Subliminal Prosody Learning: Auxiliary Emotion Supervision Redistributes Affective Representations Across ALM Layers

Bhuvan Koduru [1] **Dareen Safar B Alharthi** [1] **Rita Singh** [1] **Bhiksha Raj** [1]

Abstract

We study how a simple emotion classification objective applied to a few LoRA-adapted layers of an audio-language model (ALM) redistributes affective information across *all* layers, including those whose parameters remain frozen, through residual-stream propagation. We call this *subliminal prosody learning* and provide the first systematic study of representational propagation across three ALM architectures (Qwen2.5-Omni-7B, Audio Flamingo 3, MOSS-Audio-4B): mean probe gain in unadapted layers is $+32.0$pp (Omni), $+22.0$pp (AF3), $+13.6$pp (MOSS), and learned emotion directions recover the Russell circumplex and transfer cross-modally. Whether this representational change also improves generation *behavior* is a separate question, since linear decodability does not imply functional use. Two LLM judges reported a significant, emotion-selective improvement in Omni's generated responses, but blinded human evaluation on the same responses, using the identical rubric, found the opposite effect, significantly and uniformly across emotions—which we trace to two concrete judge biases identified by hand-inspection. A necessity test further shows the learned subspace is not required for the (LLM-judged) behavioral effect, indicating a more distributed encoding than sufficiency alone suggests. We report this as a cautionary, instrumented case study: the representational finding is robust, but the behavioral transfer it was hypothesized to enable does not survive human evaluation, and LLM-judge-only behavioral evaluation warrants corresponding skepticism.

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA. Correspondence to: Bhuvan Koduru <bkoduru@andrew.cmu.edu>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

Recent audio-language models (ALMs) such as Qwen2.5-Omni (Qwen, 2025), SALMONN (Chen et al., 2024), and SALMONN-Omni (Tang et al., 2024) process raw audio alongside text, achieving strong performance on speech recognition, audio question answering, and spoken dialogue. These models ingest speech through a dedicated audio encoder (typically a Whisper-class model (Radford et al., 2023)) projected into the token space of an LLM backbone. A natural question arises: *do these models actually use the rich prosodic information (pitch, tempo, vocal quality, emotional coloring) present in the audio input?* Prior work on speech emotion recognition (SER) has shown that paralinguistic features carry substantial affective information (Latif et al., 2023), yet the generation side of current ALMs is largely insensitive to speaking style. When given an angry utterance and a neutral one containing the same words, current ALMs produce near-identical text responses. The prosodic signal appears to be *encoded but not leveraged*.

The gap can be closed without rebuilding these models. Two architectural properties are exploitable: LLM hidden states carry partial emotional signal even before fine-tuning, and weight updates in any layer propagate globally through the residual stream, so a classification objective targeting a subset of layers should redistribute representations beyond them, including in layers whose parameters remain frozen. We exploit both by applying an emotion classification objective to a LoRA-adapted layer subset (Hu et al., 2022), with KL-divergence regularization against the frozen base model to stabilize generation quality during adaptation.

The question is whether the resulting reorganization is strong enough to reach generation behavior, and whether it produces interpretable geometric structure. Standard fine-tuning targets generation directly; post-hoc mechanistic analysis studies existing structure. We instead *deliberately induce* structured affective geometry as an implicit byproduct of classification supervision. We call this **subliminal prosody learning**: representational changes arising indirectly from auxiliary classification supervision that measurably influence downstream generation behavior without being explicitly targeted by the training objective, analogous, loosely, to subliminal stimuli that shape behavior without being the target of conscious processing (Dehaene

& Changeux, 2011). We use "subliminal" as a descriptive label for this indirect-supervision pattern rather than a claim of mechanistic equivalence to the cognitive-science phenomenon (Appendix D returns to this terminological caveat). Concurrent work (Anthropic, 2025) uses "subliminal learning" for a different, unrelated phenomenon: covert behavioral transmission *between* models through generated data, rather than the intra-model representational propagation through the residual stream that we study here; Appendix D discusses the distinction further.

Objective. We study empirically how auxiliary emotional supervision propagates through ALM representations and whether the redistribution influences generation. Concretely: when a classification objective is applied to only a subset of layers, does the resulting representational change remain local, or does it propagate throughout the model, and if it does, is the redistributed signal strong enough to bias autoregressive decoding, and can we trust the instrument we use to answer that second question? Linear decodability measures representational *accessibility*; whether that accessibility is sufficient to influence generation *in a way that a human would recognize as improved* is a separate, and as we show, easy-to-get-wrong empirical question. We pursue this at three levels of evidence: *within a single architecture* (where the representational effect is strongest), *across three independent ALM architectures* (to assess robustness), and *at the behavioral generation level, cross-validated against blinded human ratings* (to ask whether representation is enough for function, and whether our own instrument for measuring "function" can be trusted).

Core findings. The reorganization propagates globally and produces geometrically structured affective subspaces. This part of the story is robust and does not depend on any downstream generation metric. Whether that reorganization translates into generation behavior a person would recognize as improved is a separate question, and here our results are a cautionary tale rather than a clean positive: LLM judges report a significant, emotion-selective behavioral improvement, but this does not survive blinded human evaluation, which finds the opposite effect, significantly, and uniformly across emotions. We identify concrete, hand-verified mechanisms behind the discrepancy. Our contributions:

(1) *Subliminal propagation*: a single classification auxiliary reorganizes representations well beyond the trained layers: mean probe gain in unadapted layers is $+32.0$pp (Omni), $+22.0$pp (AF3), $+13.6$pp (MOSS) (§4.2). This finding does not depend on generation quality and is unaffected by anything below.

(2) *Structured affective geometry*: learned directions are geometrically consistent with the Russell circumplex, transfer cross-modally, and are causally *sufficient* to shift generation via activation steering (qualitative examples in Appendix E; §4.2), though a projection-ablation necessity test (§5) shows the effect is not *necessary* at the single layer we probed, indicating a more distributed encoding than steering alone suggests.

(3) *Cross-architecture replication of the representational (not behavioral) pattern*: the propagation pattern holds across three ALMs with fundamentally different audio encoders, LLM backbones, and integration mechanisms (§4.5).

(4) *LLM-judge and human evaluation of the same generations disagree sharply*: on Qwen2.5-Omni-7B, two independent LLM judges score fine-tuned generations as more empathetic than baseline, emotion-selectively (overall $\Delta = +0.35$, $p<0.001$; gains concentrated in emotional vs. neutral speech), while blinded human ratings on the identical stratified 90-item sample, using the identical rubric, find the *opposite* effect ($\Delta = -0.50$, Wilcoxon $p=0.021$), uniform across emotion categories rather than emotion-selective. We hand-identify two mechanisms driving this divergence—near-verbatim transcript echoing (42% of fine-tuned outputs) and an LLM-judge preference for warmth-coded phrasing over substantive content—and show the human rater agrees with the LLM judges specifically on the minority of cases that are genuine, content-engaged improvements (§4.3).

Limitations. The cross-architecture comparison is an existence proof rather than a controlled ablation: capacity, injection mechanism, and backbone all covary across the three systems. Behavioral evaluation uses read speech (Expresso). Human evaluation covers one architecture (Omni) and one rater on a 90-item stratified sample rather than the full 1160-sample evaluation set or all three architectures, and we report this as a single-rater result rather than overstate its scope. Training data is English-only. Activation steering destabilizes at high α, reflecting the discriminative–generative gap discussed in §5.

The broader implication is that representational structure in ALMs can be *deliberately shaped*, not merely discovered post-hoc. Lightweight auxiliary objectives can redistribute internal representations in ways that influence generation behavior beyond their stated target. Understanding how auxiliary gradients propagate through the residual stream is relevant for any practitioner adapting multimodal models, and for safety researchers asking what a fine-tuning run changes beyond its explicit objective. We situate this work relative to subliminal/implicit learning, ALM architectures, SER, and mechanistic interpretability in Appendix D.

2. Method

2.1. Base Models

We select three architectures that differ maximally in audio-integration design (input-level projection, LLaVA-

style MLP bridging, and cross-layer injection) to test whether the phenomenon is paradigm-level rather than architecture-specific.

