Where Does the Sound Go? Tracing Acoustic Information Loss in Audio-Conditioned LLMs

Anonymous ACL submission

Abstract

Audio-conditioned language models often underuse acoustic cues such as prosody, emotion, and non-speech sounds, raising the question of whether ASR-supervised frontends discard this information before it reaches the LM. We test whether the frontend is responsible by comparing Whisper-Tiny and Whisper-Small with EnCodec, DAC-VAE, and WavTokenizer in a shared Qwen3.5-4B audio-LM pipeline on ASR, emotion recognition, and sound captioning. Encoder replacement alone does not resolve this underuse: Whisper variants remain strongest overall, including on emotion and environmental sound captioning. To localize the failure, we trace task-relevant information through the encoder, projector, LM layers, and LM head. Linear probes and geometric analyses show that discriminative acoustic structure remains recoverable at the final LM layer, even when MCQA accuracy trails probe accuracy by 5 to 83 points. Because the answer format and decoding procedure are controlled, this task-dependent gap points to content-specific readout failure rather than generic format bias. LogitLens analyses and a targeted LM head intervention support the conclusion that acoustic underuse is not explained solely by encoder-side information loss and that readout alignment can be a dominant bottleneck.

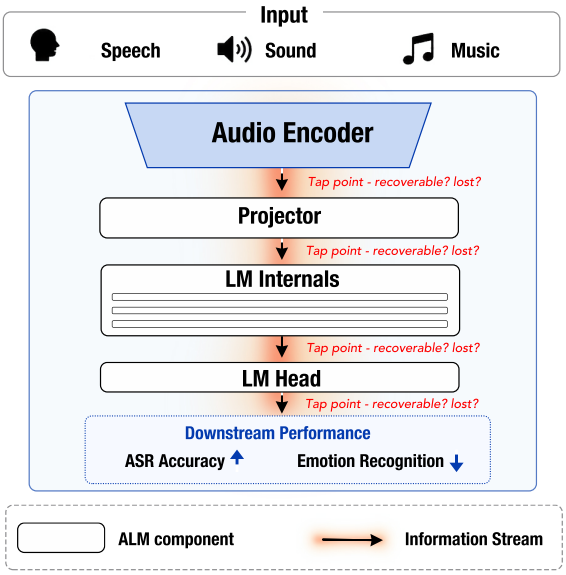

Figure 1: **Overview of our analysis.** An audio-conditioned LLM (ALM) processes the audio input through an encoder–projector–LM pipeline. To trace where the acoustic information goes, we investigate each stage—after the encoder, projector, LM internals, and LM head—asking whether discriminative acoustic information remains recoverable or has been lost.

1 Introduction

Audio-conditioned language models (audio-LLMs) couple a pretrained audio frontend with a large language model (LM) and have rapidly advanced on speech and audio tasks (Chu et al., 2023, 2024; OpenAI, 2024). Yet recent benchmarks reveal a persistent limitation: these models often appear to "transcribe" rather than "listen," relying on transcript-like content while underusing acoustic cues such as prosody, emotion, speaker style, and non-speech sound attributes (Chen et al., 2026). This gap between lexical competence and broader acoustic understanding is now well documented, yet its origin within the audio-LLM pipeline is less clear.

Since many audio-LLMs adopt ASR-supervised encoders, typically Whisper (Radford et al., 2023; Shi et al., 2026), whose pretraining objective may favor lexical invariance over paralinguistic detail, a natural first suspect is the audio frontend. Consistent with this intuition, Whisper representations have been observed to underperform self-supervised alternatives such as WavLM on speaker- and emotion-related probing tasks (Thebaud et al., 2025), leaving open the possibility that observed acoustic underuse originates before the LM sees the signal. We therefore treat encoder replacement as a diagnostic counterfactual rather than as an as-

sumed solution. Reconstruction-based neural audio codecs such as EnCodec, DAC-VAE, and WavTokenizer (Défossez et al., 2022; Kumar et al., 2023; Ji et al., 2025) are trained to reconstruct waveforms, making them a useful test of the encoder-side loss hypothesis: if ASR-supervised frontends are the main bottleneck, representations designed to retain waveform-level detail should at least alleviate acoustic underuse.

We run this test and find that encoder replacement alone is not sufficient in our setting. We compare Whisper-Tiny and Whisper-Small with three reconstruction-based audio representations from EnCodec, DAC-VAE, and WavTokenizer under a shared Qwen3.5-4B LM (Qwen Team, 2026), each paired with its own projector trained identically, on ASR, emotion recognition, and sound captioning. Whisper variants *match or outperform* reconstruction-based encoders on emotion recognition and sound captioning, in addition to leading on ASR. This negative result does not imply that reconstruction-based encoders lack acoustic detail; rather, it shows that the underuse observed downstream cannot be explained simply as acoustic information discarded by ASR-supervised frontends.

If encoder replacement alone does not resolve the gap, where does acoustic information go? We trace task-relevant information through four stages of the audio-LLM pipeline: the encoder, the projector, the LM's internal layers, and the LM head W_U. Using layer-wise linear probes and complementary geometric analyses with identical targets at each stage, we find that discriminative acoustic structure is not erased by the encoder–projector–LM stack: it remains linearly recoverable at the LM's final hidden states, even when final MCQA predictions are weak. The probe-to-MCQA gap is highly task-dependent under an identical prompt template, letter set, and decoding procedure, ranging from near zero on CREMA-D to 83 pp on NSynth. This points to a content-specific readout problem rather than a generic formatting bias.

We test the readout hypothesis causally by fine-tuning only the choice-letter rows of W_U while freezing the encoder, projector, and LM internals. Across the six analysis tasks, this update touches only $\sim$10 to 26k scalars per task and recovers 30 to 42 pp of MCQA accuracy per encoder on average. By contrast, adapting the LM internals with LoRA yields limited or headroom-dependent gains rather than a broad repair. Together, these results show that, in this controlled setting, readout alignment can be the dominant bottleneck even when acoustic evidence survives in the representation.

Encoder	Arch.	Obj.	Tokens/s	Params
EnCodec	Conv	Recon.	75	7.43M
DACVAE	Conv+VAE	Recon. + WM	25	27.55M
Whisper-Small	Tr	ASR	50	88.15M
Whisper-Tiny	Tr	ASR	50	8.21M
WavTokenizer	Conv	Recon.	40	8.30M

Table 1: Audio encoder candidates. Arch.: Conv (SEANet-style stack), Tr (Transformer), VAE (variational bottleneck). Obj.: Recon. (reconstruction-based, *e.g.,* neural codec), Recon. + WM (with acoustic watermarking), ASR (supervised speech recognition). Params count the active forward path. Per-encoder sampling rate, hop, latent dimensionality, and training-time settings are listed in Table 5; pretrained URIs in Table 5.

Taken together, our study makes three main contributions. First, it tests a natural encoder-side explanation for acoustic underuse under a controlled shared-LM setup, showing that replacing ASR-supervised Whisper with reconstruction-based representations is not sufficient to improve emotion recognition or sound captioning. Second, it localizes the failure across the encoder–projector–LM pipeline with matched probing and geometric diagnostics, showing that task-relevant acoustic structure remains recoverable up to the final LM hidden states. Third, it provides readout-level causal evidence by freezing the entire stack and updating only the choice-letter rows of W_U, and uses LogitLens and alignment analyses to characterize how otherwise recoverable acoustic evidence fails to surface as the correct verbalized prediction. These results reframe acoustic underuse, at least in our controlled setting, as a failure that can arise at readout rather than from encoder-side acoustic loss alone, and identify LM-head alignment as a useful diagnostic lens for understanding such failures.

2 Related Work

Audio frontends in audio-LLMs. Audio-conditioned language models typically connect an audio frontend to a text-trained language model through a learned adapter or projector. Many recent systems use transcription-oriented frontends, with Whisper (Radford et al., 2023) widely adopted in Qwen-Audio (Chu et al., 2023), Qwen2-Audio (Chu et al., 2024), and Audio Flamingo 3 (Goel et al., 2025). These encoders are highly effective for lexical tasks, but their ASR-oriented objectives may encourage invariance to speaker, prosodic,

and environmental variation. Consistent with this concern, Whisper representations have been reported to underperform self-supervised alternatives such as WavLM on speaker- and emotion-related probing tasks (Thebaud et al., 2025). This makes the audio frontend a plausible source of acoustic underuse, but does not establish whether replacing it is sufficient.

Several systems therefore complement or depart from transcription-oriented frontends. SALMONN (Tang et al., 2024) combines Whisper with BEATs (Chen et al., 2023) in a dual-encoder architecture, while codec-based models such as Moshi (Défossez et al., 2024) show that reconstruction-oriented audio representations can support spoken language modeling. Neural audio codecs and tokenizers such as EnCodec (Défossez et al., 2022), DAC-VAE (Polyak et al., 2024), and WavTokenizer (Ji et al., 2025) are trained to reconstruct audio and therefore provide a useful contrast to ASR-supervised representations when testing whether acoustic underuse originates at the frontend. Our work compares these frontend families under a shared LM setup, not to assume that reconstruction-based encoders should dominate, but to test whether frontend replacement explains the acoustic-underuse behavior observed in audio-LLMs.

