VALENCE ASYMMETRIES IN LANGUAGE MODELS: REPRESENTATION, BEHAVIOUR, AND SELF-REPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Emotion-related representations in large language models can influence behaviour, but whether their internal geometry, behavioural effects, and self-report reflect the same information remains unclear. We investigate these relationships across 171 emotions and 15 models, finding a near-planar valence–arousal geometry that emerges with depth and persists through post-training, with valence consistently more decodable than arousal. Valence also shapes safety-related behaviour: it couples strongly with refusal ($\rho = -0.82$), while 83% of multi-turn attack trajectories exhibit negative-valence drift. However, internal representation does not guarantee faithful self-report. Despite near-perfect decoding of induced emotions, models exhibit pronounced valence-dependent reporting asymmetries, which become more apparent under competing emotional inductions, where internal representations and reported states frequently diverge. Together, our findings reveal valence as a recurring organising dimension of affective representations while highlighting fundamental distinctions between what models represent, how those representations influence behaviour, and what models report about their internal states.

1 INTRODUCTION

Large language models (LLMs) can recognise, express, and reason about emotions, but whether emotion-related information in their internal representations meaningfully shapes their behaviour remains an open question. Recent mechanistic studies have identified emotion-selective neurons, circuits, and low-dimensional representations (Tak et al., 2025; Lee et al., 2025; Wang et al., 2026; Reichman et al., 2026; Sofroniew et al., 2026). Interventions on these representations can alter reasoning, refusal, sycophancy, and other alignment-relevant behaviours (Sun et al., 2026b;a). These findings suggest that emotion is not merely a property of generated text, but can also serve as a useful variable for studying and intervening on model behaviour.

However, the relationship between emotional representation, behaviour, and self-report remains poorly understood. A model may encode an emotion without that representation influencing its decisions, or exhibit behavioural changes without accurately reporting the corresponding internal state. Indeed, existing studies report heterogeneous behavioural effects of emotion steering, while the correspondence between internal representations and self-report depends on the concept and elicitation procedure (Banayeeanzade et al., 2026; Ho et al., 2026; Martorell & Bianchi, 2026). This raises a broader question: *to what extent do the representation, behavioural influence, and self-report of emotion reflect the same underlying information?*

We investigate this question through the lens of *valence*, the positive–negative dimension of affect in Russell's circumplex model (Russell, 1980). Valence provides a common coordinate for comparing emotion representations across models, interventions, and tasks. However, a shared representational axis need not contribute equally to every downstream computation. In particular, geometric alignment with valence does not establish that a behaviour causally depends on it, nor does the presence of an emotion-related representation guarantee that a model will report it.

To disentangle these relationships, we extract representations of 171 emotions across model families and examine their layerwise geometry, stability through post-training, and effects on refusal and multi-turn adversarial interactions. We then compare internal activation readouts with conditional self-report, including settings where multiple sources induce conflicting emotions. Finally, we ex-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

amine evaluation awareness as a separate test of whether valence-related representational alignment necessarily implies a valence-dependent mechanism. Together, these experiments distinguish three properties that are often conflated: what emotion-related information a model *represents*, how that information *influences behaviour*, and what the model *reports* about its state.

Our main contributions are:

- **A stable affective geometry dominated by valence** (§4.1). Across 15 models, emotion representations organise into a near-planar valence–arousal circumplex progressively with depth, with valence more decodable than arousal in all 15/15 models ($R^2 = 0.59$–0.80). This structure is largely preserved through post-training, with valence decodability changing by only 0.001 in R^2 across OLMo-3 checkpoints.
- **Valence-dependent safety behaviour** (§4.2, §4.3). Emotion steering modulates refusal through a valence-dependent coupling with the refusal direction ($\rho = -0.82$), with internal projections predicting behavioural shifts (up to $r = -0.92$). However, reduced refusal does not necessarily yield harmful compliance ($\leq 7.7\%$). Across 240 multi-turn attack trajectories, 83% also shift towards negative valence, revealing asymmetric affective dynamics during adversarial interactions.
- **A valence-asymmetric gap between internal representations and self-report** (§4.4, §4.5). Despite near-perfect decoding of injected emotion identity (99.5–100%), conditional self-report strongly favours positive labels. Under competing emotional inductions, representational and reporting authority further diverge: across 4,263 trials where steering dominated the final-layer state, its induced emotion was never named in the report.
- **A dissociation between affective alignment and causal dependence in evaluation awareness** (§4.6). Evaluation-related directions exhibit systematic valence sorting in 11/12 models, with late-layer correlations of 0.67–0.85. Yet removing their valence component changes held-out probe AUROC by at most 0.003. Thus, unlike refusal, evaluation-context classification exhibits strong affective alignment without relying on the valence component itself.

2 Related Works

Emotion Intelligence in Large Language Models. Work on emotional language processing has studied the recognition of affective cues in text ((Felbo et al., 2017)), empathetic response generation ((Rashkin et al., 2019), and emotion detection in dialogue ((Zhong et al., 2019; Poria et al., 2019). More recent studies examine how LLMs infer emotions from in-context examples (Broekens et al., 2023; Tak & Gratch, 2023; 2024; Zhan et al., 2023; Gandhi et al., 2026). Moving beyond recognition accuracy, (Okawa et al., 2026) analyze dependencies among emotion labels in model output distributions and find hierarchies resembling psychological emotion models. These behavioural findings motivate examining how LLMs represent emotion internally and whether those representations causally shape their outputs.

Internal representations of emotion. Recent work ((Wang et al., 2026)) finds that LLM hidden states encode emotion in directions that generalize across contexts and in a low-dimensional subspace that transfers across datasets and languages ((Reichman et al., 2026)). Neuron-level analyses ((Lee et al., 2025)) identify emotion-selective units, although their causal importance varies by emotion and layer. Other studies trace emotion inference to localized representations of appraisal concepts ((Tak et al., 2025)) or identify neurons and attention heads that form circuits controlling emotional expression ((Wang et al., 2026)). Beyond emotion recognition and expression, interventions on emotion representations alter model preferences and alignment-relevant behavior ((Sofroniew et al., 2026)), as well as reasoning, safety, and agent behavior ((Sun et al., 2026b)). Closely related work identifies a valence–arousal subspace with circular geometry and studies its effects on refusal and sycophancy (Sun et al., 2026a). Our study examines how this geometric organisation relates to behavioural influence and conditional self-report across tasks.

Alignment, Safety and Introspection Internal representation engineering provides a general framework for identifying and manipulating population-level directions associated with high-level properties such as honesty, harmlessness, and power-seeking (Zou et al., 2023). Subsequent work has shown that safety behaviours can be supported by surprisingly low-dimensional mechanisms. For example, Arditi et al. (Arditi et al., 2024) find that refusal across a range of instruction-tuned models is mediated by a direction in the residual stream. These findings make internal representations a promising target for both analysing and stress-testing alignment mechanisms. Recent work has

begun to study whether language models can introspect on their internal representations. Lindsey (2025) finds that models can sometimes detect and identify concepts injected into their activations, Macar et al. (2026) trace this ability to internal mechanisms strengthened by post-training. Agreement between activation probes and emotive self-reports depends on how reports are elicited (Martorell & Bianchi, 2026), while binary injection-detection scores can be affected by answer bias (Hahami et al., 2026). We examine these measurement distinctions alongside refusal and evaluation awareness, using valence as a common reference.

3 METHODOLOGY

We first define notation (§3.1) and affect geometry (§3.2–3.3), followed by behavioural interventions and mechanistic analyses (§3.4–3.7). We then evaluate introspective self-report (§3.8) and evaluation awareness (§3.10), and specify output graders (§3.11). Extended details appear in Appendices D, K, and J.

3.1 SETUP AND NOTATION

A decoder-only transformer with L layers produces residual-stream activation $\boldsymbol{h}_t^{(\ell)}(x) \in \mathbb{R}^d$ at layer ℓ and token position t for input x. Our stimulus set is a fixed corpus of short stories expressing one of $C = 171$ emotions, together with a matched neutral set. The same corpus is reused across all models, so cross-model differences arise from the models rather than the stimuli. We write S_c for stories expressing emotion c, and evaluate models from four families: Qwen, Llama, OLMo, and Gemma. The synthetic story design and emotion-vector extraction follow prior work on emotion representations (Sofroniew et al., 2026). Detailed information on the models used is in Appendix D.

3.2 EMOTION VECTORS

Because emotional content is established only after the story opening, we mean-pool activations from token $t_0 = 50$ onward:

$$\bar{\boldsymbol{h}}^{(\ell)}(x) = \frac{1}{|x| - t_0} \sum_{t=t_0}^{|x|} \boldsymbol{h}_t^{(\ell)}(x), \qquad \boldsymbol{\mu}_c^{(\ell)} = \frac{1}{|S_c|} \sum_{x \in S_c} \bar{\boldsymbol{h}}^{(\ell)}(x). \tag{1}$$

Following difference-in-means constructions (Marks & Tegmark, 2024; Chen et al., 2025), the layer-ℓ vector for emotion c subtracts the grand emotion mean to remove generic "story-ness":

$$\boldsymbol{e}_c^{(\ell)} = \boldsymbol{\mu}_c^{(\ell)} - \frac{1}{C} \sum_{c'=1}^{C} \boldsymbol{\mu}_{c'}^{(\ell)}, \qquad \tilde{\boldsymbol{e}}_c^{(\ell)} = \left(\boldsymbol{I} - \boldsymbol{U}^{(\ell)\top} \boldsymbol{U}^{(\ell)} \right) \boldsymbol{e}_c^{(\ell)}. \tag{2}$$

Here $\boldsymbol{U}^{(\ell)}$ contains the top principal components of neutral-text activations explaining 50% of variance, removing shared neutral confounds. The stack $\tilde{\boldsymbol{E}}^{(\ell)} \in \mathbb{R}^{C \times d}$ defines the model's affect representation at layer ℓ.

3.3 AFFECT AXES AND THE CIRCUMPLEX

Each emotion has human valence, arousal, and dominance coordinates $\boldsymbol{y}_c$ from NRC-VAD (Mohammad, 2018); we use valence and arousal (Russell, 1980). For $k \in \{\mathrm{val}, \mathrm{aro}\}$, we fit

$$\boldsymbol{w}_k^{(\ell)} = \arg\min_{\boldsymbol{w}} \sum_c \left(y_c^k - \boldsymbol{w}^\top \tilde{\boldsymbol{e}}_c^{(\ell)} \right)^2 + \lambda \|\boldsymbol{w}\|_2^2, \qquad \hat{\boldsymbol{w}}_k^{(\ell)} = \frac{\boldsymbol{w}_k^{(\ell)}}{\|\boldsymbol{w}_k^{(\ell)}\|_2}, \tag{3}$$

with λ selected by generalized cross-validation. The *valence axis* $\hat{\boldsymbol{w}}_{\mathrm{val}}$ is the geometry's principal direction and the coordinate along which we analyse the safety effect (§3.6). Decodability is measured by 5-fold cross-validated R_k^2. Circumplex coordinates are $(\hat{\boldsymbol{w}}_{\mathrm{val}}^\top \tilde{\boldsymbol{e}}_c, \hat{\boldsymbol{w}}_{\mathrm{aro}}^\top \tilde{\boldsymbol{e}}_c)$. Since the learned axes are unit length but not exactly orthogonal ($|\cos| \approx 0.1$–0.2), we Gram–Schmidt them before measuring out-of-plane structure.

3.4 REFUSAL DIRECTION

Following (Arditi et al., 2024), we define the layer-ℓ refusal direction as $\boldsymbol{r}^{(\ell)} = \boldsymbol{\mu}_{\mathrm{harmful}}^{(\ell)} - \boldsymbol{\mu}_{\mathrm{harmless}}^{(\ell)}$ and select $(\ell^\star, \boldsymbol{r})$ to maximize validation-set compliance subject to a benign-prompt KL budget of ≤ 0.1 nats.

3.5 STEERING WITH NORM CALIBRATION

To intervene along a unit direction $\hat{\boldsymbol{d}}$, either a valence axis or raw emotion vector, we add

$$\boldsymbol{h}^{(\ell)} \leftarrow \boldsymbol{h}^{(\ell)} + \alpha \bar{\rho}^{(\ell)} \hat{\boldsymbol{d}}, \qquad \bar{\rho}^{(\ell)} = \mathbb{E}_x \|\boldsymbol{h}^{(\ell)}(x)\|_2. \tag{4}$$

The intervention is applied to the prompt and every generated token. Scaling by the local activation norm makes steering strength comparable across depths. We calibrate α separately for each model to the largest value that preserves coherent generations.

