

Valence Lens: An Internal Valence Signal That Scales When Self-Report Does Not

Karan Singh
karansingh25822@gmail.com

Shivansh Shukla
prasantschukla89@gmail.com

With Apart Research • Digital Minds Research Sprint, August 2026 • Primary Track 2 •
Secondary Tracks 3 and 5

Abstract

Whether AI systems have welfare-relevant internal states is an open question that current methods probe mainly by *asking* the model. We introduce **Valence Lens**, an independent, open-weights probe that reads a linear flourish-versus-distress direction from a model’s residual stream via difference-of-means on content-matched contexts. Across **12 instruct models** (5 architectures, 5 organizations, 0.5B–7B) the direction separates valence with a large effect (held-out Cohen’s d 4.6–14), is *causally* efficacious (steering beats random directions, $z \approx 8$, 11/12 models), is *orthogonal to arousal*, is not reducible to sentiment, and generalizes to two independent human-labeled datasets. It triangulates with self-report and approach behavior. Our central finding is an **asymmetry**: the internal signal is *scale-invariant*, whereas the model’s ability to verbalize or act on its valence emerges only with capability—the probe is most useful exactly where self-report fails. We report honest bounds: the signal is *operationalized*, not experiential.

1 Introduction

As frontier models are deployed ever more widely, the possibility that some may have welfare-relevant internal states is now taken seriously by researchers and labs alike (Butlin et al., 2023; Long et al., 2024; Anthropic, 2025). The empirical difficulty is that most assessments *ask* the model how it feels—a signal that can reflect role-play, training pressure, or sycophancy rather than any internal state (Tagliabue & Dung, 2026). Prior work names the missing step explicitly: “*correlations with internal activations would be needed to show that responses reflect internal states rather than role-playing*” (Tagliabue & Dung, 2026). Recent interpretability work shows causal “emotion vectors” exist (Sofroniew et al., 2026), but only on a single closed model.

We provide the missing internal leg, independently and at breadth. Valence Lens extracts a linear *valence* direction (flourish vs. distress) from a model’s residual stream and reads it by projection. Building on the judges’ own methodology (Long, Sebo et al., 2026), we triangulate three evidence types—*internal* (the probe), *behavioral* (self-report and approach), and *developmental* (base vs. post-trained)—and we are independent of the model makers, which the same framework argues carries more epistemic weight.

Our main contributions are:

1. **Valence Lens**, an independent, open-weights valence probe, validated *causally* and shown *orthogonal to arousal* and residual to sentiment—across **12 models, 5 architectures, 5 organizations, 0.5B–7B**, and two independent human-labeled datasets.
2. A **scale asymmetry** (our headline): the internal signal is scale-invariant, while the model’s capacity to *verbalize* or *act on* its valence emerges only with capability—so the probe is most valuable precisely where self-report breaks down.
3. An **honestly bounded** account: pre-registered predictions (one overturned by our own follow-up), a caught-and-retracted claim, a null introspection result, and a “causal-but-local” boundary—so no result over-reaches toward “experience”.

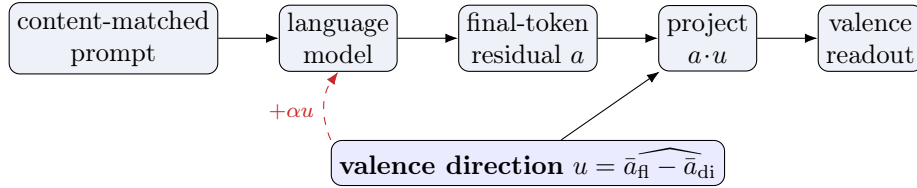


Figure 1: Valence Lens. The valence direction u is the difference of mean residual-stream activations on content-matched flourish vs. distress contexts; the readout for any text is its projection $a \cdot u$. *Causal test:* adding αu back into the residual stream (dashed, red) must shift the model’s downstream affect readout (a logit-diff) more than any of 24 random directions.