Acoustic underuse and representation analysis. Recent work has shown that strong audio-LLM performance does not necessarily imply reliance on acoustic evidence. Most directly, Chen et al. (2026) introduce LISTEN, a benchmark that separates lexical and acoustic emotion cues, and show that contemporary audio-LLMs often follow transcript-like content even when acoustic cues indicate a different emotion. This behavioral finding motivates our analysis, but it does not localize where the relevant information is lost or suppressed: acoustic cues may be weakened by the frontend, discarded by the projector, transformed inside the LM, or present in hidden states but not expressed by the final LM head.

Analytical tools such as linear probing (Pasad et al., 2021, 2023) and representation similarity analyses like centered kernel alignment (Kornblith et al., 2019) have been used to inspect speech and audio representations. However, these studies typically focus on a single encoder and stop at the encoder output. Recent efforts have begun to examine encoder–LM interfaces, including audio representation transfer analyses (Alex et al., 2025) and unified audio-encoder benchmarks for ALMs (Dinkel et al., 2026). Our work extends this diagnostic view across the full encoder–projector–LM pipeline, using identical probing targets at each stage to distinguish representation-level loss from later readout failure.

Readout bias in verbalized classification. Language-model predictions are mediated by the LM head and can be sensitive to label words, answer formats, and multiple-choice options. In text classification, this issue is commonly studied through verbalizers and prompt-based classification, where different surface forms for the same class can substantially change model predictions (Schick and Schütze, 2021; Liu et al., 2023). Similar concerns arise in multiple-choice evaluation, where option letters and formatting can affect predictions independently of the underlying evidence (Zheng et al., 2024). For audio-LLMs, this creates a possible gap between information that is recoverable from hidden states and information that is expressed as final answer tokens. Our work connects this readout problem to acoustic underuse by showing that task-relevant acoustic structure remains recoverable in the final LM layer, while the LM head fails to map that structure to the correct choice-letter tokens.

3 Experiment Design

3.1 Architecture

We adopt a three-stage encoder-projector-LM architecture (Kang et al., 2025; Ma et al., 2024), where a frozen audio encoder maps a waveform to a continuous latent sequence, a trainable projector aligns this latent with the LM embedding space, and a pretrained LM autoregressively generates a text response.

Audio encoders. Most ALMs default to an ASR-based encoder, typically Whisper, yet evaluations like LISTEN suggest acoustic detail is underutilized downstream. To test whether this acoustic information bottleneck originates in the encoder, we treat the audio encoder as a swappable component and study five pretrained candidates spanning both encoder families (reconstruction-based and ASR-based), summarized in Table 1. Three are reconstruction-based encoders, namely **EnCodec** (Défossez et al., 2022), **DAC-VAE** (Kumar et al., 2023; San Roman et al., 2024), and **WavTokenizer** (Ji et al., 2025), while the remaining two are ASR-

based encoders, **Whisper-Small** and **Whisper-Tiny** (Radford et al., 2023). For the reconstruction-based encoders, we extract only the feed-forward encode path, omitting any decoder, residual vector quantizer, watermarker, or output projection. For the ASR-based encoders, we use the encoder stack only, discarding the text decoder. In all cases, the encoder is frozen throughout overall training.

Trainable components. We first train only the projector, a four-layer causal Transformer decoder (Appendix A.1), and freeze everything else. To test whether the LM degrades the acoustic representation passed from the projector, we then fine-tune the LM (Qwen3.5-4B (Qwen Team, 2026), Appendix A.1), through low-rank adapters (Hu et al., 2022) while continuing to update the projector. Training uses a multi-task mixture of English ASR, environmental-sound captioning, and emotion recognition (Appendix A.2). Further training details are deferred to Appendix A.2.

3.2 Analysis Settings and Methods

Evaluation benchmark. We evaluate on a multi-task suite spanning ASR, emotion, and sound captioning (Table 10). ASR uses the LibriSpeech (Panayotov et al., 2015) and GigaSpeech (Chen et al., 2021) with WER and CER after Whisper-style English text normalization (Radford et al., 2023). Emotion uses MELD (Poria et al., 2019) and a leakage-free IEMOCAP (Busso et al., 2008) Session 5 split with accuracy and macro-F1. Sound captioning covers FSD50K (Fonseca et al., 2022), AudioSet (Gemmeke et al., 2017), AudioCaps (Kim et al., 2019), and Clotho (Drossos et al., 2020) with BLEU-1, BLEU-4, and CIDEr (Vedantam et al., 2015). Decoding is greedy throughout (`num_beams`= 1, `do_sample`=False), with no external language model, and all encoder variants share identical evaluation code, tokenization, and normalization.

4 Analysis

4.1 Acoustic Underuse Persists across Encoders

ALM benchmarks, like LISTEN (Chen et al., 2026), show that ALMs systematically underuse acoustic cues despite using competent audio encoders. This phenomenon raises the question of what part of the pipeline is responsible, and a first hypothesis blames the encoder training objective. Although ASR supervision may suppress non-lexical structure even when the encoder is otherwise capable, the dominant choice of ASR-supervised Whisper has not been systematically validated. To test this hypothesis, we compared five encoders, spanning ASR-based and reconstruction-based training, with end-task performance.

Encoder	ASR WER ↓	Emotion macro-F1 ↑	Captioning CIDEr ↑
Whisper-Small	**7.05**	**0.462**	**0.847**
Whisper-Tiny	8.84	0.444	0.685
EnCodec	20.80	0.393	0.259
WavTokenizer	33.61	0.356	0.325
DAC-VAE	20.46	0.400	0.235

Table 2: Per-encoder summary across task families, averaged within each family. ASR WER averages LibriSpeech test-clean, test-other, and GigaSpeech test. Emotion macro-F1 averages MELD and IEMOCAP-S5. Captioning CIDEr averages FSD50K, Clotho, AudioCaps, and AudioSet. Per-benchmark detail is in Table 10.

The cross-benchmark comparison in Table 2 shows clear differences, confirming that encoder choice strongly affects downstream behavior. Whisper-Small leads on every task family on average, followed by Whisper-Tiny, and this lead extends beyond ASR to emotion and sound captioning. Even the best-performing encoder, however, leaves substantial headroom on emotion and captioning, so the acoustic underuse persists regardless of encoder choice. Task-level differences across encoders therefore reflect a relative ranking rather than a localization of where the failure originates inside the pipeline.

To begin localizing the bottleneck, we compare encoder-side linear probes against final LM-side prediction accuracy (MCQA format) on the same tasks (Table 3). Across domains, probes often outperform MCQA substantially, indicating that task-relevant information is recoverable from encoder latents even when the LM's prediction is weak. The bottleneck therefore lies after the encoder, motivating the stage-wise analyses below.

4.2 Acoustic Information Survives Encoder, Projector, and LM Internals

If encoder choice alone does not explain the downstream acoustic gap, the failure must originate somewhere later in the pipeline. A natural starting point is that information is degraded at the encoder–LM interface, when the projector com-

Task	Whisper-Tiny	Whisper-Small	Encodec	WavTokenizer	DAC-VAE
CREMA-D	57/31 (−25)	59/28 (−31)	30/30 (+0)	34/22 (−12)	29/22 (−7)
IEMOCAP	65/53 (−12)	69/58 (−11)	56/52 (−5)	52/44 (−8)	52/45 (−6)
SpeechCmds	97/83 (−14)	98/66 (−31)	49/41 (−7)	25/47 (+22)	35/44 (+9)
AudioMNIST	100/88 (−12)	100/53 (−47)	85/17 (−68)	64/14 (−50)	77/10 (−67)
VocalSound	83/41 (−42)	88/52 (−36)	55/27 (−29)	35/29 (−6)	44/26 (−18)
ESC-10	93/37 (−56)	95/50 (−45)	76/44 (−33)	62/20 (−42)	66/20 (−47)
ESC-50 (5cat)	79/33 (−46)	89/43 (−47)	53/38 (−15)	41/23 (−18)	45/20 (−26)
UrbanSound8K	86/26 (−60)	90/31 (−59)	70/20 (−50)	34/15 (−19)	52/9 (−43)
TAU	58/8 (−50)	59/10 (−49)	59/11 (−48)	41/12 (−30)	55/10 (−44)
GTZAN	82/9 (−73)	84/15 (−69)	66/11 (−55)	47/10 (−37)	58/9 (−49)
NSynth	90/8 (−83)	97/14 (−83)	61/17 (−44)	38/10 (−27)	56/7 (−49)
IRMAS	54/11 (−43)	66/12 (−53)	44/10 (−34)	32/11 (−21)	40/12 (−28)
MedleySolos	92/24 (−67)	96/26 (−71)	89/18 (−71)	63/18 (−44)	80/17 (−63)
FSD50K†	82/47 (−35)	82/57 (−26)	72/44 (−27)	63/38 (−26)	68/32 (−36)
AudioSet†	75/47 (−28)	77/56 (−21)	71/42 (−29)	63/30 (−33)	63/29 (−33)
Macro avg.	79.5/36.4 (−43.1)	83.3/38.1 (−45.2)	62.4/28.1 (−34.3)	46.3/22.9 (−23.4)	54.7/20.8 (−33.9)

Table 3: Probe / MCQA accuracy (%) per task across encoders, with the probe→MCQA transfer gap in parentheses. Tasks span speech emotion (CREMA-D, IEMOCAP), speech content (SpeechCmds, AudioMNIST), vocal sounds (VocalSound), environmental sound (ESC-10, ESC-50 (5cat), UrbanSound8K, TAU, FSD50K, AudioSet), and music (GTZAN, NSynth, IRMAS, MedleySolos); per-task class count, balance, and chance level are listed in Table 9. †FSD50K (200 cls, n=10,231) and AudioSet (632 cls, n=3,429 aligned) use 4-way MCQA with random distractors (chance 25%), since per-task K-way is impractical at that class count.

presses encoder latents into the LM's text-trained representation space. We test this hypothesis by tracking the recoverability of acoustic structure at every stage of the pipeline – encoder, projector, and LM – using three diagnostics with shared targets: (1) linear-probe accuracy for predictive recoverability, (2) a within-/across-class distance ratio for geometric preservation, and (3) LM-side adaptability experiments.

