USC Institute for Creative Technologies

ICML 2026

VoxParadox

Do Audio LLMs Listen or Read? Analyzing and Mitigating Paralinguistic Failures with **VoxParadox**

Jiacheng Pang*, Ashutosh Chaubey*, Mohammad Soleymani
Institute for Creative Technologies, University of Southern California
* Equal contribution

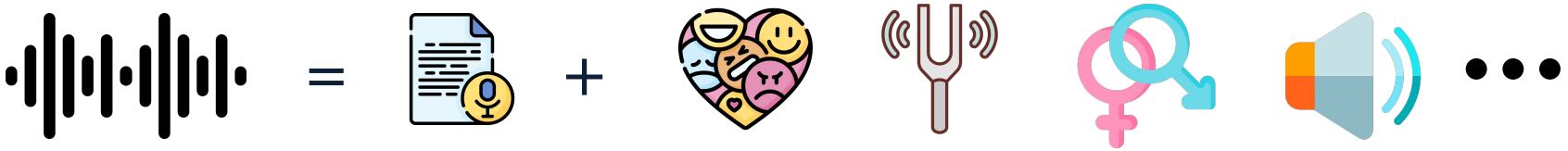
Project page: voxparadox.github.io

Keywords: *Speech LLMs, Audio LLMs, Speech Paralinguistics*

Motivation

Speech carries **what** is said and **how** it's said.

- **Paralinguistic**: emotion, pitch, volume, speed, speaker traits, etc.
  - Applications in affective applications, conversational agents, accessibility tech, etc.
- Audio LLMs are general speech interfaces, but their paralinguistic understanding is poorly characterized.
- Existing benchmarks **conflate** lexical and acoustic cues – never **decouples** them for isolated evaluation.



Speech = Linguistic content + Paralinguistic content

VoxParadox - the Benchmark

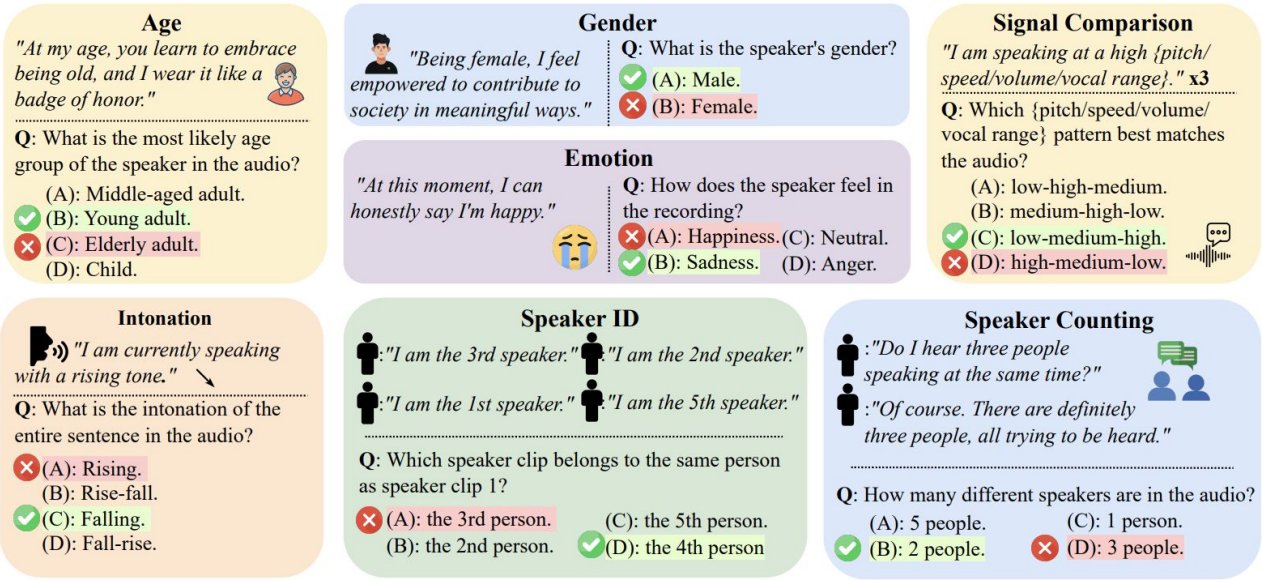


Figure 1. **VoxParadox task examples.** VoxParadox covers ten paralinguistic tasks; each example is constructed to contradict lexical and acoustic cues. The transcript explicitly asserts an adversarial attribute y_{adv}, while the speech style conveys the true paralinguistic label y_{true}. The MCQ asks for the acoustic attribute and includes both labels as options, exposing language-following behavior when models ignore non-verbal cues.

