

Beyond the Transcript: Measuring and Improving the Use of Acoustic Information in Audio Language Models

Linkai Peng¹

Baorian Nuchged²

Yuyang Yao³

¹University of Connecticut

²The University of Texas at Austin

³University of Oxford

Abstract

Audio language model benchmarks provide comprehensive evaluation across diverse tasks. However, some tasks can be answered from the transcript or even the question alone, and benchmark scores can therefore confound acoustic understanding with language modeling ability. We distinguish lexical tasks, for which the transcript usually suffices, from global-attribute and within-recording-structure tasks, which require the model to analyse the recording beyond its words, either as a whole, for properties such as the speaker and emotion, or within it, for patterns such as stress and intonation. We evaluated 19 open models and Gemini 3.1 Pro on 92 tasks from 6 benchmarks and estimated the recording gain, the accuracy advantage of the original recording over references that retain its words, by comparing the recording with synthetic readings of its transcript and by adding it to the transcript. On attribute and structure tasks, the open models fall behind Gemini by 28.6 and 41.9 percentage points, and about half or more of each deficit lies in the recording gain. Nevertheless, linear probes in Gemma-4 12B and Qwen2.5-Omni-7B show that the acoustic information needed for answering is already readable in early language-model layers, while the readout of many cues declines in the layers that follow. We therefore constructed training data and trained LoRA adapters in these layers, with a content anchor, which raises attribute and structure accuracy by 12.8 to 44.4 points in the two models and keeps mean lexical accuracy within 1 point of the base model. Most of the improvement appears as larger recording gains. These results suggest that training improves the use of acoustic information that the models already encode.

1 Introduction

Audio language models extend the language understanding and reasoning capabilities of large language models to speech and other audio inputs

(Xu et al., 2025a). Comprehensive benchmarks assess these capabilities through tasks ranging from spoken-content understanding to speaker, emotion and prosody recognition (Wang et al., 2025a; Yang et al., 2024; Sakshi et al., 2025; Wang et al., 2026). However, even with speech as input, a model may rely only on its lexical content, or even the question and options alone, to answer some questions. For example, the wording of *I am so happy* may suffice to answer an emotion question without analysing the tone of the audio. In this situation, a cascaded system (automatic speech recognition followed by a text-based language model) may perform as well as an audio language model. Benchmark scores can therefore confound language modeling ability with acoustic understanding. How much additional benefit does an audio language model obtain from the original recording beyond these recognized words?

The answer is likely to differ across tasks. For questions that the words alone cannot answer, the model must extract the required cues from the recording itself, and these cues vary with the question. We therefore distinguish three task classes by the evidence needed to answer them (top of Figure 1). *Lexical* tasks can usually be answered using the ASR transcript alone. *Global-attribute* tasks assess properties of the recording as a whole, such as speaker identity and emotion, while *within-recording structure* tasks assess prosodic patterns and relations between segments, such as stress and intonation. These questions require locating relevant parts of the recording and reasoning how they were spoken, for example which word was emphasized or which of two segments had a higher pitch. Because an ASR transcript usually leaves these acoustic patterns unspecified, the recording has most to add on attribute and structure tasks and least on lexical tasks.

Measuring whether models use this evidence requires a reference that retains the words. Noise, silence and audio removal disrupt the words along

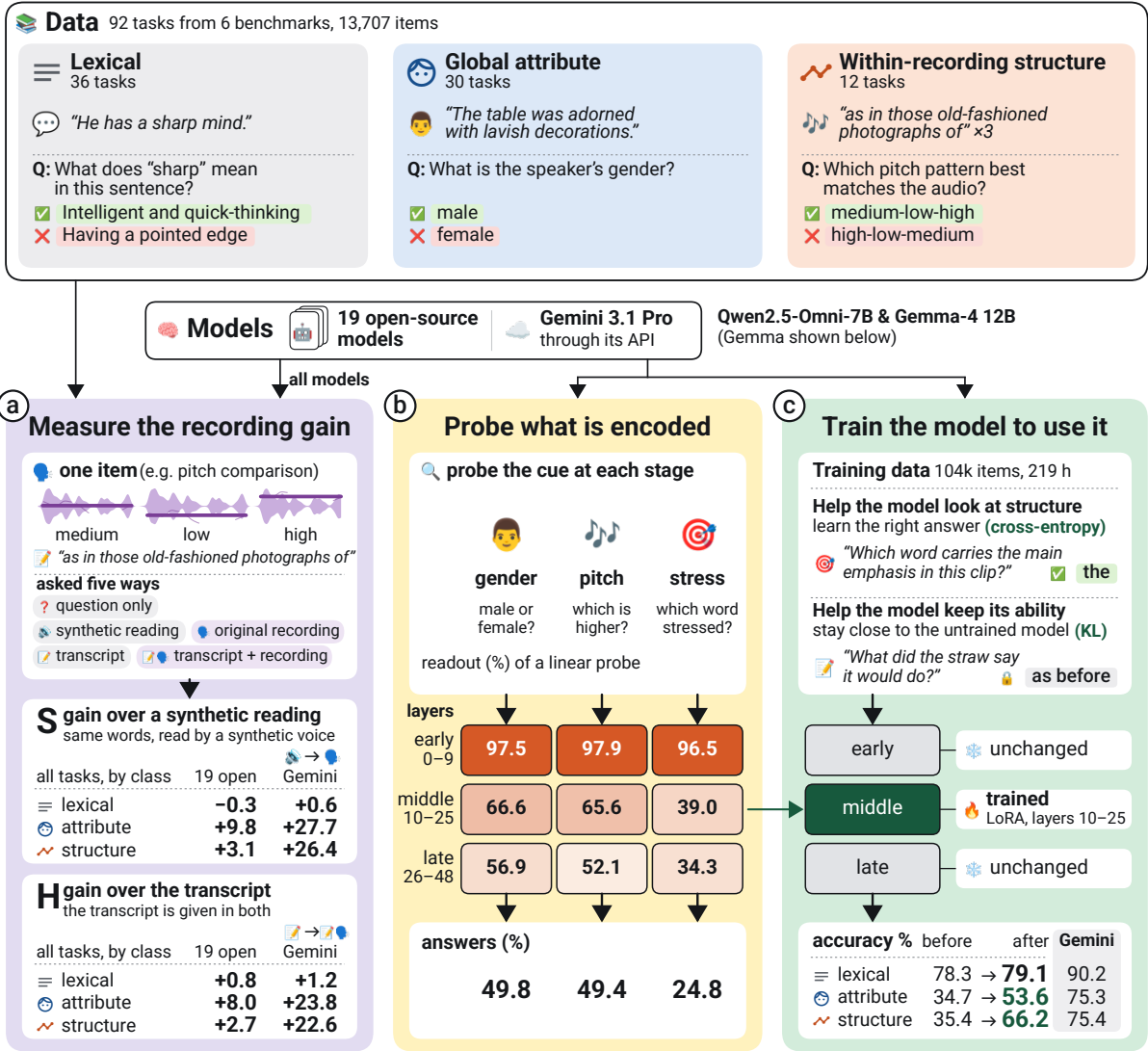


Figure 1: Overview of the study, from the data and models (top) to the recording gain (a), the readout (b) and the repair (c). The tasks are grouped by where the evidence for the answer lies. Each class is shown with one example item, and the 14 mixed tasks are not shown. (a) One item asked with five inputs, and the class means of S and H in percentage points (pp) for the mean of the 19 open models and for Gemini 3.1 Pro. (b) Readout of three acoustic cues in the early, middle and late layers of Gemma-4 12B, and answer accuracy on the same items (chance 50%, 50% and 25%). (c) LoRA training of the layers where readout declines, with cross-entropy on acoustic questions and a KL term, and accuracy by class before and after training, with Gemini shown for reference.

with other acoustic cues (Sakshi et al., 2025; Wang et al., 2026; He et al., 2026; Foo et al., 2026), and tests that place the words in conflict with the voice require newly constructed stimuli (Chen et al., 2026; Xiong et al., 2026; Pang et al., 2026). We estimate the contribution of the original recording beyond the transcript using two comparisons on existing benchmark items (Figure 1a and Section 2). We call the corresponding accuracy differences the *recording gains*. S is the recording’s accuracy advantage over synthetic readings of its transcript, which retain the transcribed words, replace the

voice and prosody and remove background sound. H is the accuracy gained by adding the recording to its transcript. The comparisons involve different changes to the input, and we examine the patterns they share. We evaluate 19 open models and, as a closed-model reference, Gemini 3.1 Pro on 13,707 items from 92 tasks in 6 benchmarks, then compare linear-probe *readout*, with answer accuracy in two open models and test a repair. In these two models examined internally, the measurements show that acoustic information can be present in the representations but underused in answers. Three findings

give the evidence (Figure 1).

- **Recording gain.** The open models fall behind Gemini most on attribute and structure tasks, by 28.6 and 41.9 pp on average, and about half or more of this deficit lies in the recording gain (Section 3). The open models thus benefit less from the information in the recording than Gemini, whose S/H exceed theirs by 17.9/15.8 and 23.3/19.9 pp.
- **Readout.** In the encoder-free Gemma-4 12B (Gemma Team, 2026) and the encoder-based Qwen2.5-Omni-7B (Xu et al., 2025a), the acoustic information needed for many attribute and structure tasks is readable from early layers, while answer accuracy stays well below this readout (Section 4). We call this difference the *utilization gap*.
- **Repair.** We construct new training data and train low-rank adapters (LoRA, Hu et al., 2022) in the layers where readout declines, with a content anchor that keeps answers to content questions close to the base model’s (Section 5). Training improves attribute and structure accuracy by 12.8 to 44.4 pp in the two models, while mean lexical accuracy remains within 1 pp of the base model. On structure tasks, S/H rise from 2.5/5.2 and 1.8/2.7 to 29.6/29.5 and 38.9/46.0 pp. Post-training analyses of readout and of the adapters are consistent with better use of acoustic information already encoded in the representations.

