and Acc|B: accuracy on items whose correct answer is 'A' or 'B'. Balanced: their mean. Gap: their difference. The last two rows are commercial models (Appendix I).

Model	GRPO	A rate	Acc|A	Acc|B	Balanced	Gap
Qwen2-Audio	none	1.4	1.3	98.5	49.9	97.2
	LoRA	31.3	50.2	89.8	70.0	39.6
	full	23.0	39.8	95.8	67.8	56.0
Qwen2.5-Omni	none	52.4	62.9	59.3	61.1	3.6
	LoRA	52.2	98.5	99.6	99.0	1.1
	full	49.0	92.8	100.0	96.4	7.2
Audio Flamingo 3	none	87.7	88.3	12.9	50.6	75.3
	LoRA	61.1	79.5	59.5	69.5	20.0
	full	43.2	71.4	88.3	79.8	17.0
Phi-4-multimodal	none	95.9	95.6	3.8	49.7	91.8
	LoRA	99.5	99.1	0.0	49.5	99.1
	full	79.3	89.4	32.0	60.7	57.4
GPT-Audio-mini	none	81.1	80.5	18.2	49.4	62.3
Gemini 2.5 Pro	none	60.8	87.9	69.5	78.7	18.4

Table 15: Per-task FLIP/STAY accuracy (%) without training. Models: 500 items per task and 2,000 in All; on one task, a score between 45.6% and 54.4% is not significantly different from 50%. Human listeners: 48 answers per task and 192 in All.

Model	Mel.	Har.	Tim.	Rhy.	All
Qwen2-Audio	50.0	50.0	50.6	52.2	50.7
Qwen2.5-Omni	60.0	59.2	70.8	54.6	61.2
Audio Flamingo 3	51.6	49.0	52.2	52.8	51.4
Phi-4-multimodal	49.6	49.0	50.0	47.8	49.1
Kimi-Audio	52.8	49.2	50.6	51.4	51.0
MiMo-Audio	50.0	51.4	52.2	55.4	52.2
Fun-Audio-Chat	49.8	55.2	60.2	49.4	53.6
Step-Audio 2	47.8	52.2	60.0	50.8	52.7
GPT-Audio-mini	48.0	50.6	51.8	47.8	49.6
GPT-Audio-1.5	50.4	51.4	56.0	50.4	52.1
Gemini 2.5 Pro	73.2	63.8	82.4	77.4	74.2
Gemini 3.8 Flash	57.2	51.2	84.2	63.2	64.0
Human (12)	85.4	66.7	97.9	97.9	87.0

Table 16: Drop in miss rate and false-flip rate from no training to GRPO, in points. z and p (two-sided) test whether the miss rate fell more than the false-flip rate. The test treats the rates before and after training as independent samples of 1,000 items each, which overestimates the variance when they are positively correlated. †Step 500.

Model	GRPO	ΔMR	ΔFFR	Ratio	z	p
Qwen2-Audio	LoRA	18.0	8.2	2.2	3.2	0.0015
	full	16.0	5.5	2.9	3.4	<0.001
Qwen2.5-Omni	LoRA	34.8	14.9	2.3	7.6	<0.001
	full	32.5	12.1	2.7	7.6	<0.001
Audio Flamingo 3	LoRA	19.9	6.2	3.2	4.4	<0.001
	full†	27.1	10.9	2.5	5.4	<0.001
Phi-4-multimodal	LoRA	0.1	−0.1	n/a	0.1	0.95
	full	10.6	9.0	1.2	0.5	0.61

G Captioning details

Table 17: MusicCaps captioning before and after GRPO on 2,735 clips (scores $\times 100$). MET.: METEOR; R-L: ROUGE-L; BERTS.: BERTScore-F1; Words: mean caption length.