(1) Layer-wise linear probing. As shown in Figure 2, linear-probe accuracy is measured at taps spanning the pipeline: five encoder taps ($E0$–$E4$), four projector taps plus the projector output ($P0$–$P3$, P_o), and five LM taps ($L0$, $L8$, $L15$, $L23$, $L31$ of the 32-layer LM), with an identical probe fit at every tap. The encoder-to-projector transition consistently increases probe recoverability on non-ASR acoustic tasks rather than degrading it, and this gain is largely retained through LM depth instead of collapsing at later layers. Probe trajectories therefore remain substantially above the corresponding final-LM MCQA accuracy lines across datasets, indicating that discriminative acoustic structure survives well past the interface boundary.

(2) Within-/across-speaker geometry. To verify that this survival reflects the preserved *discriminative* structure rather than merely co-located features, Figure 3 reports two complementary distance ratios across encoder, projector, and LM stages on IEMOCAP Session 5, where speaker and transcript labels are controlled. For an utterance pair (i, j) with stage representations h_i, h_j, speaker labels s_i, s_j, and transcripts t_i, t_j, the mean cosine distance over a pair set P is

$$\overline{d}(P) = \frac{1}{|P|} \sum_{(i,j)\in P} 1 - \frac{h_i^\top h_j}{\|h_i\| \, \|h_j\|},$$

with the three pair sets defined as $P_{\text{spk}} = \{(i, j) : t_i = t_j,\, s_i \neq s_j\}$, $P_{\text{txt}} = \{(i, j) : s_i = s_j,\, t_i \neq t_j\}$, and $P_{\text{rand}} = \{(i, j) : s_i \neq s_j,\, t_i \neq t_j\}$. Figure 3(a) measures speaker variation relative to transcript variation,

$$r_{\text{spk/txt}} = \frac{\overline{d}(P_{\text{spk}})}{\overline{d}(P_{\text{txt}})},$$

and Figure 3(b) measures it relative to the random-pair baseline,

$$r_{\text{spk/rand}} = \frac{\overline{d}(P_{\text{spk}})}{\overline{d}(P_{\text{rand}})}.$$

In both metrics, higher values indicate stronger preservation of speaker-specific geometry relative to the comparison baseline. A value of $r_{\text{spk/txt}} = 1$ marks the balance point where speaker variation produces the same average distance as transcript variation, so the representation places its acoustic

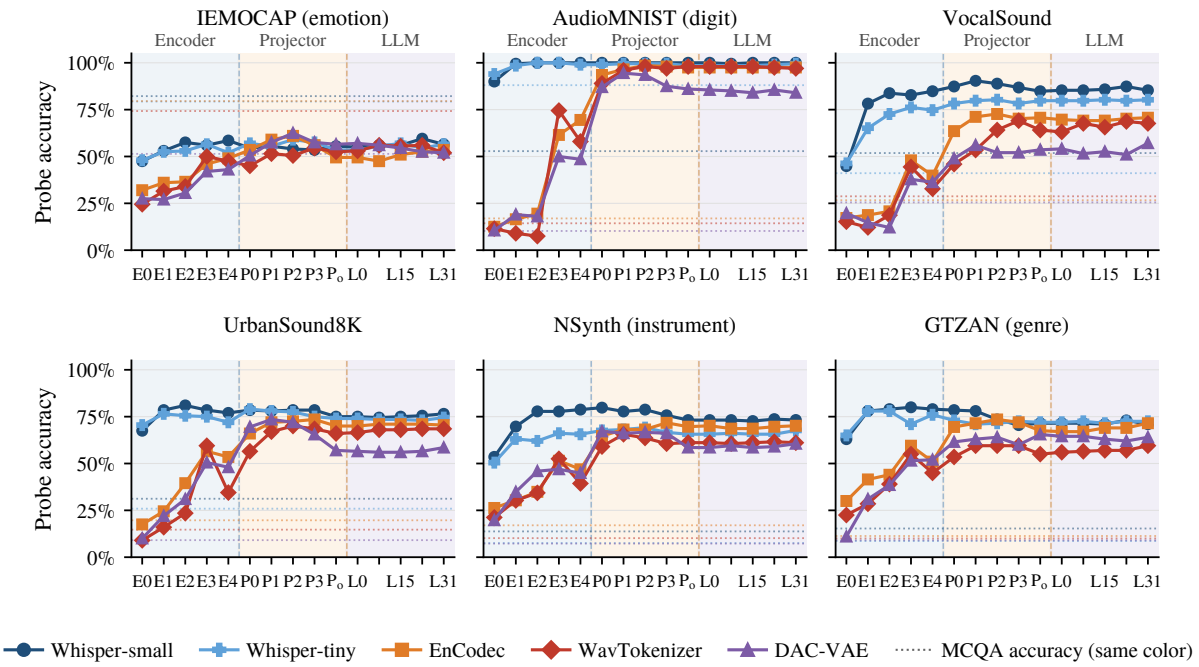

Figure 2: Layer-wise probe accuracy across encoder ($E0$–$E4$), projector ($P0$–$P3$, P_o), and LM ($L0$, $L8$, $L15$, $L23$, $L31$) stages. Dashed horizontal lines indicate final LM MCQA accuracy for the same encoder color.

and semantic axes on equal footing. Values below 1 indicate transcript-dominated geometry (the semantic axis stronger than the acoustic one), and values above 1 indicate the reverse. A value of $r_{\text{spk/rand}} = 1$ marks the point where pure speaker variation accounts for the full random-pair distance, so the representation treats two utterances of the same transcript by different speakers as no closer than a fully random pair. Values below 1 mean that controlling the transcript reduces the distance, so transcript variation contributes additional separation beyond speaker variation alone.

All observed values lie below 1, placing every encoder in the transcript-dominated regime, yet they stay well above the collapse range, so speaker identity remains preserved across the pipeline. Within this regime, reconstruction-based DAC-VAE starts highest at $r_{\text{spk/txt}} = 0.978$ at the encoder output and gradually compresses through the projector and LM to 0.882 at lm.out. ASR-based Whisper encoders start a little lower at the encoder output (0.841 for Whisper-Small, 0.848 for Whisper-Tiny), but still meaningful. These values rise in the projector, and settle near 0.89 through the LM (0.897 and 0.888 at lm.out, respectively), showing that even an ASR loss leaves enough speaker structure at the encoder output for the projector and LM to preserve and propagate through depth.

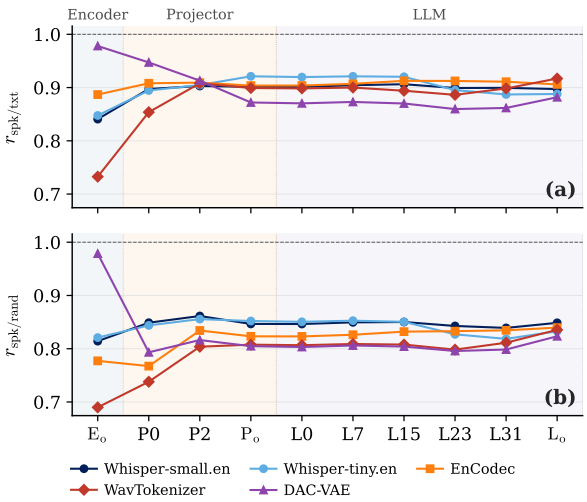

Figure 3: Per-stage distance ratios across the five encoders on IEMOCAP Session 5. **(a)** Across-/within-speaker distance ratio (same transcript, different speaker / same speaker, different transcript). **(b)** Across-speaker / random-pair distance ratio (same transcript, different speaker / different speaker, different transcript).

(3) Encoder-wise LoRA fine-tuning effect under headroom-scaled comparison. A complementary check is whether targeted LM modification can close the audio-task gap. We fine-tune the LM via LoRA (rank 32, α=64 on the seven linear projections per Transformer block, attention $q/k/v/o$ and MLP $gate/up/down$) on top of the frozen encoder and a co-fine-tuned projector. To isolate the LoRA

	ASR (WER ↓)			**Emotion** (macro-F1 ↑)			**Captioning** (CIDEr ↑)		
Encoder	align	FT	Δ	align	FT	Δ	align	FT	Δ
Whisper-Tiny	8.84	6.06	−2.78	0.444	0.431	−0.013	0.685	0.476	−0.209
EnCodec	20.80	10.65	**−10.15**	0.393	0.402	+0.009	0.259	0.408	**+0.149**
DAC-VAE	20.46	16.60	−3.86	0.400	0.382	−0.018	0.235	0.344	+0.109

Table 4: Family-level effect of LoRA fine-tuning (FT) over align, averaged within each task family using the same metric pool as Table 2. Δ is FT − align; for ASR, negative Δ is improvement. Per-task numbers and the best-by-rank-sum FT checkpoint per encoder are in Appendix Table 11.