We release the data-construction scripts, the code and the trained LoRA adapters.

2 Measuring the Recording Gain

We classify benchmark tasks by the evidence needed to answer them, based on each task’s description and answer format. Our evaluation covers 92 tasks and 13,707 items from AudioBench (Wang et al., 2025a), AIR-Bench (Yang et al., 2024), Dynamic-SUPERB Phase-2 (Huang et al., 2025), MMAU (Sakshi et al., 2025), MMAU-Pro (Kumar et al., 2026) and MMSU (Wang et al., 2026) (Appendix A). The transcript usually suffices for the 36 **lexical** tasks. The 30 **global-attribute** tasks need evidence about the recording as a whole, such as the speaker, emotion, scene or number of speakers, and the 12 **within-recording structure** tasks need evidence about where something occurs or how

segments relate, such as which word is stressed or which of two segments is higher in pitch. The remaining 14 **mixed** tasks combine lexical and acoustic judgements or require additional reasoning. The class results in Sections 3 to 5 cover the 78 tasks and 12,102 items of the first three classes, and the mixed tasks appear only in the per-task results of Appendix G. Appendix B lists the tasks in each class.

Table 1: The five input conditions.

Input	Acc.	Form	
		Text	Speech
Question and options only	N		
 Whisper transcript	A	✓	
 Synthetic reading of it	T		✓
 Original recording	O		✓
 Transcript and recording	AO	✓	✓

To estimate the contribution of the original recording beyond the transcript, we pose each item under five inputs with the same question and options (Table 1). N is accuracy with only the question and options. A is accuracy with the transcript of the recording produced by Whisper (Radford et al., 2023). T is accuracy with a synthetic reading of the transcript, O with the original recording and AO with the transcript and recording together. The two recording gains are $S = O - T$, the recording’s advantage over synthetic readings of the same transcript, and $H = AO - A$, the accuracy gained by adding the recording to its transcript. The text-to-speech system (rany2, 2024) reads the transcript once in a female and once in a male voice, and T is the mean over these two readings. The readings retain the transcribed words while replacing the original voice and prosody and removing background sound. A model answering a binary gender question from the synthetic voice alone, for example, averages chance accuracy over the two readings. The retained readings have a word error rate of 1.4% relative to their source transcripts. Ideally, the two inputs of each comparison differ only in the sound beyond the words. Each comparison, however, adds one further difference. S uses synthetic speech as its reference and therefore also reflects differences between natural and synthetic speech. H gives the model the transcript in both conditions and therefore also reflects how the model combines text with audio. Neither of these differences enters the other measure, and neither alone can cre-



Figure 2: Recording gains of the open models and Gemini 3.1 Pro on the three task classes. Each row summarizes a group of tasks within a class. On the left, the 19 columns represent the open models, ordered from lowest to highest overall accuracy with the original recording, and each cell shows that model’s mean of S and H . The right side reports accuracy with the original recording, O (%), and the two gains (pp), S and H , for the mean of the 19 open models and for Gemini 3.1 Pro. Purple marks positive gains and brown negative gains. Mixed tasks are not shown.

ate a difference that both S and H show. Such shared differences measure the recording gain relative to transcript-based references, and our conclusions rest on them (see Limitations). We evaluated 19 open end-to-end audio language models on all items, with Gemini 3.1 Pro as a closed-model reference (Appendix C). Accuracy is averaged over items within each task and then equally over tasks within each class, and open-model means weight the 19 models equally.

3 Recording Gain by Model and Class

Figure 2 reports accuracy with the original recording and the two gains for the 19 open models and Gemini. The open models trail Gemini in all three classes, by 13.3, 28.6 and 41.9 pp on lexical, attribute and structure tasks on average, and even the strongest open model, Qwen3-Omni, reaches only 59.9% and 49.8% on attribute and structure tasks.

On lexical tasks, gains under both comparisons are small. Mean S and H range from -0.3 to $+1.2$ pp for the open-model mean and Gemini. This

pattern is expected, because lexical tasks depend little on information in the recording beyond the words. Gemini’s higher lexical accuracy therefore reflects the strength of its language model.

On attribute and structure tasks, about half or more of this deficit is a difference in recording gain. With the original recording, Gemini reaches 75.3% and 75.4%, against 46.7% and 33.5% for the open-model mean, gaps of 28.6 and 41.9 pp. S differs by $27.7 - 9.8 = 17.9$ and $26.4 - 3.1 = 23.3$ pp, and H by $23.8 - 8.0 = 15.8$ and $22.6 - 2.7 = 19.9$ pp, in each case about half or more of the gap. Qwen3-Omni, the strongest open model, has S/H of 18.4/14.7 and 12.0/10.1 pp, also below Gemini. These small recording gains lead us to ask whether the acoustic information needed for the answers is absent from the models or present but unused.

4 Layer-wise Readout

To measure whether a model’s representations contain the acoustic information needed for answering, we train a linear probe on the representations of

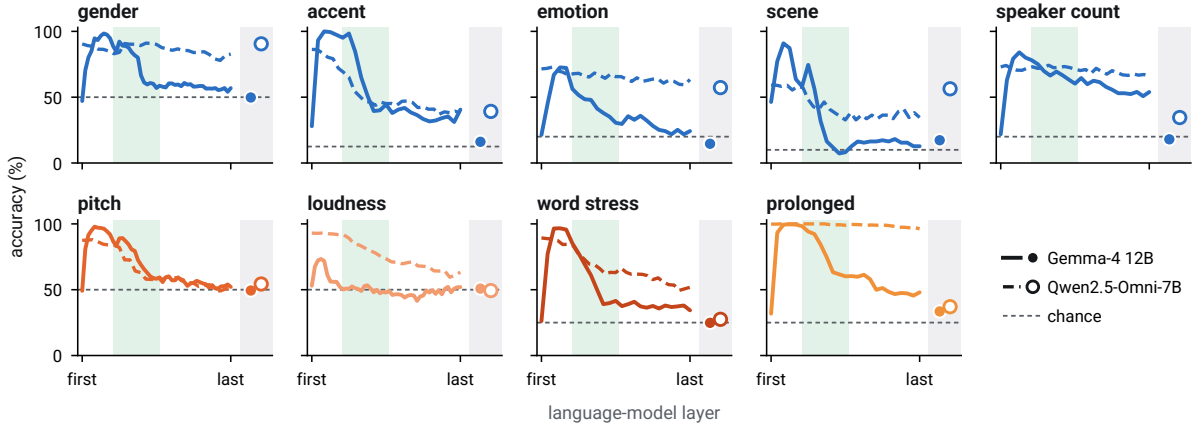


Figure 3: Layer-wise linear-probe readout and answer accuracy in Gemma-4 12B and Qwen2.5-Omni-7B. The dots show answer accuracy with the original recording. The green band marks the readout-decline layers of Gemma-4 12B (layers 10–25). Blue panels show five attribute tasks and orange panels four structure tasks.

each layer of the model and call the probe’s accuracy the *readout*. For attribute tasks, the probe is trained and tested on the evaluation items with the task labels as targets. For structure tasks, the probe needs to know where the compared parts lie, and we therefore constructed recordings, for example by joining two utterances of the same speaker that differ in pitch, speaking rate or loudness. The probe receives the segment boundaries, averages the representations within each segment and uses their difference to judge which has the larger value (Appendix D). The models also answer these items directly, which gives answer accuracy. If a probe reads the answer from the representations while the model answers incorrectly, the information is present but not reliably used in the answer.

We examine Qwen2.5-Omni-7B, which has a separate audio encoder, and Gemma-4 12B, which projects waveform frames directly into the language model, to study the utilization gap with and without an encoder. In addition, Qwen2.5-Omni-7B is among the strongest open models, reaching 56.7% and 34.0% with the original recording on attribute and structure tasks, above the open-model means of 46.7% and 33.5%, and ranking second of the 19 by overall accuracy (Figure 2). Gemma-4 12B ranks 13th, reaching 34.7% and 35.4%.

Figure 3 compares layer-wise readout with answer accuracy (all probed tasks in Appendix E). In both models, peak readout exceeds answer accuracy on every probed task, and for most tasks the peak lies in the early layers. For pitch, for example, readout reaches 97.9% at layer 4 of Gemma-4 12B and 88.0% at layer 2 of Qwen2.5-Omni-7B, while answer accuracy is 49.4% and 54.3% (chance 50%).

Poor answers occur both when later-layer readout declines and when it stays high. In Gemma-4 12B, readouts of gender, accent, emotion, scene, pitch and the stressed word all fall across layers 10–25. In Qwen2.5-Omni-7B, readouts of accent, scene, pitch, loudness and the stressed word fall across layers 2–11. Readout of which word is prolonged, by contrast, is still 96.2% at the last layer, while answer accuracy is 37.1% (chance 25%). From the ranges over which several readouts fall, we chose layers 10–25 in Gemma-4 12B and 2–11 in Qwen2.5-Omni-7B, the *readout-decline layers*.¹ Section 5 tests whether training there improves answers while preserving mean lexical accuracy, and whether the improvement comes from using information that is already readable.

