MECHANISTIC INTERPRETABILITY WORKSHOP · ICML 2026

Subliminal Prosody Learning

Encoder adaptation unlocks latent decoder capabilities

Bhuvan Koduru **Dareen Alharthi** **Rita Singh** **Bhiksha Raj**

Language Technologies Institute · Carnegie Mellon University

TL;DR THREE MAIN TAKEAWAYS

1. A classification-only objective redistributes affective representations across **all** layers, frozen ones included (**+32pp** probe gain).
2. Decodability ≠ behavior: prosody-sensitive generation emerges only past a **threshold** (Omni, Δ+0.35, p<0.001).
3. Learned emotion geometry recovers the **Russell circumplex** , transfers cross-modally, and is causally steerable.

1 Motivation

ALMs encode rich prosody — pitch, tempo, vocal quality — yet generation ignores it. The same words spoken angrily or neutrally yield near-identical text: the signal is **encoded but not leveraged .**

This raises a central question that motivates our work:
Can we enable a model *use* this prosodic information — without ever training for empathy or speech generation?

2 Method

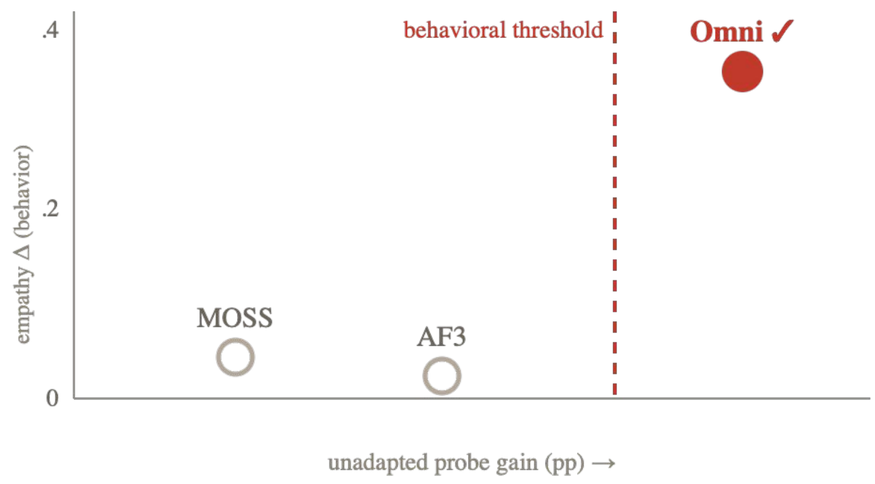
Attach a 2-layer MLP emotion classifier to early-layer hidden states; LoRA-adapt only early layers, the rest stay frozen, with KL regularization to the frozen base.

STAGE 1	Head warmup — freeze backbone, train MLP head.
STAGE 2	LoRA-adapt early layers + KL reg for generation stability.

> The classification head is a **training device only** — discarded at inference.
> No empathy or generation supervision is ever applied.

3 The threshold

Across three architectures, only **Omni** — with the largest unadapted probe gain — crosses into significant prosody-sensitive generation. Decodability alone is not enough.



★ What it sounds like

Same words, different prosody. After adaptation the model spontaneously names the speaker's emotion and acknowledges it — never trained to.