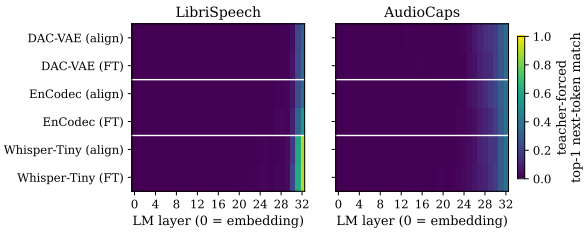

Figure 4: Layer-wise LogitLens heatmap of teacher-forced top-1 next-token match across the 32 LM blocks plus the input embedding ($x = 0$), for three encoders (Whisper-Tiny / EnCodec / DAC-VAE) in two conditions (align, fine-tuning).

contribution, we zero only the LoRA B matrices while keeping the projector intact and re-evaluate the nine downstream tasks.

Across encoders, the LoRA effect varies in both magnitude and sign (Table 4), with the unifying factor being the post-align starting headroom rather than any single LM-internal circuit. The encoder with the strongest align baseline (Whisper-Tiny) shows only modest ASR gains, near-flat emotion, and a captioning regression, while the encoders with weak align baselines (EnCodec, DAC-VAE) recover substantial captioning headroom and, for EnCodec, substantially reduce ASR WER. Per-encoder details are in Appendix C.

The same headroom-scaling pattern shows up in LM-internal decoding under a logit lens (Figure 4). On LibriSpeech, Whisper-Tiny's final-layer top-1 accuracy moves only from 0.873 align to 0.882 after fine-tuning (no headroom), whereas EnCodec moves from 0.331 to 0.520 (large headroom). The AudioCaps flatness across align → fine-tuning is consistent with the LM-internal trajectory not gaining additional captioning structure from LoRA. That the LibriSpeech shift concentrates at the last few LM layers and that neither task shows broad LM-internal change suggests that the gap concentrates in the readout-level alignment rather than in LM internal capacity.

All three probes converge on the same verdict. r stays below the equality line yet is far from collapse, probe accuracy remains substantially above final-LM MCQA accuracy at every depth, and LoRA fine-tuning leaves LM internal capacity untouched, concentrating its effect at the readout-level alignment. The encoder–projector–LM stack therefore does not erase acoustic information, shifting the failure hypothesis from *complete information loss* toward *under-utilization at the LM prediction stage* and motivating the readout-level analyses below.

4.3 The LM Readout, Not Its Representations, Drives MCQA Failure

If acoustic information survives the encoder–projector–LM stack, the remaining failure must occur when the LM converts hidden states into a discrete answer letter through the output projection W_U. A natural first hypothesis is that the LM still uses this surviving evidence at prediction time, so audio-conditioned MCQA accuracy should rise above a text-only baseline by an amount that tracks each encoder's audio-task strength.

(1) Audio effect. The *audio effect* tests this hypothesis: for each sample, a teacher-forced MCQA prompt of the form "`{question}\nChoices: A)` c_1 `. . . K)` c_K`\nAnswer with the letter.`" yields logits at the final prompt position, restricted to the candidate-letter tokens $\{A, \ldots, K\}$; *rank-1 letter accuracy* is the fraction of samples where the gold letter has the highest restricted logit, and the audio effect is the increase in this rank-1 accuracy between the audio-on condition and a text-only baseline that sets t_{audio}=0 while keeping the prompt text byte-identical (200-sample subset per task across six tasks, identical prompts across the five encoder families). As shown in Figure 5, a sharp split emerges: speech tasks (emotion, digits, vocal sounds) gain 27 to 74 pp, while music tasks (NSynth, GTZAN) show near-zero or negative gains; however, this split is consistent across

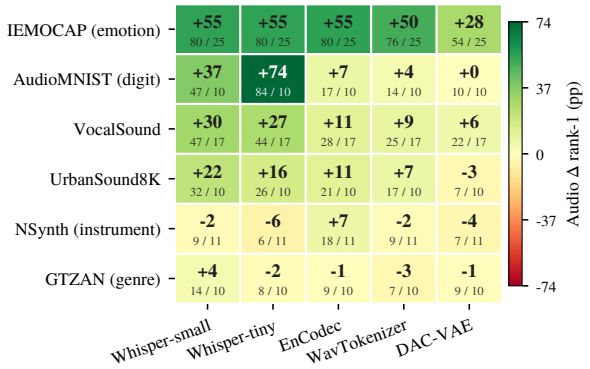

Figure 5: Rank-delta heatmap of the audio effect: the increase in rank-1 letter accuracy (%) when audio is present versus a text-only baseline. Rows are six tasks; columns are five encoders.

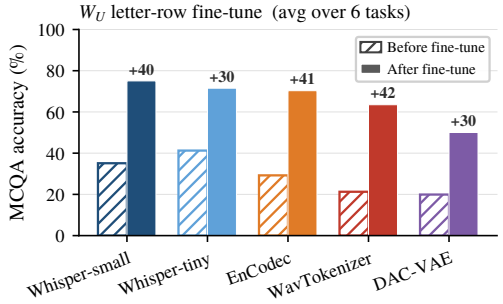

Figure 6: W_U letter-row surgical recovery, averaged over the six analysis tasks. Bars compare MCQA accuracy before and after fine-tuning only the n_c rows of W_U indexing the choice letters ($n_c \times 2560 \approx$ 10–26k parameters per task), across five encoders; the $+\Delta$ label above each After bar is the per-encoder absolute gain (mean across tasks).

encoders, arguing against a purely encoder-side attribution and pointing to the LM's answer mapping as the failure site.

(2) W_U letter-row surgical recovery. This answer-mapping failure points to the LM's output projection W_U, the layer that converts final hidden states into candidate-letter logits. To verify this causally, every weight is frozen except the n_c rows of W_U corresponding to the choice letters, and those rows are fine-tuned with cross-entropy on saved $L31$ asst-newline hidden states (80/20 stratified split, full-batch Adam, lr 10^{-3}, 200 steps; held-out test accuracy reported). This update of only $n_c \times 2560$ scalars per task ($\approx$ 10–26k, depending on the class count) recovers 30 to 42 pp of MCQA accuracy on average across the five encoders and six tasks (Figure 6; per-task breakdown in Appendix D.2). To characterize not only *that* W_U fails but also *how* it fails, we report LogitLens decoding in Appendix D.1.

MCQA failures are therefore driven by readout-level mismatch rather than absence of acoustic evidence: the LM contains usable acoustic structure at $L31$ (probes are accurate), W_U fails to route it toward the correct letter, and a tiny update to its letter rows recovers most of the MCQA gap.

5 Conclusion

We asked whether audio-LLMs' underuse of paralinguistic and non-speech information arises from their ASR-supervised encoders, which are often suspected of prioritizing transcription over acoustic detail, or from later stages of the encoder, projector, and LM pipeline. Replacing Whisper with reconstruction-based representations did not improve emotion recognition or sound captioning, arguing against the encoder as the sole source of the loss. Instead, discriminative acoustic structure remains linearly recoverable at the LM's final hidden states, well above multiple-choice question answering accuracy. The failure lies at the readout: fine-tuning a small subset of W_U rows recovers accuracy across all five encoders, while LoRA adaptation of the LM internals yields limited or headroom-dependent gains.

More broadly, our analysis shows the value of localizing where acoustic information is lost rather than assuming the encoder is at fault. Tracing information stage by stage separates what the model fails to encode from what it fails to use, revealing that the evidence is already present and that the limiting factor lies downstream. These findings carry a clear implication for audio-LLM design: effective design follows from diagnosis, by locating the stage that bottlenecks acoustic understanding and targeting it directly.

Limitations

While our analysis localizes the dominant failure mode of audio-conditioned MCQA, several dimensions warrant further investigation. Our experiments rely on a single base LM (Qwen3.5-4B), leaving the generalization of the readout-bottleneck pattern across LM families and scales open. Our stage-wise readout analysis is restricted to MCQA-format classification, and our speaker-geometry analysis is restricted to IEMOCAP Session 5; extending the same localization framework to generative tasks such as ASR and free-form sound captioning would give a fuller picture. Method-

ologically, probing establishes that acoustic information is *recoverable* but not that the model *uses* it, so we treat probe accuracy as an upper bound and rely on the causal W_U intervention to bridge the two; relatedly, the letter-row update is a post-hoc diagnostic rather than a deployment recipe, and turning it into a training-time method—e.g., answer-format-aware alignment or label-direction regularizers—is a natural next step. Finally, our use of off-the-shelf encoders makes the cross-encoder ranking an end-to-end comparison rather than a controlled attribution; future work should isolate capacity and pretraining differences and incorporate self-supervised encoders (HuBERT, WavLM, wav2vec 2.0) and semantic-aligned hybrid codecs (Mimi, SpeechTokenizer).

References

Tony Alex, Wish Suharitdamrong, Sara Atito, Armin Mustafa, Philip JB Jackson, Imran Razzak, and Muhammad Awais. 2025. Pal: Probing audio encoders via llms-audio information transfer into llms. *arXiv preprint arXiv:2506.10423*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

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 *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5848–5877, Rabat, Morocco. Association for Computational Linguistics.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2023. Beats: Audio pre-training with acoustic tokenizers. In *Proceedings of the International Conference on Machine Learning (ICML)*.