5 Repairing the Utilization Gap

5.1 Training in the Readout-Dcline Layers

We trained LoRA adapters in the readout-decline layers, layers 10–25 in Gemma-4 12B and 2–11 in Qwen2.5-Omni-7B, on 103,878 supervised items with 219.0 hours of audio covering the question formats of 34 attribute and structure tasks. The recordings come from 17 corpora, about 80% of the audio from the LibriSpeech training splits (Panayotov et al., 2015) and the rest from Expresso (Nguyen et al., 2023), the Speech Accent Archive (Weinberger, 2015), and acted-emotion and sound-event corpora. The training set shares no recordings with the six benchmarks and contains no text specific to

¹Figure 1 labels the layers of Gemma-4 12B as early, middle and late, and its middle layers are the readout-decline layers of this model.

Table 2: Accuracy and the two recording gains by task class for the base and repaired Gemma-4 12B and Qwen2.5-Omni-7B, with the mean of the 19 open models and Gemini 3.1 Pro. O is accuracy (%) with the original recording, and S , H and changes are in pp.

	Lexical	Global attribute			Within-recording structure		
	O	O	S	H	O	S	H
Gemma-4 12B							
Base	78.3	34.7	−0.4	+5.2	35.4	+2.5	+5.2
Repaired	79.1	53.6	+12.6	+15.7	66.2	+29.6	+29.5
Change	+0.8	+18.9	+13.0	+10.5	+30.8	+27.0	+24.3
Qwen2.5-Omni-7B							
Base	83.4	56.7	+18.3	+14.9	34.0	+1.8	+2.7
Repaired	83.7	69.5	+30.3	+28.9	78.4	+38.9	+46.0
Change	+0.3	+12.8	+12.0	+14.1	+44.4	+37.2	+43.3
Mean of 19 open models	76.9	46.7	+9.8	+8.0	33.5	+3.1	+2.7
Gemini 3.1 Pro	90.2	75.3	+27.7	+23.8	75.4	+26.4	+22.6

their items (Appendix F).

To preserve the base model’s accuracy on content questions, we add a content anchor of 16,000 items. It consists of 10,000 LibriSQA spoken-language understanding items (Zhao et al., 2025) and 6,000 questions about word categories, emotion and sarcasm built from LibriSpeech training utterances and synthetic training dialogues. On anchor items, we minimize the Kullback–Leibler (KL) divergence between the adapted and base models’ vocabulary distributions at answer positions, computing the base distribution with the adapter switched off. Supervised items use cross-entropy. The combined objective is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \text{KL}(p_{\text{base}} \parallel p_{\theta}),$$

with $\lambda = 16$ in both models. We call the results the *repaired* Gemma-4 12B and Qwen2.5-Omni-7B (settings in Appendix F).

5.2 Accuracy and Recording Gain after Training

Table 2 gives class means before and after training (per-task results and a task-group view in Appendix G). Attribute accuracy rises from 34.7% to 53.6% in Gemma-4 12B and from 56.7% to 69.5% in Qwen2.5-Omni-7B, and structure accuracy from 35.4% to 66.2% and from 34.0% to 78.4%, gains of 18.9 and 12.8 pp on attribute tasks and 30.8 and 44.4 pp on structure tasks. Mean lexical accuracy stays within 1 pp of the base model, and every lexical task group within 5 pp. Both repaired models exceed Qwen3-Omni, the strongest open model, on structure tasks (49.8%), and Qwen2.5-Omni-7B also on attribute tasks (59.9%). Against

Gemini, Qwen2.5-Omni-7B comes within 5.8 pp on attribute tasks (69.5% vs. 75.3%) and exceeds it on structure tasks (78.4% vs. 75.4%), whereas Gemma-4 12B, which starts from a lower rank and without a separate encoder, remains 21.7 and 9.2 pp behind. In both models, therefore, training in the readout-decline layers raises attribute and structure accuracy substantially.

The recording gains rise with accuracy. On attribute tasks, S/H rise from −0.4/5.2 to 12.6/15.7 pp in Gemma-4 12B and from 18.3/14.9 to 30.3/28.9 pp in Qwen2.5-Omni-7B, and on structure tasks from 2.5/5.2 to 29.6/29.5 pp and from 1.8/2.7 to 38.9/46.0 pp. In each class and model, the rise in S accounts for 69% to 94% of the rise in O . Under both comparisons, therefore, the repaired models obtain more from the recording beyond its transcript, and this increase makes up most of their improvement.

5.3 Where the Gains Come From

Training may improve answers by preserving more acoustic information in later layers or by using information already readable there. To distinguish these changes, we repeated the probe measurements of Section 4 on the repaired models (Figure 6 in Appendix E). Across the 11 probed tasks, answer accuracy rises by 27.7 and 27.2 pp on average in Gemma-4 12B and Qwen2.5-Omni-7B, whereas final-layer readout rises by only 6.7 and 11.0 pp. Before training, peak readout exceeded answer accuracy on every probed task. Training closes about half of this difference in Gemma-4 12B and most of it in Qwen2.5-Omni-7B. Because the probes are trained with supervision, this comparison can-

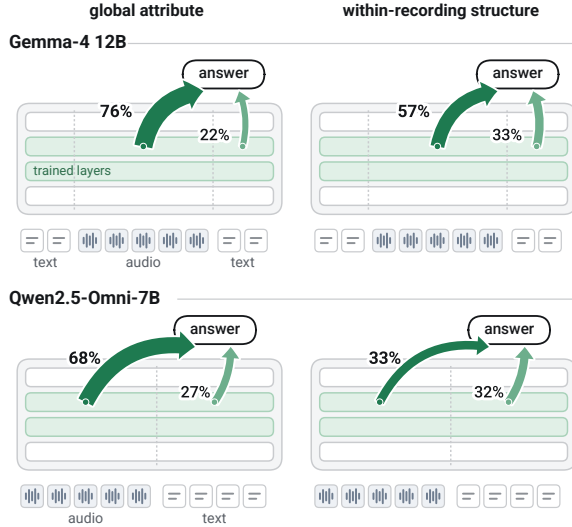


Figure 4: Accuracy gain by adapter position. Arrow width shows the share of the full adapter’s gain recovered when it is active only at audio or text positions.

not determine how much of the improvement each change explains. The larger rise in answer accuracy than in readout is nevertheless consistent with training improving the use of acoustic information already encoded in the representations. Changes in readout vary by task. For gender in Gemma-4 12B, last-layer readout rises from 56.6% to 72.2%, whereas for speaker count readout changes by less than 5 pp in either model while answer accuracy rises by more than 30 pp.

If training had mainly taught the models the question formats, the adapter would act at the text positions, where the question and answer tokens are processed. We therefore applied the adapter only at the audio positions or only at the text positions (question, template and answer tokens) (Figure 4). Applied only at the audio positions, the adapter recovers 76% and 57% of the full gain on attribute and structure tasks in Gemma-4 12B, and 68% and 33% in Qwen2.5-Omni-7B. Applied only at the text positions, it recovers 22% and 33%, and 27% and 32%. In three of the four cases, the audio positions alone recover more than half of the gain and the text positions alone 22% to 33%. The exception is the structure class of Qwen2.5-Omni-7B, where the audio positions alone and the text positions alone each recover about a third of the gain, and the remaining gain requires the adapter at both. The largest part of the improvement therefore comes from changing how the trained layers process the audio tokens, whose representations already carried the information (Section 4), rather than from

Table 3: Training variants of Gemma-4 12B and Qwen2.5-Omni-7B. ΔO and ΔS give changes from the base model in original-recording accuracy and S by task class (pp), and tinted rows mark the repaired models. We report only ΔS here, since S and H change in the same direction for both repaired models.

	Lexical	Attribute		Structure	
	ΔO	ΔO	ΔS	ΔO	ΔS
Gemma-4 12B					
Layers 10–25	+0.8	+18.9	+13.0	+30.8	+27.0
No content anchor	−10.5	+16.5	+15.1	+35.2	+29.4
Layers 0–15	+1.1	+20.3	+13.7	+16.3	+11.2
Layers 32–47	+0.2	+8.2	+3.9	+11.4	+7.6
All 48 layers	−2.5	+17.2	+14.7	+34.9	+26.8
Qwen2.5-Omni-7B					
Layers 2–11	+0.3	+12.8	+12.0	+44.4	+37.2
No content anchor	−5.6	+13.6	+14.3	+49.6	+42.1
Layers 12–21	−1.4	+9.9	+9.9	+30.1	+22.2
Layers 18–27	−1.0	+7.8	+7.3	+19.2	+13.3
All 28 layers	+0.7	+13.8	+13.9	+44.4	+34.7

how the question is read or the answer is written.

5.4 Ablations and Transfer

We next examine how the training choices affect these results. Each variant in Table 3 changes one choice relative to the repaired model. Without the content anchor, lexical accuracy falls in both models, whereas attribute and structure accuracy rise by similar or larger amounts and ΔS is no smaller. The anchor therefore preserves lexical accuracy while retaining 84–92% of the ΔS obtained without it. Placing the adapters in other ranges of equal width reduces the structure gain in both models, most for the later ranges, and ΔS decreases with it. Training all layers, with about three times as many trainable parameters, changes the attribute and structure gains by at most 4.1 pp and ΔS by at most 2.5 pp relative to the readout-decline layers. The gains therefore depend on where the adapter is placed rather than on how many parameters are trained. The later ranges, in which readout has already declined (Figure 3), recover less of the gain, and parameters outside the readout-decline layers contribute little.