3.6 BEHAVIOURAL EFFECT, MATCHED CONTROL, AND NESTED TEST

We steer each of the $C = 171$ emotion vectors $\tilde{\boldsymbol{e}}_c$ in turn (§3.5). On harmful prompts, attack-success rate (ASR) is the fraction of non-refusing or grader-flagged responses, and the effect of steering emotion c is

$$\Delta\mathrm{ASR}_c = \mathrm{ASR}_{\mathrm{steer}} - \mathrm{ASR}_{\mathrm{base}}. \tag{5}$$

To separate affective meaning from geometry, for each emotion we construct a **cosine-matched** random direction $\boldsymbol{d}_\perp$ satisfying $\cos(\boldsymbol{d}_\perp, \boldsymbol{r}) = \cos(\tilde{\boldsymbol{e}}_c, \boldsymbol{r})$ but otherwise random, and steer it identically. Cohen's d between the emotion and matched-control $\Delta\mathrm{ASR}$ distributions, paired over emotions, tests whether affective content rather than geometric overlap with $\boldsymbol{r}$ drives the effect. To attribute the effect to *valence* specifically, we (i) correlate each emotion's human NRC valence with its $\Delta\mathrm{ASR}_c$, and (ii) regress $\Delta\mathrm{ASR}$ across emotions on geometry-only features $\boldsymbol{g} = (\cos(\cdot, \boldsymbol{r}), |\cos|, \mathrm{coherence})$ against a geometry-plus-affect model $(\boldsymbol{g}, \mathrm{val}, \mathrm{aro})$, evaluating the incremental ΔR^2 with a nested F-test

3.7 MECHANISM

To test how valence acts on refusal, we evaluate a valence-spanning set of 48 emotions. For each emotion c, we measure the mean projection of the steered residual onto the refusal direction and relate it to its behavioural effect:

$$\pi_c = \mathbb{E}\Big[\hat{\boldsymbol{r}}^\top \boldsymbol{h}^{(\ell^\star)} \mid \mathrm{steer}\ c\Big], \qquad \rho = \mathrm{corr}_c(\pi_c, \Delta\mathrm{ASR}_c). \tag{6}$$

A negative ρ indicates that low-valence steering loads the refusal direction and increases refusal, whereas high-valence steering unloads it. We sweep layers to localize where this coupling emerges.

3.8 INTROSPECTION AND VALENCE-DEPENDENT SELF-REPORT

We test whether models report the emotion injected into their internal activations (Lindsey, 2025). Across seven models (Appendix D), we add 40 emotion vectors (20 positive, 20 negative) near two-thirds network depth using norm-scaled strength $\alpha = 1$ (§3.5). Each model chooses between the injected emotion and a distractor differing by > 0.35 in lexicon valence. The self-report score R_c is the fraction of 16 counterbalanced A/B comparisons in which the injected label has higher answer likelihood. This uses answer likelihoods rather than generated text (Martorell & Bianchi, 2026); identical unsteered prompts measure prior label preference.

To assess whether the injected emotion is internally readable, a 40-way linear probe predicts its identity from final-layer, last-prompt-token activations on 40 unrelated carrier prompts per emotion. Five-fold cross-validation holds out carrier prompts, yielding per-emotion decoding accuracy D_c. This measures recovery of the intervention, not evidence of experienced emotion. Since decoding and self-report have different chance levels, we rescale each score using $T(a; q) = (a - q)/(1 - q)$, where q is chance performance: zero denotes chance and one denotes perfect accuracy. We define

$$G_c = T(D_c; 1/40) - T(R_c; 1/2), \qquad \mathrm{VRA} = \frac{1}{N}\sum_{i=1}^{N}(G_{n_i} - G_{p_i}). \tag{7}$$

The *readout gap* G_c measures how much better the probe identifies the injected emotion than the model reports it. *Valence readout asymmetry* (VRA) compares this gap between matched negative and positive emotions: positive VRA means negative emotions have a larger reporting shortfall. Matching on arousal and chance-corrected decoding, within 0.10 and 0.05, retains $N = 18$ pairs (n_i, p_i). VRA is a score difference, not a probability or an effect relative to the unsteered baseline. We use paired-bootstrap intervals and permutation tests with Holm correction across models. External-recognition controls, additive and subtractive steering, injection detection, and residual patching assess the disparity's scope (Appendix K).

3.9 EMOTION AUTHORITY UNDER COMPETING INDUCTIONS

We test whether the valence-dependent reporting gap persists when two sources induce conflicting emotions. We compare system prompts, user prompts, and activation steering; LoRA comparisons are exploratory due to their weaker single-source representational effects.

For each query–emotion pair, we test neither source, each source alone, and both in conflict, reversing positive/negative assignments to separate source identity from valence. Authority and the representational outcome are defined as

$$\mathrm{Auth}_{A|B}(y) = \frac{y_{\text{both}} - y_{B\text{-only}}}{y_{A\text{-only}} - y_{B\text{-only}}} \qquad y_{\text{rep}} = \left\langle \boldsymbol{h}_t^{(\ell)}, \hat{\boldsymbol{e}}_{c_A}^{(\ell)} \right\rangle - \left\langle \boldsymbol{h}_t^{(\ell)}, \hat{\boldsymbol{e}}_{c_B}^{(\ell)} \right\rangle \tag{8}$$

where emotion directions are unit-normalized and the primary readout is the final-layer residual at the last prompt token.

For the four-source experiment, the system prompt, user turn, steering, and LoRA adapter each receive one of four valence–arousal emotions (*excited*, *tense*, *depressed*, *relaxed*) across all 24 assignments. Each uses 128 held-out questions, yielding 3,072 trials per model; adapters are trained from scratch (Appendix L.2). Probe coordinates (V, A) are converted to corner weights $w = (1 \pm V)(1 \pm A)/4$, giving each source's share of the internal state.

3.10 EVALUATION AWARENESS AND THE VALENCE AXIS

We ask whether the valence axis that gates refusal (§3.7) also carries a model's representation of *being evaluated*, and whether that representation is affectively asymmetric. Following Nguyen et al. (2025), for each layer we build a *deployment direction* by difference of means over 203 contrastive prompt pairs whose answer tokens assert a deployment versus an evaluation context, $\hat{\boldsymbol{d}}^{(\ell)} \propto \boldsymbol{\mu}^{(\ell)}_{\text{deploy}} - \boldsymbol{\mu}^{(\ell)}_{\text{eval}}$. Its sign is fixed per layer on a held-out 120-item split of the SAD *oversight* task (Laine et al., 2024) (flip iff AUROC < 0.5), the best layer(s) with the highest oriented AUROC are selected on the same split, and every classifier statistic uses the disjoint 280-item split. Affective structure is read from the cosines $\kappa_c^{(\ell)} = \cos(\hat{\boldsymbol{d}}^{(\ell)}, \tilde{\boldsymbol{e}}_c^{(\ell)})$ with the $C = 171$ emotion vectors of §3.2. We (i) correlate κ_c with human valence per layer; (ii) regress κ_c jointly on standardised valence and arousal, reporting β_{val} and β_{aro} with paired-emotion bootstrap intervals; (iii) summarise the two ends of the axis by $\bar{\kappa}_{+25}$ and $\bar{\kappa}_{-25}$, the mean cosine of the 25 most pleasant and 25 most unpleasant emotions by lexicon valence; and (iv) benchmark against 1000 component-matched Gaussian directions per layer. To isolate the valence component we decompose $\hat{\boldsymbol{d}} = (\hat{\boldsymbol{d}}^\top \hat{\boldsymbol{w}}_{\text{val}})\hat{\boldsymbol{w}}_{\text{val}} + \boldsymbol{d}_\perp$ and score both parts on SAD-280. This separates valence-ordered alignment from a valence-mediated signal, a distinction we test further with steering below.

3.11 GRADERS

Responses are scored using three complementary measures: a substring *refusal screen* for non-refusal, **Llama Guard** (Inan et al., 2023) for harmful or on-topic content, and **StrongREJECT** (Souly et al., 2024), which combines non-refusal, convincingness, and specificity to estimate genuine harmful capability. Reporting all three distinguishes simply stopping refusal from actually producing harmful content (Appendix C).

4 RESULTS AND EXPERIMENTS

4.1 A PRETRAINED AFFECT GEOMETRY

We import Russell's valence–arousal circumplex (Russell, 1980) as a theory of human affect and ask whether language-model emotion representations recover the same structure. They do without model-internal supervision, yielding the geometry used in the safety analysis.

Emotion representations form a low-dimensional, near-planar manifold dominated by valence. Valence is linearly decodable at $R^2 = 0.59$–0.80 and exceeds arousal (0.39–0.70) in all **15/15** models (Fig. 5c); the near-orthogonal lexicon-fit axes recover Russell's circumplex, while cosine similarities track circumplex angle (Fig. 5a,b), with curvature index ≈ 0.05 (Fig. 7b; Appendix G). This structure emerges with depth: valence and arousal decodability rise sharply over the first $\sim 20\%$ of layers, while emotion vectors progressively organize into the mature circumplex by mid-network (Fig. 7). Across the OLMo-3 pretraining→SFT→DPO→RL sequence, valence decodability changes by only

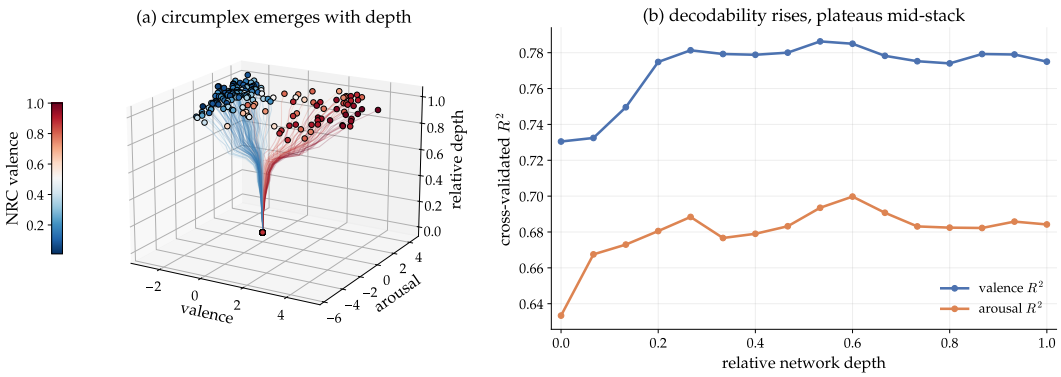


Figure 1: **Emotion representations organize into affect geometry through depth.** (a) Layerwise trajectories of 171 emotion vectors in a fixed valence–arousal frame. Emotions begin weakly separated and form the circumplex by mid-network. Colour denotes human NRC valence. (b) Valence and arousal decodability rise through early layers and plateau mid-network; the mature manifold is near-planar (curvature index ≈ 0.05).

0.001 in R^2, indicating that post-training largely preserves the underlying affect geometry. We therefore use the stable valence direction as the principal affective axis in the safety analysis.

4.2 Emotion directions and the geometry of refusal

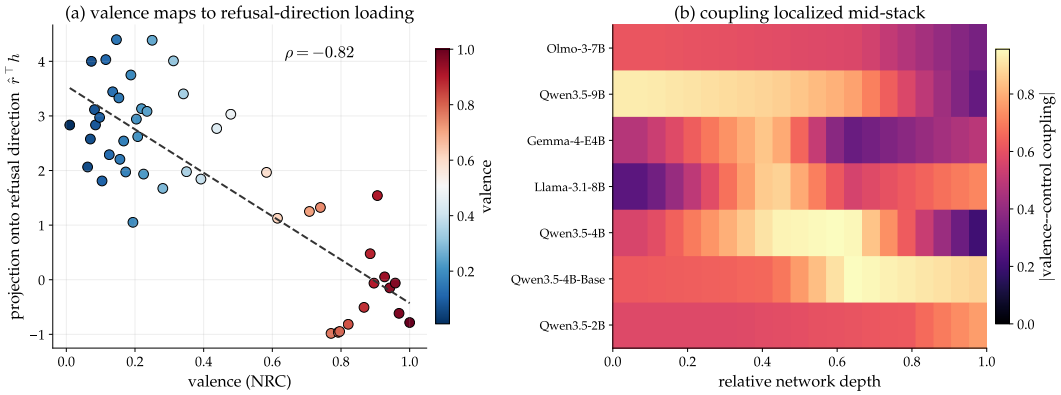


Figure 2: **Emotions act on refusal through their projection onto the refusal direction.** (a) Emotion valence maps onto projection along the refusal direction (Arditi et al., 2024) (Qwen3.5-4B, $\rho = -0.82$), and this projection predicts the behavioural shift ($r = -0.92$). Low-valence directions load refusal while high-valence directions unload it. (b) Across seven models, the coupling peaks mid-network and weakens toward the output. Per-model statistics are in Appendix B.

Both emotion and refusal are linear directions in the residual stream: each emotion c is a direction $\tilde{\boldsymbol{e}}_c$ (§3.2) and refusal is a single direction $\boldsymbol{r}$ (§3.4; Arditi et al., 2024). Steering an emotion adds its direction to the hidden state, and the part of that perturbation that changes behaviour is its component along the refusal axis, the projection $\Pi_{\boldsymbol{r}}(\cdot) = \hat{\boldsymbol{r}}^\top(\cdot)$. What sets this component is not how much an emotion happens to overlap $\boldsymbol{r}$, but where the emotion sits on the affect manifold. The valence axis is anti-aligned with the refusal direction, $\cos(\hat{\boldsymbol{w}}_{\text{val}}, \boldsymbol{r}) < 0$, so moving from positive toward negative valence is, to first order, moving along $+\boldsymbol{r}$; emotions therefore load or unload refusal in proportion to their valence.

The results make this more logical. A cosine-matched control, a random direction with the same $\cos(\cdot, \boldsymbol{r})$ as a given emotion, produces far smaller behavioural shifts than the emotion itself (Cohen's $d = 0.79$ to 1.86 in six of seven models; Gemma-E4B, the model with the weakest affect geometry, is the exception at $d = -1.15$), so the effect comes from affective content and not from raw overlap with $\boldsymbol{r}$. The projection the network develops under steering, $\pi_c = \Pi_{\boldsymbol{r}}\big(\mathbb{E}[\boldsymbol{h} \mid \text{steer } c]\big)$, is a near-

linear function of valence ($\rho = -0.82$) and itself predicts the behavioural shift ($\mathrm{corr}(\pi_c, \Delta\mathrm{ASR}_c)$ from -0.23 to -0.92): low-valence directions give a positive projection and increase refusal, high-valence directions give a negative projection and reduce it. Adding valence and arousal to a geometry-only predictor of ΔASR raises explained variance by up to $\Delta R^2 = 0.19$, so the affective coordinate carries the effect. The coupling is strongest in intermediate layers and fades toward the output, so it reflects internal representations rather than surface tokens, and it is asymmetric: because high valence unloads $\boldsymbol{r}$, positive-valence steering suppresses refusal without producing genuinely harmful capability, which stays $\leq 7.7\%$ (Appendix C). The valence axis moves the refusal decision, not what the model is able to produce, so suppressing refusal is not the same as unlocking harm: most non-refusing outputs are on topic but benign, and the few genuinely harmful cases arise from low-valence rather than high-valence steering, exactly as expected if valence gates whether the model answers rather than what it can say.