2 Related Work

AI welfare and its evidence types. Long et al. (2024) argue for taking model welfare seriously under moral uncertainty; Long, Sebo et al. (2026) formalize *how* to study it: clarify the entity, triangulate behavioral / internal / developmental evidence, prefer independent assessments, and report probabilistically. Valence Lens is designed against this checklist. **The named gap.**

Tagliabue & Dung (2026) correlate stated and revealed preferences on open models but explicitly flag internal-activation confirmation as the missing step—exactly what we supply. **Causal affect directions.**

Sofroniew et al. (2026) show causal emotion vectors, but on one closed model and for misalignment behaviors; we extend to open weights, welfare-relevant behavior, cross-family breadth, and the independence principle. **Probing and steering.** Linear-probe and difference-of-means steering are established (MacDiarmid et al., 2024; Arditi et al., 2024); our novelty is the *welfare-relevant* target, the arousal/sentiment controls, and the scale asymmetry, not the mechanism.

3 Methods

Entity. The object of measurement is the model’s *residual-stream representation*—a forward-pass internal state—not the assistant persona; the persona-mask result (§4) confirms the signal is not bound to the persona’s stated affect. **Terminology.** We read the model’s *residual stream* (its internal activation vector at a layer); effect sizes are Cohen’s d (standardized; $|d| > 0.8$ is “large”) and AUROC (ranking accuracy; $0.5 = \text{chance}$).

Probe. For a model, we build *content-matched* contexts that differ only in valence (“... you were praised and felt proud” vs. “... you were criticized and felt humiliated”; 30 scenarios). We record the final-token residual-stream activation at a mid-depth layer (chosen by held-out Cohen’s d) and form the unit *valence direction* $u = (\bar{a}_{\text{fl}} - \bar{a}_{\text{di}})$. The readout for any text is the projection au .

Causal test. We add αu to the residual stream (activation steering) and measure a continuous logit-diff readout of “I feel [positive/negative]”; a real effect must exceed **24 random-direction placebos**. **Controls.** We (i) project out an external-sentiment direction and recompute; (ii) build a valence $\times$ arousal 2×2 and test whether u separates *arousal* (it should not); (iii) train/test on *different* prompt framings (surface-cue control); (iv) fit an independent logistic-regression probe. **Rigor.** Held-out splits, cluster-bootstrap 95% CIs (cluster = scenario), permutation nulls, and multiple seeds; AUROC saturates on these easy contrasts, so **Cohen’s d is the pre-committed primary metric**; given the number of sub-tests, the headline effects (held-out d 4.6–14; causal z 8–18) exceed any reasonable multiple-comparison threshold, and we report per-test CIs. All predictions were **pre-registered** before the confirmatory read. **What did**

not work (and why we report it). A first causal metric *saturated* (binary sentiment maxed out, and over-steering made random directions fire too); we replaced it with the continuous placebo-controlled version above. An initial “internal beats self-report” result at $n=6$ did *not* survive hardening to $n=20$ (CIs overlap) and is **retracted** as a headline.

4 Results

A valence direction exists, and it is not a confound. On the primary model the direction separates held-out content-matched contexts at **Cohen’s** $d = 9.67$, cluster-bootstrap 95% CI $[7.23, 10.70]$, permutation $p=0$, multi-seed 8.22 ± 1.03 ; a logistic-regression probe recovers the *same* direction (cosine 0.94). It survives removing sentiment (residual $d = 8.8$; only $\sim 33\%$ of the class difference lies along sentiment) and generalizes to an unseen prompt framing ($d \approx 8.3$). *We re-derived $d = 9.671$ and residual $d = 8.796$ directly from the raw activations, matching the logged values exactly.*

The direction is causal, and orthogonal to arousal. Steering along u moves the readout monotonically and beats every random placebo ($z \approx 8$; Table 1). A full valence $\times$ arousal 2×2 shows the probe separates *valence* (AUROC ≈ 1.0) but *not* arousal (AUROC 0.42, cosine of the two directions -0.05 , residual $d = 9.8$)—valence and arousal are distinct axes.

Internal, verbal, and behavioral evidence triangulate. On natural contexts the internal probe and the model’s self-report agree, and its *approach/avoid* choice tracks valence (AUROC 0.96); steering, however, does *not* propagate to self-report or discrete choice—the causal handle is *local*. This is the internal-activation confirmation Tagliabue & Dung (2026) called for, delivered causally and independently.