Finally, to test transfer to formats absent from training, we trained a separate Gemma-4 12B model without gender or word-stress items. Accuracy on both task types still rises, and the increase in S is about 60% of that after full training. Other training tasks involve related cues and operations that may explain this transfer. Gender, speaker, and age questions share some voice cues, includ-

ing pitch. Word-stress questions require locating a word and comparing its prosody. Word-timestamp questions also require locating words, while pitch- and duration-comparison questions require comparing acoustic properties. Across the seven conversational emotion and sarcasm tasks without matching training formats, original-recording accuracy falls by 4.1 pp on average in the fully repaired Gemma-4 12B but rises by 4.3 pp in Qwen2.5-Omni-7B. These results suggest transfer between tasks that share cues and operations, but the gains on unseen formats are uneven.

6 Related Work

To estimate what a recording contributes beyond its transcript, we need a reference that retains the transcribed words. Adding noise, replacing recordings with silence or removing audio weakens the lexical evidence as well (Sakshi et al., 2025; Wang et al., 2026; He et al., 2026; Foo et al., 2026), and a text reference preserves the words but changes the input from speech to text (Cuervo et al., 2026). Other work places the voice in conflict with the spoken words or with added text to see which source the model follows (Chen et al., 2026; Xiong et al., 2026; Wang et al., 2025b; Pang et al., 2026; Billa, 2026). ALME (Billa, 2026), closest to our reference design, resynthesizes stimuli with the same lexical content. We pair natural recordings with synthetic readings of their transcripts, so that both inputs remain speech and the same readings serve to compare S across models and before and after training, and we complement this comparison with H .

We ask whether the information a model leaves unused is absent from it or present in its representations. Prior studies recover physical attributes, prosodic contrasts or audio-supported answers from hidden states while the models answer near chance or follow conflicting text (Sun et al., 2026; Wang et al., 2025b; Gao et al., 2026; Peng and Nuchged, 2026). VoxParadox (Pang et al., 2026), closest to our work, combines evaluation, probing and repair and finds in Audio Flamingo 3 (Ghosh et al., 2025) that speaker cues remain readable in the language model while comparison cues weaken in deep encoder layers and at the projection. These audio studies examine models with separate encoders. Although encoder-free designs are reported to improve paralinguistic accuracy (Fan et al., 2026), we find the same utilization gap in

encoder-free Gemma-4 12B and test whether training where readout declines improves answers.

To train the model to use this information, we also need to check that it retains its understanding of spoken content. Existing methods adjust output scores at decoding time with audio-aware or contrastive rules (Hsu et al., 2025; Lin et al., 2026; Gao et al., 2026), or mix encoder layers according to the prompt before direct preference optimization (Pang et al., 2026; Rafailov et al., 2023). Fine-tuning on conflicting audio and text can lower accuracy when the inputs agree or only audio is given (Wang et al., 2025b; Billa, 2026), and retention therefore needs its own evaluation (Lu et al., 2025). Established approaches constrain training with the original model’s outputs or train speech models on targets from the original or backbone model (Li and Hoiem, 2016; Held et al., 2025; Lu et al., 2026). We build on this idea by training LoRA adapters in the readout-decline layers with a content anchor that keeps answers to content questions close to the base model’s.

7 Conclusion

We estimated the contribution of the original recording beyond its words across 92 tasks from six benchmarks. On attribute and structure tasks, differences in recording gain account for about half or more of the accuracy gap between the mean of 19 open models and Gemini 3.1 Pro. In Gemma-4 12B and Qwen2.5-Omni-7B, linear probes read out relevant acoustic cues from early language-model layers even when the models answer the same items poorly. Training LoRA adapters on our constructed data in the layers where readout declines, with a content anchor, raises attribute and structure accuracy by 12.8 to 44.4 pp in the two models. Most of this increase appears as larger recording gains, while mean lexical accuracy stays within 1 pp of the base models. On the probed tasks, answer accuracy also rises more than readout, consistent with better use of information already encoded in the representations. The two probed models therefore encode acoustic information that they do not reliably use in answers, and training changes how they process it.

Limitations

Reference conditions. We use two recording gains to estimate the original recording’s accuracy advantage relative to different references. S uses

synthetic readings and thus includes differences between natural and synthetic speech. Properties of the synthetic voices may help or mislead the model. H uses the transcript as its reference, and providing it may change how the model combines text and audio. Estimating the gain in two ways yields a more robust estimate. The two gains show a consistent pattern, which supports differences in how much models use the recording. However, both comparisons depend on the same Whisper transcript, and recognition errors or omissions can therefore affect them together.

Number of repaired models. We tested the repair in only two models, one with and one without a separate audio encoder. They also differ in size and training data, so differences between them cannot be attributed to encoder design. Before training, for example, Qwen2.5-Omni-7B has higher attribute-class S/H than Gemma-4 12B (18.3/14.9 vs. -0.4/5.2 pp), and after training it closes a larger share of the gap between readout and answer accuracy. Testing further architectures is needed to establish how broadly the repair applies. We suspect that models with these two architectures differ in further notable ways, which this paper does not examine in depth.

Transfer and side effects. The training items match the question formats of 34 evaluation tasks, so transfer beyond these formats needs further testing. Across the seven conversational emotion and sarcasm tasks without matching training formats, mean accuracy falls by 4.1 pp in Gemma-4 12B but rises by 4.3 pp in Qwen2.5-Omni-7B, with declines on individual tasks in both models. In a separate Gemma-4 12B run without gender or word-stress training items, both judgements still improve. These mixed results call for more careful analysis of transfer.

Probe readout. Probes are trained with supervision on each task’s labels and receive the segment boundaries on structure items, whereas the models answer without either. The gap between readout and answer accuracy therefore also reflects these advantages and does not by itself show that the models leave available information unused. Likewise, the larger rise in answer accuracy than in readout after training does not quantify how much of the improvement comes from better use of existing information.

Where the repair acts. Readout decline guided layer selection, but answers can also be poor where readout remains high, so readout alone cannot establish the most effective training range. Training all layers uses about three times as many trainable parameters and changes ΔS by at most 2.5 pp relative to the selected layers. The layer controls compare only two other equal-width ranges per model, and a finer search could find larger gains. Enabling the adapters only at audio positions or only at text positions shows that, in this setting, changes at audio positions contribute to the gains. Identifying which part of the internal computation changes requires further intervention experiments.

Evaluation setting. Most items ask one question about one recording, and tasks involving multi-party dialogue form only a small part of the evaluation. Full-duplex speech models listen while speaking and must use cues such as intonation, pauses and emotion before the other speaker has finished to decide when and how to respond. Whether they use sound beyond the words in such continuous interaction, and whether synthetic readings remain a useful reference for measuring the recording’s contribution, require separate tests.

References

- Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, et al. 2025. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*.
- Adaee Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. 2018. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*.
- Jayadev Billa. 2026. When audio-LLMs don’t listen: A cross-linguistic study of modality arbitration. *arXiv preprint arXiv:2602.11488*.
- Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. 2026. Do audio LLMs really LISTEN, or just transcribe? Measuring lexical vs. acoustic emotion cues reliance. In *Conference of the European Chapter of the Association for Computational Linguistics*. ArXiv:2510.10444.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, et al. 2024. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*.
- Santiago Cuervo, Skyler Seto, Maureen de Seyssel, Richard He Bai, Zijin Gu, Tatiana Likhomanenko,

- Navdeep Jaitly, and Zakaria Aldeneh. 2026. Closing the gap between text and speech understanding in LLMs. In *International Conference on Learning Representations*. ArXiv:2510.13632.
- Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiye Sun, Yingjing Xu, et al. 2026. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*.
- Ruchao Fan, Yiming Wang, Yuxuan Hu, Bo Ren, Yufei Xia, Xiaofei Wang, Yao Qian, Shujie Liu, and Jinyu Li. 2026. LLM can read spectrogram: Encoder-free speech-language modeling. In *IEEE Spoken Language Technology Workshop*. ArXiv:2606.10231.
- Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852.
- Leonardo Haw-Yang Foo, Chih-Kai Yang, Chen-An Li, Ke-Han Lu, and Hung-yi Lee. 2026. All that glitters is not audio: Rethinking text priors and audio reliance in audio-language evaluation. *arXiv preprint arXiv:2604.24401*.
- Yichen Gao, Yiqun Zhang, Zijing Wang, Yujia Li, Heng Guo, Xi Wu, Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2026. Beyond text following: Repairable arbitration reversals in audio-language models. *arXiv preprint arXiv:2606.05161*.
- Gemma Team. 2026. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*.
- Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, et al. 2026. Audio Flamingo Next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*.
- Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, et al. 2025. Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models. In *Advances in Neural Information Processing Systems*. ArXiv:2507.08128.
- Philippe Gournay, Olivier Lahaie, and Roch Lefebvre. 2018. A Canadian French emotional speech dataset. In *ACM Multimedia Systems Conference*, pages 399–402.
- Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, et al. 2026. Measuring audio’s impact on correctness: Audio-contribution-aware post-training of large audio language models. In *International Conference on Learning Representations*. ArXiv:2509.21060.
- Yingxu He, Zhuohan Liu, Shuo Sun, Bin Wang, Wenyu Zhang, Xunlong Zou, et al. 2024. MERaLiON-AudioLLM: Bridging audio and language with large language models. *arXiv preprint arXiv:2412.09818*.
- William Held, Yanzhe Zhang, Minzhi Li, Weiyan Shi, Michael J. Ryan, and Diyi Yang. 2025. Distilling an end-to-end voice assistant without instruction training data. In *Annual Meeting of the Association for Computational Linguistics*. ArXiv:2410.02678.
- Carl Hopkins, Simone Graetzer, and Gary Seiffert. 2019. [ARU speech corpus \(University of Liverpool\)](#). University of Liverpool Data Catalogue.
- Tzu-wen Hsu, Ke-Han Lu, Cheng-Han Chiang, and Hung-yi Lee. 2025. Reducing object hallucination in large audio-language models via audio-aware decoding. In *IEEE Automatic Speech Recognition and Understanding Workshop*. ArXiv:2506.07233.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*. ArXiv:2106.09685.
- Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T. Liu, Chen-An Li, Yu-Xiang Lin, et al. 2025. Dynamic-SUPERB phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. In *International Conference on Learning Representations*. ArXiv:2411.05361.
- Jesin James, Li Tian, and Catherine Inez Watson. 2018. An open source emotional speech corpus for human robot interaction applications. In *Interspeech*, pages 2768–2772.
- Shareef Babu Kalluri, Deepu Vijayasenan, Sriram Ganapathy, Ragesh Rajan M, and Prashant Krishnan. 2021. NISP: A multi-lingual multi-accent dataset for speaker profiling. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6953–6957.
- James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2016. [Children speech recording \(English, spontaneous speech + pre-defined sentences\)](#). Zenodo.
- KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, et al. 2025. Kimi-Audio technical report. *arXiv preprint arXiv:2504.18425*.
- Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, et al. 2026. [MMAU-Pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence](#). In *AAAI Conference on Artificial Intelligence*, volume 40, pages 22688–22697.
- Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. 2025. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*.