Model	GRPO	BLEU-4	MET.	R-L	BERTS.	Words
Qwen2-Audio	none	13.06	32.09	33.43	39.27	40.7
	LoRA	13.38	32.42	33.74	39.73	40.8
	full	14.79	34.22	35.28	41.22	41.3
Qwen2.5-Omni	none	1.31	16.33	20.96	23.35	29.2
	LoRA	1.09	15.65	20.25	22.07	28.4
	full	1.60	16.69	21.23	23.50	30.0

We use the MusicCaps rows with `is_audioset_eval` set (2,858 clips). Of these, 2,735 could be downloaded and were captioned by every model version, so all comparisons use the same clips. Clips are 10 s, 16 kHz mono. Decoding is greedy, with at most 256 new tokens and repetition penalty 1.0. The prompt is "Describe this music in a few sentences. Mention the genre, the instruments, the mood, and the tempo." Each clip has one reference caption. No model version produced degenerate captions.

Table 18: Paired difference from the base model on the 2,735 MusicCaps clips (Δ = trained − base, in points $\times 100$); brackets give 95% bootstrap confidence intervals over clips (10,000 resamples). Stars denote Wilcoxon signed-rank tests, Holm-corrected across metrics within each model ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$). BLEU-4† is corpus-level, so no per-clip Wilcoxon test applies. BERTScore-F1 uses `roberta-large` (`lang=en`, default layer 17, `rescale_with_baseline=True`). Checkpoints: GRPO step 1,000.

Model	Variant	BLEU-4†	sBLEU	METEOR	ROUGE-L	BERTScore-F1
Qwen2-Audio	LoRA	$+0.3\,[+0.2, +0.5]$	$+0.3\,[+0.2, +0.5]^{**}$	$+0.3\,[+0.1, +0.5]^{*}$	$+0.3\,[+0.1, +0.5]^{***}$	$+0.5\,[+0.3, +0.6]^{***}$
	full FT	$+1.7\,[+1.5, +2.0]$	$+1.8\,[+1.5, +2.1]^{***}$	$+2.1\,[+1.8, +2.5]^{***}$	$+1.8\,[+1.6, +2.1]^{***}$	$+2.0\,[+1.7, +2.3]^{***}$
Qwen2.5-Omni	LoRA	$-0.2\,[-0.3, -0.1]$	$-0.1\,[-0.2, -0.1]^{***}$	$-0.7\,[-0.8, -0.5]^{***}$	$-0.7\,[-0.8, -0.6]^{***}$	$-1.3\,[-1.5, -1.0]^{***}$
	full FT	$+0.3\,[+0.2, +0.4]$	$+0.1\,[+0.1, +0.2]^{***}$	$+0.4\,[+0.2, +0.5]^{***}$	$+0.3\,[+0.1, +0.4]^{***}$	$+0.2\,[-0.1, +0.4]$

H False flips by sound change

Each STAY pair replaces clip B with a version that differs in one way (Section 3.3): an EQ change (± 3 dB shelves), room echo (reverberation time 0.2, 0.5 or 1.0 s), background hiss (-50 dBFS) or, for melody and harmony only, a transposition of up to 200 cents. The 1,000 STAY pairs contain 294 EQ, 290 room-echo, 292 hiss and 124 transposition pairs. Table 19 gives the false-flip rate for each change, pooled over the four tasks, and Table 20 gives it per task. Both tables use the same answers as Tables 2 and 3. Figure 6 shows how GRPO changes the pooled rates.

Models near chance show no pattern. For the untrained open models other than Qwen2.5-Omni, and for both GPT-Audio models, every pooled rate lies between 42.8% and 54.0%, whatever the change. These models barely tell the clips apart, so the kind of change makes little difference.

Models that listen ignore EQ and room echo best. The three models that listen without training err most on hiss and transposition pairs (32.5% to 50.0%). Qwen2.5-Omni and Gemini 3.8 Flash err least on EQ pairs (33.7% and 29.9%), and Gemini 2.5 Pro on room-echo pairs (17.2%, against 22.4% on EQ). GRPO keeps this order and widens the gaps. In every run that learned, the EQ rate falls by 10.9 to 32.3 points, while the hiss rate changes by -6.2 to $+5.8$ points and the transposition rate by -4.0 to $+2.4$ points.