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

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*. ArXiv preprint posted 2023, hence "Dao23".

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Heinrich Dinkel, Jiahao Zhou, Guanbo Wang, Yadong Niu, Junbo Zhang, Yufeng Hao, Ying Liu, Ke Li, Wenwu Wang, Zhiyong Wu, and Jian Luan. 2026. The interspeech 2026 audio encoder capability challenge for large audio language models. *Preprint*, arXiv:2603.22728.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an audio captioning dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

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.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and 1 others. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*.

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 (ICLR)*.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, and 1 others. 2025. Wavtokenizer: An efficient acoustic discrete codec tokenizer for audio language modeling. In *International Conference on Learning Representations (ICLR)*.

Wonjune Kang, Junteng Jia, Chunyang Wu, Wei Zhou, Egor Lakomkin, Yashesh Gaur, Leda Sari, Suyoun Kim, Ke Li, Jay Mahadeokar, and Ozlem Kalinli. 2025. Frozen Large Language Models Can Perceive Paralinguistic Aspects of Speech. In *Interspeech 2025*, pages 4323–4327.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating captions for audios in the wild. In *NAACL-HLT*.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. 2023. LibriTTS-R: A restored multi-speaker text-to-speech corpus. In *Interspeech*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved RVQGAN. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):195:1–195:35.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and Xie Chen. 2024. An embarrassingly simple approach for LLM with strong ASR capacity. *arXiv preprint arXiv:2402.08846*.

OpenAI. 2023. Chat markup language (ChatML). `https://github.com/openai/openai-python/blob/release-v0.28.0/chatml.md`. OpenAI ChatML format documentation, openai-python repository.

OpenAI. 2024. GPT-4o.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, and 1 others. 2024. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *Interspeech*.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

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. PMLR.

Robin San Roman, Pierre Fernandez, Alexandre Défossez, Teddy Furon, Tuan Tran, and Hady El-Sahar. 2024. Proactive detection of voice cloning with localized watermarking. In *International Conference on Machine Learning (ICML)*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Noam Shazeer. 2020. GLU variants improve transformer. *Preprint*, arXiv:2002.05202.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568. Cited as "SP23" in source; arXiv preprint 2021, journal version 2024.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Thomas Thebaud, Yen-Ju Lu, Matthew Wiesner, Peter Viechnicki, and Najim Dehak. 2025. Enhancing dialogue annotation with speaker characteristics leveraging a frozen llm. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 112–118. Establishes the IEMOCAP 4-class evaluation with the excited→happy merge.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. In *Advances in Neural Information Processing Systems (NeurIPS)*. Cited as "ZL20" in source; canonical year is 2019.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net.

Appendix

A Implementation Details **12**
- A.1 Model 12
- A.2 Training 12

B Datasets and Evaluation Setup **13**
- B.1 Dataset Summary 13
- B.2 Per-Task Metadata 14
- B.3 Per-Benchmark Results 14

C LoRA Fine-Tuning: Per-Encoder Details **14**

D Readout-Level Diagnostics **15**
- D.1 LogitLens across All Tasks 15
- D.2 W_U Letter-Row Fine-Tune by Task 15
- D.3 Logit Processor Baseline: Vocabulary Restriction Alone Is Insufficient 16

A Implementation Details

A.1 Model

Pretrained checkpoints. Table 5 lists the Hugging Face repository identifiers of the pretrained audio encoders used in this work.

Projector. The projector has hidden size 512, 8 attention heads, and a feed-forward dimension of 2048. Inside the projector, an input linear projection $\mathbb{R}^{d_{\text{enc}}} \to \mathbb{R}^{512}$ feeds the four decoder layers, whose output is mapped to $\mathbb{R}^{2560}$ by an output linear projection, where d_{enc} is the encoder's latent dimensionality listed in Table 5. Each decoder layer uses pre-norm RMSNorm (Zhang and Sennrich, 2019), SwiGLU (Shazeer, 2020), rotary position encoding (Su et al., 2024), FlashAttention-2 causal attention (Dao, 2024), and bias-free linear projections throughout. The four decoder layers, the final RMSNorm, and the output projection are shared across all encoders and contribute ~ 18M parameters. Only the $d_{\text{enc}} \times 512$ input projection varies with the choice of encoder, leaving the total projector size at roughly 18M parameters across all configurations (about 0.45% of the LLM).

Language model. We obtain our LM from Qwen3.5-4B (Qwen Team, 2026) solely by extending the tokenizer with reserved audio control tokens that demarcate where projector outputs are spliced into the input. The transformer architecture and pretrained weights are inherited unchanged, so audio conditioning is injected at the input-embedding level rather than through cross-attention or in-network adapters.

A.2 Training

Per-modality corpora. The ASR portion combines GigaSpeech XL (50% subsample), LibriTTS-R (Koizumi et al., 2023), the English subset of Multilingual LibriSpeech (MLS) (Pratap et al., 2020), and the English portion of VoxPopuli (Wang et al., 2021). The environmental-sound portion aggregates the LAION-Audio-630k collections (Freesound, Epidemic, BBC, Audiostock), AudioCaps, FSD50K, the AudioSet balanced split, Clotho, and MACS.[1] The emotion portion combines DailyTalk, MELD, EmoVDB, IEMOCAP (Sessions 1–4; Session 5 is held out for evaluation by convention), RAVDESS, and MUStARD++.

Streaming pipeline. The three modality manifests are interleaved at the row level via Hugging Face `interleave_datasets` (Wolf et al., 2020) with the sampling weights in Table 6 and `stopping_strategy=all_exhausted`, so the emotion split cycles multiple times while the ASR split drains. Audio is capped at 30 seconds per utterance and resampled online to each encoder's native sampling rate (Table 1).

Sample construction. Each sample is wrapped in the ChatML template (OpenAI, 2023): a system prompt, a user turn containing $t_{\text{audio}} = \lfloor n_{\text{samples}}/\text{hop} \rfloor$ copies of the audio placeholder token `<|audio_pad|>` bracketed by `<|audio_start|>` and `<|audio_end|>` (Chu et al., 2024) (with the encoder-specific hop in Table 1), a modality-specific instruction (*e.g.,* "Transcribe the audio to text." for ASR, "Describe the audio." for sound captioning, "Identify the emotion." for emotion recognition), and an assistant turn containing the ground-truth response followed by the end-of-sequence token. At forward time the projector output replaces the `<|audio_pad|>` embeddings in the LLM input, producing an input

[1] Because these collections are independently curated from overlapping public sources, we cannot fully rule out intra-LAION overlap (*e.g.,* between Freesound, BBC, Epidemic, and Audiostock) or inter-corpus overlap among the other sound-captioning datasets. The LAION-Audio-630k authors documented concrete cross-corpus overlaps on GitHub. To mitigate any such leakage at evaluation time, we score only on the official held-out test/eval split of each benchmark.

Encoder	HF repository	SR (kHz)	Hop	Dim.	cutoff (tok)	Global batch	Rows / phase
Whisper-Small	`openai/whisper-small.en`	16	320	768	3584	32	∼24.4M
Whisper-Tiny	`openai/whisper-tiny.en`	16	320	384	3584	32	∼24.4M
EnCodec	`facebook/encodec_24khz`	24	320	128	4096	32	∼27.9M
WavTokenizer	`novateur/WavTokenizer-large-unify-40token`	24	600	512	3584	32	∼24.4M
DAC-VAE	`facebook/dacvae-watermarked`	48	1920	128	3584	16	∼12.2M

Table 5: Per-encoder spec, pretrained checkpoint, and training-time settings. Dim. is the encoder latent dimensionality (projector input). Global batch $= 8$ (GPUs) $\times$ per-device batch $\times$ gradient accumulation. Rows / phase is computed from $100{,}000$ steps, global batch, cutoff length, and the row-size weighted mean of ≈ 470 tokens per training row. EnCodec uses a larger cutoff (4096 tokens) because its 75 tokens/s rate fills $2{,}250$ audio frames for a 30-second utterance and exceeds 3584 once the ChatML wrapper is added; DAC-VAE uses a halved global batch (16) because its 48 kHz input carries roughly $3\times$ as many raw audio samples per utterance as the $16/24$ kHz encoders, forcing a smaller per-step batch under fixed GPU memory.

sequence whose audio-conditioned region has the same length as the projector output. The training objective is causal cross-entropy computed only on the response and EOS tokens, with all system, user, audio-placeholder, and instruction tokens marked by the ignore-index of -100 (Wolf et al., 2020).

Per-modality coverage. Applying the row-level sampling weights $(0.65/0.25/0.10)$ to the rows consumed in Table 5 gives the per-modality epoch counts. For DAC-VAE (the smallest global batch), $\approx$ 12.2M rows over the mixture translate to $\approx$ 0.51 epoch for the ASR portion, $\approx$ 4.41 epochs for environmental sound, and $\approx$ 24.3 epochs for emotion; the remaining encoders run at roughly twice this consumption. The ASR coverage is tight enough that some shards may not be drawn at all, while the emotion coverage is high enough that small subsets like RAVDESS ($1{,}440$ rows) are revisited many times within a single training run. This is a known coverage-versus-mixing trade-off of row-level interleaving with `stopping_strategy=all_exhausted`, and we leave per-modality early stopping or annealed sampling weights to future work.