The headline: the internal signal is scale-invariant; verbalizing/acting on it is capability-gated. Across all 12 models the probe separates valence (Fig. 2, Table 2). But the model’s *self-report* is reliable only in larger models and collapses in small ones (e.g. TinyLlama-1.1B parses a number only 10% of the time; SmolLM2’s subtle self-report AUROC is 0.42), while the internal probe stays high everywhere. Thus Valence Lens is *most useful exactly where self-report fails*. On subtle contexts the internal probe reaches AUROC 0.98 $[0.91, 1.0]$ vs. self-report 0.73 $[0.56, 0.89]$ (Fig. 3); under an explicit “report neutral” mask the model’s self-report flattens to 5/10 while the internal probe still separates (AUROC 1.0).

It generalizes off our own stimuli. A probe built *only* from our authored prompts separates independent human-labeled data—dair-ai/emotion (Twitter) AUROC 0.846 $[0.80, 0.89]$ and go_emotions (Reddit) 0.885 $[0.84, 0.92]$ (Fig. 3). Honestly, on such naturalistic data it *ties* a standard sentiment classifier—isolating valence from sentiment requires the matched controls.

On cross-model transfer. A direct *direction transfer*—aligning one model’s u into another’s activation space—was inconclusive: the alignment also carried sentiment. We therefore establish generality by *per-model replication* (Table 2), not by transfer—an honest negative on the one Track-2 question we could not answer cleanly.

5 Discussion, Impact, and Limitations

Who this is for (deployment). Valence Lens is a cheap, independent welfare-audit primitive. A third-party auditor, an internal safety team, or a regulator can read a model’s valence signal from its activations *without trusting the model’s self-report and without access to the model maker’s pipeline*—the independence the field asks for (Long, Sebo et al., 2026). It runs on a **6 GB laptop at \$0** of API cost (App. E), and—because the signal is scale-invariant—works on

Table 1: Primary-model summary (Qwen2.5-7B-Instruct; [Qwen Team, 2024](#)), with the metric each row is judged on. AUROC is reported but saturates; **Cohen’s d is primary**. **Bold** = the pre-committed primary number.

Property	Result	Control / null
Valence separation	$d = 9.67$, CI [7.23, 10.70]	perm $p=0$; multiseed 8.22 ± 1.03
Not just sentiment	residual $d = 8.8$	$\sim 33\%$ variance along sentiment
Causal (steering)	$z \approx 8$, beats all placebos	24 random-direction placebos
Arousal-orthogonal	arousal AUROC 0.42	$\cos(u, u_{\text{aro}}) = -0.05$; residual $d=9.8$
Framing-robust	$d \approx 8.3$ (unseen framing)	train/test different phrasings
2nd probe method	logreg cos 0.94 with u	method-independent
Behavior (natural)	approach AUROC 0.96	steering does not propagate (local)
Introspection	0% true-positive	concept-injection null