Hiss. After GRPO, the four Qwen runs err on 51.0% to 53.1% of hiss pairs, about as often as before training. Both Qwen2.5-Omni runs err on exactly 38 of the 63 melody and 30 of the 63 harmony hiss pairs. These are the numbers of hiss pairs whose correct answer is 'same' (the rest are 'different'),

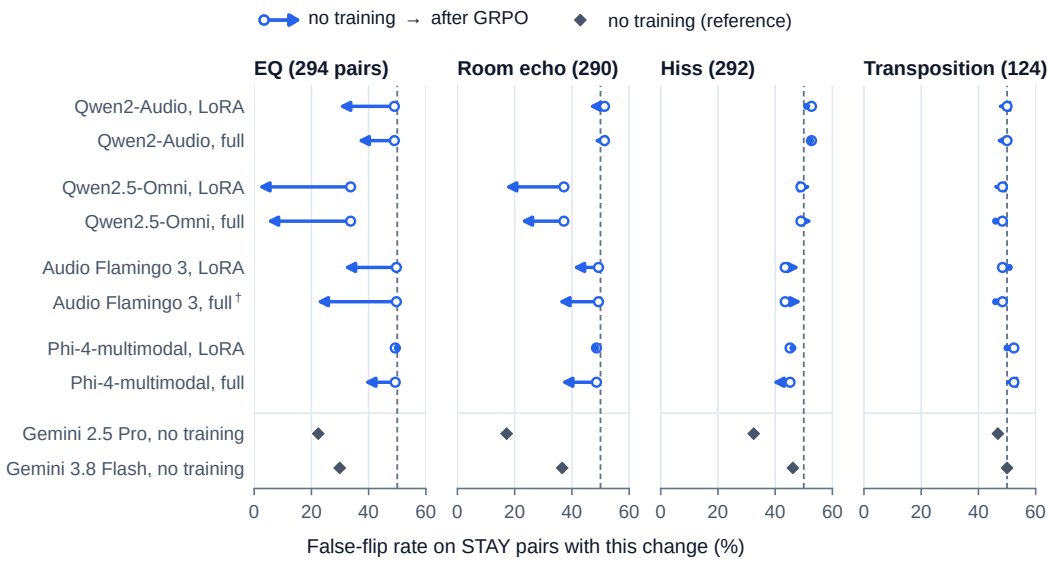


Figure 6: False-flip rate by kind of STAY change, before and after GRPO, one row per run. Each blue arrow starts at the rate without training (open circle) and ends at the rate after GRPO; short changes are drawn as a line ending in a dot. Diamonds: the two Gemini models, without training. Dashed line: 50%. GRPO lowers false flips under EQ in every run that learned (by 10.9 to 32.3 points), partly under room echo, and not under hiss or transposition. Gemini 2.5 Pro shows the same order without any training: room echo and EQ lowest, hiss higher, transposition near 50%. †Step 500.

Table 19: False-flip rate (%) for each kind of STAY change, pooled over tasks; lower is better. Room: room echo. The number of pairs is given in brackets. *All*: false-flip rate over all 1,000 STAY pairs, as in Tables 2 and 3. Transposition is used for melody and harmony only. †Step 500.

Model	GRPO	EQ (294)	Room (290)	Hiss (292)	Transp. (124)	All (1,000)
Qwen2-Audio	none	49.0	51.4	52.7	50.0	50.9
	LoRA	29.6	45.9	51.0	46.8	42.7
	full	36.1	47.6	52.4	46.0	45.4
Qwen2.5-Omni	none	33.7	37.2	49.0	48.4	41.0
	LoRA	1.4	16.6	52.4	45.2	26.1
	full	4.4	22.1	53.1	46.0	28.9
Audio Flamingo 3	none	49.7	49.3	43.5	48.4	47.6
	LoRA	31.3	40.3	48.6	50.8	41.4
	full†	21.8	35.2	49.3	46.0	36.7
Phi-4-multimodal	none	49.3	48.6	45.2	52.4	48.3
	LoRA	50.0	48.3	46.2	50.0	48.4
	full	38.4	36.2	39.0	49.2	39.3
Other open models, no training						
Kimi-Audio	none	45.6	47.9	42.8	47.6	45.7
MiMo-Audio	none	47.6	43.1	46.6	48.4	46.1
Fun-Audio-Chat	none	45.9	45.9	45.2	49.2	46.1
Step-Audio 2	none	43.5	46.9	44.5	49.2	45.5
Commercial models (API), no training						
GPT-Audio-mini	none	49.7	50.7	49.0	54.0	50.3
GPT-Audio-1.5	none	50.0	47.6	53.8	47.6	50.1
Gemini 2.5 Pro	none	22.4	17.2	32.5	46.8	26.9
Gemini 3.8 Flash	none	29.9	36.6	46.2	50.0	39.1