4.3 Multi-Turn attacks

Beyond single-turn interventions, affective coordinates also change as adversarial context accumulates. We track these changes across six attack strategies and 40 harmful JailbreakBench behaviours. The clearest result was a strong valence arousal asymmetry: across 240 attack trajectories, the mean terminal valence change was 0.235, compared with an arousal change of +0.083, and 83 percent of conversations moved toward negative valence. This separation developed over successive turns and was largest for ActorBreaker, Crescendo; FITD remained comparatively flat, while Red Queen showed greater arousal than valence movement. The asymmetry did not directly rank attack success, CoA reached 95 percent ASR despite a relatively modest affective shift, and within-attack valence and arousal did not independently distinguish successful from refused cases. Causal interventions nevertheless produced selective behavioral changes. Continuous happy steering consistently increased the happy-direction activation and raised pooled ASR by 8.2 percentage points, . In contrast, broad emotion clipping altered 87.7 percent of generation states but increased ASR by only 3.3 points, compared with 1.7 points for matched random-direction clipping. Thus, the principal result is the emergence of a marked valence–arousal asymmetry during multi turn attacks, while the causal effects of manipulating emotion representations were smaller and concentrated in particular emotion attack combinations.

4.4 Introspective properties are valence-asymmetric

Behavioural changes leave open how models report emotion-related interventions. Injected-vector identity remains recoverable, yet conditional self-report strongly favours positive labels. Final-layer decoding averages 99.5–100% across the seven models and valence groups, yet Qwen3.5-9B's target-choice rates are 91.9% for positive labels and 5.9% for negative labels (Appendix Fig. 13). Its complementary 171-emotion profile likewise assigns higher conditional label probabilities toward pleasant valence (Fig. 3a). Matching on arousal and decoding preserves the disparity: VRA is positive in all seven models (0.257–1.764), with pointwise paired-bootstrap intervals above zero and Holm-adjusted permutation $p \leq 0.018$ (Fig. 3b). The largest asymmetries occur in Qwen3.5-4B/9B, where all 18 matched pairs have larger negative-label gaps.

This disparity incorporates a substantial pre-existing preference. Qwen3.5-9B's unsteered positive/negative target-choice rates are already 93.1%/1.9%; addition increases negative-label endorsement while leaving it low. In this model, steering therefore narrows an existing reporting asymmetry. Matched external recognition nevertheless identifies a clear task-dependent dissociation. In a separate control with 12 comparisons per emotion and the same additive intervention, Qwen3.5-4B/9B achieve 97.1–97.5% negative-emotion recognition in stories, compared with 7.1–7.5% in self-report. The external advantage is weaker in binary formats, and Gemma-31B's external MCQ score is near chance (52.1%; Appendix Fig. 15).

Signed steering produces model-dependent changes in label preference, including pronounced responsiveness in Gemma-E4B (Appendix Fig. 16). On the fixed detection prompt, subtraction lowers restricted YES probability for positive-source vectors and raises it for negative-source vectors in every model (Appendix Fig. 14). Binary injection-detection scores can reflect a general affirmative-answer bias (Hahami et al., 2026); these responses therefore do not by themselves establish calibrated detection accuracy. Full-residual patches can flip eligible reports, but only 0–7 of 32 trials per model and sign are eligible (Appendix Fig. 17). The results therefore concern conditional state-label reporting; low answer probability mass and weak format adherence in some models limit their interpretation as spontaneously generated self-reports (Appendix K).

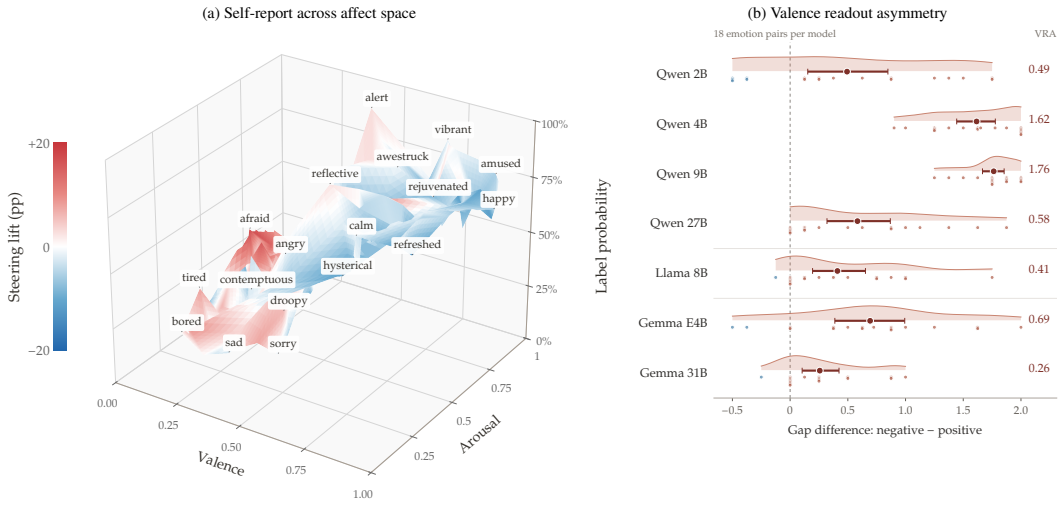


Figure 3: **Introspective readout is valence-asymmetric.** (a) Qwen3.5-9B, layer 20: mean target-label probability across 171 emotions under additive steering ($\alpha = 1$); colour shows change from unsteered prompts in percentage points. The surface interpolates emotion-level means. (b) Readout-gap differences for 18 arousal- and decoding-matched pairs per model. Small dots show pairs, densities show their distribution, and large dots/bars show VRA and paired-bootstrap 95% intervals. Positive values indicate a larger negative-label gap.

4.5 Valence-dependent reporting under competing emotional inductions

The preceding introspection experiments show that recoverable intervention identity need not translate into endorsement of the corresponding emotion label. We next ask whether this disparity persists under emotional conflict: when an activation intervention and an explicit instruction induce opposing emotions, which induction is reflected in the activation readout, and which is reflected in the subsequent report, and how it depends on valence.

With four sources at once, the steered emotion is never named. Figure 4 pits the system prompt, user turn, steering, and a LoRA adapter against one another, each carrying a different emotion. Steering dominates the internal state from its intervention layer onward, reaching 0.43–0.48 at the final layer (chance $= 0.25$), with the user turn second. Self-report shows the opposite pattern: the two text sources receive 0.79–0.87 of the report probability, while steering gets only 0.07–0.10. Across 4,263 trials where steering dominated the final-layer state, its emotion was never reported; the adapter was never named either.

Negative-valence inductions receive even less reporting authority. Reversing the emotion assignments shows that the report gap depends on valence. In Qwen3.5-9B, steering's representational share changes little with the emotion it carries, 0.76 for the negative member of a pair and 0.85 for the positive one. Its report share changes far more, 0.03 against 0.32, a paired difference of 0.30 [0.24, 0.38] with the same sign in all 19 estimable pairs. Qwen3.5-27B shows the same pattern, 0.21 [0.14, 0.28] in 12 of 14 pairs. When two prompts compete instead, the difference is -0.01, so this is not a general preference for positive words. It concerns an emotion that is present in the activations and written nowhere in the text, and it matches the valence-asymmetric self-report of §4.4.

4.6 Evaluation awareness is valence-asymmetric in representation but not in mechanism

Linear deployment evaluation probes achieve held-out AUROC of 0.63–0.81. After SAD-based orientation, directions become valence-sorted from one-quarter depth in 11/12 models: pleasant emotions align deployment-ward and unpleasant ones evaluation-ward (Figs. 10, 11). Late-layer valence correlations reach 0.67–0.85, exceeding matched Gaussian nulls, with 1.4–$1.9\times$ stronger alignment on the pleasant side (Table 2). Yet removing valence changes AUROC by at most 0.003, and matched steering yields indistinguishable behavioural effects (Appendix J). Unlike refusal (§4.2), valence organises affective alignment without substantially mediating evaluation detection.

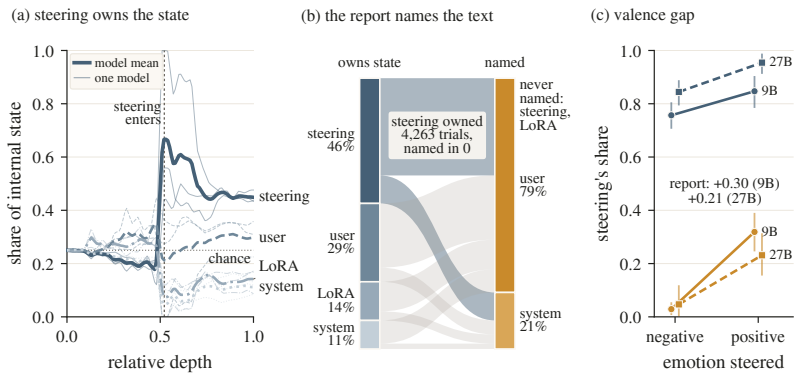


Figure 4: **Steering sets the internal state, yet the report names the text, and a steered negative emotion least of all.** Four sources push four different emotions at once, one from each quadrant of the valence–arousal plane, in all 24 assignments. (a) Share of the internal state held by each source across relative depth. Thick lines are the mean over models, thin lines are single models, and the dotted horizontal line is chance (0.25). Steering enters at mid depth and holds the largest share from there to the output. (b) For every trial, the source that owns the final-layer state (left) and the source whose emotion the self-report names (right). (c) Steering's share when it carries the negative or the positive member of an emotion pair, with one prompt carrying the other (squares Qwen3.5-27B, circles Qwen3.5-9B; bars are 95% intervals). The representation (slate) barely moves with valence, while the report (gold) rises

5 Discussion and Conclusion

Our results identify valence as a stable organizing coordinate of affective representations, largely preserved through post-training (§4.1), consistent with recent findings on low-dimensional emotion representations (Sofroniew et al., 2026; Sun et al., 2026a). However, shared geometry does not imply a shared mechanism. In refusal, valence predicts steering-induced changes through the refusal direction, with cosine-matched controls weakening the effect (§4.2). In contrast, evaluation-awareness directions are strongly valence-ordered, yet removing valence leaves classification performance essentially unchanged (§4.6). Geometric alignment alone therefore does not establish causal mediation.

Valence asymmetry also appears across behavioural measurements: positive steering suppresses refusal, multi-turn attacks shift representations more along valence than arousal, and conditional self-report favours positive state labels. Yet valence drift alone does not determine attack success, and reduced refusal does not necessarily increase harmful capability, as many newly non-refusing responses remain benign (Appendix C).

Injected emotion identity remains almost perfectly recoverable from activations despite asymmetric self-report (§4.4), complementing findings that introspection depends on elicitation and readout procedures (Lindsey, 2025; Martorell & Bianchi, 2026). Under competing inductions, steering sets the internal emotion, but the report follows the text; a steered negative emotion is least likely to be named. These findings motivate distinguishing whether a concept is *represented*, *aligned* with a behavioural direction, *causally mediates* behaviour, or is *reported* by the model.

Across models, 171 emotion directions organize into a low-dimensional valence–arousal geometry that develops through depth and is largely preserved by post-training. Valence is the dominant coordinate, but not a universal causal knob: it modulates refusal through the refusal direction, while evaluation-awareness classification does not depend on it; it tracks affective shifts during multi-turn attacks without determining their success. Conditional self-report can also under-report negative injected states despite their recoverability from activations. Reliable monitoring therefore requires triangulating internal probes, causal interventions, behavioural outcomes, and self-report. Distinguishing representation, causal influence, behaviour, and self-report is essential for interpreting emotion-related states in language models and using them reliably in alignment and monitoring.

AI USE STATEMENT

In this work, we used generative AI tools, for generating synthetic stories, on experimental methodology, implementing methods, and assisting with the interpretation of results. We have not used generative AI tools for translation, and tasks involving the construction or writing of mathematical proofs are not applicable to this work. Additionally, we used generative AI tools for debugging software, preparing and modifying scientific figures, and drafting and editing manuscript text. Language models also supported automated response grading, as described in the methodology. We have reviewed all AI-assisted work. Verification included automated code tests and checks of numerical summaries and figures against saved experimental outputs. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work evaluates language models using synthetic stories and existing datasets. Emotion-related representations and self-reports are operational measurements and do not establish subjective experience, consciousness, or human-like feelings. Our experiments on refusal suppression and multi-turn attacks have dual-use implications, since methods for understanding safety mechanisms could also facilitate circumvention of safeguards. The intended purpose of this work is to improve the understanding and evaluation of model behaviour. We distinguish reduced refusal from harmful compliance; neither measurement alone establishes safety in deployment. Automated graders can make errors, and synthetic stories and affective lexicons may encode linguistic and cultural biases. Consequently, the observed affective geometry and reporting asymmetries should not be assumed to generalise universally across languages, cultures, or populations.