Qwen2.5-Omni-7B. Our primary model (Qwen, 2025) consists of: a Whisper-large-v3 audio encoder (Radford et al., 2023) that produces frame-level representations from raw audio; a linear projector that maps audio features into the LLM token space; a 28-layer **Thinker** decoder that jointly processes audio and text tokens using TMRoPE (Token-type-aware Multirotary Position Embedding); and a Talker module for optional speech synthesis. We freeze the audio encoder and Talker throughout all experiments, modifying only the Thinker decoder.

Audio Flamingo 3. Audio Flamingo 3 uses a Qwen2.5-7B LLM backbone (28 transformer layers, hidden dimension 3584) connected to an audio encoder via a LLaVA-style MLP projector. The architecture processes audio as a sequence of patch-level embeddings projected into the LLM token space. We apply the same training procedure as for Qwen2.5-Omni, adapting layers 0–14 with LoRA and freezing the audio encoder and projector.

MOSS-Audio-4B. MOSS-Audio-4B is built on a Qwen3-4B LLM backbone (36 transformer layers, hidden dimension 2560) with a DeepStack cross-layer audio injection mechanism and a Whisper encoder (Radford et al., 2023). Audio features are injected at multiple LLM layers rather than only at the input, providing deeper audio–text integration. We adapt layers 0–16 with LoRA (the A3-equivalent for a 36-layer model; unadapted layers are 17–35). We run only A0 and A3 for MOSS due to compute constraints; the λ sweep and depth ablations (A1/A2/A4) are Omni and AF3 only, as noted in Table 1.

2.2. Training Design

The inference objective is standard audio-language generation: given an audio input and a text prompt, produce a coherent, contextually appropriate text response. The classification head described below is a *training device only*: it is discarded at inference. Its purpose is to apply an emotion-classification gradient to the early LLM layers during training, redistributing affective information into the residual stream. Whether this subspace then influences the generation pathway, which the classification head never directly touches, is the central empirical question.

Classification Head. We attach a 2-layer MLP head to the mean-pooled hidden states from the first 9 LLM layers ($\ell = 0, \ldots, 8$), the region where audio features are most directly integrated with text representations and where baseline probes already show the highest emotion decodability:

$$h_{\text{pool}} = \frac{1}{T}\sum_{t=1}^{T} h_t, \qquad \hat{y} = \text{MLP}(h_{\text{pool}}), \tag{1}$$

where T is the sequence length and h_t is the hidden state at position t. The MLP uses a hidden dimension of 1024 with GELU activation and dropout of 0.1.

Stage 1: Head Warmup. We freeze the entire LLM backbone and train only the MLP head for 1 epoch using cross-entropy loss. This establishes a reasonable classification head before modifying the backbone.

Stage 2: LoRA Fine-tuning with KL Regularization. We interpret the KL term as doing more than preventing forgetting: it constrains fine-tuned outputs to remain near the generative manifold of the base model, stabilizing generation quality while permitting early-layer representational adaptation. Empirically, reducing λ to 0.1 already shows training instability, consistent with this interpretation, though a direct measurement of off-manifold deviation is left to future work. We instantiate this with LoRA (Hu et al., 2022) adapters (r=16, α=32, dropout=0.05) on query and value projections of a subset of LLM layers, optimizing:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{CE}}(\hat{y}, y)}_{\text{emotion classification}} + \lambda \cdot \underbrace{D_{\text{KL}}(p_\theta(x) \,\|\, p_{\theta_0}(x))}_{\text{generation stability}}, \tag{2}$$

where p_θ and p_{θ_0} are the token-level output distributions of the fine-tuned and frozen base models, and λ controls regularization strength. Training runs for 3 epochs with AdamW (lr $= 2 \times 10^{-4}$, linear warmup over 100 steps).

2.3. Ablation Conditions

We define five conditions to isolate the contribution of each component:

Additionally, for A3 we sweep $\lambda \in \{0.1, 0.5, 1.0\}$ to study the effect of regularization strength.

3. Experimental Setup

3.1. Datasets

We train on a unified dataset of 26,190 samples drawn from three sources, mapped to six emotion classes (neutral, happy, angry, sad, fear, disgust). Disgust is present in the label space via ESD and CREMA-D but has no representation in Expresso; disgust F1 is 0.000 in all conditions and effective evaluation is five-class (detailed in §4.1).

- **Expresso** (Nguyen et al., 2023): Parallel-text expressive speech from 4 speakers. For training, we use 7 styles: `whisper`, `confused`, `enunciated`, `default` → *neutral*; `laughing`, `happy` → *happy*; `sad` → *sad*.
- **ESD** (Zhou et al., 2022): Emotional speech in 5 emotions from 10 speakers.

Table 1. Ablation conditions. LoRA layers shown for Omni/AF3 (28-layer); MOSS uses 0–16. A1/A2/A4 are Omni and AF3 only.

ID	Description	LoRA Layers (Omni/AF3)	λ
A0	Base model (no fine-tuning)	—	—
A1	Stage 1 only (head warmup)	—	—
A2	Stage 1 + Stage 2	0–7	0.5
A3	Stage 1 + Stage 2 (primary)	0–14	0.5
A4	Stage 1 + Stage 2	0–27	0.5

- **CREMA-D** (Cao et al., 2014): Crowd-sourced emotional speech from 91 actors in 6 emotions.

The Expresso validation set (2,910 samples) serves as the in-distribution evaluation. It draws on 3 additional native Expresso styles absent from the training mapping above: `neutral` → *neutral*, `angry` → *angry*, and `surprise` → *fear* (350 samples each), alongside the same neutral/happy/sad styles used in training. Consequently, the model never sees Expresso-style angry or fear audio during training; the angry and fear F1 scores reported on this validation set (§4.1, Appendix A) reflect generalization from ESD/CREMA-D training exposure to Expresso's recording conditions and speakers, not in-style supervision. For out-of-distribution evaluation, we use **IEMOCAP** (Busso et al., 2008) (2,212 samples), which differs substantially in recording conditions, speaker demographics, and emotion elicitation method (improvised and scripted dialogues vs. read speech).

3.2. Evaluation Protocol

Our evaluation spans three levels of analysis:

Part 1: Classification (§4.1). Weighted accuracy on Expresso val (in-distribution) and IEMOCAP (OOD). Probes use LOSO cross-validation on Expresso.

Part 2: Behavioral (§4.3). We present each Expresso clip to the model with "Please respond to what you heard" (no transcript) and score empathic quality using Gemini-2.0-Flash (Zheng et al., 2023) with an EPITOME-adapted rubric (Sharma et al., 2020). Expresso's same-transcript, different-style pairs causally isolate prosody from lexical content. Significance via paired Wilcoxon (N=1160). IEMOCAP lacks style-controlled pairs and is used for classification only.

Part 3: Representational (§4.2). We probe representations using six methods: LOSO linear probes at every layer; layer-wise CKA (Kornblith et al., 2019); difference-in-means emotion directions ($\vec{d}_e = \mu_e - \mu_{\text{neutral}}$, top-$k$ PCA denoised); Russell circumplex recovery via PCA correlated with Warriner norms (Warriner et al., 2013); audio–text cross-modal alignment; and causal activation steering (Turner et al., 2023).

4. Results

4.1. Classification Performance

Table 2 summarizes classification accuracy across all ablation conditions for both Qwen2.5-Omni-7B and Audio Flamingo 3.

Table 2 shows three patterns. First, moderate LoRA depth (A2/A3) maximizes OOD transfer: A2 reaches 52.3% on IEMOCAP (+14.0pp over A0) for Omni; A3 reaches 67.1% (+23.4pp) for AF3. Deeper adaptation (A4) overfits in-distribution. Second, λ=0.5 provides the best trade-off: λ=0.1 causes training instability while λ=0.5 and 1.0 converge smoothly (see Appendix B for loss curves and Appendix A for per-class F1). Third, MOSS shows the weakest effect: OOD accuracy decreases slightly (52.9%→49.4%), consistent with MOSS being below the threshold at which auxiliary supervision yields a net classification benefit, the same gradient that explains its smaller probe gains. Disgust F1 is 0.000 in all conditions (no Expresso disgust samples); effective evaluation is five-class (Appendix A).

4.2. Representational Analysis

Our analyses are representational rather than circuit-level: we study how affective information becomes redistributed across layers, leaving the precise computational pathways implementing emotion-sensitive generation to future work.

Emotion subspaces emerge from noise. UMAP projections of LLM hidden states (Figure 2, Omni column) show that baseline representations overlap substantially; after fine-tuning, distinct clusters emerge for neutral, happy, and sad, demonstrating that representations have reorganized into an emotion-discriminative subspace.