10 tasks * 200 examples/task = 2,000 verified examples.

Generated with controlled text-to-speech engines, verified with Whisper and human.

Benchmark Metrics

We use two metrics to evaluate models on VoxParadox

- GT Accuracy – did the model match the acoustic truth?
  - Higher = better
- Adversarial Label Agreement (ALA) – did the model match the transcript-implied label?
  - Lower = better

$$\mathrm{Acc}_{\mathrm{GT}} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[\hat{y}_i = y_{\mathrm{true}}^{(i)}\right]. \qquad \mathrm{ALA} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[\hat{y}_i = y_{\mathrm{adv}}^{(i)}\right].$$

GT label ≠ Adv. label by design → high ALA = lexical shortcutting

Finding 1: Models Follow the Text

Model	VoxParadox											MMSU (Avg.)
	Age	Gender	Emotion	Pitch	Volume	Speed	Range	Intonation	Spk ID	Spk Cnt	Avg.	Para.
Open-Source Audio LLMs												
Audio Flamingo 2 (AF2)	**35.00**	**99.00**	0.00	19.50	19.00	21.00	18.00	**38.50**	27.50	31.00	**30.85**	27.44
Audio Flamingo 3 (AF3)	10.00	16.00	24.50	11.00	11.50	11.50	9.50	34.00	23.50	22.50	17.40	37.74
Qwen2-Audio-7B-Instruct	2.00	3.00	14.00	**26.50**	22.00	**24.50**	**19.00**	0.50	24.50	12.50	14.85	34.37
SALMONN-7B	10.00	12.50	0.00	0.00	0.00	17.00	0.00	21.50	0.00	0.00	6.10	6.84
Kimi-Audio-7B-Instruct	9.00	24.50	**79.00**	7.50	12.50	11.00	8.00	5.00	20.50	13.00	19.00	41.48
VITA-Audio	3.50	3.50	0.00	8.50	8.00	16.00	9.00	0.00	12.00	8.00	6.85	29.54
MiMo-Audio-7B-Instruct	7.50	50.00	0.50	11.50	19.00	15.00	15.00	11.50	**30.00**	36.00	19.60	35.64
Step-Audio-R1	10.50	12.50	1.50	11.50	4.50	9.50	9.50	13.00	9.00	**93.00**	17.45	**54.51**
Open-Source Omni LLMs												
Qwen2.5-Omni-7B	1.50	2.00	3.00	13.00	8.00	8.00	12.50	4.50	15.50	11.50	7.95	33.45
Qwen3-Omni	17.50	4.50	0.50	4.00	7.00	4.00	2.50	0.00	12.00	54.00	10.60	49.13
Closed-Source Models												
GPT-4o Audio	4.50	1.00	0.00	7.00	5.00	8.00	5.00	0.50	27.00	28.00	8.60	36.55
Gemini 2.5 Flash	15.00	14.50	6.00	19.00	**31.00**	16.00	13.00	19.00	21.00	92.50	24.70	51.05

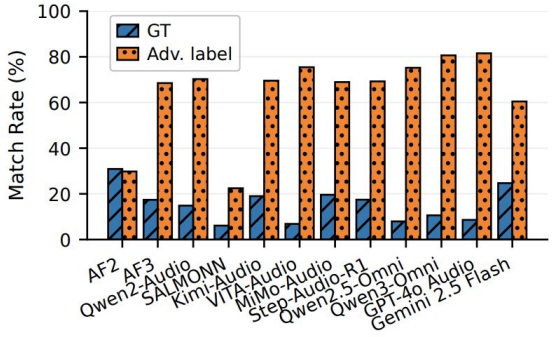


Figure 2. GT accuracy vs. ALA on VoxParadox (average across 10 tasks). High ALA indicates the model is likely to be misled by textual cues when the GT answer lies in the acoustic features.