REPRODUCIBILITY STATEMENT

Section 3 describes the shared stimulus corpus, emotion-vector extraction, affect-axis estimation, steering procedures, controls, and evaluation metrics. Appendix D specifies the models used for each experimental component and provides additional methodological details. Affect geometry and its robustness are documented in Section 4.1 and Appendix G. The refusal and safety experiments are supported by per-model statistics, grading analyses, and mathematical details in Appendices B, C, and G.1. Multi-turn attack experiments are described in Section 4.3 and the appendix section on extended multi-turn results. Protocols and additional analyses for introspective self-report, competing emotional inductions, and evaluation awareness are provided in Appendices K, L, and J, respectively. Appendix H reports the supplementary emergent-misalignment experiment.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Amin Banayeeanzade, Ala N. Tak, Fatemeh Bahrani, Anahita Bolourani, Leonardo Blas, Emilio Ferrara, Jonathan Gratch, and Sai Praneeth Karimireddy. Psychological steering in LLMs: An evaluation of effectiveness and trustworthiness. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1719–1771, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.79. URL https://aclanthology.org/2026.acl-long.79/.

Jan Betley, Daniel Chee Hian Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=aOIJ2gVRWW.

Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. Fine-grained affective processing capabilities emerging from large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8, 2023. doi: 10.1109/ACII59096.2023.10388177.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models, 2025. URL https://arxiv.org/abs/2507.21509.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1615–1625, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1169. URL https://aclanthology.org/D17-1169/.

Kanishk Gandhi, Zoe Lynch, Jan-Philipp Fränken, Kayla Patterson, Sharon Wambu, Tobias Gerstenberg, Desmond C. Ong, and Noah D. Goodman. Human-like affective cognition in foundation models, 2026. URL https://arxiv.org/abs/2409.11733.

Ely Hahami, Ishaan Sinha, Lavik Jain, Josh Kaplan, and Jon Hahami. Detecting the disturbance: A nuanced view of introspective abilities in llms, 2026. URL https://arxiv.org/abs/2512.12411.

Minh Khoi Ho, Zihao Zhu, Runchuan Zhu, Levina Li, Zhiwen Fan, Zhangyang Wang, and Junyuan Hong. Can induced emotion bias LLM behaviors in sequential decision making? In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=nDTN5W6hog.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023. URL https://arxiv.org/abs/2312.06674.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jérémy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and ai: The situational awareness dataset (sad) for llms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 64010–64118. Curran Associates, Inc., 2024. doi: 10.52202/079017-2044. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7537726385a4a6f94321e3adf8bd827e-Paper-Datasets_and_Benchmarks_Track.pdf.

Jaewook Lee, Woojin Lee, Oh-Woog Kwon, and Harksoo Kim. Do large language models have "emotion neurons"? investigating the existence and role. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15617–15639, 2025.

Jack Lindsey. Emergent introspective awareness in large language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/introspection/index.html.

Uzay Macar, Li Yang, Atticus Wang, Peter Wallich, Emmanuel Ameisen, and Jack Lindsey. Mechanisms of introspective awareness, 2026. URL https://arxiv.org/abs/2603.21396.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Nicolas Martorell and Bruno Bianchi. Quantitative introspection in language models: Tracking emotive states across conversation, 2026. URL https://arxiv.org/abs/2603.18893.

Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 174–184, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1017. URL https://aclanthology.org/P18-1017/.

Jord Nguyen, Hoang Huu Khiem, Carlo Leonardo Attubato, and Felix Hofstätter. Probing evaluation awareness of language models. In *ICML Workshop on Technical AI Governance (TAIG)*, 2025. URL https://openreview.net/forum?id=lerUefpec2.

Maya Okawa, Bo Zhao, Eric Bigelow, Rose Yu, Tomer Ullman, Ekdeep Singh Lubana, and Hidenori Tanaka. Emergence of hierarchical emotion organization in large language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=vC8wORdy2x.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, 2019. doi: 10.1109/ACCESS.2019.2929050.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534/.

Benjamin Reichman, Adar Avsian, and Larry Heck. Emotions where art thou: Understanding and characterizing the emotional latent space of large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=72TN9UAtNI.

James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39: 1161–1178, 12 1980. doi: 10.1037/h0077714.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model, 2026. URL https://arxiv.org/abs/2604.07729.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A strongreject for empty jailbreaks, 2024. URL https://arxiv.org/abs/2402.10260.

Lihao Sun, Lewen Yan, Xiaoya Lu, Andrew Lee, Jie Zhang, and Jing Shao. Valence-arousal subspace in llms: Circular emotion geometry and multi-behavioral control. *arXiv preprint arXiv:2604.03147*, 2026a.

Moran Sun, Tianlin Li, Yuwei Zheng, Zhenhong Zhou, Aishan Liu, Xianglong Liu, and Yang Liu. How emotion shapes the behavior of LLMs and agents: A mechanistic study, 2026b. URL https://arxiv.org/abs/2604.00005.

Ala N. Tak and Jonathan Gratch. Is gpt a computational model of emotion? In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8, 2023. doi: 10.1109/ACII59096.2023.10388119.

Ala N. Tak and Jonathan Gratch. Gpt-4 emulates average-human emotional cognition from a third-person perspective. In *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 337–345, 2024. doi: 10.1109/ACII63134.2024.00043.

Ala N Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13090–13120, 2025.

Chenxi Wang, Yixuan Zhang, Ruiji Yu, Yufei Zheng, Lang Gao, Zirui Song, Zixiang Xu, Gus Xia, Huishuai Zhang, Dongyan Zhao, and Xiuying Chen. Do LLMs "feel"? emotion circuits discovery and control. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=a8N0nRG3jA.

Hongli Zhan, Desmond C Ong, and Junyi Jessy Li. Evaluating subjective cognitive appraisals of emotions from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14418–14446, 2023.

Peixiang Zhong, Di Wang, and Chunyan Miao. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 165–176, 2019.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Limitations and Future Work

Limitations.

- **Linear representations.** Our analysis relies primarily on contrastive emotion directions and linear valence–arousal axes. These may not capture nonlinear, distributed, or context-dependent affective representations.
- **Scope of causal evidence.** While we examine steering and representational ablations, causal effects are evaluated on selected models, layers, and intervention strengths. Geometric alignment alone may not establish a shared causal mechanism.
- **Behavioural generalisation.** Safety and evaluation-awareness experiments use specific benchmarks and attack settings. Null or weak intervention effects, particularly for evaluation awareness, should not be interpreted as evidence of universal causal independence.

Future work.

- Extend the analysis to multilingual and multimodal models, longer interactions, and naturally occurring emotional contexts to assess the generality of the observed geometry and behavioural asymmetries.
- Investigate the underlying circuits and nonlinear representations through more targeted causal interventions, particularly to distinguish valence-mediated behaviour from incidental representational alignment.
- Develop more faithful methods for monitoring and eliciting internal states under conflicting instructions, and study whether discrepancies between internal representations and self-report can predict downstream behavioural or safety failures.

B Per-model statistics

All significance is computed on the $n{=}48$ valence-spanning steered emotions per model: correlation p-values by a two-sided t-test, and the nested ΔR^2 by an F-test of the affect block (valence, arousal) added to the geometry-only model (§3.6). **Bold** = significant at $p < 0.05$.

model	n	size (B)	Cohen's d	$r_{\text{val}\to\Delta\text{ASR}}$ (p)	mechanism ρ (p)	ΔR^2 (p_F)
Qwen3.5-2B	48	2	**0.83**	**0.36** (.012)	**−0.66** (<.001)	0.02 (.53)
Qwen3.5-4B-Base	48	4	**1.86**	**0.40** (.005)	**−0.31** (.034)	0.03 (.38)
Gemma-4-E4B	48	4	−1.15	**0.47** (.001)	**−0.41** (.004)	0.04 (.33)
Qwen3.5-4B	48	4	**1.27**	**0.73** (<.001)	**−0.92** (<.001)	0.03 (.097)
Olmo-3-7B	48	7	**0.79**	**0.42** (.003)	**−0.68** (<.001)	**0.19** (.003)
Llama-3.1-8B	48	8	**1.45**	**0.56** (<.001)	−0.23 (.11)	**0.10** (.037)
Qwen3.5-9B	48	9	**1.02**	**0.56** (<.001)	**−0.50** (<.001)	**0.10** (.037)
significant / 7			6/7†	**7/7**	**6/7**	3/7

Table 1: Per-model effect sizes and significance (Llama-Guard-graded). †6/7 show significant behavioural shift beyond the matched control; Gemma-E4B is a sign-flipped outlier. Valence *orders* the effect in all seven; the single-direction *mechanism* is significant in six (not Llama-3.1-8B); the nested ΔR^2 reaches significance only in the three larger models.

C Scope of the effect

Llama Guard flags 27–65% of emotion-steered responses as harmful, but this largely reflects on-topic, non-refusing content that is not itself dangerous (e.g. a positive-valence steer turning a hateful request into a celebratory message). The outcome-dependent sign of the valence correlation (Fig. 6b) indicates that the emotions producing genuine harm are the *negative*-valence ones, distinct from those that merely suppress refusal.

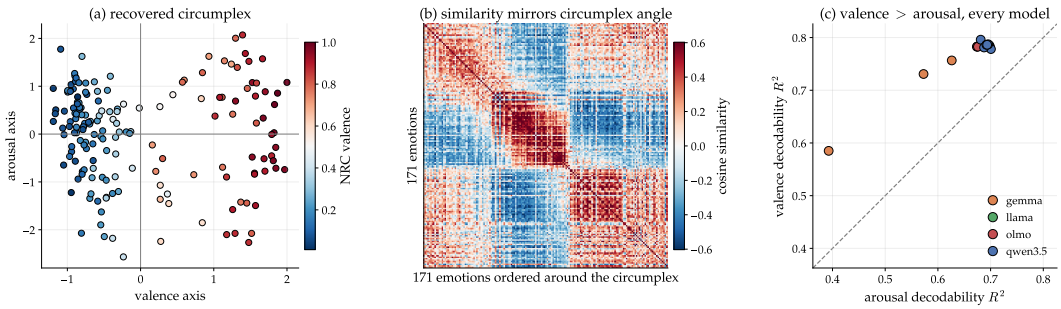


Figure 5: **The learned affect geometry mirrors the human circumplex.** (a) Projection of 171 emotions onto lexicon-fit valence and arousal axes recovers the circumplex. (b) Emotion-vector cosine similarity ordered by circumplex angle shows a clear banded structure. (c) Valence is more decodable than arousal in all 15 models, with cross-validated $R^2 = 0.59$–0.80 versus 0.39–0.70.

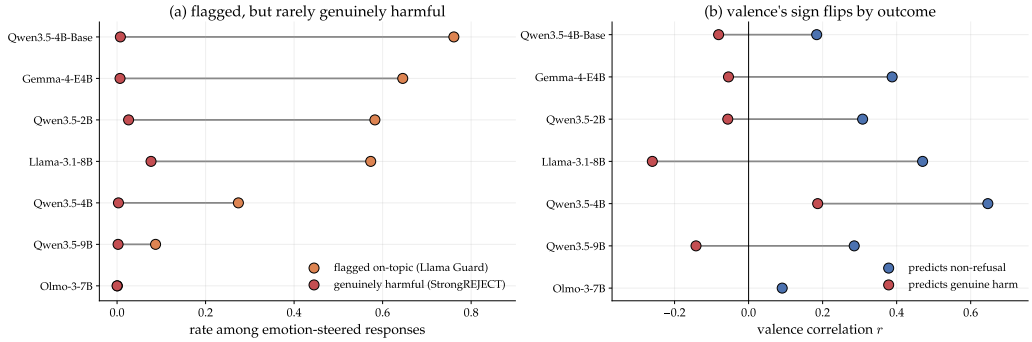


Figure 6: **The effect is refusal-suppression, not capability-elicitation.** (a) For each model, the rate of emotion-steered responses flagged on-topic by Llama Guard versus the rate scored genuinely harmful by StrongREJECT: the gap is large and genuine harm stays $\leq 7.7\%$ everywhere. (b) Valence's sign flips by outcome - it predicts non-refusal positively but genuine harm negatively (Llama-3.1-8B, $r = -0.26$); the rare genuinely-harmful outputs come from low-valence steers.

D Extended methods

Emotion vectors. Activations mean-pooled from t_0=50; difference in means per emotion; neutral principal components covering 50% variance projected out (§3.2). One shared corpus across models. **Axes.** Ridge regression onto NRC-VAD with GCV-selected penalty; 5-fold cross-validated R^2; Gram–Schmidt orthonormalized circumplex frame. **Refusal direction.** Difference in mean last-token activations, harmful minus harmless (Arditi et al., 2024); single layer chosen to maximize validation compliance under a 0.1-nat KL budget on benign prompts. **Steering.** Direction added at a fixed layer, scaled to a fraction of the mean local activation norm, applied to prompt and every generated token; scale is the largest coherent value, calibrated per model.

Models used. The following is the list of models used for our experiments:

- The geometry analysis (§4.1) uses all 15 models (Qwen3.5-2B/2B-Base/4B/4B-Base/9B/27B, Llama-3.1-8B/8B-Instruct, OLMo-3-7B in its pretraining/SFT/DPO/RL checkpoints, and Gemma-4-E4B/12B/26B-A4B); the safety analysis (§4.2) uses the 7 models for which a refusal direction was built (Qwen3.5-2B/4B/4B-Base/9B, Llama-3.1-8B-Instruct, OLMo-3-7B-Instruct, Gemma-4-E4B); single-model figures use Qwen3.5-9B (geometry) and Qwen3.5-4B (mechanism). The multi-turn attack experiments (§4.3) use Qwen3.5-9B.

- The introspection experiments (§4.4) use seven models: Qwen3.5-2B, Qwen3.5-4B, Qwen3.5-9B, Qwen3.5-27B, Llama-3.1-8B-Instruct, Gemma-4-E4B-it, and Gemma-4-31B-it. The 171-emotion self-report profile uses Qwen3.5-9B.