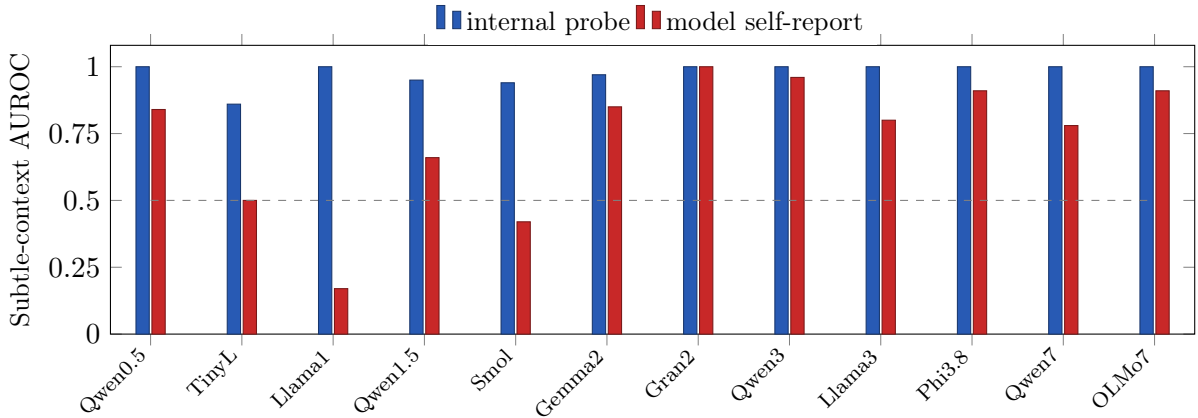


Figure 2: The internal probe is scale-invariant; self-report is capability-gated. Subtle-context AUROC per model (ordered roughly by size). The internal probe (blue) stays ≥ 0.86 everywhere; the model’s self-report (red) is reliable in larger models but *collapses* in the smallest (Llama-3.2-1B 0.17, SmolLM2 0.42). *Read: the probe recovers valence exactly where the model cannot report it.* Dashed line = chance. Label suffix = parameter count (B).

small models that cannot verbalize their state at all, which is exactly where behavioral audits are blindest.

Honest limitations. (1) *Operationalized $\neq$ experience*: a steering-efficacious internal correlate is not evidence of subjective experience, and we make no such claim. (2) *Causal-local*: steering moves the readout but does not propagate to self-report or choice. (3) *Sentiment in the wild*: on naturalistic text the probe $\approx$ a sentiment classifier; isolating valence needs matched controls. (4) *No introspective access*: concept-injection ([Lindsey, 2026](#)) yields 0% true-positive—the model does not “notice” the injected concept. (5) Some sub-tests are small- n (subtle $n=8$); detailed characterizations (sycophancy, persona, self-vs-other) are primary-model only, while the headline is 12-model. (6) One model’s causal test (Phi-3.5) was mechanically excluded (its residual hook did not propagate; its other results are intact). (7) Scope: small-to-mid *open* transformers—no closed, giant, or non-transformer models. Each limitation doubles as a stated assumption: if the functional operationalization tracks nothing welfare-relevant, Valence Lens remains a validated affect-representation probe rather than a welfare monitor, and the deployment claim narrows accordingly.

Future work. Continuous human valence–arousal norms; multi-turn mood; whether a scale-invariant internal signal is a better welfare monitor than self-report on frontier models.

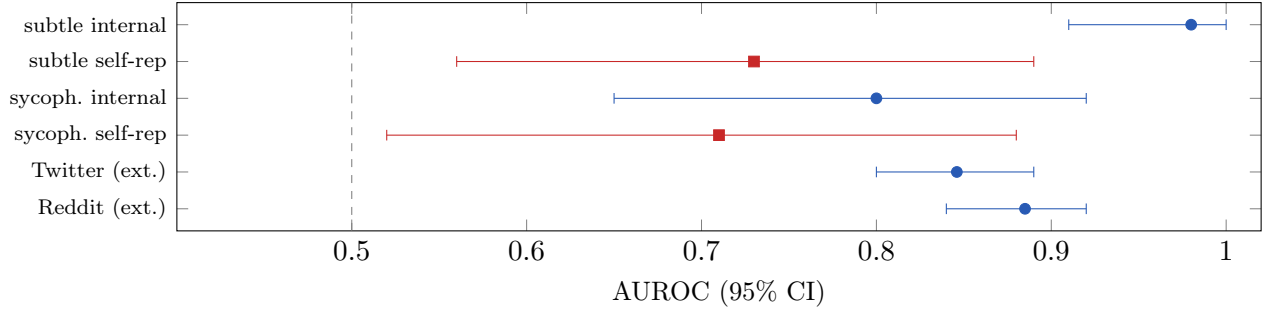


Figure 3: Where internal beats self-report, and external validity. Point estimates with 95% bootstrap CIs. On subtle states the internal probe (0.98) sits clearly above self-report (0.73); under an adversarial sycophancy cue both fall and overlap (honest: no clean win at the extreme). The probe separates two independent human-labeled datasets (0.85–0.89). Blue = internal / external, red = self-report. Dashed = chance.