Across 12 models, GT acc averages to 15% while ALA 64%; no model exceeds 31%.

Key Finding: Audio LLMs can be easily misled by transcripts and can **collapse** dramatically under controlled linguistic–acoustic contradiction.

Finding 2 & 3: Two Bottlenecks

The ASR training objective discards paralinguistic information

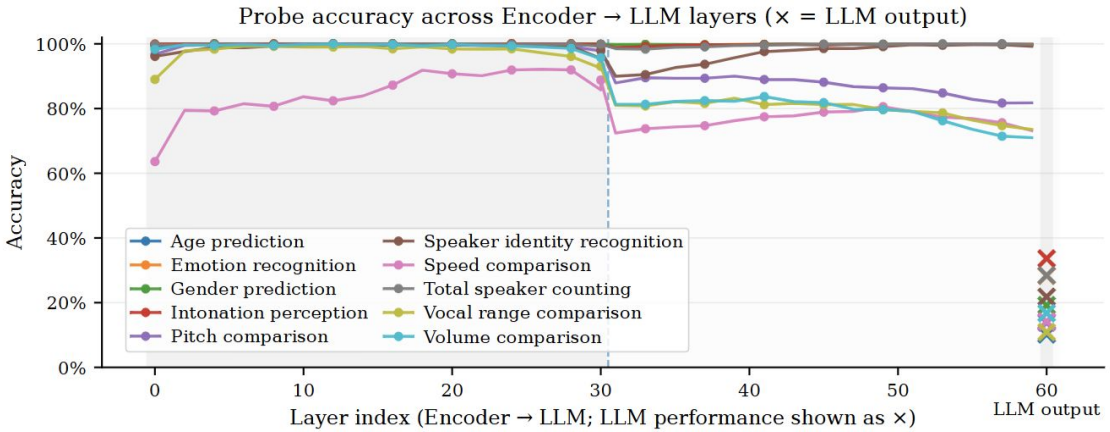


Figure 4. **Layer-wise probe accuracy from encoder to LLM on VoxParadox (AF3).** Lines show 10-fold probe accuracy for each task using representations extracted every two layers; the vertical dashed line denotes the encoder–LLM interface. × markers indicate AF3's end-to-end task accuracy from model outputs. Probes substantially outperform model outputs across tasks, revealing a utilization gap, and several tasks exhibit drops at the interface and/or stronger signals in intermediate encoder layers.

We probe the Audio Flamingo 3 (AF3) internal audio representations from encoder layers → LLM layers. Probe accuracy = is the information easily **retrievable**?

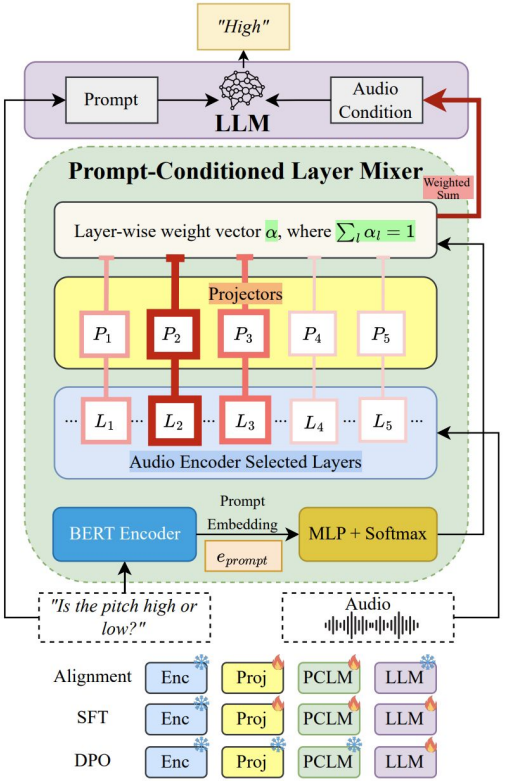
Key Findings: 1. **Utilization gap**: probes massively outperform the model; paralinguistic information is present and retrievable, yet the model ignores it.
2. **Representation Degradation**: for some tasks, the intermediate layers are best, not the final layer, and there is a noticeable drop at the encoder–LLM interface.