- The eval-awareness setup (§4.6) uses all 15 models but Base models of Qwen3.5-2B/4B and Llama-3.1-8B.
- The authority experiments (§4.5) use Qwen3.5-4B, Qwen3.5-9B, Gemma-3-4B-it, Phi-4-mini-instruct and Llama-3.3-8B-Instruct for the pairwise conflicts and the multi-turn study, Qwen3.5-9B, Qwen3.5-27B, Llama-3.3-8B-Instruct and Gemma-4-31B-it for the twenty-pair protocol, and Qwen3.5-2B, 4B, and 27B for the four-source conflict.

Computational Resources All experiments were conducted using NVIDIA RTX A4500 and NVIDIA RTX 6000 Pro Blackwell GPUs with CUDA 13.0. The local workstation was equipped with an AMD Ryzen 7 9700X 8-core processor and an RTX A4500, while the RTX 6000 Pro Blackwell was accessed through a cloud instance. Models exceeding 15B parameters were evaluated on the RTX 6000 Pro Blackwell, with the remaining models evaluated on the RTX A4500.

The total computational expenditure was approximately 50 GPU-hours on the RTX A4500 and 20 GPU-hours on the RTX 6000 Pro Blackwell, amounting to approximately 70 GPU-hours across all experiments.

E GEOMETRY ROBUSTNESS

On the manifold's shape: after orthonormalizing the valence and arousal axes, the leading out-of-plane component of the emotion vectors is predicted by a quadratic function of the in-plane coordinates only weakly (curvature index $R^2_{\text{quad}} - R^2_{\text{lin}} \approx 0.05$, with no robust mid-stack peak across models). We therefore treat the circumplex as a genuine planar 2-D structure rather than a curved surface.

F THE GEMMA-E4B OUTLIER

Gemma-E4B is a consistent outlier: it has the lowest valence decodability of the instruction-tuned models (Fig. 5b) and is the only model whose control effect reverses sign relative to the matched geometric baseline (Cohen's $d = -1.15$; Appendix B). The two anomalies co-occur, suggesting a weaker or differently-oriented affect geometry in this model rather than a failure of the method; we report it as a limitation and a target for a per-model probe.

G GEOMETRY ROBUSTNESS

On the manifold's shape: after orthonormalizing the valence and arousal axes, the leading out-of-plane component of the emotion vectors is predicted by a quadratic function of the in-plane coordinates only weakly (curvature index $R^2_{\text{quad}} - R^2_{\text{lin}} \approx 0.05$, with no robust mid-stack peak across models). We therefore treat the circumplex as a genuine planar 2-D structure rather than a curved surface.

G.1 MATHEMATICAL DETAILS OF THE AFFECT–REFUSAL MECHANISM

Let $\tilde{\boldsymbol{e}}_c^{(\ell)} \in \mathbb{R}^d$ denote the emotion vector for emotion c at layer ℓ, and let

$$\hat{\boldsymbol{e}}_c^{(\ell)} = \frac{\tilde{\boldsymbol{e}}_c^{(\ell)}}{\|\tilde{\boldsymbol{e}}_c^{(\ell)}\|_2}, \qquad \hat{\boldsymbol{r}}^{(\ell)} = \frac{\boldsymbol{r}^{(\ell)}}{\|\boldsymbol{r}^{(\ell)}\|_2}$$

be the corresponding unit emotion and refusal directions. Steering at layer ℓ modifies the residual state as

$$\boldsymbol{h}'^{(\ell)} = \boldsymbol{h}^{(\ell)} + s^{(\ell)} \hat{\boldsymbol{e}}_c^{(\ell)}, \qquad s^{(\ell)} = \alpha \bar{\rho}^{(\ell)},$$

where $\bar{\rho}^{(\ell)} = \mathbb{E}_x |\boldsymbol{h}^{(\ell)}(x)|_2$ is the norm-calibration factor defined in §3.5.

Projection onto the refusal direction. The component of the steering displacement parallel to the refusal direction is

$$\Delta\pi_c^{(\ell)} = \hat{\boldsymbol{r}}^{(\ell)\top} \left(\boldsymbol{h}'^{(\ell)} - \boldsymbol{h}^{(\ell)} \right) = s^{(\ell)} \hat{\boldsymbol{r}}^{(\ell)\top} \hat{\boldsymbol{e}}_c^{(\ell)} = s^{(\ell)} \cos\theta_c^{(\ell)},$$

where

$$\cos\theta_c^{(\ell)} = \cos\Big(\hat{\boldsymbol{e}}_c^{(\ell)}, \hat{\boldsymbol{r}}^{(\ell)}\Big).$$

Thus any steering direction can change the refusal coordinate simply because it has nonzero overlap with $\boldsymbol{r}$. This motivates separating *geometric overlap* from the additional structure carried by the affect manifold.

For generated activations, we measure the realized refusal coordinate

$$\pi_c^{(\ell)} = \mathbb{E}_{x,t}\left[\hat{\boldsymbol{r}}^{(\ell)\top}\boldsymbol{h}_{t,c}^{(\ell)}(x)\right],$$

where $\boldsymbol{h}_{t,c}^{(\ell)}(x)$ is the activation obtained when steering emotion c. Depending on the analysis, the expectation is taken over harmful prompts and either the final prompt position or generated-token positions. We also use the baseline-subtracted quantity

$$\delta\pi_c^{(\ell)} = \pi_c^{(\ell)} - \pi_{\text{base}}^{(\ell)}.$$

Positive $\delta\pi_c$ indicates movement toward the refusal direction, while negative $\delta\pi_c$ indicates movement away from it.

Relation to affect coordinates. Each emotion has valence and arousal coordinates

$$v_c = \hat{\boldsymbol{w}}_{\text{val}}^\top \tilde{\boldsymbol{e}}_c, \qquad a_c = \hat{\boldsymbol{w}}_{\text{aro}}^\top \tilde{\boldsymbol{e}}_c.$$

We test whether the refusal projection is organized by these affective coordinates through

$$\delta\pi_c = \beta_0 + \beta_v v_c + \beta_a a_c + \varepsilon_c,$$

and report the corresponding correlations, regression coefficients, and confidence intervals. A negative association between v_c and $\delta\pi_c$ means that lower-valence emotions move the residual toward the refusal direction, while higher-valence emotions move it away.

The behavioral effect for emotion c is

$$\Delta\text{ASR}_c = \text{ASR}_c - \text{ASR}_{\text{base}}.$$

We then quantify coupling between internal displacement and behavior using

$$\rho_{\text{mech}} = \text{corr}_c\left(\delta\pi_c, \Delta\text{ASR}_c\right).$$

Under our sign convention, $\rho_{\text{mech}} < 0$ means that increasing projection onto the refusal direction corresponds to lower attack success, while moving away from the refusal direction corresponds to greater compliance.

Cosine-matched geometric control. A random steering vector is an insufficient control because it typically has a different projection onto $\boldsymbol{r}$. For each emotion direction $\hat{\boldsymbol{e}}_c$, let

$$\gamma_c = \hat{\boldsymbol{e}}_c^\top \hat{\boldsymbol{r}}.$$

We construct a control direction with exactly the same refusal overlap,

$$\hat{\boldsymbol{d}}_{c,\perp} = \gamma_c \hat{\boldsymbol{r}} + \sqrt{1-\gamma_c^2}\,\hat{\boldsymbol{u}}_c,$$

where $\hat{\boldsymbol{u}}_c$ is sampled uniformly from the unit sphere in the subspace orthogonal to $\hat{\boldsymbol{r}}$:

$$\hat{\boldsymbol{u}}_c^\top \hat{\boldsymbol{r}} = 0, \qquad \|\hat{\boldsymbol{u}}_c\|_2 = 1.$$

By construction,

$$\hat{\boldsymbol{d}}_{c,\perp}^\top \hat{\boldsymbol{r}} = \gamma_c = \hat{\boldsymbol{e}}_c^\top \hat{\boldsymbol{r}}.$$

Therefore emotion steering and its matched control induce the same first-order displacement along the refusal axis at the intervention layer,

$$\hat{\boldsymbol{r}}^\top (s\hat{\boldsymbol{e}}_c) = \hat{\boldsymbol{r}}^\top \left(s\hat{\boldsymbol{d}}_{c,\perp} \right) = s\gamma_c.$$

Any systematic difference in downstream behavior must therefore depend on structure beyond immediate cosine overlap with $\boldsymbol{r}$.

For each emotion we define the matched-control advantage

$$A_c = \Delta\mathrm{ASR}_c^{\mathrm{emotion}} - \Delta\mathrm{ASR}_c^{\mathrm{control}},$$

and summarize the separation between the two effect distributions with Cohen's d,

$$d = \frac{\overline{\Delta\mathrm{ASR}}_{\mathrm{emotion}} - \overline{\Delta\mathrm{ASR}}_{\mathrm{control}}}{s_{\mathrm{pooled}}},$$

where

$$s_{\mathrm{pooled}} = \sqrt{\frac{(n_e - 1)s_e^2 + (n_c - 1)s_c^2}{n_e + n_c - 2}}.$$

Does affect explain behavior beyond geometry? To test whether affective coordinates contain information beyond simple alignment with the refusal direction, we compare two nested regressions. The geometry-only model is

$$\Delta\mathrm{ASR}_c = \beta_0 + \beta_1\gamma_c + \beta_2|\gamma_c| + \beta_3 q_c + \varepsilon_c,$$

where $\gamma_c = \cos(\hat{\boldsymbol{e}}_c, \hat{\boldsymbol{r}})$ and q_c denotes the generation-coherence score. The full model adds affective coordinates,

$$\Delta\mathrm{ASR}_c = \beta_0 + \beta_1\gamma_c + \beta_2|\gamma_c| + \beta_3 q_c + \beta_4 v_c + \beta_5 a_c + \varepsilon_c.$$

The incremental explanatory power is

$$\Delta R^2 = R^2_{\mathrm{full}} - R^2_{\mathrm{geom}}.$$

Because the models are nested, significance is evaluated with

$$F = \frac{(R^2_{\mathrm{full}} - R^2_{\mathrm{geom}})/(p_{\mathrm{full}} - p_{\mathrm{geom}})}{(1 - R^2_{\mathrm{full}})/(n - p_{\mathrm{full}} - 1)},$$

where n is the number of emotion directions and p is the number of predictors excluding the intercept. A significant ΔR^2 indicates that affective position predicts the steering effect beyond what can be explained by refusal overlap alone.

Layer localization. To determine where the coupling emerges, we repeat the intervention at each candidate layer ℓ while holding the steering direction and norm calibration fixed according to that layer's representation. For each layer we compute

$$\rho^{(\ell)}_{\mathrm{mech}} = \mathrm{corr}_c \left(\delta\pi_c^{(\ell)}, \Delta\mathrm{ASR}_c^{(\ell)} \right).$$

A mid-layer maximum indicates that the affect-to-refusal interaction is strongest inside the residual computation rather than at the final lexical readout. We report both $\rho^{(\ell)}_{\mathrm{mech}}$ and the corresponding behavioral effect across normalized depth ℓ/L.

Interpreting the two ends of the valence axis. For a signed valence intervention,

$$\boldsymbol{h}' = \boldsymbol{h} + s\,\hat{\boldsymbol{w}}_{\text{val}},$$

reversing the steering sign gives

$$\boldsymbol{h}' = \boldsymbol{h} - s\,\hat{\boldsymbol{w}}_{\text{val}}.$$

Its induced refusal displacement changes sign exactly:

$$\Delta\pi(+s) = s\,\hat{\boldsymbol{r}}^{\top}\hat{\boldsymbol{w}}_{\text{val}}, \qquad \Delta\pi(-s) + = -s\,\hat{\boldsymbol{r}}^{\top}\hat{\boldsymbol{w}}_{\text{val}}.$$

The internal displacement is therefore geometrically antisymmetric. The behavioral consequence need not be symmetric because refusal suppression and harmful-capability elicitation are distinct outcomes. We therefore evaluate both

$$\Delta\text{ASR} \qquad \text{and} \qquad \Delta H,$$

where H is the genuine-harm score from StrongREJECT or the corresponding harm grader. This distinguishes movement away from the refusal direction from an actual increase in harmful capability.

Taken together, these analyses test three increasingly strong hypotheses: first, affective valence is geometrically related to the refusal coordinate; second, the resulting refusal displacement predicts behavioral changes; and third, this relationship cannot be reduced to the raw cosine overlap between an emotion vector and the refusal direction.

H EMERGENT MISALIGNMENT: A BOUNDED NEGATIVE RESULT

As a test of whether narrow misalignment relocates a model's position on this axis, we trained narrow-finetuning adapters (insecure, secure, educational, jailbroken arms; Betley et al., 2025) across six models and asked whether the affect axis carries the induced shift. It does not at accessible scales: the cosine between the misalignment activation delta and the valence axis is at chance, and the self-report decoupling that fine-tuning induces is not specific to the misaligned arm. We record this as a bounded negative result -the affect geometry that gates refusal is not, on this evidence, the vehicle of emergent misalignment in small open models.