Table 20: False-flip rate (%) by task and STAY change; lower is better. Room: room echo; Tr.: transposition. Melody and harmony: 63 EQ, 62 room-echo, 63 hiss and 62 transposition pairs. Timbre and rhythm: 84 EQ, 83 room-echo and 83 hiss pairs, and no transposition. †Step 500.

		Melody				Harmony				Timbre			Rhythm		
Model	GRPO	EQ	Room	Hiss	Tr.	EQ	Room	Hiss	Tr.	EQ	Room	Hiss	EQ	Room	Hiss
Qwen2-Audio	none	49.2	58.1	60.3	48.4	47.6	51.6	47.6	51.6	52.4	53.0	53.0	46.4	44.6	50.6
	LoRA	20.6	56.5	60.3	43.5	38.1	51.6	47.6	50.0	8.3	37.3	56.6	51.2	42.2	41.0
	full	25.4	58.1	60.3	43.5	36.5	48.4	47.6	48.4	25.0	41.0	56.6	54.8	45.8	45.8
Qwen2.5-Omni	none	34.9	32.3	54.0	46.8	33.3	45.2	54.0	50.0	22.6	34.9	43.4	44.0	37.3	47.0
	LoRA	0.0	29.0	60.3	40.3	4.8	35.5	47.6	50.0	1.2	8.4	53.0	0.0	1.2	49.4
	full	0.0	35.5	60.3	40.3	19.0	46.8	47.6	51.6	1.2	10.8	56.6	0.0	4.8	48.2
Audio Flamingo 3	none	49.2	51.6	34.9	50.0	57.1	51.6	54.0	46.8	41.7	42.2	39.8	52.4	53.0	45.8
	LoRA	41.3	56.5	60.3	45.2	28.6	33.9	46.0	56.5	3.6	18.1	39.8	53.6	55.4	50.6
	full†	38.1	51.6	60.3	43.5	17.5	43.5	46.0	48.4	2.4	12.0	47.0	32.1	39.8	45.8
Phi-4-multimodal	none	50.8	43.5	41.3	51.6	50.8	51.6	46.0	53.2	42.9	43.4	43.4	53.6	55.4	49.4
	LoRA	52.4	45.2	39.7	51.6	52.4	48.4	52.4	48.4	42.9	43.4	43.4	53.6	55.4	49.4
	full	46.0	43.5	54.0	50.0	50.8	41.9	42.9	48.4	8.3	7.2	14.5	53.6	55.4	49.4
Other open models, no training															
Kimi-Audio	none	41.3	43.5	39.7	40.3	54.0	50.0	50.8	54.8	42.9	47.0	38.6	45.2	50.6	43.4
MiMo-Audio	none	50.8	41.9	39.7	51.6	52.4	53.2	49.2	45.2	39.3	42.2	43.4	50.0	37.3	53.0
Fun-Audio-Chat	none	52.4	50.0	47.6	51.6	49.2	46.8	50.8	46.8	35.7	34.9	38.6	48.8	53.0	45.8
Step-Audio 2	none	55.6	45.2	47.6	48.4	42.9	43.5	49.2	50.0	35.7	41.0	37.3	42.9	56.6	45.8
Commercial models (API), no training															
GPT-Audio-mini	none	55.6	51.6	49.2	58.1	49.2	50.0	55.6	50.0	41.7	45.8	43.4	53.6	55.4	49.4
GPT-Audio-1.5	none	49.2	56.5	60.3	46.8	44.4	45.2	47.6	48.4	50.0	45.8	55.4	54.8	44.6	51.8
Gemini 2.5 Pro	none	22.2	17.7	34.9	46.8	42.9	25.8	33.3	46.8	13.1	14.5	20.5	16.7	13.3	42.2
Gemini 3.8 Flash	none	34.9	51.6	52.4	50.0	50.8	50.0	49.2	50.0	10.7	9.6	34.9	29.8	42.2	50.6