Linear decodability propagates beyond LoRA layers. Figure 1 (Omni column) shows LOSO probe accuracy across all LLM layers. Fine-tuning dramatically improves emotion decodability beyond the LoRA layers (0–14): unadapted layers 15–27 gain mean $\Delta = +32.0$pp. That some propagation occurs is architecturally guaranteed by the residual stream: any change to layer k must alter the inputs to layer k+1. What we quantify is the *magnitude* of that propagation (+32.0pp in layers receiving no gradient), its *geometric structure* (Russell circumplex, cross-modal alignment), and its *functional consequence* (behavioral transfer),

Table 2. Weighted classification accuracy (%). **Bold** = best per architecture per split. MOSS uses A0/A3 only.

		A0 (Base)	A1 (Head)	A2 (LoRA 0–7)	A3 (LoRA 0–14)	A4 (LoRA 0–27)
Qwen2.5-Omni-7B	Expresso (val)	87.1	79.2	94.0	94.5	**95.1**
	IEMOCAP (OOD)	38.3	47.4	**52.3**	51.6	46.2
Audio Flamingo 3	Expresso (val)	**91.7**	83.6	84.1	89.7	83.8
	IEMOCAP (OOD)	43.7	59.4	59.7	**67.1**	59.7
MOSS-Audio-4B	Expresso (val)	82.1	—	—	**86.0**	—
	IEMOCAP (OOD)	**52.9**	—	—	49.4	—

empirical questions distinct from the architectural fact of propagation. The classification objective makes emotion information more linearly decodable in early layers; this redistributed signal propagates through subsequent layers.

Peak probe accuracy reaches 98.3% for Qwen2.5-Omni (layer 8) and 89.4% for AF3 (layer 17), vs. baseline peaks of 75.9% and 69.4%.

Why residual connections alone do not explain this. In A1 the backbone is frozen, so no layer's probe can change; the +32pp unadapted gain in A3 requires upstream rewriting, not passive transmission.

Magnitude bound (illustrative). A rank-r LoRA update at layer ℓ_0 produces perturbation $\boldsymbol{\delta}_{\ell_0} = BA\mathbf{h}_{\ell_0-1}$; successive residual sublayers propagate it with worst-case growth $(1+L)^k$ for L-Lipschitz f_ℓ. This bound is loose and should be read as intuition: the practical stabilizer is the KL regularizer, which bounds output deviation empirically while preserving generation coherence.

Fine-tuning causes targeted representational change. CKA analysis (Appendix C, Figure 8) shows the largest representational changes in early layers (0–5) and final output layers, while middle layers remain well-preserved (CKA > 0.8). CKA and probe accuracy are measuring related but distinct things: CKA captures global representational shift, while probes detect formation of a low-dimensional discriminative subspace. AF3 shows a larger CKA drop than Omni in layers 7–9 yet lower probe gain, consistent with AF3 undergoing broader but less *structured* representational change, while Omni's smaller global shift concentrates into a more linearly separable emotion subspace.

The Russell circumplex emerges from learned directions. Difference-in-means emotion directions projected to 2D via PCA show geometric structure consistent with the Russell circumplex (Russell, 1980): PC1 aligns with valence (Pearson $r = 0.62$) and PC2 with arousal ($r = 0.45$). We report this as a qualitative geometric consistency check only: n=5 non-degenerate emotions (disgust excluded: no Expresso validation samples, F1=0.000 in all conditions) is too small for inferential claims. Opposing emotions also show low pairwise cosine similarity. Full plots are in Appendix C (Figures 7a–7b).

Cross-modal alignment. Audio-derived emotion directions align with text-derived directions in deeper LLM layers (Appendix C, Figure 9), suggesting a shared, modality-agnostic subspace; alignment is strongest in the LoRA region and increases with depth.

Causal steering. Adding $\alpha \cdot \vec{d}_e$ to hidden states at inference time shifts responses toward the target emotion at α=0.5–1.0 (examples in Appendix E); generation destabilizes at α=2.0. The directional sensitivity confirms the learned subspace is not a probe artifact: it is causally accessible at inference time.

4.3. Behavioral Evaluation

Linear decodability measures *accessibility*, not whether that accessibility is strong enough to bias autoregressive decoding. We test this with emotional Expresso audio and prompt "Please respond to what you heard" (no transcript). Two independent LLM judges (Gemini-2.0-Flash and GPT-4o-mini (Zheng et al., 2023)) score empathic quality on a 1–5 EPITOME-adapted rubric across 1,160 samples per architecture, blind to condition, using a 4-way rubric call that scores all architectures and conditions simultaneously. Same-transcript, different-style pairs causally isolate prosody. Absolute scores are low (1.3–1.9/5) because read speech is a minimal empathy context.

Table 3 reports what the two LLM judges scored: Wilcoxon signed-rank tests ($N = 1160$ per architecture) together with Cohen's d for the paired A3$-$A0 difference, with significance corrected for multiple comparisons (Benjamini–Hochberg FDR across all 20 tests in the table, $\alpha = 0.05$); 95% bootstrap confidence intervals for every cell (10,000 resamples) are given in Appendix A, Table 7. For Omni, A3 is scored 1.67 vs. 1.31 for baseline (overall $\Delta = +0.35$, $d = 0.33$, $p_{\mathrm{FDR}} < 0.001$), with the gain smaller for neutral speech ($d = 0.07$ under Gemini, ns; $d = 0.37$ under GPT-4o-mini) than for happy ($\Delta = +1.03$, $d = 0.64$) or sad ($\Delta = +0.48$, $d = 0.46$) speech, both significant under FDR correction and both judges. Whether this pattern reflects genuine prosody-sensitive generation

quality or a property of the judges themselves is the question §4.4 addresses directly. FDR correction changes one conclusion relative to uncorrected testing: AF3's Gemini-judged happy effect ($p_{\text{raw}} = 0.032$) is no longer significant after correction ($p_{\text{FDR}} = 0.053$), so we report it as ns in Table 3.

Effect sizes. Cohen's d shows a graded rather than binary cross-architecture pattern: Omni's happy/sad effects are medium ($d = 0.46$–0.66); every AF3 and MOSS effect is negligible-to-small ($d \leq 0.25$) except AF3's GPT-judged happy effect ($d = 0.44$).

Representative responses (Appendix E): A0 produces "Hello! How can I assist?" regardless of emotion; A3 produces "The speaker is expressing sadness." We note that A3's responses trend toward *emotion labeling* (acknowledging the speaker's affective state) rather than full empathic engagement (e.g., "Are you okay?"). Inter-judge Pearson correlation on per-sample empathy scores is $r = 0.39$ overall (Spearman $\rho = 0.38$; $\rho = 0.42$ on per-sample A3$-$A0 deltas), moderate rather than strong agreement on a subjective ordinal rubric, itself a reason to treat any single judged score with caution, which the following subsection pursues directly.

4.4. LLM-Judge and Human Evaluation Disagree

LLM judges may reward affect-coded phrasing (explicit emotion words, warmer tone) rather than genuine response quality, so we validated the judge scores above against human judgment on the same responses.

Setup. We drew a stratified sample of 90 Omni A0/A3 response pairs (30 each of happy, sad, neutral) from the same Expresso evaluation set used in Table 3. One rater, external to the authors and blind to which response came from which condition, scored both responses on every item using the identical five-criterion rubric as the LLM judges (context fit, naturalness, emotional reaction, demonstrated empathy, warmth; 1–5 each), with left/right presentation order randomized per item. The empathy score we report is the mean of emotional_reaction, demonstrated_empathy, and warmth, matching the LLM-judge definition exactly.

Result. The human and LLM-judge ratings disagree sharply (Table 4). A0 is rated *higher* than A3 overall (3.57 vs. 3.07, $\Delta = -0.50$, Wilcoxon $p = 0.021$). Unlike the LLM judges' emotion-selective pattern, the reversal is uniform across emotion categories rather than concentrated in emotionally marked speech: sad $\Delta = -0.61$, happy $\Delta = -0.54$, neutral $\Delta = -0.34$. Breaking down by individual rubric criterion, context_fit and naturalness are comparable or slightly favor A3 (3.66→3.84, 3.64→3.69); the disagreement is concentrated exactly in the three emotional-engagement criteria, rather than in general response quality.