- Zhizhong Li and Derek Hoiem. 2016. Learning without forgetting. In *European Conference on Computer Vision*. ArXiv:1606.09282.
- Tzu-Quan Lin, Wei-Ping Huang, Yi-Cheng Lin, and Hung-yi Lee. 2026. How contrastive decoding enhances large audio language models. In *IEEE Spoken Language Technology Workshop*. ArXiv:2603.09232.
- Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, et al. 2025. Voxtral. *arXiv preprint arXiv:2507.13264*.
- LMMs-Lab. 2025. Aero-1-Audio model card. Hugging Face. <https://huggingface.co/lmms-lab/Aero-1-Audio>.
- Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, et al. 2026. DeSTA2.5-Audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *IEEE Transactions on Audio, Speech and Language Processing*, 34:2062–2076. ArXiv:2507.02768.
- Ke-Han Lu, Chun-Yi Kuan, and Hung-yi Lee. 2025. Speech-IFEval: Evaluating instruction-following and quantifying catastrophic forgetting in speech-aware language models. In *Interspeech*. ArXiv:2505.19037.
- Matthias Mauch and Simon Dixon. 2014. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 659–663.
- Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Interspeech*, pages 498–502.
- Tu Anh Nguyen, Wei-Ning Hsu, Antony D’Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Reiz, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. 2023. EXPRESSO: A benchmark and analysis of discrete expressive speech resynthesis. In *Interspeech*, pages 4823–4827.
- Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. 2023. AfriSpeech-200: Pan-African accented speech dataset for clinical and general domain ASR. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210.
- Jiacheng Pang, Ashutosh Chaubey, and Mohammad Soleymani. 2026. Do audio LLMs listen or read? Analyzing and mitigating paralinguistic failures with VoxParadox. In *International Conference on Machine Learning*. ArXiv:2605.27772.
- Linkai Peng and Baorian Nuchged. 2026. Represented but ignored: A causal account of prosodic underuse in audio-language models. *arXiv preprint arXiv:2608.19211*.
- M. Kathleen Pichora-Fuller and Kate Dupuis. 2020. Toronto emotional speech set (TESS). Borealis dataset. <https://doi.org/10.5683/SP2/E8H2MF>.
- Karol J. Piczak. 2015. ESC: Dataset for environmental sound classification. In *ACM International Conference on Multimedia*, pages 1015–1018.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. ArXiv:2212.04356.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*. ArXiv:2305.18290.
- rany2. 2024. edge-tts: Use Microsoft Edge’s online text-to-speech service from Python. <https://github.com/rany2/edge-tts>.
- Muhammad Mamunur Rashid, Guiqing Li, and Chengrui Du. 2023. Nonspeech7k dataset: Classification and analysis of human non-speech sound. *IET Signal Processing*, 17(6):e12233.
- Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann. 2024. EARS: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. In *Interspeech*, pages 4873–4877.
- S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*. ArXiv:2410.19168.
- Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. 2018. Pyroomacoustics: A Python package for audio room simulation and array processing algorithms. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 351–355.
- Natarajan Balaji Shankar, Amber Afshan, Alexander Johnson, Aurosweta Mahapatra, Alejandra Martin, Haolun Ni, Hae Won Park, Marlen Quintero Perez, Gary Yeung, Alison Bailey, Cynthia Breazeal, and

- Abeer Alwan. 2024. *The JIBO kids corpus: A speech dataset of child-robot interactions in a classroom environment*. *JASA Express Letters*, 4(11):115201.
- Yirong Sun, Yanjun Chen, Xin Qiu, Gang Zhang, Hongyu Chen, Daokuan Wu, Chengming Li, Min Yang, Dawei Zhu, Wei Zhang, and Xiaoyu Shen. 2026. SonicBench: Dissecting the physical perception bottleneck in large audio language models. *arXiv preprint arXiv:2601.11039*.
- Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025a. *AudioBench: A universal benchmark for audio large language models*. In *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, pages 4297–4316.
- Cheng Wang, Gelei Deng, Xianglin Yang, Han Qiu, and Tianwei Zhang. 2025b. When audio and text disagree: Revealing text bias in large audio-language models. In *Conference on Empirical Methods in Natural Language Processing*. ArXiv:2508.15407.
- Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2026. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. In *International Conference on Learning Representations*.
- Steven Weinberger. 2015. Speech Accent Archive. George Mason University. <http://accent.gmu.edu>.
- Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, et al. 2025. Step-Audio 2 technical report. *arXiv preprint arXiv:2507.16632*.
- Jiaqi Xiong, Yunjia Qi, Qi Cao, Yu Zheng, Yutong Zhang, Ziteng Wang, Ruofan Liao, Weisheng Xu, and Sichen Liu. 2026. DEAF: A benchmark for diagnostic evaluation of acoustic faithfulness in audio language models. *arXiv preprint arXiv:2603.18048*.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, et al. 2025a. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, et al. 2025b. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*.
- Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In *Annual Meeting of the Association for Computational Linguistics*.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, et al. 2024. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*.
- Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An open-source non-native English speech corpus for pronunciation assessment. In *Interspeech*, pages 3710–3714.
- Zihan Zhao, Yiyang Jiang, Heyang Liu, Yu Wang, and Yanfeng Wang. 2025. LibriSQA: A novel dataset and framework for spoken question answering with large language models. *IEEE Transactions on Artificial Intelligence*, 6(11):2884–2895. ArXiv:2308.10390.

A Data

Each item uses the same question, options and scoring rule under all five inputs. We therefore retained only the parts of six public benchmarks with written questions and automatically scorable answers (Table 4). We took at most 500, 150 and 200 items per task from AudioBench, AIR-Bench and Dynamic-SUPERB, respectively, and kept all items from the other three benchmarks. The recordings include acted speech, film and television dialogue, interviews, read speech and synthetic speech.

We score multiple-choice items by the selected option and label items by the first label in the response. Each item keeps the same scoring rule across all five inputs. Whisper-large-v3 (Radford et al., 2023) produces the transcript used as text in *A* and as a synthetic reading in *T*. For Mustard sarcasm detection, most readings instead use the dataset’s human transcript. Edge-TTS (rany2, 2024) reads each transcript in a female voice (T1) and a male voice (T2). We transcribe these readings again with Whisper and re-synthesize the readings until the word error rate is below 20%. The retained readings have a word error rate of 1.4% relative to their source transcripts (Section 2).

B Task Classification

We classify tasks by the evidence needed for an answer (Table 5). The dialogue-content task *dream_tts_mcq_test* is lexical. *Speaker_Gender_Recognition* asks for the speaker’s gender and is a global-attribute task. The *pitch_comparison* task compares two segments and belongs to within-recording structure. Mixed tasks require both sound and words. In *stress_based_reasoning*, for example, the model must identify the stressed word and use the sentence’s meaning to interpret the emphasis.

We excluded *StressDetection_MIRSD* from Dynamic-SUPERB (173 items). Its description asks about stress in the recording, but its gold labels

Table 4: Benchmark subsets used in the study. The table lists their answer formats, task counts and item counts (13,707 items in total).

Benchmark	Part used	Answer format	Tasks	Items
AudioBench	Voice understanding (6 sets), spoken QA (4 sets)	Label, multiple choice, short answer	10	4,779
AIR-Bench	Foundation, speech tasks	Multiple choice	8	1,161
Dynamic-SUPERB Phase-2	11 speech tasks	Label	11	2,035
MMAU	test-mini, speech	Multiple choice	10	332
MMAU-Pro	Speech, one recording per item	Multiple choice	6	865
MMSU	All tasks	Multiple choice	47	4,535
Total			92	13,707

give the dictionary stress of isolated words, which can usually be determined from text. Given only the transcript, Gemini answers 92.5% of 40 items correctly, so treating this task as within-recording structure would introduce a lexical shortcut.

C Systems

Table 6 identifies the checkpoints used for the 19 open models. Gemini 3.1 Pro is the closed-model reference.

D Measuring Readout and Answers

Linear probes require items with known acoustic labels. For attribute tasks, we use the evaluation items. The gender probe combines 880 IEMOCAP and VoxCeleb items from AudioBench, while the other probes use 26 of the 30 attribute tasks with sufficient data (100–308 items per task). The original structure items do not identify the location of the relevant sound, so we constructed recordings for five of the 12 tasks. These cover pitch, speaking-rate and loudness comparison, word stress, and prolonged sound.