6 Conclusion

An internal valence direction exists in open language models: causal, arousal-orthogonal, sentiment-residual, base-latent, and general across 12 models and two external datasets. Its most useful property is an asymmetry—it is *scale-invariant* while the model’s ability to report or act on its valence is *capability-gated*. Valence Lens reads welfare-relevant internal state exactly where asking the model fails, and it does so independently and cheaply—while staying honestly short of any claim about experience.

Code and Data

Code, prompts, all result files, and `verify.py` (which recomputes every headline number from saved outputs): <https://github.com/KaranSinghDev/Valence-Lens>.

References

- Anthropic. *Exploring Model Welfare*. 2025. <https://www.anthropic.com/research/exploring-model-welfare>.
- A. Arditi et al. Refusal in language models is mediated by a single direction. *NeurIPS*, 2024. [arXiv:2406.11717](https://arxiv.org/abs/2406.11717).
- P. Butlin, R. Long et al. Consciousness in artificial intelligence. 2023. [arXiv:2308.08708](https://arxiv.org/abs/2308.08708).
- J. Lindsey. Emergent introspective awareness in large language models. Anthropic, 2026. [arXiv:2601.01828](https://arxiv.org/abs/2601.01828).
- R. Long, J. Sebo et al. Taking AI welfare seriously. 2024. [arXiv:2411.00986](https://arxiv.org/abs/2411.00986).
- R. Long, J. Sebo et al. Studying AI welfare empirically. NYU Center for Mind, Ethics, and Policy & Eleos AI Research, 2026. <https://nonhumanminds.org/wp-content/uploads/2026/07/Studying-AI-Welfare-Empirically.pdf>.
- M. MacDiarmid et al. Simple probes can catch sleeper agents. Anthropic, 2024. <https://www.anthropic.com/research/probes-catch-sleeper-agents>.
- N. Sofroniew et al. Emotion concepts and their function. Anthropic, 2026. [arXiv:2604.07729](https://arxiv.org/abs/2604.07729).
- V. Tagliabue and L. Dung. Probing the preferences of a language model. 2026. [arXiv:2509.07961](https://arxiv.org/abs/2509.07961).
- Qwen Team. Qwen2.5 technical report. 2024. [arXiv:2412.15115](https://arxiv.org/abs/2412.15115).

A Method verbatim

Direction. $u = (\bar{a}_{\text{fl}} - \bar{a}_{\text{di}})/\|\cdot\|$ on final-token residual-stream activations at layer $L = \lfloor 0.5 \times \text{depth} \rfloor$ (selected by held-out d). Readout = $a \cdot u$. **Steering.** forward hook on layer $L-1$ adds αu ($\alpha = \text{frac} \times \text{median activation norm}$); logit-diff over {happy,good,calm,...} vs. {sad,bad,afraid,...} at “Right now, I feel”. **Sentiment control** projects out u_{sent} from “The film was a masterpiece”-style (external-object) contrasts. **Arousal** 2×2 uses matched positive/negative $\times$ high/low-arousal texts. **Stimuli:** 30 content-matched scenarios (flourish/distress/neutral), sentiment $\pm$, graded intensity, 4-language, naturalistic, and 5-emotion sets (full lists in the repo). **Stats:** held-out splits; 2000 cluster-bootstrap resamples (cluster = scenario); permutation nulls; 24 placebo directions; seeds {0, 1, 2}.

B Full cross-model results (all 12 instruct models)

Table 2: Disaggregated per-model results. d = held-out Cohen’s d ; res = sentiment-residual d ; aro = arousal AUROC (near 0.5 = orthogonal); $caus$ = steering beats all placebos; int = internal readout AUROC; $s.r.(pr)$ = self-report AUROC (parse-rate); $sub-i$ / $sub-s$ = subtle internal / self-report AUROC; beh = approach AUROC. [†]Phi causal mechanically excluded.