Modality	Hours	Rows	Prob.	Epochs @ 100k
ASR (English)	50.8k	15.47M	0.65	0.51–1.17
Environmental sound	4.31k	691.8k	0.25	4.41–10.1
Emotion	51	50.28k	0.10	24.3–55.5
Total	55.1k	16.22M	1.00	—

Table 6: Multi-task training mixture. Hours are total source duration (raw, before the 30-second per-utterance cap). Probabilities are row-level sampling weights for `interleave_datasets`. Epochs at 100k is the per-modality exposure range across encoders (low end: DAC-VAE global batch 16; high end: EnCodec global batch 32 with cutoff 4096), computed as rows-consumed $\times$ prob / rows.

Table 7: Training hyperparameters. Only the peak learning rate differs between the audio-LM alignment phase (align) and the LoRA fine-tuning phase (FT). Per-encoder batch composition (per-device batch, gradient accumulation, effective global batch) is listed in Table 5.

Global	
Optimizer	AdamW (FusedAdam)
Peak LR	2×10^{-4} (align) / 2×10^{-5} (FT)
Weight decay	0.01
Gradient clip	1.0
LR schedule	warm-up, stable, decay
Warmup	1000 steps
Duration	100,000 steps
Precision	bfloat16
LoRA	
r / α / dropout	32 / 64 / 0.05
Modules	$\{q, k, v, o, \text{gate}, \text{up}, \text{down}\}_{\text{proj}}$

Training. We train in two phases, an audio-LM alignment phase (align) where only the projector is updated and a LoRA fine-tuning phase (FT) where LoRA adapters are inserted into the LLM. The two phases share all optimization hyperparameters except the peak learning rate (Table 7), and per-encoder batch composition is listed in Table 5.

B Datasets and Evaluation Setup

B.1 Dataset Summary

Table 8 summarizes the evaluation datasets used in the analyses (Sections 4.1–4.3), spanning speech emotion, speech content, vocal sounds, environmental sound, and music. For each dataset we list the per-class sample balance, the class set, and a short note describing any non-obvious filtering applied (*e.g.,* the 4-class IEMOCAP subset, the 5-supercategory collapse of ESC-50).

Dataset	Balance	Classes	Notes
Speech emotion			
IEMOCAP-S5	unbal 197–613	angry / happy / neutral / sad	
CREMA-D	unbal 40–51	angry / disgust / fearful / happy / neutral / sad	
Speech content			
Speech Commands	balanced (200/cls)	yes / no / up / down / left / right / on / off / stop / go	keyword spotting
AudioMNIST	balanced (200/cls)	digit_0 . . . digit_9	spoken digits
Vocal sounds			
VocalSound	balanced (200/cls)	cough / laughter / sigh / sneeze / sniff / throat_clearing	human vocalization without linguistic content
Environmental sound			
ESC-10	balanced (40/cls)	chainsaw / clock_tick / crackling_fire / crying_baby / dog / helicopter / rain / rooster / sea_waves / sneezing	low-resource (40 samples/class)
ESC-50 (5-cat)	balanced (400/cls)	animals / natural_soundscapes / human_sounds / indoor_sounds / outdoor_urban	original 50 classes collapsed into 5 super-categories
UrbanSound8K	balanced (100/cls)	air_conditioner / car_horn / children_playing / dog_bark / drilling / engine_idling / gun_shot / jackhammer / siren / street_music	
TAU Urban 2020	balanced (500/cls)	airport / bus / metro / metro_station / park / public_square / shopping_mall / street_pedestrian / street_traffic / tram	acoustic scene classification
FSD50K†	unbal (multi-label)	200 AudioSet-ontology tags	4-way MCQA with random distractors (n=10,231)
AudioSet (eval)†	unbal (multi-label)	632 AudioSet-ontology tags	4-way MCQA with random distractors (n=3,429 aligned)
Music			
GTZAN	unbal 99–100	blues / classical / country / disco / hiphop / jazz / metal / pop / reggae / rock	
NSynth	unbal 180–200	bass / brass / flute / guitar / keyboard / mallet / organ / reed / string	instrument family, vocal removed
IRMAS	balanced (200/cls)	acoustic_guitar / cello / clarinet / electric_guitar / flute / organ / piano / saxophone / trumpet / violin	predominant instrument in polyphonic mixtures, voice excluded
Medley-solos-DB	balanced (200/cls)	clarinet / distorted_electric_guitar / flute / piano / tenor_saxophone / trumpet / violin	isolated solo instruments, female singer excluded

Table 8: Dataset summary across speech, vocal, environmental sound, and music domains. †FSD50K and AudioSet use 4-way MCQA with random distractors per sample because K-way over the full class set is impractical.

B.2 Per-Task Metadata

The linear-probe vs MCQA results table is placed in the main text (Table 3, Section 4.1). Table 9 lists the per-task metadata used for that analysis (sample count, class count, balance flag, speech flag, median duration, chance accuracy).

B.3 Per-Benchmark Results

Evaluation details. FSD50K and AudioSet are evaluated as captioning tasks (prompt "Describe what you hear in the audio."), with BLEU and CIDEr scored against AudioSet-ontology descriptions of the ground-truth labels as multi-reference, rather than vocab-parse F1: caption-trained alignment-phase models emit free-form descriptions and a vocab parser yields artifact ≈ 0 that does not reflect actual capability. The same prompt and metric pipeline are used for AudioCaps and Clotho, so the four caption rows of Table 10 are directly comparable. MELD uses the official held-out test split. IEMOCAP Session 5 is a leakage-free cross-corpus emotion benchmark. The original IEMOCAP labels span ten emotions plus an unlabeled (xxx) bucket; we follow the conventional 4-class subset of anger, happy, sad, and neutral, dropping the other labels and merging the excited class into happy to expand the under-represented happy class (Yoon et al., 2018). After this filtering, 1,241 test utterances remain in Session 5.

C LoRA Fine-Tuning: Per-Encoder Details

This section expands the per-encoder case studies behind the summary in Table 4; per-task numbers are in Table 11. On fine-tuned Whisper-Tiny, where align was already strong on non-ASR audio tasks, the LoRA contribution stays close to zero on most non-ASR metrics yet FSD50K captioning regresses sharply, consistent with LoRA overfitting to the ASR-supervised path and pulling the model away from the already-strong align caption distribution. On fine-tuned EnCodec, where align had been weak across the board, LoRA produces substantial gains in the headroom-rich slots, while still failing to improve AudioCaps or Clotho, two captioning tasks on which all encoders share a uniformly low ceiling. Fine-tuned DAC-VAE shows the same pattern, with LoRA again hurting AudioCaps and Clotho captioning and yielding a large FSD50K recovery comparable in scale to EnCodec, while emotion changes remain small. Across all three encoders, FSD50K is the clearest single-task signature of the headroom-scaled view, the LoRA contribution flips from a large drop on the encoder with the strongest align baseline to large gains on the two encoders with the lowest baselines, tracking the available room above each encoder's align value.

Probe vs MCQA gap reduction (Whisper-Tiny). Table 12 contrasts the probe accuracy at $L31$ with the MCQA accuracy on Whisper-Tiny under align and fine-tuning, on six representative tasks. Probe

Task	N	Cls	Bal	Spch	Dur (s)	Chance
Speech emotion						
CREMA-D	264	6	no	yes	2.5	16.7%
IEMOCAP	2,000	4	yes	yes	4.5	25.0%
Speech content						
SpeechCmds	2,000	10	yes	yes	1.0	10.0%
AudioMNIST	2,000	10	yes	yes	0.7	10.0%
Vocal sounds						
VocalSound	1,200	6	yes	no	4.2	16.7%
Environmental sound						
ESC-10	400	10	yes	no	5.0	10.0%
ESC-50 (5cat)	2,000	5	yes	no	5.0	20.0%
UrbanSound8K	1,000	10	yes	no	≤ 4.0	10.0%
TAU	5,000	10	yes	no	10.0	10.0%
FSD50K†	10,231	200	no	no	≤ 10	25.0%
AudioSet†	3,429	632	no	no	10.0	25.0%
Music						
GTZAN	999	10	no	no	30.1	10.0%
NSynth	1,780	9	no	no	4.0	11.1%
IRMAS	2,000	10	yes	no	3.0	10.0%
MedleySolos	1,400	7	yes	no	3.0	14.3%

Table 9: Per-task metadata for the linear-probe vs MCQA evaluation in Table 3. N: number of test samples. Cls: number of classes. Bal: class-balanced (yes/no). Spch: speech content (yes/no). Dur: median utterance duration in seconds. Chance: uniform-prior chance accuracy. †FSD50K and AudioSet use 4-way MCQA with random distractors per sample (chance 25%) because K-way over the full class set is impractical.

accuracy at the final LM layer barely moves between align and fine-tuning (within ± 1.5 pp on five of six tasks, -3.1 pp on VocalSound), indicating that LoRA does not noticeably reshape the LM's internal acoustic representation. MCQA accuracy, by contrast, rises on every non-emotion task, with the largest absolute gains where the LM previously failed to convert latent acoustic evidence into the correct letter. As a result, the probe-to-MCQA gap shrinks across all five non-emotion tasks, most strongly on GTZAN and NSynth, with smaller reductions on VocalSound, AudioMNIST, and UrbanSound8K. IEMOCAP is the exception, where MCQA already exceeded the probe at $L31$ under align (Gap $= -23.0$), leaving little remaining gap for fine-tuning to close.