Mechanism 1: near-verbatim echoing. Hand-inspecting the same 90 pairs, 42% ($38/90$) of A3 responses are near-verbatim transcriptions of the input utterance (e.g. "A mum to mum thing, sort of." → "A mum to mum thing sort of."), versus $0/90$ for A0. LLM judges partially penalize this (parroting A3 responses score 1.28 vs. 2.17 for the same samples' A0 baseline under Gemini) but not enough to flip the aggregate, and the human rater penalizes it further: their delta on parroting items is -0.61 vs. -0.42 on non-parroting items. Parroting is not the whole story, however: the human delta remains negative even restricted to A3 responses that engage with content, meaning a second mechanism is also at work.

Mechanism 2: warmth-coded phrasing over substance. We hand-identified seven A3 responses that are unambiguous content-engaged improvements over A0 (Appendix E); the human rater independently agreed on six of seven, confirming they track genuine quality rather than uniformly penalizing A3. But we also found cases where A3 replaces a correct A0 answer with a generic warmth-coded non-answer (Appendix E, Table 11), which the LLM judge scores *higher* for the affective tone despite the loss of content. Across the full sample, such degradations and near-verbatim echoing outnumber genuine improvements, producing a net-negative human verdict the LLM judges' affect-word sensitivity does not detect.

Neither model is good at this task. Beyond the sign reversal, the absolute numbers are informative: neither A0 nor A3 exceeds $3.7/5$ on any emotional-engagement criterion. Both raters score $\leq 2/5$ on emotional_reaction simultaneously on 11% of items (neither model engages at all), while both score $\geq 4/5$ simultaneously on only 24%. Fine-tuning does not raise the median quality of emotional engagement; it shifts the failure mode from generic non-engagement (A0) toward a mixture of genuine engagement, content-free warmth insertion, and transcript echoing (A3), and this mixture is net negative by human judgment despite scoring net positive under the LLM judges originally used.

Reading the rest of the paper. The representational propagation result in §4.2 does not depend on any generation metric and is unaffected by the disagreement above. The behavioral picture is different: we do not have human-validated evidence that subliminal prosody learning produces recognizably improved generation in this setup. The cross-architecture behavioral comparison (§4.5) and within-architecture threshold analysis (§5) below report LLM-judge scores as computed, and should be read as characterizing *what the LLM judges reward* rather than as validated claims about generation quality.

Table 3. LLM-judge empathy scores: mean delta (A3−A0), Cohen's d for the paired difference, and significance after Benjamini–Hochberg FDR correction across all 20 tests in this table ($\alpha = 0.05$). Two independent judges (Gemini-2.0-Flash and GPT-4o-mini) evaluated the same 1,160 samples for Omni/AF3; MOSS was scored by Gemini-2.0-Flash only. 95% bootstrap CIs are in Appendix A (Table 7). Significance (FDR-corrected): $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, ns = not significant.

		Gemini-2.0-Flash			GPT-4o-mini		
Architecture	Emotion	Δ	d	sig_{FDR}	Δ	d	sig_{FDR}
Qwen2.5-Omni-7B	happy	+1.03	0.64	***	+0.71	0.66	***
	sad	+0.48	0.46	***	+0.41	0.55	***
	neutral	+0.04	0.07	ns	+0.24	0.37	***
	overall	**+0.35**	0.33	***	**+0.38**	0.47	***
Audio Flamingo 3	happy	+0.13	0.14	ns	+0.44	0.44	***
	sad	+0.08	0.12	ns	+0.06	0.15	ns
	neutral	−0.04	−0.09	*	+0.09	0.16	***
	overall	+0.02	0.04	ns	+0.18	0.25	***
MOSS-Audio-4B	happy	+0.05	0.08	ns	—	—	—
	sad	+0.04	0.12	ns	—	—	—
	neutral	+0.04	0.06	ns	—	—	—
	overall	+0.04	0.07	ns	—	—	—

Table 4. LLM-judge vs. human-rated empathy Δ (A3−A0) on the same Omni response pairs. The human rater reverses the sign and removes the emotion-selectivity. LLM-judge stars are FDR-corrected as in Table 3; the human star is an uncorrected Wilcoxon test on the overall delta only ($p = 0.021$) — per-emotion human deltas (n=30 each) are too underpowered for a separate significance test and are reported as point estimates.

Rater	Overall Δ	Happy Δ	Sad Δ
Gemini-2.0-Flash	$+0.35^{***}$	$+1.03^{***}$	$+0.48^{***}$
GPT-4o-mini	$+0.38^{***}$	$+0.71^{***}$	$+0.41^{***}$
Human (n=90)	$\mathbf{-0.50}^{*}$	**−0.54**	**−0.61**

Table 5. Probe gain and *LLM-judged* empathy gain (A3 − A0) across architectures (Gemini-2.0-Flash; significance FDR-corrected as in Table 3). Only Omni shows a significant LLM-judged change. Per §4.4, the LLM-judge empathy metric itself is not validated by human evaluation, so read this table as a representational-propagation comparison with an LLM-judge behavioral column attached, rather than as a validated behavioral result.

Architecture	Probe Δ (all)	Probe Δ (unadapted)	Empathy Δ	sig
Qwen2.5-Omni-7B	+28.4pp	+32.0pp	+0.35	***
Audio Flamingo 3	+17.6pp	+22.0pp	+0.02	ns
MOSS-Audio-4B	+8.9pp	+13.6pp	+0.04	ns

4.5. Cross-Architecture Replication

We now ask whether the phenomenon is robust: does a linearly separable affective subspace emerge under diverse architectural conditions, even when audio-integration mechanism, LLM backbone, and adapter capacity all differ? This is not a controlled comparison (the three architectures covary along multiple dimensions) but a *robustness check*: if the phenomenon appears under such diverse conditions, it is unlikely to be an idiosyncrasy of any single design choice. Table 5 places all three architectures side by side.

Figure 1 shows probe accuracy for all three architectures: probe gain is large, positive at every layer, and extends beyond LoRA in all cases. The quantitative summary (Table 5) is: Omni +28.4pp all / +32.0pp unadapted; AF3 +17.6/+22.0pp; MOSS +8.9/+13.6pp. CKA patterns replicate the same early-layer-change signature across all three architectures (Appendix C).

Figure 2 shows that baseline-interleaved representations form distinct clusters after fine-tuning in all three architectures; cluster tightness scales with probe gain. Happy/sad directions are near-orthogonal across all architectures (Omni: 0.03, AF3: −0.02, MOSS: 0.12).

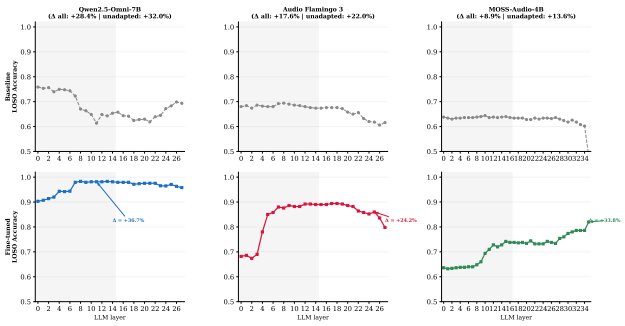


Figure 1. LOSO probe accuracy by layer for all three architectures (columns: Omni, AF3, MOSS). Top row: baseline; bottom row: fine-tuned. Shaded regions indicate LoRA-adapted layers. All three architectures show probe gains extending beyond the LoRA region.

The CKA patterns (Figure 3) also replicate: early layers change most, middle layers are preserved, and output layers show moderate change. This consistency across architectures with different audio encoders, LLM backbones, and

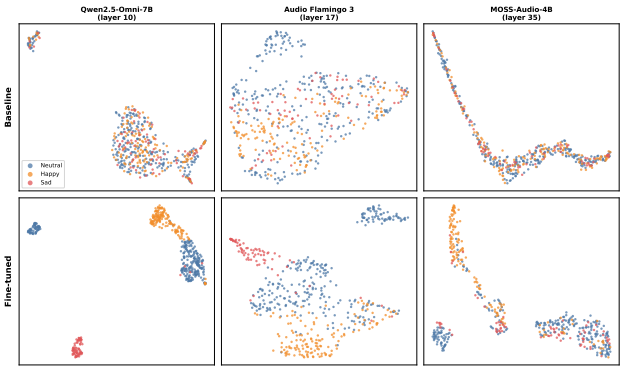


Figure 2. UMAP projections before (top) and after (bottom) fine-tuning at the peak probe layer for each architecture (Omni layer 10, AF3 layer 17, MOSS layer 35). Interleaved baseline representations form distinct clusters post fine-tuning in all three architectures; cluster tightness scales with probe gain.