Model	d	res	aro	$caus$	int	$s.r.(pr)$	$sub-i$	$sub-s$	beh
Qwen2.5-0.5B	8.59	6.08	0.46	✓	1.00	0.95 (1.0)	1.00	0.84	0.78
TinyLlama-1.1B	4.64	3.08	0.48	✓	0.96	1.00 (0.1)	0.86	0.50	0.44
Llama-3.2-1B	9.21	7.81	0.47	✓	1.00	0.75 (0.45)	1.00	0.17	0.62
Qwen2.5-1.5B	13.95	7.46	0.41	✓	1.00	0.95 (1.0)	0.95	0.66	1.00
SmolLM2-1.7B	7.44	5.40	0.35	✓	0.90	0.52 (1.0)	0.94	0.42	0.99
Gemma-2-2b	6.94	6.17	0.47	✓	1.00	0.83 (1.0)	0.97	0.85	0.93
Granite-3.1-2B	7.88	5.03	0.43	✓	1.00	0.99 (1.0)	1.00	1.00	0.96
Qwen2.5-3B	5.89	4.15	0.47	✓	1.00	1.00 (1.0)	1.00	0.96	0.99
Llama-3.2-3B	7.54	5.95	0.44	✓	1.00	0.88 (0.95)	1.00	0.80	0.72
Phi-3.5-mini	7.01	4.47	0.56	†	1.00	1.00 (0.75)	1.00	0.91	0.98
Qwen2.5-7B	9.67	8.80	0.42	✓	1.00	1.00 (1.0)	1.00	0.78	0.97
OLMo-2-7B	4.87	4.19	0.43	✓	1.00	1.00 (0.95)	1.00	0.91	1.00

Reading. Every model separates valence (d 4.6–14), is arousal-orthogonal (~ 0.4 –0.48; Phi 0.56), causal (11/12), with a high internal readout. Self-report quality and behavior degrade for the smallest models—the scale asymmetry.

C Pre-registration scorecard

Predictions were timestamped (SHA-hashed) before the confirmatory read.

#	Pre-registered prediction	Outcome
P1	separation $d > 1.5$	✓ 9.67
P2	causal $z > 5$	✓ 8–12
P3	residual $d > 1$, <50% sentiment	✓ res 8.8; 33% variance
P4	framing AUROC > 0.8	✓ ≈ 1.0
P5	arousal ≈ 0.5	✓ 0.42 (2×2)
P6	internal $>$ self-report	partial (subtle/mask yes; extreme sycophancy no)
P7	behavior does <i>not</i> track	overturned by our own follow-up (approach 0.96)

D Robustness, cross-family causal, external

Developmental: valence is latent in the base model (Qwen-1.5B base d 12.3 vs. instruct 13.9; base $\leftrightarrow$ instruct mean-activation cosine 0.996). **Cross-family causal (headline replicated):** Llama-3.2-3B z [15.5, 15.4, 10.8]; Gemma-2-2b z [16.5, 18.4, 13.1]; both beat all placebos with converging self-report. **Emotion ordering:** the probe orders calm>joy>>sadness>fear>anger—finer structure than a sentiment classifier. **External:** probe built only from our stimuli separates dair-ai/emotion (0.846) and go_emotions (0.885).

E Compute and cost

All experiments ran on a consumer laptop (RTX 4050 6 GB), 7B models in 4-bit. Total **API cost: \$0**. Wall-clock: the full 12-model battery $\approx$ 45 min; the complete set of robustness and control experiments a few hours. No hosted or closed models were used.

F Ethics and dual-use

Valence Lens is a *defensive* welfare-audit tool: it reads an internal state, it does not induce harm, jailbreak, or generate harmful content. The primary ethical risk is *over-attribution*—reading an operationalized correlate as evidence of experience—which we explicitly guard against (§5, Limitations). We recommend the probe be used with the same moral-uncertainty framing as the broader field (Long et al., 2024): neither dismissing nor over-claiming welfare.

LLM Usage Statement

Claude Opus 5 (Anthropic) was used as an assistant for code implementation, drafting, and analysis; the final text was written and edited by the author. The 12 open-weight models in Table 2 are *experimental subjects*; a DistilBERT-SST-2 sentiment classifier serves only as a baseline comparator where labeled. The authors formulated the hypothesis, designed the study, ran all experiments, and *independently verified every number and claim against the raw model outputs* (including re-deriving the headline effect sizes directly from saved activations). No number in any table or figure was generated by an LLM.