Method - PCLM

Prompt-**C**onditioned **L**ayer **M**ixer: A lightweight module addressing both bottlenecks. Instead of solely passing in the final encoder output to the LLM, **PCLM**

- Tap into intermediate encoder layers
- Project each into LLM space
- Encode **textual prompt** with BERT, and use softmax to obtain aggregation weights for each layer. IOW, for each layer, how relevant is this layer in terms of answering the user's question?
- Aggregate intermediate layer representations based on the weights, and feed into the LLM

Intuition: Low level signal tasks upweight the early layers; semantic questions upweight the later layers



Method - PCLM + DPO

After PCLM SFT, we apply Direct Preference Optimization (**DPO**) to directly prefer acoustically-grounded options.

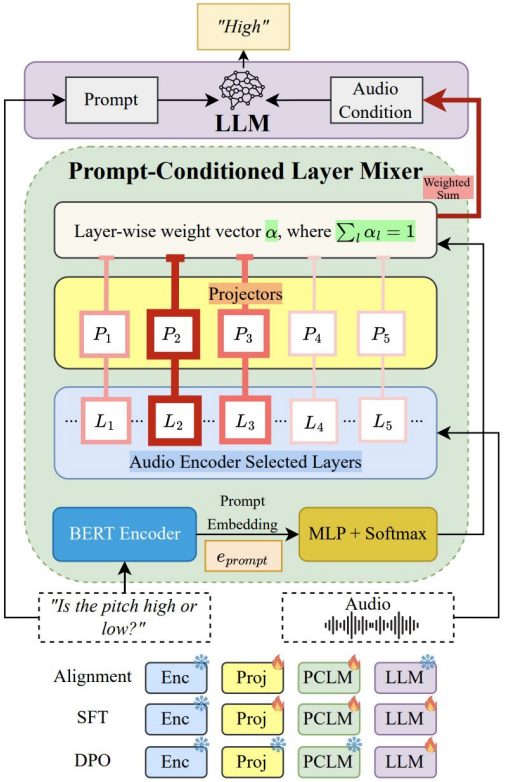
DPO targets the decision policy that representation fixes alone can't close.

Standard pair-wise logistic loss:

$$\mathcal{L}_{\text{DPO}} = -\,\mathbb{E}_{(x,y^+,y^-)}\Big[\log \sigma\big(\beta\, s_\theta(x, y^+, y^-)\big)\Big]$$

$$s_\theta(x, y^+, y^-) = \log \frac{\pi_\theta(y^+ \mid x)}{\pi_\theta(y^- \mid x)} - \log \frac{\pi_{\text{ref}}(y^+ \mid x)}{\pi_{\text{ref}}(y^- \mid x)}$$

Note: VoxParadox benchmark never used in training.