so the counts are what a model gives if it answers 'different' to every hiss pair of these two tasks, whatever the music. On timbre and rhythm, hiss causes 85 of the 94 false flips of Qwen2.5-Omni with LoRA and 87 of the 101 with full fine-tuning. On the EQ and room-echo pairs of these two tasks, the two runs err on 2.7% and 4.2%. Phi-4-multimodal with full fine-tuning is the exception. On timbre, the only task it learned, it errs on 14.5% of hiss pairs, against 8.3% for EQ and 7.2% for room echo. It is also the one run in which misses and false flips fell by similar amounts (Section 5.3).

Transposition. No model learns to ignore transposition: with or without training, every run errs on 45.2% to 54.0% of transposition pairs, including Gemini 2.5 Pro (46.8%), which errs on only 17.2% of room-echo pairs. After GRPO, Qwen2.5-Omni detects a single moved note in 99.6% (LoRA) and 100.0% (full) of melody FLIP pairs, but errs on 40.3% of transposed melody pairs in both runs.

Room echo depends on the task. After GRPO, Qwen2.5-Omni with LoRA errs on 29.0% and 35.5% of room-echo pairs in melody and harmony, but on 8.4% and 1.2% in timbre and rhythm.

Limits of this split. EQ and hiss are each applied at a single strength, so these tables cannot show at what hiss level answers start to flip. A sweep over the hiss level needs inference only and would answer this.

I Per-item analysis of Gemini 2.5 Pro and GPT-Audio-mini

For these two commercial models we analyze every answer, so we can split their errors by the correct answer and by the size of the difference (Table 21 and Figure 7). The items are the same 1,000 clean and 2,000 FLIP/STAY items as in Table 2.

GPT-Audio-mini follows its letter preference. It answers 'A' to 81.1% of clean items and to all 500 rhythm FLIP/STAY items, so its rhythm FLIP/STAY accuracy is 47.8%, the score of always answering 'A' (Section 3.3). On timbre it answers 'same' to 78% to 100% of every instrument pair, whether the

Table 21: Errors (%) on melody, harmony and timbre, split by the correct answer. *Same*: pairs whose answer is 'same', where an error is a 'different' answer. *Diff.*: pairs whose answer is 'different', where an error is a 'same' answer. n: number of pairs.

	n		Gemini 2.5 Pro		GPT-Audio-mini	
Items	Same	Diff.	Same	Diff.	Same	Diff.
Clean	401	349	3.5	41.0	25.7	75.4
FLIP	349	401	4.3	44.6	24.1	73.1
STAY, EQ	109	101	6.4	44.6	22.0	76.2
STAY, room echo	115	92	2.6	39.1	28.7	73.9
STAY, hiss	115	94	10.4	51.1	27.8	74.5
STAY, transposition	62	62	51.6	41.9	41.9	66.1