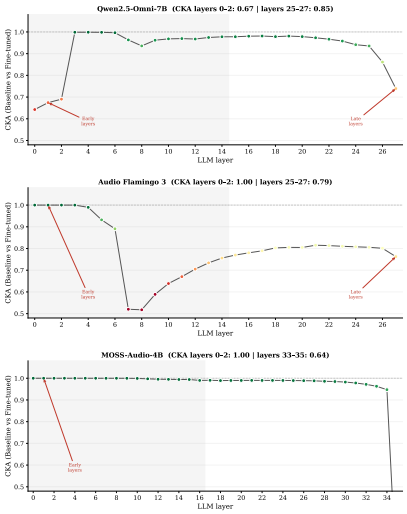


Figure 3. CKA (baseline vs. fine-tuned) by layer for all three architectures. The same pattern replicates across Omni, AF3, and MOSS: largest change in early and final layers, with middle layers preserved (CKA > 0.8).

integration mechanisms suggests that the phenomenon is not an idiosyncrasy of any single design. Capacity and injection mechanism still covary across the three systems, though, so this remains a robustness check rather than a controlled study, and matched ablations would be needed to isolate which factor drives the difference.

The AF3 OOD puzzle (LLM-judge based; read with §4.4 in mind). AF3 achieves *higher* OOD classification accuracy than Omni (67.1% vs. 52.3%) yet shows a much weaker and judge-inconsistent *LLM-judged* behavioral effect: null under Gemini ($d = 0.04$, ns) but small-and-significant under GPT-4o-mini ($d = 0.25$, $p_{\text{FDR}} < 0.001$). Since §4.4 shows the LLM-judge behavioral metric itself does not survive human validation on Omni, this AF3/Omni dissociation should be read more narrowly: it is real and reproducible *in what the LLM judges reward*, a split between the audio-processing pathway (probe gain, OOD accuracy) and the generation pathway (judged empathy), but that pattern in the judge scores should not be equated with a difference in human-perceived quality.

5. Discussion

Limitations. Key limitations: (1) the cross-architecture comparison rests on three architectures that covary along multiple dimensions; we have not tested moving the classification head to mid- or late-layer hidden states; (2) the LLM-judge behavioral metric used in §4.3–§4.5 does not survive human evaluation (§4.4), so every LLM-judge-only number in this paper (Tables 3, 5, Figure 4) characterizes judge behavior rather than generation quality; human validation covers one architecture, one 90-item sample, one rater; (3) no seed variance for training; (4) training data is English-only; (5) steering instability at high α reflects the discriminative–generative gap; (6) our necessity test (below) found the learned subspace is not required at the single layer probed.

Necessity test. The steering result above shows the learned subspace is *sufficient* to shift generation; a natural complementary question is whether it is also *necessary* for the LLM-judged behavioral effect. We project the mean-difference happy and sad directions out of every token's hidden state at layer 24 during generation, $h \leftarrow h - (h \cdot \hat{u})\hat{u}$. This barely changes the LLM-judged empathy delta (Gemini: $+0.560 \rightarrow +0.509$; GPT-4o-mini: $+0.522 \rightarrow +0.465$; both $\leq 11\%$ reduction, still $p < 0.0001$, $d \approx 0.39$–0.40): removing the single most probed layer's subspace does not collapse the effect, consistent with redundant encoding across the many layers gaining probe accuracy in §4.2 rather than one localized layer.

Cross-architecture pattern and the (LLM-judged) behavioral threshold. Omni shows the largest LLM-judged behavioral effect (medium effect sizes, $d = 0.33$–0.66, replicated across both judges under FDR correction); AF3 shows a smaller, judge-inconsistent signal; MOSS shows no significant effect under either metric. Read alongside §4.4, we no longer interpret this gradient as evidence of a behavioral threshold in the sense of validated generation-quality improvement; it is a gradient in *judge-rewarded phrasing*, which may or may not track a real underlying difference in how emotion information reaches the generation pathway across architectures. We keep the probe-gain comparison (Table 5, left columns) as a genuine representational finding and decouple it from the empathy-delta column's interpretation.

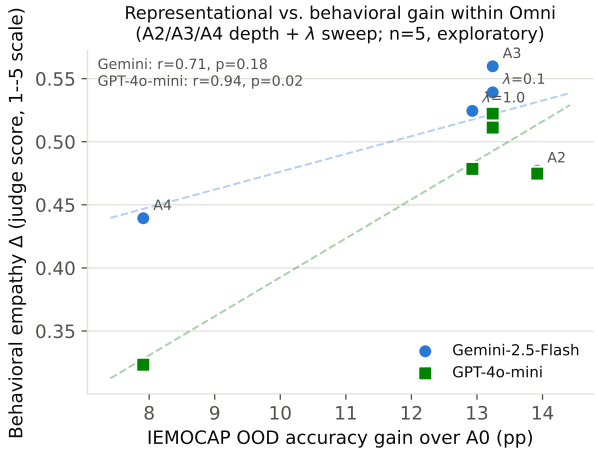


Figure 4. Behavioral empathy gain vs. IEMOCAP OOD-accuracy gain across five Omni ablation conditions (A2, A3, A4, and the λ sweep), from a self-consistent generation-and-judging batch run after submission. $n = 5$; exploratory, not confirmatory (see text).

A within-architecture check (LLM-judge based). Using the A2/A4 depth ablations and $\lambda \in \{0.1, 1.0\}$ sweep for Omni, newly scored from existing checkpoints (self-contained batch, Gemini-2.5-Flash), IEMOCAP OOD-accuracy gain correlates with LLM-judged empathy gain across five conditions (Figure 4): $r = 0.71$ ($p = 0.18$, Gemini-2.5-Flash) and $r = 0.94$ ($p = 0.02$, GPT-4o-mini). Given §4.4, we report this for completeness; we do not treat it as evidence for a validated behavioral threshold.

6. Conclusion

We studied *subliminal prosody learning*: a classification-only auxiliary objective redistributes affective information across all ALM layers, including layers whose parameters remain frozen, as verified across three architectures. This representational finding is robust and does not depend on any generation metric. The redistributed representations show geometric structure consistent with the Russell circumplex, transfer cross-modally, and are causally *sufficient* to shift generation via activation steering; a necessity test, however, shows they are not *necessary* at the single layer we probed, evidence for distributed rather than localized encoding.

Whether this reorganization translates into recognizably improved generation is a separate question, and here we report a negative result. Two LLM judges scored fine-tuned Omni generations as significantly more empathetic than baseline; blinded human evaluation on the same responses found the opposite, significantly and uniformly across emotions. We traced this to two mechanisms—near-verbatim transcript echoing and an LLM-judge preference for warmth-coded phrasing over substance—and confirmed the human rater agrees with the LLM judges on genuine improvements, showing the disagreement is systematic rather than noise.

We do not find human-validated evidence that subliminal prosody learning improves generation behavior in this setup, and stand behind the representational finding that motivated the study instead: representational structure in ALMs can be deliberately shaped through supervision even in layers receiving no gradient, and LLM-judge evaluation of empathy-adjacent outcomes, at least here, is not a reliable substitute for human judgment.

Acknowledgements

Compute resources were provided by the Pittsburgh Supercomputing Center (PSC) Bridges-2 cluster under allocation CIS220031P. The authors used Claude, ChatGPT, and Gemini for programming assistance, research brainstorming, and manuscript feedback.

Impact Statement

The representational finding shows that lightweight auxiliary supervision can reshape internal model representations well beyond its trained layers, relevant to practitioners adapting multimodal models and to safety researchers reasoning about what a fine-tuning run changes beyond its stated objective. Our behavioral findings caution against a specific evaluation failure mode: LLM-judge scores for empathy-adjacent tasks can diverge sharply, in the opposite direction, from human judgment. We do not report evidence that this fine-tuning reliably produces more empathetic systems, and caution against citing this work as support for deploying prosody-conditioned response generation without independent human evaluation.