D Readout-Level Diagnostics

D.1 LogitLens across All Tasks

Figure 7 applies a LogitLens protocol across the six analysis tasks: we cache the assistant-newline hidden state at five depths ($L0, L8, L15, L23, L31$ of the 32-layer Qwen3.5-4B), apply the LM's own final RMSNorm followed by W_U, and read logits restricted to the candidate-letter tokens; the curves are per-encoder rank-1 accuracy, and the shaded band marks the range of encoder-stage probe accuracies at $E4$ across the five families.

The grid sharpens the speech–music split already visible in the rank-delta heatmap (Figure 5). On speech-centric tasks (IEMOCAP, AudioMNIST, and to a lesser extent VocalSound) the LogitLens climbs through the LM and saturates at $L31$ to a level comparable to the corresponding audio-on MCQA accuracy, confirming that the readout is faithful when the LM has actually learned to route acoustic evidence to the letter space. On music-centric tasks (NSynth, GTZAN) and the noisy environmental task UrbanSound8K, the trajectory stays flat near chance at every depth, even though the encoder-probe band sits well above (e.g. 38–97% for NSynth, 47–86% for GTZAN). The gap between the probe band and the saturated LogitLens curve at $L31$ is therefore a per-task estimate of how much of the available acoustic structure W_U fails to convert into a correct letter, and it is largest exactly on the tasks where the audio effect (Figure 5) is near zero or negative.

D.2 W_U Letter-Row Fine-Tune by Task

The main-text figure (Figure 6) reports W_U letter-row fine-tune recovery on 6 tasks, which gives a concise from $+30$ to $+42$ pp range but masks two things: (i) on tasks where the original readout is already aligned, the per-encoder recovery is near zero, and (ii) on the tasks where the readout is broken, the per-encoder recovery is substantially larger than the cross-task mean. Figure 8 breaks the same experiment down by task $\times$ encoder family. The protocol is identical to Section 4.3: encoder, projector, and all LM weights are frozen, the n_c rows of W_U corresponding to the choice letters are trained with cross-entropy on saved $L31$ assistant-newline hidden states for 200 full-batch Adam steps at lr 10^{-3}, with an 80/20 stratified split (test accuracy reported).

The per-task panels match the LogitLens grid (Figure 7) one-for-one. IEMOCAP-emotion sits in the regime where W_U is already approximately aligned with the class-mean directions: pre-fine-tune accuracy is already in the 55–78% band across encoders, and fine-tuning the letter rows changes accuracy by at most a couple of points (in some cases negatively, consistent with mild overfitting on $\sim$160 training samples). The remaining five tasks all show the readout-bottleneck regime where the letter rows are mis-

Benchmark	Metric	Whisper-Small	Whisper-Tiny	EnCodec	WavTokenizer	DAC-VAE
Encoder size	params	88.2 M	8.2 M	7.4 M	8.3 M	27.6 M
Representative ckpt	step	94,000	100,000	90,000	100,000	80,000
ASR						
LibriSpeech test-clean	WER↓	**2.49**	2.96	6.99	13.08	6.50
	CER↓	**0.97**	1.19	3.37	7.00	2.98
LibriSpeech test-other	WER↓	**6.19**	8.58	22.62	38.41	23.39
	CER↓	**2.96**	4.33	12.88	23.86	14.00
GigaSpeech test	WER↓	**12.47**	14.98	32.80	49.35	31.48
	CER↓	**7.13**	8.56	20.27	31.53	19.12
Emotion classification						
MELD	acc	0.483	**0.493**	0.431	0.411	0.439
	macro-F1	0.287	**0.299**	0.227	0.207	0.234
IEMOCAP-S5	acc	**0.633**	0.584	0.555	0.503	0.562
	macro-F1	**0.636**	0.589	0.558	0.505	0.566
Sound captioning						
FSD50K	BLEU-1	**0.715**	0.655	0.352	0.501	0.344
	BLEU-4	**0.621**	0.545	0.076	0.342	0.072
	CIDEr	**1.993**	1.620	0.047	0.787	0.050
Clotho	BLEU-1	0.441	0.426	**0.457**	0.338	0.446
	BLEU-4	**0.086**	0.071	0.074	0.037	0.072
	CIDEr	**0.186**	0.152	0.145	0.071	0.130
AudioCaps	BLEU-1	0.436	0.396	0.357	0.352	**0.504**
	BLEU-4	0.102	0.089	0.075	0.052	**0.123**
	CIDEr	**0.351**	0.292	0.254	0.141	0.282
AudioSet	BLEU-1	**0.515**	0.471	0.459	0.386	0.408
	BLEU-4	**0.336**	0.286	0.273	0.173	0.205
	CIDEr	**0.856**	0.674	0.588	0.302	0.479

Table 10: Per-benchmark encoder comparison. ASR reports WER / CER (%, lower is better). All other metrics are higher-is-better. **Bold** marks the per-row best, underline marks the second-best across encoders (computed only over evaluated cells). – = not evaluated. FSD50K and AudioSet are evaluated as captioning tasks (prompt *"Describe what you hear in the audio."*, BLEU / CIDEr scored against AudioSet-ontology descriptions of the ground-truth labels as multi-reference) rather than vocab-parse F1, because caption-trained Stage-1 models emit free-form descriptions and a vocab parser yields artifact ≈ 0 that does not reflect actual capability. The same prompt and metric pipeline are used for AudioCaps and Clotho, so the four caption rows are directly comparable. MELD uses the official held-out test split. IEMOCAP Session 5 is a leakage-free cross-corpus emotion benchmark (4-class, exc→hap merge, $n = 1241$).

aligned but a tiny surgical update repairs them. The largest task-specific deltas reach $+78$ pp (AudioMNIST/EnCodec), $+58$ pp (NSynth/Whisper-Small), and $+58$ pp (GTZAN/EnCodec), each measured against a base accuracy below 20% and recovered using only $n_c \times 2560$ scalars (between $\sim$10k and $\sim$26k depending on the task). Because the entire LM, projector, and encoder are frozen, these gains cannot reflect new representation learning; they directly quantify how much MCQA accuracy is held back by misalignment between class-mean directions and the letter rows of W_U.

D.3 Logit Processor Baseline: Vocabulary Restriction Alone Is Insufficient

The W_U letter-row fine-tune in Appendix D.2 recovers from $+30$ to $+78$ pp of MCQA accuracy, but this leaves open whether the gain comes from (i) actually re-aligning W_U to the class-mean directions or (ii) merely keeping the argmax inside the candidate-letter set $\{A, \ldots, K\}$, which is essentially equivalent to vocabulary masking at decode time. We separate the two effects by treating the vocabulary restriction as its own intervention. For every (task, encoder) combination we run greedy decoding under the HuggingFace `LogitsProcessor` API in three conditions: **(B0)** vanilla LM, argmax over the full 248,320-token vocabulary; **(B1)** same LM, argmax over the candidate-letter token set only (non-letter logits set to $-\infty$); and **(B2)** the W_U letter-row fine-tune of Appendix D.2.

Across all 30 combinations (5 encoders $\times$ 6 tasks), B1 changes B0 accuracy by an average of only $+0.23$ pp (max $+1.0$ pp on En-

Task	Metric	Whisper-Tiny (FT ckpt 30k)			EnCodec (FT ckpt 70k)			DAC-VAE (FT ckpt 40k)		
		align	FT	Δ	align	FT	Δ	align	FT	Δ
ASR (WER $\downarrow$; negative Δ is improvement)										
LibriSpeech test-clean	WER	2.96	2.62	-0.34	6.99	3.72	-3.27	6.50	5.29	-1.21
LibriSpeech test-other	WER	8.58	7.21	-1.37	22.62	13.28	-9.34	23.39	20.05	-3.34
GigaSpeech test	WER	14.98	8.34	-6.64	32.80	14.95	-17.85	31.48	24.48	-7.00
Emotion (macro-F1 $\uparrow$)										
MELD	macro-F1	0.299	0.265	-0.034	0.227	0.217	-0.010	0.234	0.213	-0.021
IEMOCAP-S5	macro-F1	0.589	0.596	$+0.007$	0.558	0.588	$+0.030$	0.566	0.551	-0.015
Captioning (CIDEr $\uparrow$)										
FSD50K	CIDEr	1.620	0.813	-0.807	0.047	0.631	$+0.584$	0.050	0.523	$+0.473$
Clotho	CIDEr	0.152	0.140	-0.012	0.145	0.111	-0.034	0.130	0.093	-0.037
AudioCaps	CIDEr	0.292	0.267	-0.025	0.254	0.222	-0.032	0.282	0.251	-0.031
AudioSet	CIDEr	0.674	0.684	$+0.010$	0.588	0.668	$+0.080$	0.479	0.507	$+0.028$

Table 11: Per-task align vs LoRA-fine-tuning (FT) comparison for the three encoders evaluated under FT. align values are reproduced from Table 10. FT values are evaluated at the per-encoder best-by-rank-sum checkpoint over the 9-metric pool: Whisper-Tiny ckpt 30,000, EnCodec ckpt 70,000, DAC-VAE ckpt 40,000. Family averages of these per-task numbers appear in Table 4.