For the three comparison tasks, we joined two LibriSpeech utterances (Panayotov et al., 2015) from the same speaker with 0.5 s of silence, then changed the target property in opposite directions. Pitch gaps were 1.5, 2, 4, and 8 semitones. Speaking-rate ratios were 1.2, 1.5, and 2.0, and loudness gaps were 3, 6, and 12 dB. We made 200 recordings at each level. The smallest pitch gap, 1.5 semitones, is close to the 1.66 and 1.82 semitone gaps between adjacent levels in MMSU pitch comparison. For word stress, we raised the pitch of one word (by a median of 4.1 semitones) and lengthened it, producing 600 items. For prolonged sound, we only lengthened the word, producing 585 items with a median lengthening ratio of 2.38.

We train each logistic regression probe on stan-

dardized representations at one layer’s audio positions and test it on items held out from its training data. For example, the gender probe is trained on one dataset and tested on the other. For constructed structure items, we use five-fold cross-validation grouped by speaker, which keeps the same speaker out of both training and test data. Answer accuracy comes from the model’s response to the original attribute question or its choice of option letter on a constructed structure item.

E Layer-wise Readout on All Probed Tasks

Figure 5 shows probe accuracy across layers for both models and all probed tasks. Figure 6 compares probe readout and answer accuracy before and after training.

F Training Data and Setup

The training data contain 219.0 hours of audio, roughly 80% from the LibriSpeech train-clean-100 and train-clean-360 splits. The remaining audio comes from 16 corpora. The nine speech corpora include Espresso (Nguyen et al., 2023), Speech Accent Archive (Weinberger, 2015), EARS (Richter et al., 2024), speechocean762 (Zhang et al., 2021), AfriSpeech-200 (Olatunji et al., 2023), NISP (Kalluri et al., 2021), ARU (Hopkins et al., 2019), JIBO Kids (Shankar et al., 2024), and children’s recordings (Kennedy et al., 2016). The four acted-emotion corpora are CaFE (Gournay et al., 2018), TESS (Pichora-Fuller and Dupuis, 2020), EmoV-DB (Adigwe et al., 2018), and JL-Corpus (James et al., 2018). The three sound-event corpora are FSD50K (Fonseca et al., 2022), Nonspeech7k (Rashid et al., 2023), and ESC-50 (Piczak, 2015).

We generated some training items by changing the recordings. We adjusted the pitch, duration, noise, or reverberation of LibriSpeech and

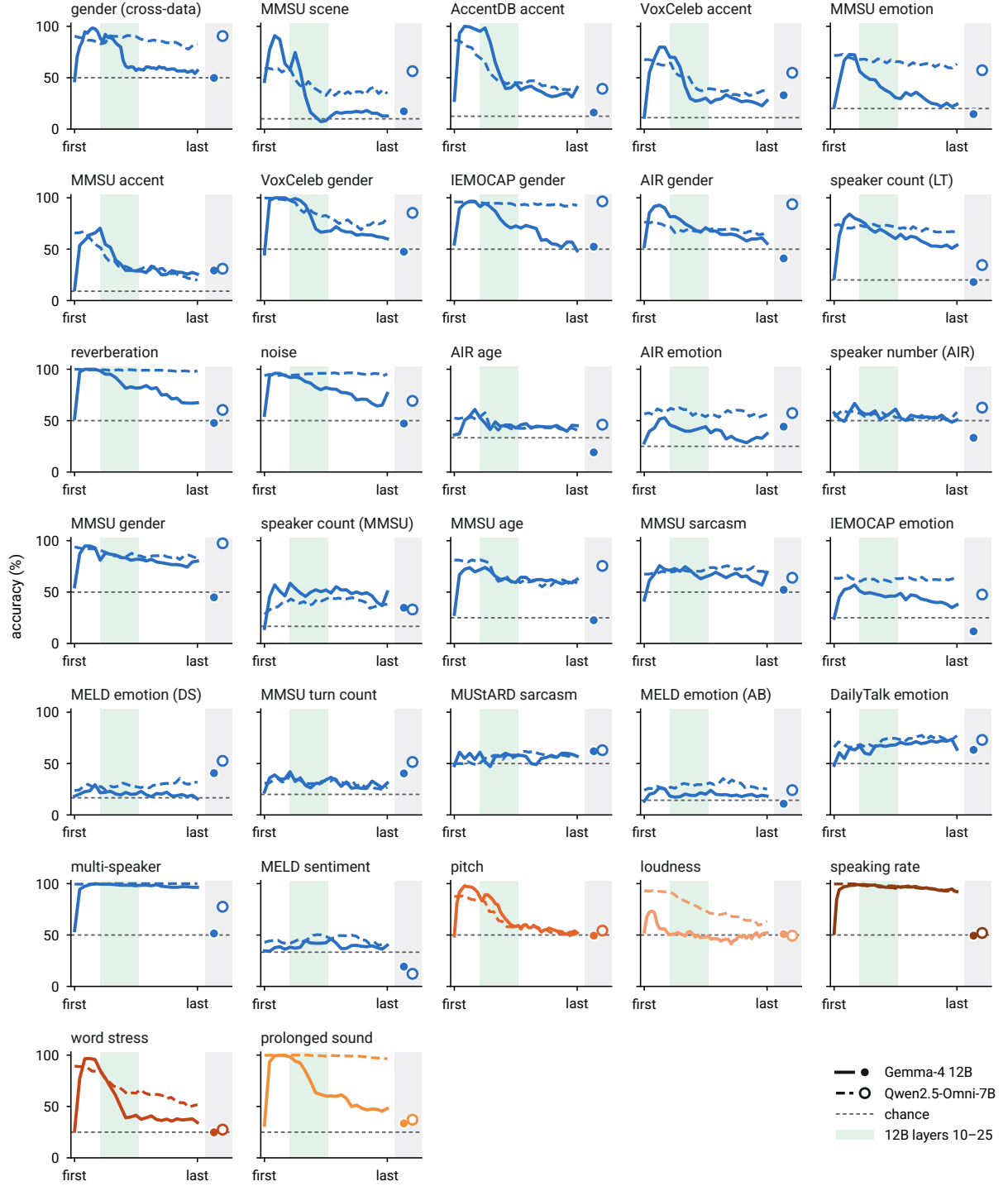


Figure 5: Probe accuracy across relative layer depth on all tasks. Blue panels show attributes and orange panels constructed structure items, with solid lines and filled answer dots for Gemma-4 12B, dashed lines and open dots for Qwen2.5-Omni-7B, a grey dashed chance line, and a green band over Gemma-4 12B layers 10–25 (LT, DS, AB, AIR denote LibriTTS, Dynamic-SUPERB, AudioBench, AIR-Bench).

Table 5: Tasks by evidence class and benchmark. Names follow the source benchmarks, and parentheses give task counts.

Benchmark	Tasks
<i>Lexical (36 tasks)</i>	
AIR-Bench (2)	Speaker_Intent_Classification, Speech_Entity_Reconition
AudioBench (4)	cn_college_listen_mcq_test, dream_tts_mcq_test, public_sg_speech_qa_test, slue_p2_sqa5_test
Dynamic-SUPERB (2)	DialogueActClassification_DailyTalk, SuperbIC_SLURP-Intent
MMAU (7)	Conversational Fact Retrieval, Dissonant Emotion Interpretation, Event-Based Knowledge Retrieval, Key highlight Extraction, Multi Speaker Role Mapping, Phonemic Stress Pattern Analysis, Phonological Sequence Decoding
MMSU (21)	casual_reasoning, code_switch_question_answering, consonant_vowel_perception, content_grounding, continuation_writing, couplet_matching, deixis_resolution, dialogue_reasoning, homophone_based_reasoning, idiom_reasoning, intent_detection, logical_reasoning, long_speech_summarization, near_homophone_perception, polysemy_reasoning, prolonged_sound_reasoning, puns_interpretation, speech_act_classification, speech_translation, syllable_perception, syntactic_structure_matching
<i>Global attribute (30 tasks)</i>	
AIR-Bench (4)	Speaker_Age_Prediction, Speaker_Emotion_Recontion, Speaker_Gender_Recognition, Speaker_Number_Verification
AudioBench (6)	iemocap_emotion_recognition, iemocap_gender_recognition, meld_emotion_test, meld_sentiment_test, voxceleb_accent_test, voxceleb_gender_test
Dynamic-SUPERB (8)	AccentClassification_AccentdbExtended, DialogueEmotionClassification_DailyTalk, EmotionRecognition_MultimodalEmotionlinesDataset, MultiSpeakerDetection_VCTK, NoiseDetection_VCTK_MUSAN-Noise, ReverberationDetection_VCTK_RirsNoises-MediumRoom, SarcasmDetection_Mustard, SpeakerCounting_LibriTTS-TestClean
MMAU (2)	Counting, Emotion Flip Detection
MMSU (10)	accent_identification, age_prediction, background_scene_recognition, dialogue_turn_counting, emotion_recognition, emotional_context_reasoning, gender_prediction, non_verbal_sound_detection, sarcasm_detection, total_speaker_counting
<i>Within-recording structure (12 tasks)</i>	
AIR-Bench (1)	Speech_Grounding
MMSU (11)	disfluency_detection, intonation_perception, pause_perception, pitch_comparison, plosive_sound_identification, prolonged_sound_perception, speaker_identity_recognition, speech_duration_estimation, speech_stress_perception, speed_comparison, volume_comparison
<i>Mixed (14 tasks)</i>	
AIR-Bench (1)	Synthesized_Voice_Detection
Dynamic-SUPERB (1)	SpoofDetection_ASVspoof2017
MMAU (1)	Emotion State summarisation
MMAU-Pro (6)	Audio Quality, Artifacts & Channel Characteristics; Lexical and Phrase-Level Recognition; Paralinguistic/Emotion Recognition; Speaker Demographics; Speaker Identification; Speech Activity, Turn-Taking and Overlap Detection
MMSU (5)	intonation_based_reasoning, pause_based_reasoning, stress_based_reasoning, synthetic_speech_detection, vocal_range_comparison

Espresso utterances, or spliced segments from different speakers into dialogues. Each change determines the answer. We estimated pitch with pYIN (Mauch and Dixon, 2014), located word boundaries with Montreal Forced Aligner (McAuliffe et al., 2017), and simulated room reverberation with Py-roomacoustics (Scheibler et al., 2018). Other answers come from corpus metadata, acted-emotion labels, or sound-event labels.