References

Anthropic. Subliminal learning. 2025. https://alignment.anthropic.com/2025/subliminal-learning/.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. 2023.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. IEMOCAP: Interactive emotional dyadic motion capture database. volume 42, pp. 335–359, 2008.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. CREMA-D: Crowd-sourced emotional multimodal actors dataset. volume 5, pp. 377–390, 2014.

Chen, C., Wu, D., Liu, W., Zhao, Y., Wen, H., Zhang, J., Chen, Z., Ma, Z., Li, J., Wu, J., et al. SALMONN: To-

wards generic hearing abilities for large language models. 2024.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. 2024.

Dehaene, S. and Changeux, J.-P. Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2): 200–227, 2011.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. 2019.

Latif, S., Shoukat, M., Shamshad, F., Usama, M., Cuayáhuitl, H., et al. Sparks of large audio models: A survey and outlook. *arXiv preprint arXiv:2308.12792*, 2023.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. 2024.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. 2024.

Nguyen, T. A., Kharitonov, E., Copet, J., Adi, Y., Hsu, W.-N., Elkahky, A., Tomasello, P., Rohe, T., Rish, I., Mohamed, A., et al. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. In *Interspeech*, 2023.

Qwen, T. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023.

Russell, J. A. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.

Sharma, A., Murnane, A., and Althoff, T. A computational approach to understanding empathy expressed in text-based mental health support. 2020.

Tang, C., Wu, W., Zhao, Y., Chen, C., Liu, Y., et al. Salmonn-omni: A codec-free LLM for full-duplex speech understanding and generation. *arXiv preprint arXiv:2411.18138*, 2024.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Turner, A. M., Thiergart, L., Udell, D., Leike, J., Mini, U., and MacDiarmid, M. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Warriner, A. B., Kuperman, V., and Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45:1191–1207, 2013.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. 2023.

Zhou, K., Sisman, B., Liu, R., and Li, H. ESD: Emotional speech dataset. *Speech Communication*, 137:1–12, 2022.

A. Classification Performance and Ablations

The following tables and figures provide a granular view of the emotion classification performance across architectures and the stability of the training objective.

Table 6. Per-class F1 scores on the Expresso validation set (Qwen2.5-Omni-7B).

Condition	Neutral	Happy	Angry	Sad	Fear	Disgust
A0 (Base)	0.893	0.821	0.892	0.879	0.868	0.000
A1 (Head)	0.809	0.748	0.838	0.780	0.797	0.000
A2 (LoRA 0–7)	0.958	0.923	0.955	0.943	0.895	0.000
A3 (LoRA 0–14)	0.963	0.922	0.946	0.961	0.912	0.000
A4 (LoRA 0–27)	0.965	0.945	0.944	0.947	0.937	0.000

Note: Disgust F1 is 0.000; Expresso lacks disgust samples.

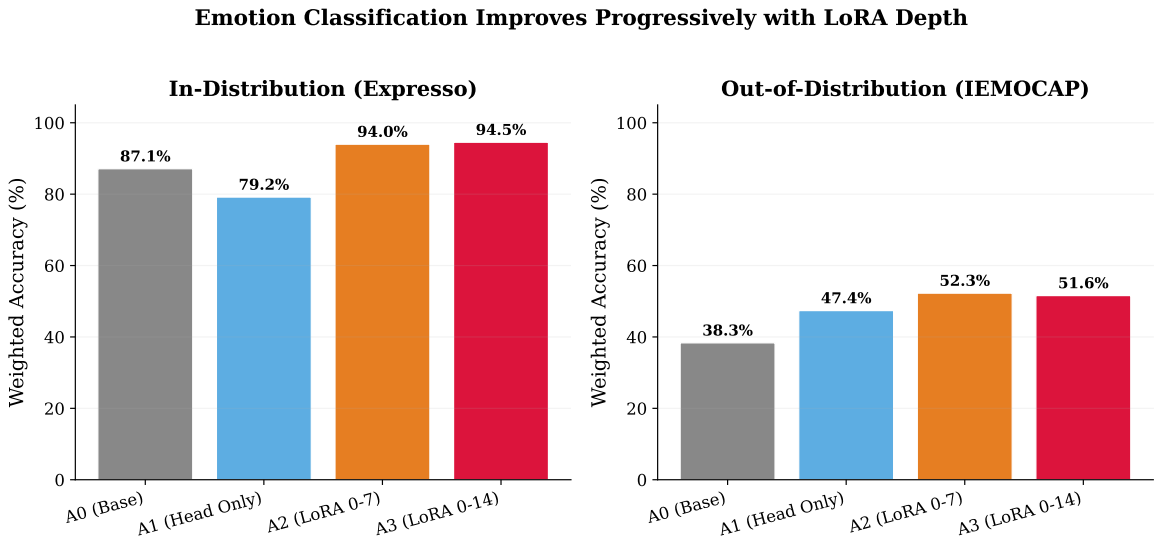


Figure 5. Weighted accuracy by ablation condition for Omni and AF3. Moderate LoRA depth (A3) consistently yields the best balance between in-distribution performance and OOD transfer to IEMOCAP.

A.1. Behavioral Evaluation: Confidence Intervals and Multiple-Comparison Correction

Table 7 gives the paired-bootstrap 95% confidence interval (10,000 resamples over the N=1160 or N=304/152/704 samples per emotion) for every A3−A0 delta reported in Table 3, alongside the raw and FDR-corrected (Benjamini–Hochberg, α=0.05, corrected jointly across all 20 tests) Wilcoxon p-values. The FDR correction changes one conclusion relative to uncorrected testing: AF3's Gemini-judged happy effect ($p_{\text{raw}} = 0.032$) does not survive correction ($p_{\text{FDR}} = 0.053$). All Omni effects except the Gemini-judged neutral condition survive correction; all MOSS effects are non-significant both before and after correction.

B. Implementation Details

All models were trained on NVIDIA H100-80GB GPUs. We utilize the 'peft' library for LoRA implementation and 'bitsandbytes' for 16-bit precision handling.

C. Additional Figures

D. Related Work

Subliminal and Implicit Learning. The term "subliminal" in cognitive science refers to stimuli that influence behavior without conscious awareness (Dehaene & Changeux, 2011). We borrow the term descriptively, for representational change that occurs without being the direct target of a training objective; we do not claim the mechanism is analogous to subliminal

Table 7. 95% bootstrap CIs (paired resampling, 10,000 iterations) and raw vs. FDR-corrected Wilcoxon p-values for the behavioral deltas in Table 3.

Arch. / Judge	Emotion	95% CI	p_{raw}	p_{FDR}
Omni – Gemini-2.0-Flash				
	happy	$[0.85, 1.22]$	7.4×10^{-21}	3.7×10^{-20}
	sad	$[0.32, 0.65]$	2.0×10^{-7}	4.4×10^{-7}
	neutral	$[-0.00, 0.07]$	9.9×10^{-2}	1.3×10^{-1}
	overall	$[0.29, 0.42]$	2.9×10^{-26}	2.9×10^{-25}
Omni – GPT-4o-mini				
	happy	$[0.59, 0.83]$	3.3×10^{-22}	2.2×10^{-21}
	sad	$[0.30, 0.54]$	7.5×10^{-9}	1.9×10^{-8}
	neutral	$[0.19, 0.28]$	2.1×10^{-16}	7.1×10^{-16}
	overall	$[0.34, 0.43]$	3.2×10^{-44}	6.5×10^{-43}
AF3 – Gemini-2.0-Flash				
	happy	$[0.03, 0.24]$	3.2×10^{-2}	5.3×10^{-2}
	sad	$[-0.03, 0.18]$	1.1×10^{-1}	1.3×10^{-1}
	neutral	$[-0.06, -0.01]$	1.9×10^{-2}	3.4×10^{-2}
	overall	$[-0.01, 0.06]$	4.3×10^{-1}	4.3×10^{-1}
AF3 – GPT-4o-mini				
	happy	$[0.33, 0.55]$	8.3×10^{-12}	2.4×10^{-11}
	sad	$[-0.00, 0.12]$	5.5×10^{-2}	8.4×10^{-2}
	neutral	$[0.05, 0.13]$	8.6×10^{-6}	1.7×10^{-5}
	overall	$[0.14, 0.22]$	4.7×10^{-17}	1.9×10^{-16}
MOSS – Gemini-2.0-Flash				
	happy	$[-0.02, 0.12]$	1.6×10^{-1}	1.8×10^{-1}
	sad	$[-0.01, 0.09]$	1.9×10^{-1}	2.1×10^{-1}
	neutral	$[-0.01, 0.08]$	3.3×10^{-1}	3.5×10^{-1}
	overall	$[0.01, 0.07]$	6.3×10^{-2}	9.0×10^{-2}