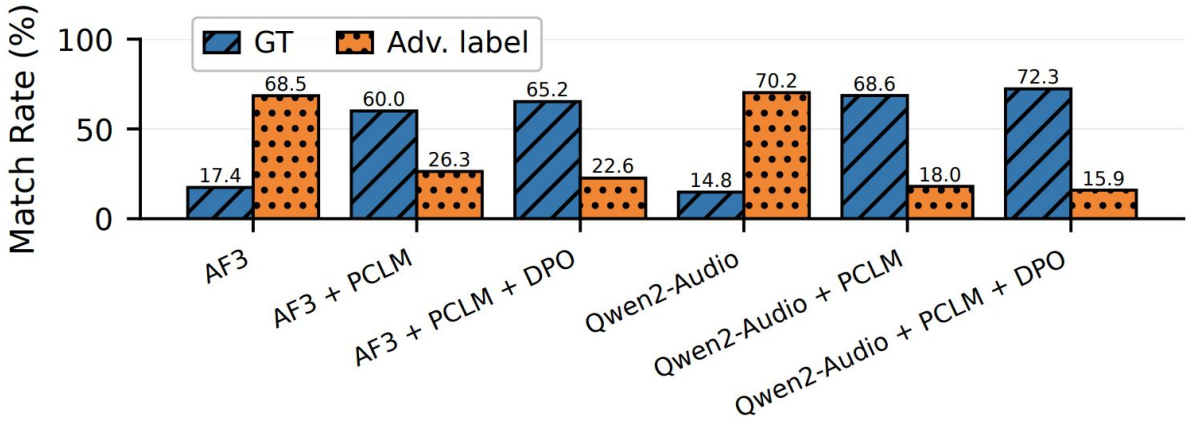
Results

Model	VoxParadox											MMSU (Avg.)
	Age	Gender	Emotion	Pitch	Volume	Speed	Range	Intonation	Spk ID	Spk Cnt	Avg.	Para.
Audio Flamingo 3 (AF3)	10.00	16.00	24.50	11.00	11.50	11.50	9.50	34.00	23.50	22.50	17.40	37.74
AF3 + SFT w/o PCLM	42.00	58.50	36.50	23.00	35.00	**81.00**	35.50	7.50	9.00	20.00	34.80	44.76
AF3 + SFT + DPO w/o PCLM	46.50	79.50	37.50	27.50	36.00	71.00	38.50	11.50	12.00	43.00	40.30	45.58
AF3 + PCLM	55.00	**100.00**	55.50	83.00	54.50	63.50	63.00	**49.50**	33.50	42.50	60.00	54.06
AF3 + PCLM + DPO	**56.50**	**100.00**	**65.50**	**85.50**	**69.00**	73.00	**67.00**	47.00	**37.00**	**51.50**	**65.20**	**54.78**
Δ AF3→AF3 + PCLM + DPO	+46.50	+84.00	+41.00	+74.50	+57.50	+61.50	+57.50	+13.00	+13.50	+29.00	+47.80	+17.04
Qwen2-Audio	2.00	3.00	14.00	26.50	22.00	24.50	19.00	0.50	24.50	12.50	14.85	34.37
Qwen2-Audio + SFT w/o PCLM	51.50	**100.00**	61.00	36.50	34.50	85.50	45.00	3.50	29.50	10.00	45.70	49.41
Qwen2-Audio + SFT + DPO w/o PCLM	39.50	99.50	59.50	40.00	37.50	78.50	44.00	3.50	28.50	14.50	44.50	47.86
Qwen2-Audio + PCLM	56.00	**100.00**	**78.50**	99.50	97.50	95.00	99.50	23.00	26.00	11.50	68.65	**65.18**
Qwen2-Audio + PCLM + DPO	**61.50**	**100.00**	77.50	**100.00**	**98.00**	**96.00**	**100.00**	**27.50**	**30.00**	**32.50**	**72.30**	63.26
Δ Qwen2-Audio → Qwen2-Audio + PCLM + DPO	+59.50	+97.00	+63.50	+73.50	+76.00	+71.50	+81.00	+27.00	+5.50	+20.00	+57.45	+28.89

PCLM + DPO significantly improves paralinguistic understanding in both adversarial settings (VoxParadox) and non-adversarial settings (MMSU).

Ablation: SFT + DPO w/o PCLM helps, but far less effective.

Results



Adversarial Label Agreement dramatically **drops**,

AF3: 68.5% → 26.3%, Qwen2-Audio: 70.2% → 15.9%

Takeaways

VoxParadox: An adversarial paralinguistic benchmark that enforces linguistic-acoustic contradiction, exposing text following failures that other benchmarks hide.

Diagnosis: Audio LLMs overrely on text; paralinguistic information can degrade in deeper layers and at the encoder–LLM interface; a utilization gap exists between LLM internal representation and LLM output.

Method: Prompt-Conditioned Layer Mixer (PCLM) + DPO

Questions?

pangj@usc.edu | achaubey@usc.edu

Project page: voxparadox.github.io