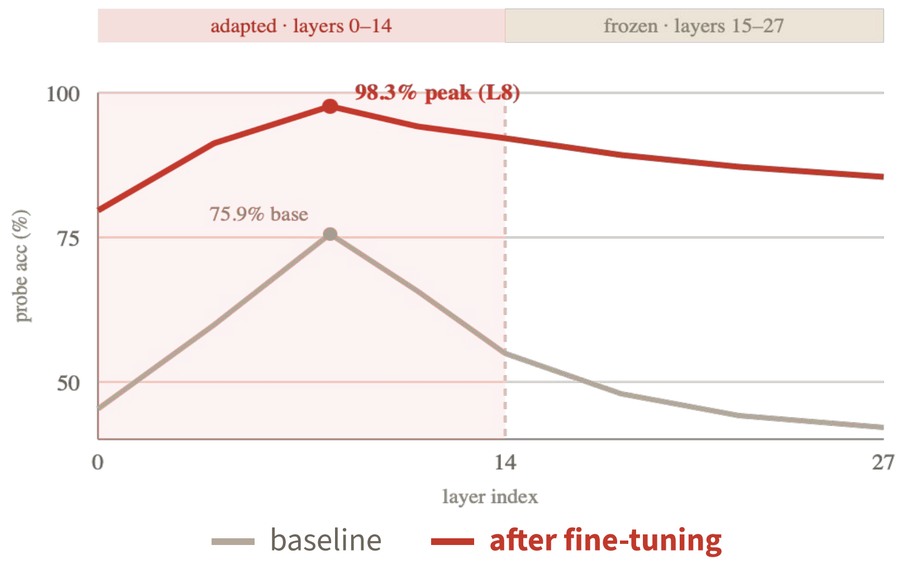
BASE MODEL
“Hello! How can I assist you today?”

OURS (sad input)
“The speaker sounds sad — I'm here for you. What's wrong?”

KEY FIGURE

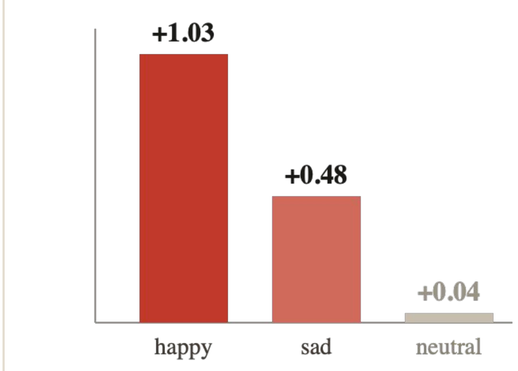
Decodability propagates beyond trained layers

Linear-probe accuracy rises across *every* layer — including frozen ones. Peak 98.3% vs. 75.9% baseline. `Qwen2.5-Omni-7B`

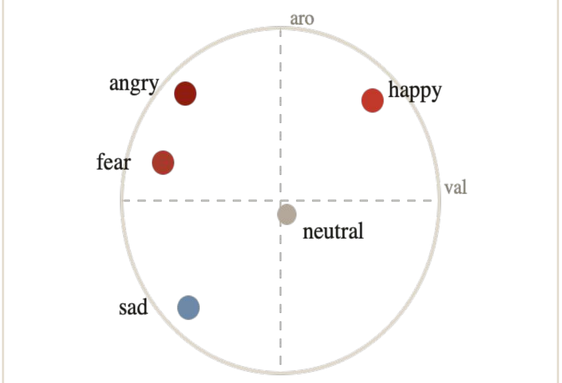


4 Emotion-selective & structured

On **Qwen2.5-Omni-7B**, empathy gains appear for **happy** (+1.03) and **sad** (+0.48) but vanish for **neutral** (+0.04, ns) — ruling out a generic verbosity confound. Learned directions recover the Russell circumplex (valence r=0.62).



empathy gain = adapted (A3) − base (A0)



learned emotion directions (PCA)

5 Transfers & steers

CROSS-MODAL
Audio- and text-derived emotion directions align in deeper layers — a shared, modality-agnostic subspace.

CAUSAL
Adding $\alpha \cdot \hat{d}$ to the residual stream steers generation toward the target emotion — not a probe artifact.

6 Limitations & discussion

The threshold rests on three architectures that co-vary along multiple dimensions (capacity, injection mechanism), so the boundary is suggestive, not causally isolated.

Still, the result reframes the field: representational structure in ALMs can be **deliberately shaped** — not merely discovered post-hoc — and lightweight auxiliary objectives reach generation beyond their stated target.

bkoduru@andrew.cmu.edu · dalharth@andrew.cmu.edu · rsingh@cs.cmu.edu · bhiksha@cs.cmu.edu