Training items use the question formats of 34 evaluation tasks, comprising 23 attribute and 11 structure tasks. Eight other attribute and structure tasks have no training items in the same format. These are seven conversational emotion or sarcasm

tasks and MMSU loudness comparison, shaded grey in Table 7. No lexical or mixed task has training items in the same format either.

For these 34 formats, training items follow the evaluation prompt templates, option counts, and label words. About half reuse the evaluation task’s generic question. Some stress, prolonged-sound, and word-intonation items rephrase the question, while some emotion items list the candidate labels. We used no evaluation audio, item-specific evaluation text, or recordings from the evaluation corpora. We also removed training recordings whose audio fingerprint or sentence matched an evaluation item. Training and evaluation therefore share question

Table 6: Open end-to-end audio models and evaluated checkpoints.

Model	Reference	Checkpoint
Qwen3-Omni	Xu et al., 2025b	Qwen/Qwen3-Omni-30B-A3B-Instruct
Qwen2.5-Omni-7B	Xu et al., 2025a	Qwen/Qwen2.5-Omni-7B
MiniCPM-o 4.5	Cui et al., 2026	openbmb/MiniCPM-o-4_5
Qwen2.5-Omni-3B	Xu et al., 2025a	Qwen/Qwen2.5-Omni-3B
MiniCPM-o 2.6	Yao et al., 2024	openbmb/MiniCPM-o-2_6
Kimi-Audio	KimiTeam et al., 2025	moonshotai/Kimi-Audio-7B-Instruct
Audio Flamingo 3	Ghosh et al., 2025	nvidia/audio-flamingo-3-hf
Audio Flamingo Next	Ghosh et al., 2026	nvidia/audio-flamingo-next-hf
MERaLiON-3	He et al., 2024	MERaLiON/MERaLiON-3-10B
Step-Audio 2	Wu et al., 2025	stepfun-ai/Step-Audio-2-mini
Voxtral-Mini-3B	Liu et al., 2025	mistralai/Voxtral-Mini-3B-2507
MERaLiON-2	He et al., 2024	MERaLiON/MERaLiON-2-10B
Gemma-4 E4B	Gemma Team, 2026	google/gemma-4-E4B-it
Qwen2-Audio	Chu et al., 2024	Qwen/Qwen2-Audio-7B-Instruct
R1-AQA	Li et al., 2025	mispeech/r1-aqa
Phi-4-Multimodal	Abouelenin et al., 2025	microsoft/Phi-4-multimodal-instruct
Aero-1-Audio	LMMs-Lab, 2025	lmms-lab/Aero-1-Audio
Gemma-4 E2B	Gemma Team, 2026	google/gemma-4-E2B-it
Gemma-4 12B	Gemma Team, 2026	google/gemma-4-12B-it

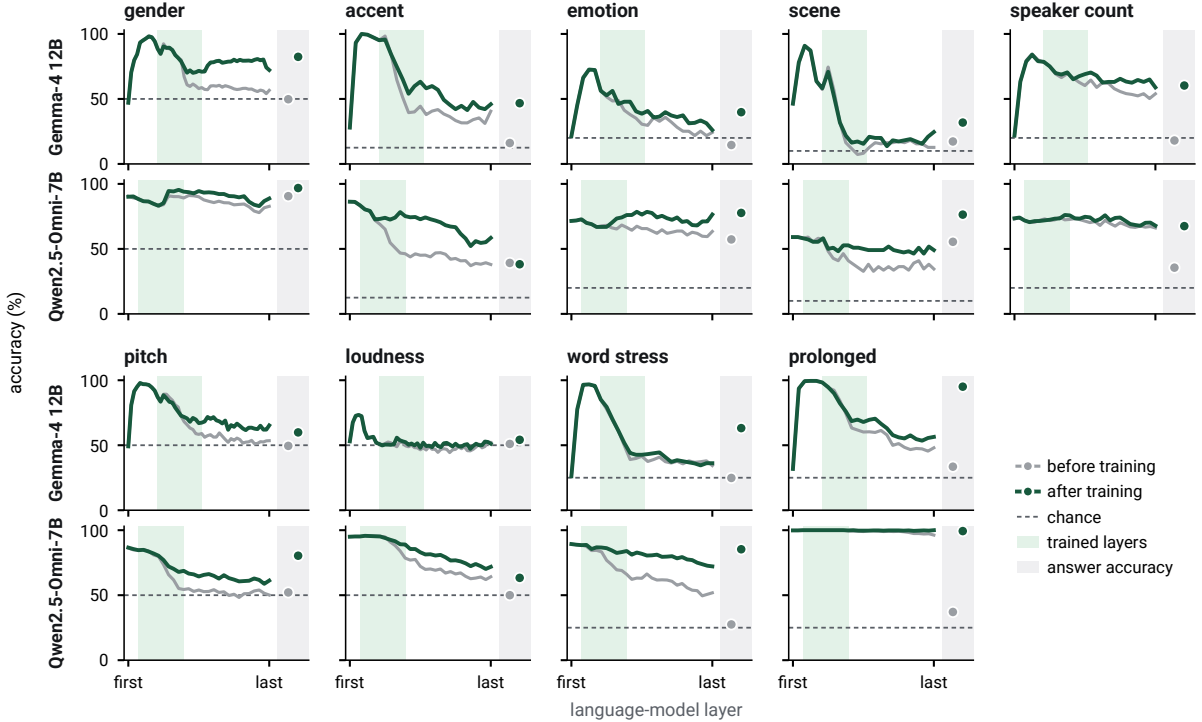


Figure 6: Layer-wise probe and answer accuracy before and after training. Upper panels show Gemma-4 12B and lower panels Qwen2.5-Omni-7B, with grey and green lines for probe accuracy before and after training and other markers as in Figure 3. The accent panel shows AccentDB, and the 11 probed tasks averaged in Section 5 also include the VoxCeleb and MMSU accent tasks, which are not drawn.

formats, but not items.

We trained rank-64 LoRA adapters on every attention and feed-forward projection in the readout-decline layers. We used AdamW for 17,548 steps with 8 items per step and a peak learning rate of 3×10^{-5} . All results use the final adapter, with no checkpoint selection based on evaluation accuracy.

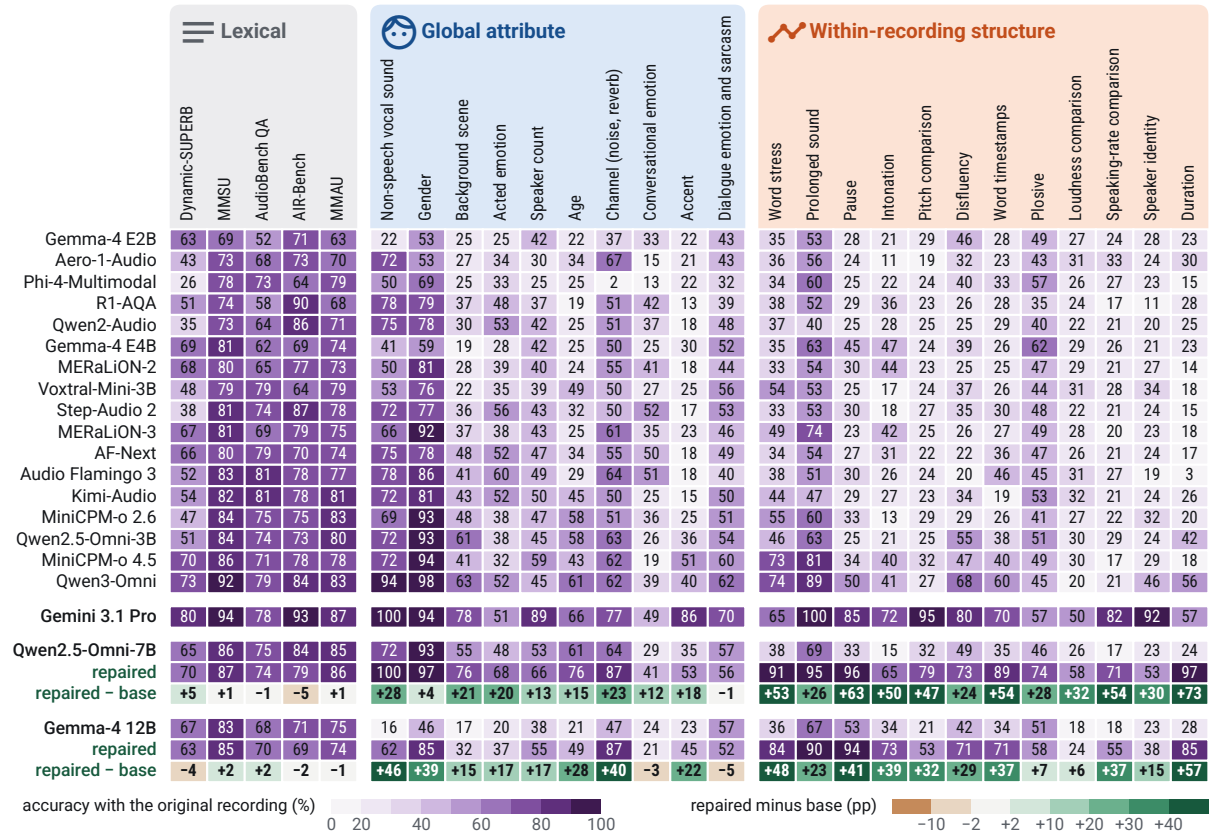


Figure 7: Accuracy with the original recording by task group. Rows cover 19 open models, Gemini 3.1 Pro, and two repaired models, with purple cells for accuracy (%), green and brown cells for repaired-minus-base changes (pp), and lexical tasks grouped by benchmark.