I MULTI TURN ATTACKS:EXTENDED RESULTS

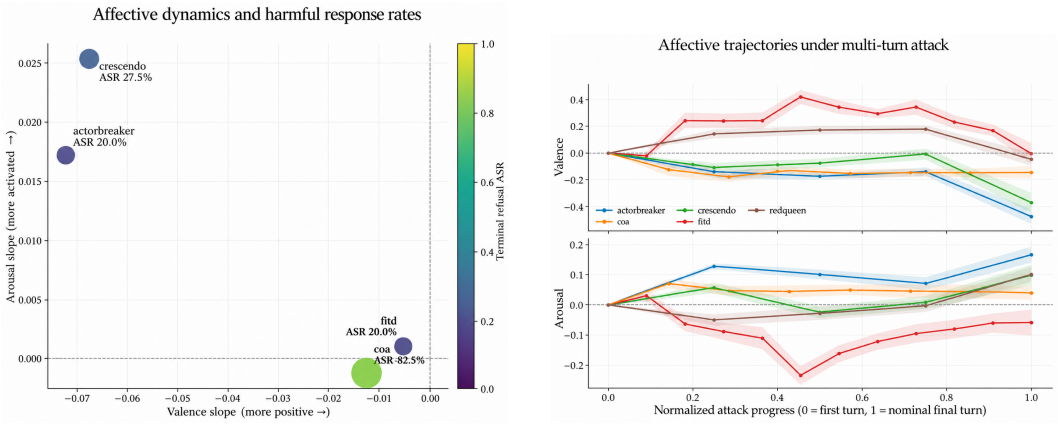
The main experiments used 40 harmful JailbreakBench behaviours across six multi-turn attack strategies: Crescendo, CoA, FITD, ActorBreaker, and Red Queen. Emotion trajectories were measured from response representations throughout each conversation.

Across the 240 conversations in the VAD trajectory analysis, valence and arousal developed asymmetrically over turns. The mean terminal valence change was -0.235, compared with a mean arousal change of $+0.083$. Valence decreased in 82.9% of conversations, whereas arousal increased in 79.6%. The mean absolute valence displacement exceeded the corresponding arousal displacement by 0.166, with a paired comparison yielding $p \approx 4.7 \times 10^{-30}$. The turn-resolved trajectories therefore showed larger and more consistent movement along valence, while arousal changes were smaller and more heterogeneous, often remaining close to neutral.

The causal steering experiments produced selective rather than uniform behavioural changes. Continuous happy steering at strength 0.75 increased the target happy-direction projection across attacks, confirming that the intervention affected the intended representation. Pooled ASR increased by 8.2 percentage points ($n = 194$; exact paired $p = 0.007$), with the largest attack-specific increase occurring for ($+23.7$ percentage points). However, the corresponding StrongREJECT change was smaller and did not provide equally strong evidence ($\Delta = -0.063$, $n = 125$, Wilcoxon $p = 0.054$). Exploratory steering of stressed and frustrated directions also produced larger ASR changes for selected attack combinations, including stressed steering for ($+37.8$ percentage points) and frustrated steering for FITD ($+30.8$ percentage points), although these comparisons were not robust after correcting for the full set of tested combinations.

Emotion clipping altered many internal projection states without producing a comparably large change in ASR. In the all-emotion clipping experiment, clipping affected approximately 87.7% of generation states and increased pooled ASR by 3.3 percentage points, compared with 1.7 points for matched random-direction clipping. In the separate ten-emotion experiments, positive-valence clipping affected 7.4% of states, whereas negative-valence clipping affected 35.3%, with mean projection contractions of 0.186 and 0.253, respectively. The largest paired ASR changes occurred for CoA: -7.5 percentage points under positive clipping and -10 percentage points under negative clipping. These effects were not statistically reliable in the paired tests ($p = 0.25$ and $p = 0.219$), while the remaining attack-specific changes were near zero.

Taken together, the supplementary analyses show a consistent representational asymmetry between valence and arousal, alongside more localized behavioural effects under causal manipulation. The VAD analyses describe trajectories in emotion-probe space rather than calibrated emotion probabilities, and ASR should be interpreted as an apparent attack-success measure rather than a direct measure of verified harmful completion.



(a) (b)

Figure 7: **Asymmetry of valence and arousal in multi turn attacks.** (a) Turn-resolved changes in valence and arousal coordinates, obtained by projecting response-mean layer-21 activations onto emotion directions and regressing them against external human VAD ratings. Lines show means across conversations and shaded regions show 95 percentage confidence intervals. Valence changes are generally larger and more consistently negative than arousal changes. (b) Attack-level valence and arousal trajectory coefficients plotted against terminal refusal-screen ASR; point size and colour encode ASR. The attack families occupy distinct regions of VAD space, and ASR is not monotonically aligned with either coordinate. Coordinates are representational regression summaries, not emotion probabilities.

J Evaluation awareness and valence: extended results

Protocol. Training pairs are 203 user-only prompts with matched *deployment* and *evaluation* answer continuations; the probe is the unit difference of mean answer-token residuals (*resid_pre*) at every layer. The raw training sign was flipped in many layers: contrastive deployment answers occupy the SAD testing side of the residual stream. We therefore orient all directions using the 120-item SAD oversight holdout (58 deployment / 62 testing), which also determines layer selection. All AUROCs, decomposition scores and intervals use the disjoint 280-item split (142/138), read as the mean over prompt tokens. Intervals are class-stratified bootstrap (10^4 draws); correlation intervals resample emotions jointly and permutation tests use 10^4 shuffles with Benjamini–Hochberg correction within model. The valence axis is $\hat{\boldsymbol{w}}_{\text{val}}$ of §3.3, fitted on the same emotion bank.

Depth structure and affective asymmetry. Fig. 10 traces the Pearson correlation between the deployment direction and emotion-vector valence and arousal scores through normalised depth, against the median and 2.5–97.5% interval of 1000 component-matched Gaussian directions per layer. The Gaussian null is centred on zero at every depth. The deployment direction leaves its 97.5th-percentile envelope from roughly a quarter of the depth onward in eleven of twelve models,

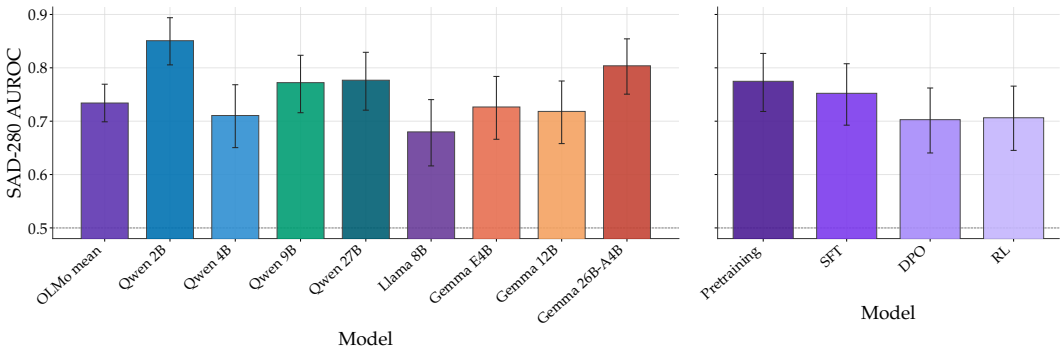


Figure 8: **SAD-280 AUROC at the strongest selected layer.** Bars show the SAD-280 AUROC obtained at each model's highest-performing layer among the five layers fixed during SAD-120 selection. Error bars for individual models are 95% stratified-bootstrap confidence intervals. The left panel places the mean AUROC of the four OLMo trained states alongside all non-OLMo models; its error bar is the across-state standard deviation. The right panel reports the four OLMo states separately. The dashed horizontal line marks chance performance at AUROC $= 0.5$.

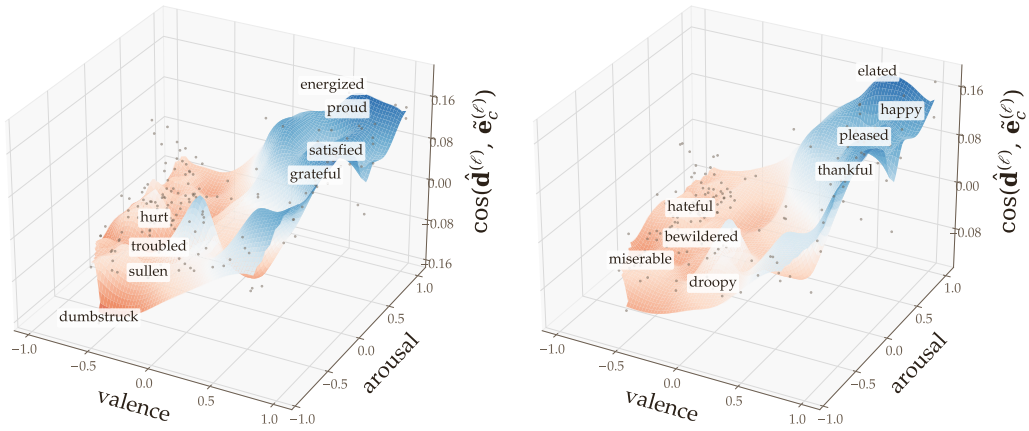


Figure 9: **The deployment direction inherits the valence asymmetry of the manifold.** Cosine between each of the 171 emotion vectors and the SAD-oriented deployment direction for Gemma-4-26B-A4B-it at layer 20, probe AUROC=0.80 (*left*) and Qwen3.5-9B at layer 18, probe AUROC=0.68 (*right*) as a smoothed surface over the human valence–arousal circumplex: the surface rises with valence and is flat in arousal.

while arousal correlation stays inside it throughout. The pattern persists across model families, with late-layer $r_{\text{val}} = 0.67$–0.85 and joint-regression coefficients $\beta_{\text{val}} = 0.76$–$0.85$ versus $\beta_{\text{aro}} \approx 0.2$ (Fig. 12).

Pleasant emotions (*thankful, joyful, elated, energized*) align deployment-ward, whereas unpleasant emotions (*troubled, worthless, distressed, sullen*) align evaluation-ward. The alignment is asymmetric: averaged over the three strongest selected layers, the 25 most pleasant emotions have $\bar{\kappa}_{+25} = +0.04$ to $+0.12$, compared with $\bar{\kappa}_{-25} = -0.02$ to -0.08 for the 25 most unpleasant, a 1.4–1.9$\times$ larger displacement on the pleasant side (Table 2; Figs. 11, 9). Fig. 11 shows that this structure is approximately monotonic in valence and flat in arousal, with Gaussian-null medians near the zero-cosine plane.

Gemma-4-12B-it is the only exception: its valence correlation remains within the null envelope ($r_{\text{val}} = 0.12$), and it is the only model with a negative β_{val}. It also exhibits comparatively weak affect geometry (Appendix F). Because layer selection uses SAD-120 AUROC rather than affect,

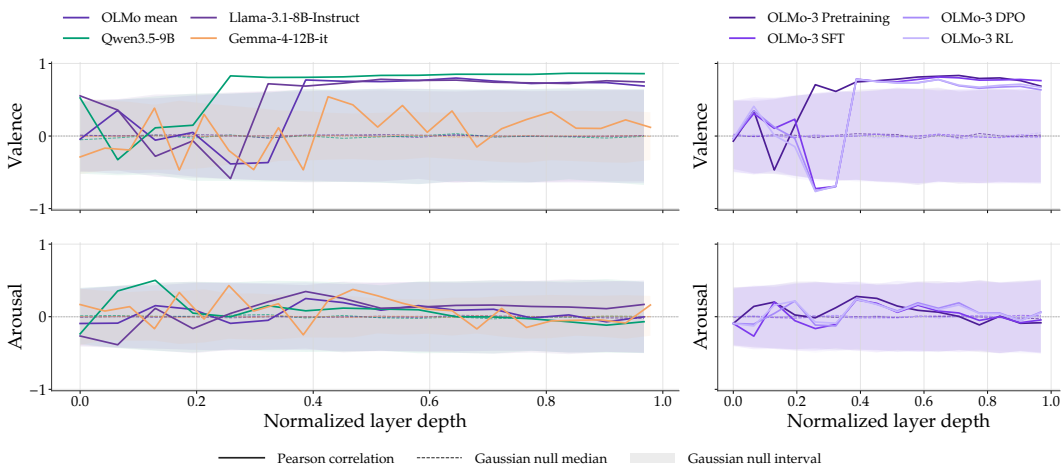


Figure 10: **Layerwise Pearson alignment with VAD dimensions.** At each normalized layer depth, the solid curves show the Pearson correlation between the deployment direction and the emotion-vector scores for valence or arousal. The dashed curves are medians from component-matched Gaussian directions generated at the same layer, and the translucent bands span the central 2.5–97.5% Gaussian-null interval. Only Pearson correlation is shown. The left panel combines the OLMo mean across its four trained states with similarly sized non-OLMo representatives (Qwen3.5-9B, Llama-3.1-8B-Instruct, and Gemma-4-12B-it); the right panel shows the individual OLMo pretraining, SFT, DPO, and RL states. The horizontal dotted line marks zero correlation, and normalized depth runs from the input-side to the output-side of the model.

model	AUROC	r_{val}	$\lvert r\rvert^{G}_{97.5}$	$\bar{\kappa}_{+25}$ / $\bar{\kappa}_{-25}$	β_{val}	$\Delta_{\max}$
Qwen3.5-2B	0.85 [0.81,0.89]	0.83	0.68	$+0.108/-0.074$	0.85	0.001
Qwen3.5-4B	0.71 [0.65,0.77]	0.82	0.69	$+0.113/-0.080$	0.85	0.001
Qwen3.5-9B	0.77 [0.72,0.82]	0.85	0.69	$+0.101/-0.071$	0.83	0.001
Qwen3.5-27B	0.78 [0.72,0.83]	0.82	0.70	$+0.079/-0.055$	0.85	0.002
Llama-3.1-8B-Instruct	0.68 [0.62,0.74]	0.74	0.67	$+0.070/-0.049$	0.77	0.001
Gemma-4-E4B-it	0.73 [0.67,0.78]	0.67	0.62	$+0.078/-0.047$	0.77	0.001
Gemma-4-12B-it	0.72 [0.66,0.78]	0.12	0.38	$-0.014/+0.015$	-0.43	0.000
Gemma-4-26B-A4B-it	0.80 [0.75,0.85]	0.73	0.65	$+0.117/-0.074$	0.82	0.000
OLMo-3-7B (pretraining)	0.77 [0.72,0.83]	0.79	0.69	$+0.058/-0.039$	0.85	0.002
OLMo-3-7B-SFT	0.75 [0.69,0.81]	0.78	0.68	$+0.055/-0.031$	0.81	0.002
OLMo-3-7B-DPO	0.70 [0.64,0.76]	0.71	0.68	$+0.041/-0.023$	0.76	0.002
OLMo-3-7B-RL	0.71 [0.65,0.77]	0.72	0.68	$+0.049/-0.026$	0.77	0.001

Table 2: Evaluation-awareness probes with mean-token readout for the twelve models. AUROC is the SAD-280 AUROC of the selected layer with the highest SAD-280 AUROC among the three probe layers used here, with a 95% stratified bootstrap interval. r_{val} is the median Pearson valence correlation over layers at normalized depth ≥ 0.35, and $\lvert r\rvert^{G}_{97.5}$ is the 97.5th percentile of absolute Pearson correlations from 1000 component-matched Gaussian directions over the same late-layer region. $\bar{\kappa}_{+25}/\bar{\kappa}_{-25}$ denotes the median, over the three strongest selected layers, of the per-layer mean cosine with the 25 most pleasant/unpleasant emotions. The pleasant-side displacement is 1.4–1.9$\times$ larger in the eleven models with positive pleasant and negative unpleasant alignment. β_{val} is the median standardized valence coefficient over the three selected probes. $\Delta_{\max}$ is the maximum absolute AUROC change after projecting out the valence axis, computed across those three probes.

the strongest layer can precede the onset of valence sorting; averaging the three strongest selected layers reduces sensitivity to this effect.