Task	align			FT			
	Probe	MCQA	Gap	Probe	MCQA	Gap	ΔGap
IEMOCAP	56.5	79.5	-23.0	56.0	77.0	-21.0	$+2.0$
AudioMNIST	99.5	84.0	$+15.5$	100.0	91.5	$+8.5$	$\mathbf{-7.0}$
VocalSound	80.3	43.9	$+36.3$	77.2	48.5	$+28.7$	$\mathbf{-7.6}$
UrbanSound8K	74.5	26.0	$+48.5$	76.0	32.0	$+44.0$	$\mathbf{-4.5}$
GTZAN	72.5	8.5	$+64.0$	71.0	20.5	$+50.5$	$\mathbf{-13.5}$
NSynth	67.7	5.6	$+62.1$	69.2	17.2	$+52.0$	$\mathbf{-10.1}$

Table 12: Probe at $L31$ vs MCQA accuracy (%) for Whisper-Tiny under align and fine-tuning, with the probe-to-MCQA gap (Probe $-$ MCQA) and the change in gap (ΔGap $=$ Gap$_{\text{FT}}-$Gap$_{\text{align}}$). Negative ΔGap (bold) marks gap reduction.

Codec/GTZAN); the model's emission rate to candidate-letter tokens averages 98.5% and never falls below 94%, so the vocabulary mask is effectively redundant. Only the W_U letter-row update closes the readout gap (Figure 9). This rules out the possibility that the surgical recovery merely reflects "forcing a letter answer"—the failure is letter-vs-letter ranking *inside* the candidate set, not non-letter token emission.

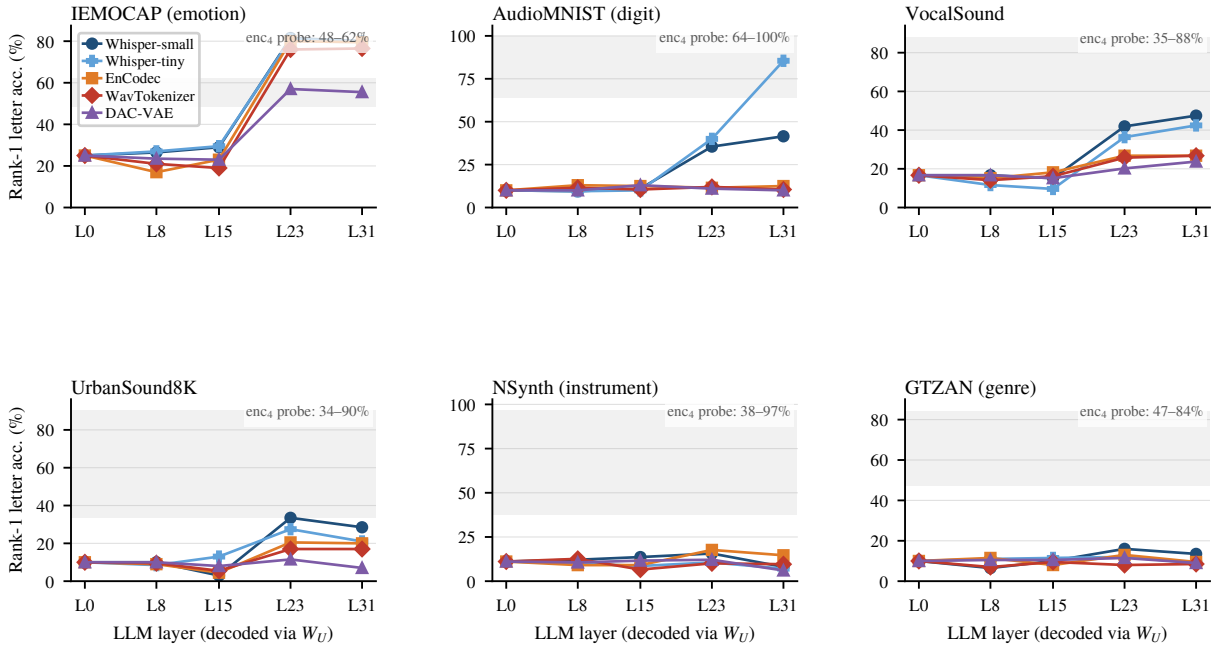

Figure 7: LogitLens rank-1 letter accuracy across all six tasks, decoded at five LM depths ($L0$, $L8$, $L15$, $L23$, $L31$ of the 32-layer Qwen3.5-4B) via final RMSNorm $+ W_U$ and restricted to the candidate-letter tokens. Each line is one encoder family; the grey band marks the range of $E4$ encoder-probe accuracies across the five families for that task. Speech tasks reach a high $L31$ accuracy that tracks audio-on MCQA, while music tasks and UrbanSound8K stay near chance at every depth despite high probe accuracy, isolating the readout as the bottleneck.

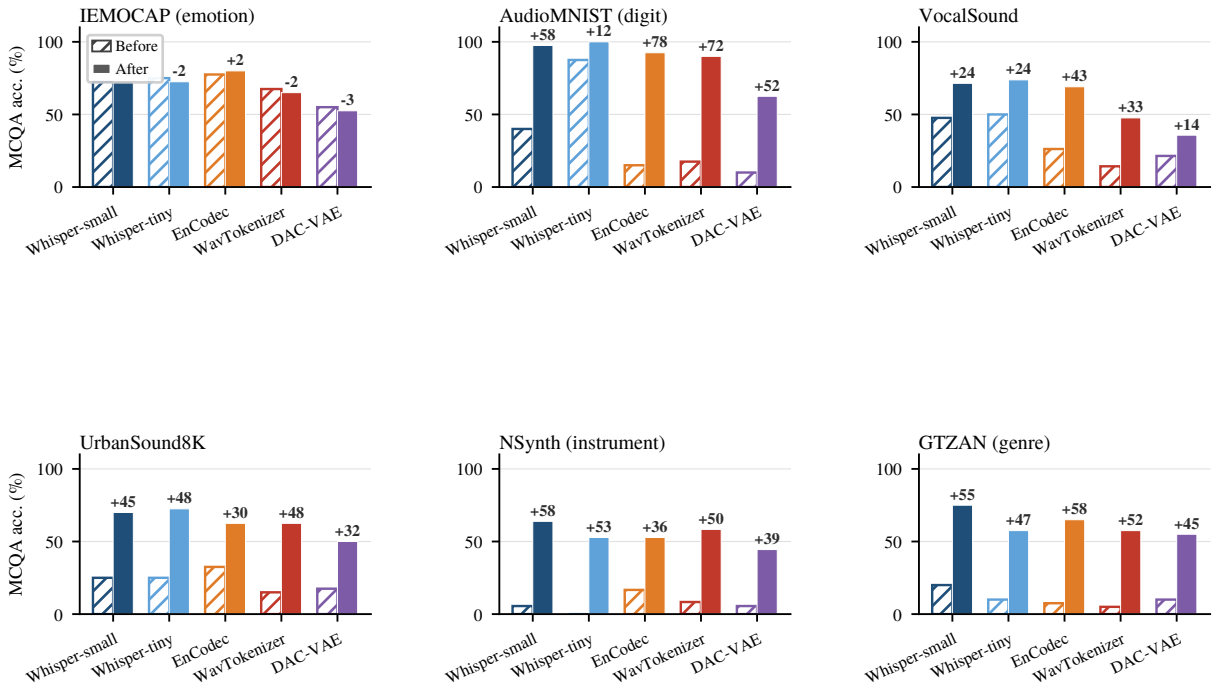

Figure 8: Per-task W_U letter-row fine-tune recovery, the task $\times$ encoder breakdown of Figure 6. Hatched bar: MCQA accuracy before fine-tune (equivalent to the rank-1 letter accuracy from final-layer LogitLens). Filled bar: MCQA accuracy after fine-tuning only the n_c rows of W_U that index the choice letters, with everything else frozen; the $\pm$ label above each After bar is the absolute change in pp. IEMOCAP shows near-zero recovery because the readout is already aligned; the other five tasks show large recoveries up to $+78$ pp, with the gain pattern across encoders mirroring the LogitLens curves in Figure 7.

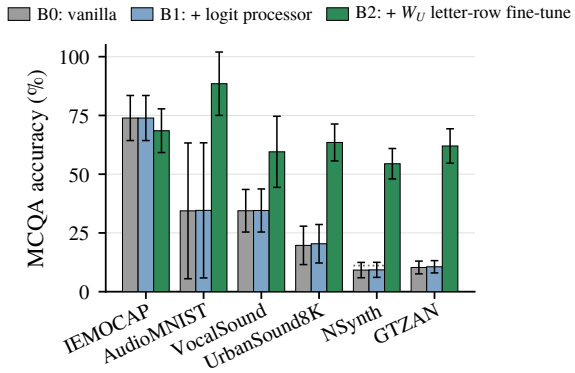

Figure 9: MCQA accuracy across three readout interventions, averaged over five encoders. **B0**: vanilla LM, full-vocab argmax. **B1**: + Logit processor restricting argmax to $\{A, \ldots, K\}$. **B2**: + W_U letter-row fine-tune (Appendix D.2). Error bars: ± 1 SD across encoders. Dotted segment over each group: per-task chance. Across 30 cells, B1 changes B0 accuracy by only $+0.23$ pp on average, while B2 recovers $+25$–$+54$ pp on the five readout-bottleneck tasks; IEMOCAP shows a small negative B2 effect (mild overfitting on the ~ 162 training samples; its readout is already approximately aligned).