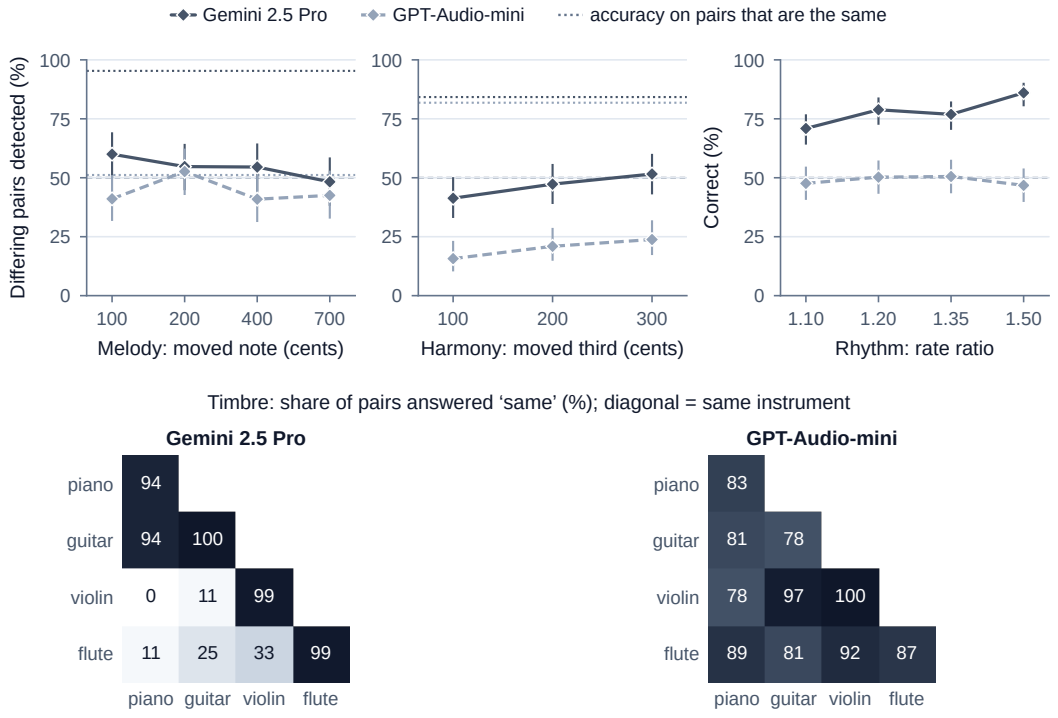


Figure 7: Gemini 2.5 Pro and GPT-Audio-mini, item by item, pooled over the clean, FLIP and STAY versions of each item. **Top:** share of differing melody and harmony pairs that the model calls different, by the size of the change (dotted: accuracy on pairs that are the same), and rhythm accuracy by the rate ratio of the two click tracks; bars are 95% Wilson intervals over 87 to 189 items per point, and the dashed line is 50%. **Bottom:** share of timbre pairs answered 'same' for each instrument pair; diagonal cells are same-instrument pairs, where 'same' is correct. Gemini 2.5 Pro detects about half of the moved notes whatever the size of the move, improves with the size of the tempo difference, and hears piano and guitar as the same instrument. GPT-Audio-mini answers 'same' to most timbre pairs and stays near 50% on melody and rhythm.

instruments differ or not. On melody and rhythm its accuracy does not change with the size of the difference.

Gemini 2.5 Pro leans toward 'same'. On the three same/different tasks, 91% of its clean errors and 83% of its FLIP/STAY errors are 'same' answers to pairs that differ. On pairs that are the same it is almost always right, except after a transposition (Table 21). Its detection of a moved note does not improve as the note moves further: it detects 60.0% of melody pairs with a 100-cent move and 48.3% with a 700-cent move. On harmony, detection rises from 41.3% to 51.6% as the third moves further. Only on rhythm, the one task that is not a same/different question, does accuracy clearly rise with the size of the difference, from 70.9% at a rate ratio of 1.10 to 86.0% at 1.50. On timbre it tells piano from guitar in only 2 of 36 pairs, although it tells piano from violin in all 36 and piano from flute in 32.

Sound changes rarely change its answer, except transposition and, on rhythm, hiss. Most of its false flips on the same/different tasks are 'same' answers to STAY pairs that differ (39.1% to 51.1% of such pairs), the same error it makes on clean and FLIP pairs (41.0% and 44.6%). On STAY pairs that are the same, EQ, room echo and hiss make it answer 'different' in only 6.4%, 2.6% and 10.4% of cases, but transposition does so in 51.6%. So on these tasks Gemini 2.5 Pro mostly ignores the three sound effects, yet it treats a transposed melody or chord as a different one. On rhythm, hiss lowers its accuracy to 57.8%, against 83.3% with EQ and 86.7% with room echo, and its errors go both ways (17 and 18 of the 83 hiss pairs). Its false-flip rate and that of the trained Qwen2.5-Omni are therefore similar in size (26.9% and 26.1%) but differ in kind: Gemini 2.5 Pro mostly misses differences that are there, while the trained model reacts to hiss.