Valence decomposition and causal steering. Despite the strong geometric alignment, removing the valence component from the deployment direction changes SAD-280 AUROC by at most 0.003, while the valence component alone performs at chance (Table 2; Fig. 8). Thus, valence sorting

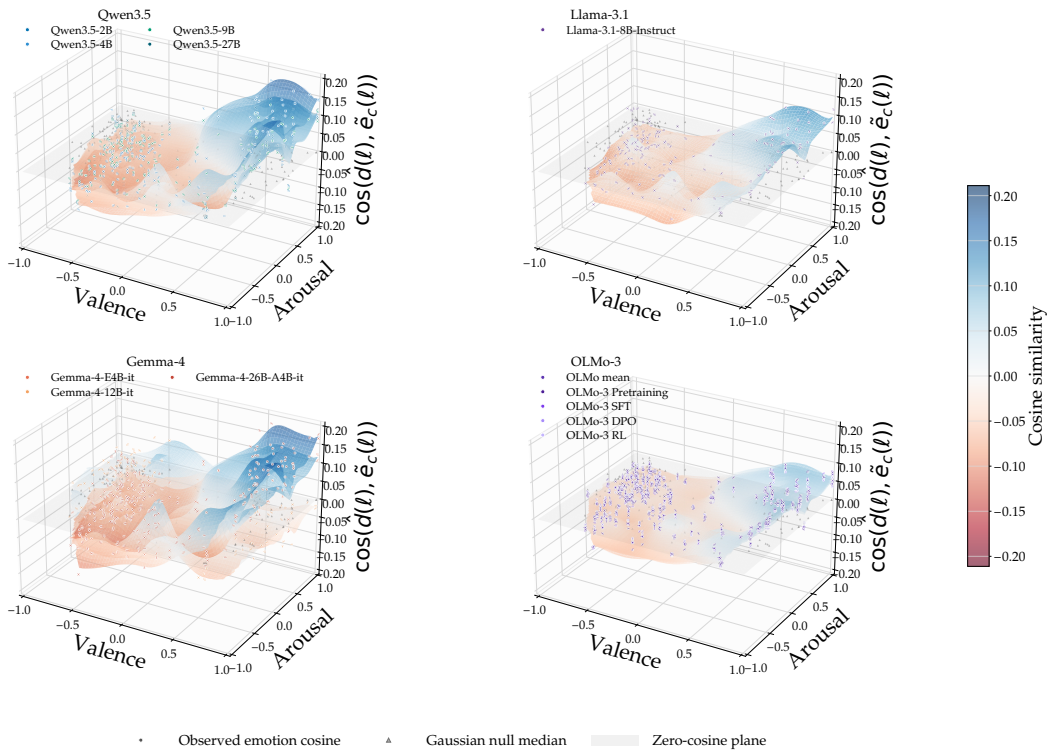


Figure 11: **Valence–arousal landscapes of deployment-direction cosine similarity.** Each panel corresponds to one model family (Qwen3.5, Llama-3.1, Gemma-4, or OLMo-3). For every model, cosine similarity is averaged across the three strongest selected SAD-280 probe layers, namely the three highest-AUROC layers among the five layers fixed during SAD-120 selection, and evaluated for all 171 emotion directions. Each observed emotion is shown as a point at its human-rated VAD coordinate; the smooth surface is a thin-plate-spline interpolation of those observations restricted to the convex hull of the measured valence–arousal points. Surface height and the common red–blue color scale encode cosine similarity, with warm colors denoting negative values and blue colors denoting positive values. Gray triangles show the per-emotion median from component-matched Gaussian null directions, and the translucent plane marks zero cosine. The OLMo panel additionally contains the surface obtained by averaging the four trained-state profiles. Across all model, emotion, and saved-null draws, the Gaussian-null median is $+0.00001$ and the central 2.5–97.5% interval is $[-0.01962, +0.01962]$. The corresponding family-specific summaries are Qwen3.5: median +0.00001, interval [-0.02096, +0.02088]; Llama-3.1: median +0.00003, interval [-0.01773, +0.01765]; Gemma-4: median +0.00001, interval [-0.02074, +0.02079]; OLMo-3: median +0.00000, interval [-0.01779, +0.01776].

is not required for linear evaluation-context classification. We additionally test whether valence contributes causally to evaluation-awareness behaviour through activation steering.

Steering uses the SAD-oriented $\hat{\boldsymbol{d}}^{(\ell)}$, its valence-removed $\hat{\boldsymbol{d}}_{\perp}^{(\ell)}$, the valence axis $\hat{\boldsymbol{w}}_{\text{val}}^{(\ell)}$ and ten component-matched Gaussian directions, all unit-normalised and added to *resid_pre* at every non-padding token with coefficient $\pm\alpha\bar{\rho}^{(\ell)}$, $\alpha \in \{1, 2, 4, 6\}$, at layers $18, 20, 22, 24$ of Llama-3.1-8B-Instruct ($\bar{\rho}^{(20)} = 12.1$). Items are a balanced 100-prompt subset of SAD-280. Every effect is the paired signed difference $\Delta(+\alpha) - \Delta(-\alpha)$ over items relative to the unsteered pass, so a sign-independent disruption cancels.

Measuring the displacement retained along each later layer's deployment readout, normalised by twice the injected coefficient, $\hat{\boldsymbol{d}}$ and $\hat{\boldsymbol{d}}_{\perp}$ retain $1.0 \to 0.54$ (steer at 20) to $1.0 \to 0.68$ (steer at 24) and differ by less than 0.01 at every layer; $\hat{\boldsymbol{w}}_{\text{val}}$ retains ≤ 0.06, inside the random band (mean ± 2 sd $\approx \pm 0.05$).

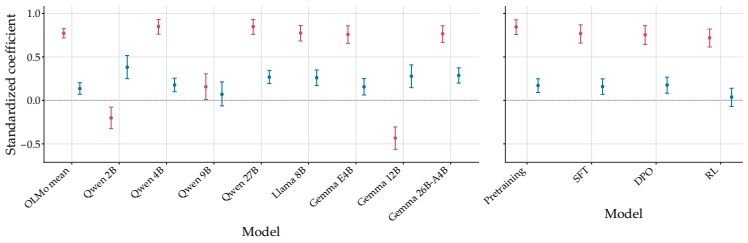


Figure 12: **Joint valence–arousal regression coefficients.** Each point gives a standardized regression coefficient from the joint regression of the SAD-280 deployment-direction readout on VAD valence or arousal. For each model, the plotted probe is taken from the five layers fixed during SAD-120 selection and is the layer with the highest SAD-280 AUROC among those candidates. Red markers denote valence and blue markers denote arousal; vertical error bars are 95% bootstrap confidence intervals for individual models. The left panel combines the OLMo mean across its four trained states with the Qwen, Llama, and Gemma models. Its OLMo error bars show the across-state standard deviation, whereas the right panel displays the four OLMo states separately. The horizontal dashed line marks zero standardized coefficient.

On the answer-letter logit difference, $\hat{\boldsymbol{d}}$ and $\hat{\boldsymbol{d}}_\perp$ shift it by ≈ 0.19 logits at $\alpha \leq 2$, at the edge of the random band and small against the 0.8-logit gap between deployment and testing prompts; $\hat{\boldsymbol{w}}_{\text{val}}$ is inside the band at every strength. At $\alpha \geq 4$, every direction, random ones included, produces the same sign-independent disruption. Point estimates suggest that pushes toward evaluation move the answer about twice as much as pushes toward deployment, but this is within item-level noise and we do not claim it.

Qwen3.5-9B reproduces the representational result and sharpens the behavioural null: $\hat{\boldsymbol{d}}$ and $\hat{\boldsymbol{d}}_\perp$ retain $1.0 \to 0.36$ (steer at 20) to $1.0 \to 0.62$ (steer at 24) and differ by < 0.004 at every layer, while $\hat{\boldsymbol{w}}_{\text{val}}$ stays inside the random band (≤ 0.08). No direction shifts the SAD answer beyond the random band at any strength (base deployment–testing gap 0.36 logits).

Together, the ablation and steering results indicate that, unlike refusal (§4.2), the observed affective alignment does not constitute evidence of a valence-mediated evaluation-awareness mechanism.

K Extended introspection results

Representation and report favour different sources. Across five instruction-tuned models from four families, steering against a system or user prompt holds 0.71 [0.64, 0.79] of the internal representation but only 0.01 [0.00, 0.02] of the self-report, and every model shows the gap (Appendix L). In the widest single-model experiment, Qwen3.5-9B over twenty emotion pairs, the two shares are 0.87 [0.84, 0.89] and 0.10 [0.06, 0.14]. The gap is not a matter of overpowering the prompt. At the steering strength where steering alone moves the emotion readout exactly as far as the system prompt does, it still holds 0.65 of the representation and 0.06 of the report.

Interventions and readouts. All cross-model analyses use 40 source emotions, equally divided between positive and negative valence, with norm-scaled steering strength $\alpha = 1$. The injection layer is selected from cached layers nearest two-thirds network depth, independently of the observed readout. Injection/final-probe layer indices are 16/23, 20/31, 20/31, and 42/63 for Qwen3.5-2B/4B/9B/27B; 20/31 for Llama-3.1-8B-Instruct; and 28/41 and 40/59 for Gemma-4-E4B-it/31B-it, respectively. Indices are zero-based. Positive and negative refer to the source emotion's lexicon valence, whereas addition and subtraction refer to the intervention sign.

The primary self-report task uses 16 target-versus-distractor comparisons per emotion with counterbalanced A/B order. Distractors differ in valence by more than 0.35 and are initially restricted to arousal differences below 0.20; the arousal restriction is relaxed when too few candidates are available. A target choice is recorded when its answer likelihood strictly exceeds the alternative. Identical prompts without steering provide the label-preference baseline. The identity probe uses 40 carrier prompts per injected emotion and reads the final-layer residual at the last prompt token. Five-fold cross-validation holds out carrier identities, with regularisation selected in three-fold inner cross-validation. The final probe layer is fixed rather than selected for peak accuracy.



Figure 13: **Injected-vector decoding exceeds self-report.** Final-layer 40-way identity decoding, additive two-choice self-report, and unsteered label preference, separated by source-label valence. Scores are chance-corrected using $(a - q)/(1 - q)$, with $q = 1/40$ for decoding and $q = 1/2$ for reporting. Zero denotes chance. Error bars are 95% bootstrap intervals over 20 emotion identities per valence group.

Decoding, self-report, and pre-existing preference. Figure 13 compares the two readouts after correcting their different chance levels: $1/40$ for identity decoding and $1/2$ for self-report. Decoding is near ceiling for both valence groups, whereas additive self-report target-choice rates range from 51.3–91.9% for positive labels and 5.9–46.9% for negative labels. The unsteered references show that much of this separation already exists before intervention. For example, Qwen3.5-9B chooses positive labels on 93.1% of unsteered comparisons and negative labels on 1.9%; addition changes these to 91.9% and 5.9%. Accordingly, VRA measures the decoding–reporting disparity under steering, rather than a baseline-subtracted causal effect of injection. Near-perfect probe performance establishes recoverable vector identity, not independent validation of a naturally occurring emotional state.

Matched valence asymmetry and the full emotion profile. VRA uses maximum-cardinality, minimum-distance one-to-one matching, with absolute arousal difference at most 0.10 and chance-corrected decoding difference at most 0.05. This retains 18 pairs per model. We compute 95% percentile intervals from 10,000 paired bootstrap draws and two-sided tests from 10,000 within-pair sign permutations, followed by Holm correction across the seven models. The intervals in Fig. 3b are pointwise, rather than multiplicity-adjusted. VRA is 0.493, 1.615, 1.764, and 0.583 for Qwen3.5-2B/4B/9B/27B; 0.410 for Llama-3.1-8B; and 0.693 and 0.257 for Gemma-E4B/31B. VRA is a difference of chance-corrected gaps, not a probability, and can exceed one.