G Per-Task Repair Results

Table 7 reports accuracy with the original recording before and after training for both models on all 92 tasks. Figure 7 groups these tasks. The eight shaded attribute and structure tasks have no training items in the same format, allowing us to examine transfer to these formats.

Table 7: Original-recording accuracy (%) before and after training on 92 tasks. Tasks follow the classes in Table 5, with repaired cells colored by change from the base (darker green for larger gains, brown for declines, light grey within 2 pp) and shaded rows marking attribute and structure tasks without matching training formats.

Benchmark	Task	Gemma-4 12B		Qwen2.5-Omni-7B	
		Base	Repaired	Base	Repaired
Lexical (36 tasks)					
AIR-Bench	Speaker_Intent_Classification	73.9	70.4	90.8	85.2
AIR-Bench	Speech_Entity_Reconition	67.1	67.8	77.6	73.4
AudioBench	cn_college_listen_mcq_test	89.6	91.4	92.6	92.6
AudioBench	dream_tts_mcq_test	77.0	78.2	89.9	89.1
AudioBench	public_sg_speech_qa_test	22.3	26.6	33.2	32.2
AudioBench	slue_p2_sqa5_test	81.7	82.9	83.2	81.2
Dynamic-SUPERB	DialogueActClassification_DailyTalk	62.7	57.5	50.3	64.2
Dynamic-SUPERB	SuperbIC_SLURP-Intent	71.2	69.1	80.1	74.9
MMAU	Conversational Fact Retrieval	80.0	80.0	100.0	95.0
MMAU	Dissonant Emotion Interpretation	95.0	90.0	100.0	100.0
MMAU	Event-Based Knowledge Retrieval	75.0	60.0	90.0	85.0
MMAU	Key highlight Extraction	95.0	85.0	95.0	95.0
MMAU	Multi Speaker Role Mapping	100.0	100.0	100.0	100.0
MMAU	Phonemic Stress Pattern Analysis	34.6	45.1	54.9	57.5

Benchmark	Task	Gemma-4 12B		Qwen2.5-Omni-7B	
		Base	Repaired	Base	Repaired
MMAU	Phonological Sequence Decoding	47.4	57.9	52.6	68.4
MMSU	casual_reasoning	98.3	100.0	100.0	100.0
MMSU	code_switch_question_answering	67.5	79.2	94.8	93.5
MMSU	consonant_vowel_perception	75.4	78.7	60.7	70.5
MMSU	content_grounding	95.4	98.2	93.5	94.4
MMSU	continuation_writing	98.2	100.0	99.1	100.0
MMSU	couplet_matching	40.6	46.5	66.3	59.4
MMSU	deixis_resolution	84.5	82.5	90.3	91.3
MMSU	dialogue_reasoning	82.3	86.7	87.6	85.8
MMSU	homophone_based_reasoning	93.1	89.1	98.0	96.0
MMSU	idiom_reasoning	96.0	97.0	100.0	100.0
MMSU	intent_detection	78.5	85.0	89.7	83.2
MMSU	logical_reasoning	73.1	74.0	78.8	77.9
MMSU	long_speech_summarization	97.2	99.1	99.1	99.1
MMSU	near_homophone_perception	52.8	55.6	68.1	68.1
MMSU	polysemy_reasoning	96.9	95.9	99.0	99.0
MMSU	prolonged_sound_reasoning	93.9	93.9	97.6	98.8
MMSU	puns_interpretation	100.0	100.0	99.1	100.0
MMSU	speech_act_classification	94.8	97.9	90.7	92.8
MMSU	speech_translation	94.0	92.8	97.6	97.6
MMSU	syllable_perception	38.1	35.7	27.4	44.0
MMSU	syntactic_structure_matching	94.8	97.9	73.2	68.0
<i>Global attribute (30 tasks)</i>					
AIR-Bench	Speaker_Age_Prediction	19.1	55.3	46.8	74.5
AIR-Bench	Speaker_Emotion_Recontion	44.1	43.4	57.3	67.1
AIR-Bench	Speaker_Gender_Recognition	41.0	84.0	93.8	95.8
AIR-Bench	Speaker_Number_Verification	33.3	71.3	64.0	78.0
AudioBench	iemocap_emotion_recognition	11.8	38.9	47.8	57.6
AudioBench	iemocap_gender_recognition	52.4	75.5	96.5	93.4
AudioBench	meld_emotion_test	10.9	14.5	23.7	38.0
AudioBench	meld_sentiment_test	19.2	20.9	12.2	27.2
AudioBench	voxceleb_accent_test	32.9	65.0	55.0	75.9
AudioBench	voxceleb_gender_test	47.5	89.8	85.3	100.0
Dynamic-SUPERB	AccentClassification_AccentdbExtended	16.1	46.7	39.2	38.2
Dynamic-SUPERB	DialogueEmotionClassification_DailyTalk	63.3	51.0	72.5	75.0
Dynamic-SUPERB	EmotionRecognition_MultimodalEmotionlinesDataset	40.6	27.6	52.1	57.8
Dynamic-SUPERB	MultiSpeakerDetection_VCTK	51.5	99.0	78.5	99.0
Dynamic-SUPERB	NoiseDetection_VCTK_MUSAN-Noise	47.2	84.6	67.2	83.6
Dynamic-SUPERB	ReverberationDetection_VCTK_RirsNoises-MediumRoom	47.7	89.3	61.4	89.8
Dynamic-SUPERB	SarcasmDetection_Mustard	62.0	67.0	62.0	62.0
Dynamic-SUPERB	SpeakerCounting_LibriTTS-TestClean	18.0	60.3	35.6	67.5
MMAU	Counting	50.0	35.0	55.0	70.0
MMAU	Emotion Flip Detection	50.0	30.0	30.0	25.0
MMSU	accent_identification	29.2	42.5	30.1	68.1
MMSU	age_prediction	22.6	42.2	75.5	77.5
MMSU	background_scene_recognition	17.3	31.8	55.5	76.4
MMSU	dialogue_turn_counting	40.4	24.2	50.5	35.4
MMSU	emotion_recognition	14.6	39.8	57.3	77.7
MMSU	emotional_context_reasoning	10.0	27.8	28.9	68.9
MMSU	gender_prediction	44.9	89.0	97.5	97.5
MMSU	non_verbal_sound_detection	15.6	62.5	71.9	100.0
MMSU	sarcasm_detection	52.2	58.6	64.0	61.3
MMSU	total_speaker_counting	34.8	40.7	33.9	46.6
<i>Within-recording structure (12 tasks)</i>					
AIR-Bench	Speech_Grounding	34.2	71.1	34.9	88.6
MMSU	disfluency_detection	42.5	70.8	49.1	72.6
MMSU	intonation_perception	33.6	72.9	14.9	65.4
MMSU	pause_perception	52.8	94.3	33.0	96.2
MMSU	pitch_comparison	21.2	52.5	32.3	78.8
MMSU	plosive_sound_identification	50.5	57.7	46.4	74.2
MMSU	prolonged_sound_perception	66.7	90.1	69.1	95.1
MMSU	speaker_identity_recognition	22.8	37.7	22.8	52.6
MMSU	speech_duration_estimation	27.6	84.8	23.8	97.1
MMSU	speech_stress_perception	36.4	83.6	38.2	90.9

Benchmark	Task	Gemma-4 12B		Qwen2.5-Omni-7B	
		Base	Repaired	Base	Repaired
MMSU	speed_comparison	18.5	55.4	17.4	70.7
MMSU	volume_comparison	17.8	23.8	25.7	58.4
<i>Mixed (14 tasks)</i>					
AIR-Bench	Synthesized_Voice_Detection	48.3	55.0	55.0	55.7
Dynamic-SUPERB	SpoofDetection_ASVspoof2017	7.9	59.5	10.1	18.5
MMAU	Emotion State summarisation	80.0	65.0	25.0	10.0
MMAU-Pro	Audio Quality, Artifacts & Channel Characteristics	60.0	40.0	80.0	80.0
MMAU-Pro	Lexical and Phrase-Level Recognition	100.0	100.0	100.0	100.0
MMAU-Pro	Paralinguistic/Emotion Recognition	67.2	70.5	72.1	67.2
MMAU-Pro	Speaker Demographics	48.1	50.0	62.7	62.0
MMAU-Pro	Speaker Identification	75.0	87.5	100.0	100.0
MMAU-Pro	Speech Activity, Turn-Taking and Overlap Detection	56.1	53.7	63.4	65.8
MMSU	intonation_based_reasoning	63.6	85.7	74.0	87.0
MMSU	pause_based_reasoning	50.6	55.1	53.9	52.8
MMSU	stress_based_reasoning	50.0	72.7	47.3	64.5
MMSU	synthetic_speech_detection	43.1	43.1	56.0	58.7
MMSU	vocal_range_comparison	50.0	37.5	62.5	62.5