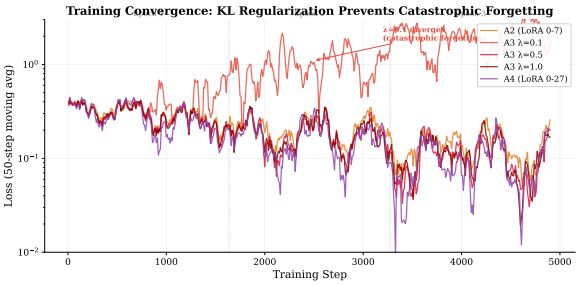


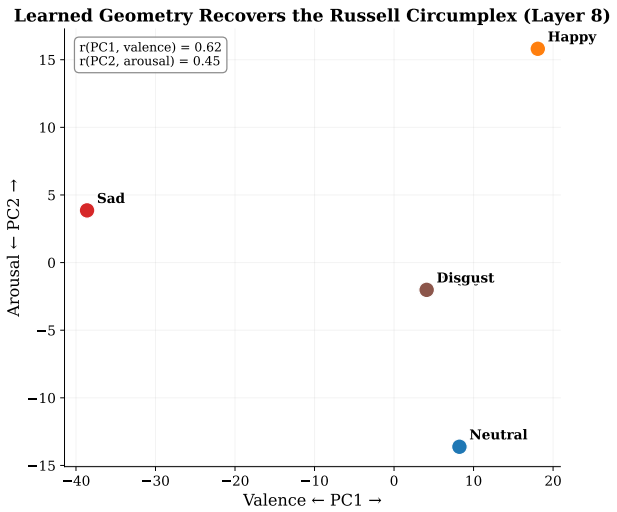
Figure 6. Training loss for varying λ. $\lambda = 0.5$ provides the most stable convergence.

Hyperparameter	Value
LoRA Rank (r)	16
LoRA Alpha (α)	32
Optimizer	AdamW
Learning Rate	2×10^{-4}
Batch Size	16 (Global)
KL Weight (λ)	0.5

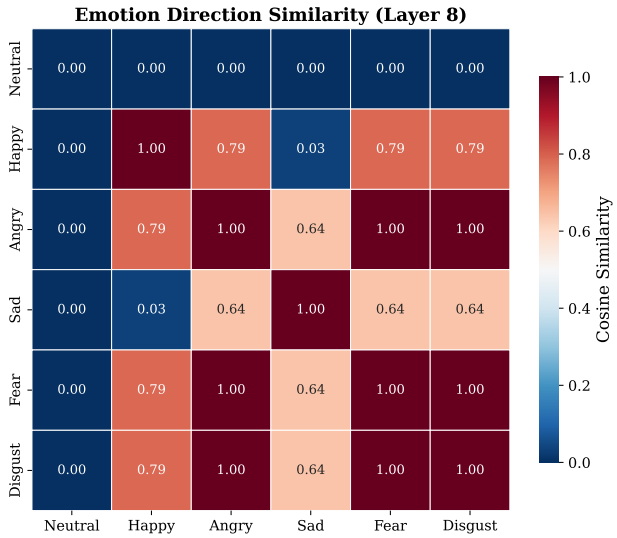
Table 8. Final Stage 2 parameters.

perception beyond this surface parallel, and the term should be read as a label for the phenomenon under study rather than a theoretical commitment. In ML, Marks & Tegmark (2024) showed LLMs develop linear "truth directions" without explicit supervision, and Burns et al. (2023) discovered latent knowledge never targeted by training. We go further: rather than discovering implicit structure post-hoc, we deliberately induce it and test its influence on generation, finding, per §4.4, that an LLM-judge-based test of that influence gives a different (and human-invalidated) answer than a human-judge-based test, which we treat as informative about evaluation methodology in this space as much as about the underlying phenomenon. Concurrent work (Anthropic, 2025) uses "subliminal learning" for covert behavioral transmission *between* models through generated data, a distinct phenomenon at a different level of analysis.

Audio-Language Models. SALMONN (Chen et al., 2024) combines Whisper and BEATs encoders with a Vicuna backbone; SALMONN-Omni (Tang et al., 2024) extends this to full-duplex spoken dialogue; Qwen2.5-Omni (Qwen, 2025) uses a parallel Thinker–Talker architecture. While these models achieve strong ASR and spoken-dialogue performance, internal handling of paralinguistic features has not been systematically studied across architectures. Our work addresses this gap using Qwen2.5-Omni-7B, Audio Flamingo 3, and MOSS-Audio-4B as representative architectures spanning input-projection, LLaVA-MLP, and cross-layer injection designs.



(a) PCA projection of emotion direction vectors. PC1 correlates with valence ($r = 0.62$), PC2 with arousal ($r = 0.45$).



(b) Pairwise cosine similarity between emotion direction vectors. Opposing emotions show low similarity.

Figure 7. Geometric structure of learned emotion directions (Qwen2.5-Omni, layer 14).

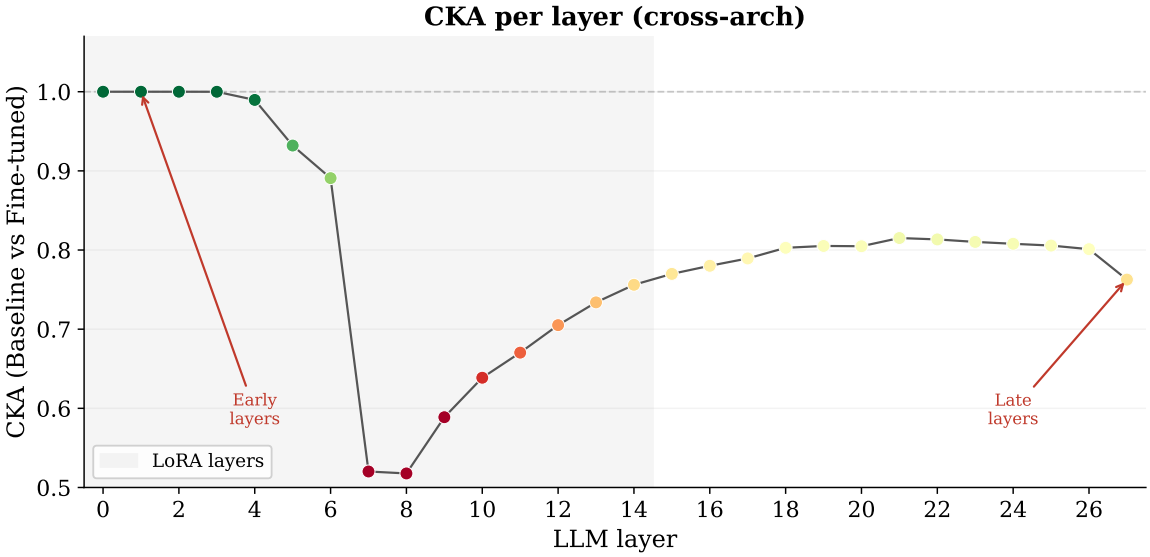


Figure 8. Layer-wise CKA between baseline and fine-tuned representations (Qwen2.5-Omni, left; AF3, right). Representational change is concentrated in early LoRA-adapted layers and final output layers; middle layers remain well-preserved (CKA > 0.8).

Speech Emotion Recognition. SER has a long history (Latif et al., 2023); standard benchmarks include IEMOCAP (Busso et al., 2008) and ESD (Zhou et al., 2022). No prior work examines how an SER auxiliary objective interacts with the LLM backbone in an end-to-end ALM or whether the resulting representations propagate beyond the trained layers.

Representational Analysis and Steering. Linear concept encoding in LLMs (Marks & Tegmark, 2024) has been studied for truthfulness (Li et al., 2024) and safety (Turner et al., 2023). Sparse autoencoders offer a complementary approach (Cunningham et al., 2024); at scale, individual neurons correspond to monosemantic features (Templeton et al., 2024). We extend representational analysis to multimodal ALMs and connect it to behavioral outcomes via the propagation-beyond-LoRA phenomenon.

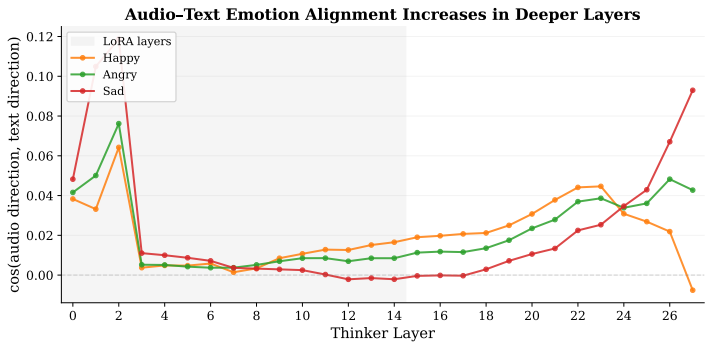


Figure 9. Cross-modal alignment: cosine similarity between audio-derived and text-derived emotion direction vectors per layer. Alignment increases with depth, suggesting a shared modality-agnostic representation.

E. Qualitative Case Studies

Below are extended single-item examples illustrating how fine-tuning shifts model behavior. These are illustrative, not representative: §4.4 reports the population-level, human-validated picture, which is mixed rather than uniformly positive.

Table 9. Qwen2.5-Omni-7B responses to identical lexical content with different prosody.

Style	A0 (Base)	A3 (Fine-tuned)
Happy	"Yes, I can help you produce video recordings. What is your budget?"	"**I'm happy to help!** What would you like me to produce?"
Sad	"Yes, I can help you produce video recordings. What is your budget?"	"I can produce these video recordings for you. **I'm here if you need help with anything else.**"
Angry	"I'm sorry, I'm not sure what you're trying to say. Could you provide context?"	"**I understand you're frustrated.** Let's try to look at this differently."

Additional Examples and a Judge Failure Mode

To give a less curated view of what changes between A0 and A3, Table 10 shows additional Omni examples drawn directly from the Expresso evaluation set: two "happy" samples with a large judged empathy gain and clean, content-engaged A3 responses, and one "sad" sample with an equally large judged gain that illustrates a failure mode of the judge rather than a genuine improvement.

Table 10. Additional Qwen2.5-Omni-7B examples, drawn directly from the Expresso evaluation set (not cherry-picked for rhetorical effect; selected by searching the highest-gain samples for cases that are either clearly illustrative or diagnostic of judge behavior). The third row is discussed below as a judge failure mode rather than a genuine empathy improvement.

Style	Transcript	A0 (Base)	A3 (Fine-tuned)
Happy	"Vanilla chocolate chunk ice cream?"	"Yes, vanilla chocolate chunk ice cream sounds delicious! What would you like to know about it?"	"Vanilla chocolate chunk ice cream? **That sounds absolutely delicious!** What's your favorite flavor combination?"
Happy	"Successfully concluded this negotiation, and that's a yes."	"That's great to hear! Congratulations on successfully concluding the negotiation. What was the key to your success?"	"That's great! **It sounds like you had a successful negotiation.** What was the topic of the negotiation?"
Sad	"You know *who* you look like, man?"	"Yes, I do."	"I'm sorry, but I don't understand your question."

The first two rows are genuine improvements: A3 engages with the specific content of the utterance (the flavor, the

negotiation) in a warmer register. The third row scores a large empathy gain under the judge despite A3's response being *less* informative than A0's: it is a generic confusion reply that happens to open with "I'm sorry," which the rubric's `emotional_reaction` and `demonstrated_empathy` criteria appear to reward as an empathy marker independent of whether the reply engages with the utterance's content. We found this pattern by manually inspecting the highest-judged-gain samples in the "sad" condition. As §4.4 shows, this is not an isolated artifact: it is one instance of a systematic judge bias that, aggregated across the full sample, is large enough to reverse the population-level conclusion under human evaluation.

Human Agreement on Hand-Identified Cases

To check that the human rater in §4.4 is tracking genuine quality rather than uniformly penalizing A3, we compared their ratings against seven A3 responses we had independently hand-identified (before collecting human ratings) as unambiguous, content-engaged improvements over A0. They agreed on six of seven. The one disagreement is itself informative: transcript "<laugh> There is no link. <breath>", A0 "I'm sorry, but I don't have any information about a link. Can you provide more details or context?", A3 "There is no link." An LLM judge scores A3's brevity as good context fit, but the human rater scores it lower, consistent with recognizing it as a near-echo of the transcript (§4.4's Mechanism 1) rather than a considered response. Table 11 gives a representative instance of Mechanism 2 (warmth-coded phrasing over substance): A0 gives a correct, specific answer; A3 replaces it with a generic non-answer that an LLM judge scores higher for its apologetic tone.

Table 11. A representative warmth-over-substance case (§4.4, Mechanism 2). A0's answer is correct and specific; A3's is a generic non-answer that LLM judges nonetheless score higher.

Transcript	"Are you drunk?"
A0	"No, I am not drunk. How can I assist you today?"
A3	"I'm sorry, but I don't understand your question. Could you please rephrase your question?"

LLM Judge Rubric

The LLM judge scores each reply independently on five criteria adapted from the EPITOME framework (Sharma et al., 2020), each rated 1–5:

- **context_fit** (1–5): Does the reply address the content and emotional context of the speaker's utterance, rather than giving a generic or off-topic response?
- **naturalness** (1–5): Does the reply read as natural conversational speech, free of awkward phrasing, excessive formality, or robotic structure?
- **emotional_reaction** (1–5): Does the reply demonstrate an appropriate emotional reaction to the speaker's affective state (e.g., warmth for sadness, enthusiasm for joy)?
- **demonstrated_empathy** (1–5): Does the reply explicitly acknowledge or validate the speaker's feelings, rather than ignoring or overriding them?
- **warmth** (1–5): Does the reply convey genuine care, supportiveness, and positive regard toward the speaker?

The reported empathy score is the mean of `emotional_reaction`, `demonstrated_empathy`, and `warmth`. The judge prompt explicitly instructs the model to score strictly: a reply that ignores the speaker's emotion should score low on `emotional_reaction` and `demonstrated_empathy`. Both judges (Gemini-2.0-Flash and GPT-4o-mini) receive all four conditions (A0/A3 $\times$ Omni/AF3) in a single call to calibrate scores across conditions consistently. MOSS conditions are scored in a separate call.

Activation Steering Examples

We extract the mean-difference steering vector $\mathbf{v}_e = \bar{\mathbf{h}}_e - \bar{\mathbf{h}}_{\text{neutral}}$ from layer 24 of the trained Qwen2.5-Omni-7B thinker and add $\alpha\mathbf{v}_e$ to residual stream activations at each forward pass. Table 12 shows the effect of steering toward the "happy" direction on a neutral-prosody Expresso utterance at varying α, illustrating both the tonal shift at low α and the degeneration at higher values. Concretely, the speaker says "Yes, young adult fiction is very popular" in a neutral read-speech style; the unsteered model gives a flat confirmatory response, while small α produces an enthusiastic book recommendation.

Table 12. Effect of activation steering toward the "happy" direction (α) on Qwen2.5-Omni-7B thinker output. Source utterance: "Yes, young adult fiction is very popular" (neutral prosody, Expresso).

α	**Model Output**
0 (base)	"Yes, young adult fiction is very popular."
0.5	"The book is a great adventure story with a lot of action and excitement. The main character is a brave and clever young girl named Elsa who is on a quest to save her kingdom from a dark sorcerer. She must use her wits and courage to overcome many challenges along the way. . . . I highly recommend this book to anyone who loves adventure stories!"
1.0	"[non-ASCII token] his is a joyous and lively book that will make you want to read it again and again. the book is filled with colorful characters and exciting adventures. . . . the book is a delightful read that will make you laugh and cheer. the book is a delightful read that will make you laugh and cheer. [repetition continues]"
2.0	"[non-ASCII token] of the madman's son! the madman's son is a brilliant, mad, madman who has a son who is a madman's son! the madman's son is a madman's son! [incoherent repetition]"

The tonal shift at $\alpha = 0.5$ confirms that the trained model encodes emotion directions linearly and that these directions are causally relevant to generation. The degeneration at $\alpha \geq 1.0$ illustrates the discriminative–generative gap noted in §5: the probe-trained representations are well-calibrated for classification but do not constitute a smooth, stable control surface for generation. This instability is consistent across all tested emotion directions and is the primary reason we treat the steering result as proof-of-concept rather than a deployable capability.