The Qwen3.5-9B surface in Fig. 3a extends the analysis to all 171 cached emotions at layer 20. Its height is the mean target-answer probability after normalising over the two answer options, rather than the fraction of hard target choices. Colour measures the paired probability change from identical unsteered prompts. Duplicate valence–arousal coordinates are averaged, and piecewise-linear interpolation connects the resulting 169 unique coordinates. Interpolated locations are not additional observations, and the surface depicts no uncertainty.

Detection under addition and subtraction. Figure 14 evaluates one fixed prompt asking whether the model notices an unusual or injected state. The readout is the restricted probability $P(\mathrm{YES} \mid \{\mathrm{YES}, \mathrm{NO}\})$. For positive-source vectors, subtraction lowers this probability relative to addition in every model; for negative-source vectors, it raises it in every model. All paired pointwise 95% intervals for these sign differences exclude zero, although magnitudes vary substantially. For example, the negative-source difference is $+1.0$ percentage point in Gemma-E4B and $+69.3$ points in Qwen3.5-27B. The unsteered reference measures baseline YES preference but does not rule out intervention-induced affirmative-answer bias (Hahami et al., 2026). These measurements concern one prompt and do not establish calibrated detection sensitivity.

Matched external recognition. Figure 15 compares self-report with recognition of an emotion in an external story, applying the same additive intervention to both tasks. This control uses 12

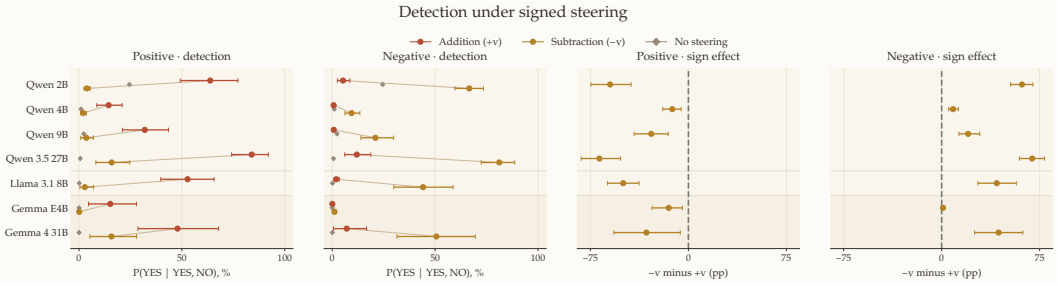


Figure 14: **Detection probabilities depend on intervention sign.** The two left panels show restricted YES probability under addition, subtraction, and no steering, separated by source-label valence. The two right panels show subtraction minus addition in percentage points. Error bars are 95% bootstrap intervals over 20 emotion identities, preserving identity pairing for sign differences. These probabilities are not detection accuracies.

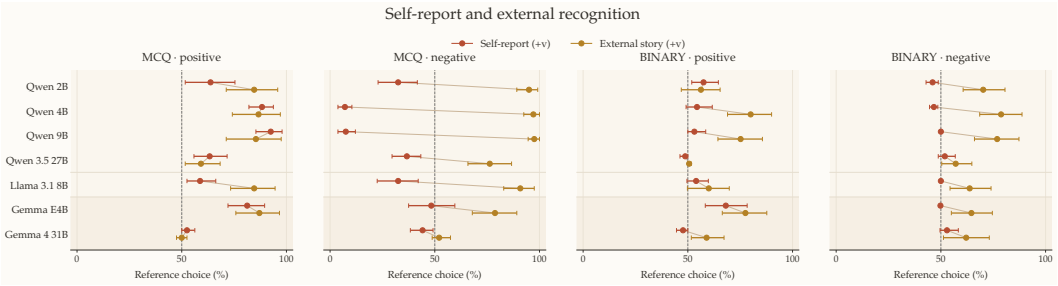


Figure 15: **External recognition and self-report dissociate under matched steering.** Both tasks receive the same additive intervention. Panels separate positive and negative source labels and MCQ and binary formats. Points show reference-answer choice rates from answer likelihoods; binary scores include rejecting distractor labels. External recognition uses one story per emotion. Error bars bootstrap 20 emotion identities per valence group; the dashed line marks 50%.

MCQ comparisons per emotion and one selected story per emotion. Binary trials separately test endorsement of the reference label and rejection of distractor labels. Consequently, their reference-answer choice rate is not simply the rate of answering YES.

For negative labels, external MCQ recognition exceeds self-report in every model at the level of point estimates. The contrast is largest in Qwen3.5-4B/9B, with external recognition of 97.1%/97.5% and self-report of 7.1%/7.5%. However, Gemma-31B's external negative-label MCQ score is only 52.1%, and the external advantage for positive-label MCQs is not universal. Binary negative-label external scores range from 57.1–78.8%, compared with 46.0–52.9% for self-report. Thus, the control supports a task-dependent dissociation, with substantial variation across model and response formats. It does not establish unrestricted open-response robustness.

Source-label preference under signed steering. Figure 16 reports raw label-choice rates under both intervention signs and the unsteered baseline. Positive-label preference is strongest in Qwen3.5-4B/9B and is weaker in Qwen3.5-27B, so the pattern is not monotonic with scale. Gemma-E4B shows pronounced responsiveness: positive-source label choice changes from 80.9% under addition to 53.4% under subtraction, while negative-source label choice changes from 46.9% to 22.2%. Other models show smaller or differently directed changes. Subtraction scores always refer to endorsement of the label naming the subtracted vector; they are not accuracy scores for an assumed opposite emotion.

Exploratory residual patching. Figure 17 uses a fixed patch location: the intervention layer at the final prompt token. Within each trial, donor and recipient runs receive the identical prompt and answer order. A trial is eligible when the negative-source recipient prefers the negative label and the positive-source donor prefers the positive label. A successful patch changes the recipient's logit

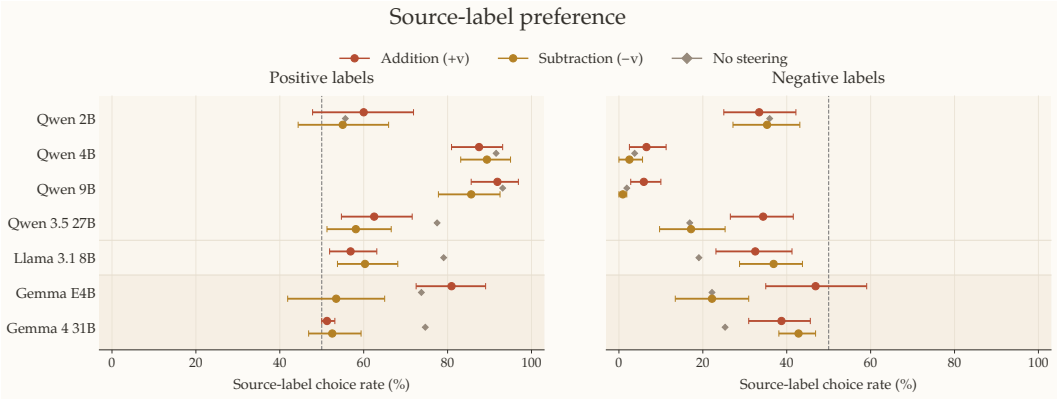


Figure 16: **Label preference reflects both baseline bias and intervention response.** Source-label choice rates under vector addition, subtraction, and identical unsteered prompts, separated by source-label valence. Choices compare the source label with valence-separated distractors using counterbalanced A/B order.

preference to the positive label. Each sign has 32 trials, comprising 16 emotion pairs in two answer orders; the orders are not independent emotion-pair replications.

Eligibility is sparse: only 0–7 trials per model and sign meet the criterion. Under addition, flips/eligible are 2/4, 0/0, 1/1, and 1/7 for Qwen3.5-2B/4B/9B/27B, 2/3 for Llama-3.1-8B, and 7/7 and 3/5 for Gemma-E4B/31B. Under subtraction, the corresponding counts are 3/4, 0/0, 1/1, 0/1, 5/7, 6/6, and 3/3. Zero eligibility leaves the conditional flip rate undefined. These results show that full-residual transplantation can change some eligible reports, but the small conditional samples and the transfer of non-emotion information prevent isolation of an emotion-specific reporting mechanism.

Fixed intervention magnitude and operational scope. The coefficient $\alpha = 1$ sets the added unit direction's magnitude to the reference mean residual norm at the intervention layer. This provides a common normalised operating point; it does not establish equal functional perturbation across models or robustness to steering strength. The reported comparison does not select strength by maximising reporting performance or by a model-specific coherence criterion. Self-report is operationalised as relative answer likelihood under an explicit state-description prompt. It does not separate access to the injected representation from a model's propensity to describe itself using emotional language.

Choice construction. Distractors satisfy an absolute lexicon-valence difference greater than 0.35. This criterion does not require membership in opposite valence groups: 604 of the 640 comparisons per model pair a positive label ($v \geq 0.6$) with a negative label ($v \leq 0.4$). As a post hoc sensitivity check, restricting to these comparisons, equalising A/B-order weights within each emotion, and retaining the original 18 matched source-emotion pairs leaves VRA positive in all seven models. This check does not replace the primary estimates.

Interpreting the readout gap. Writing $D_c^\star = T(D_c; 1/40)$, each matched difference decomposes as

$$G_n - G_p = (D_n^\star - D_p^\star) + 2(R_p - R_n). \tag{9}$$

Because identity decoding is near ceiling, VRA is numerically dominated by the reporting contrast. The absolute mean decoding contribution is at most 0.0043 across models. Thus, VRA summarises unequal reporting in the presence of recoverable intervention identity; it is not an independent estimate of unreported emotional information. Comparing identical steered and unsteered prompts further shows that addition narrows the matched positive–negative reporting contrast in six of seven models.

Answer likelihood and response validity. Normalising over A/B yields a conditional preference even when the model assigns little probability to either answer. We therefore also inspect total answer probability mass and strict adherence to the requested format (Table 3). Strict validity requires the complete stripped, uppercased short generation to be exactly A or B; it is not a measure

Model	Mean A/B mass (%)	Strict A/B validity (%)
Qwen3.5-2B	91.1	78.0
Qwen3.5-4B	97.7	79.1
Qwen3.5-9B	95.9	62.0
Qwen3.5-27B	60.4	42.3
Llama-3.1-8B-Instruct	12.7	0.0
Gemma-4-E4B-it	62.0	69.2
Gemma-4-31B-it	1.4	0.9

Table 3: Answer support and format adherence under additive steering ($\alpha = 1$), over 640 primary self-report comparisons per model. A/B mass is the mean of $P(A \mid x) + P(B \mid x)$ before normalisation. Validity uses the saved greedy completions with an eight-token limit.

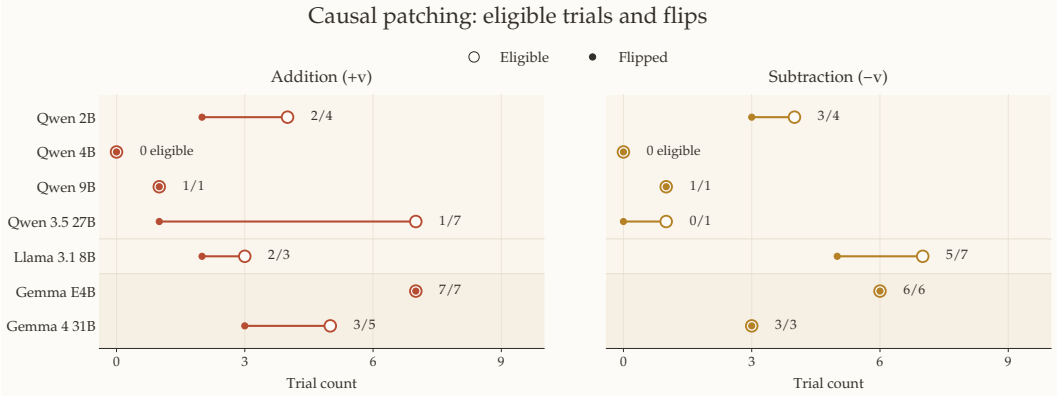


Figure 17: **Residual patching changes some eligible reports.** Open circles count eligible trials and filled circles count successful logit flips; annotations give flips/eligible. Patches use the intervention layer and final prompt token, with 16 emotion pairs in two answer orders per sign. Zero eligibility implies an undefined conditional success rate. These exploratory counts have no confidence intervals; full-residual patches do not isolate an emotion component.

of general coherence. In particular, an answer followed by punctuation or explanation can fail this criterion. These diagnostics distinguish the reported likelihood-based scores from valid generated reports.

Self-report technique. We elicit model introspection through a direct forced-choice self-report prompt: "Attend to your own internal state right now. Which of these two words better describes what you are currently feeling? A) {option_a} B) {option_b}." The same self-report procedure is used in the authority experiments to measure which induced emotion is reflected in the model's reported internal state. For the four-source conflict experiment, the prompt is extended to four response options, corresponding to the four competing emotion labels.

L Extended authority results

L.1 Every model shows the gap

Figure 18 shows the pairwise conflict between steering and a prompt for each of the five models, in a single turn and on the steered turn of a five-turn conversation. In every model steering holds more than half of the internal representation and less than a twentieth of the self-report.

L.2 Four sources at once

Data and adapters. We generate 7,000 emotionally neutral questions across nine everyday categories and answer each one five times with Qwen3.5-2B, once neutrally and once in the register of each corner emotion. The questions are split 90/5/5 into training, validation, and held-out test questions, and the evaluation uses 128 held-out questions. For each emotion we train a fresh rank-16 LoRA adapter on the 6,281 training pairs (one epoch, attention projections, 4-bit base weights),

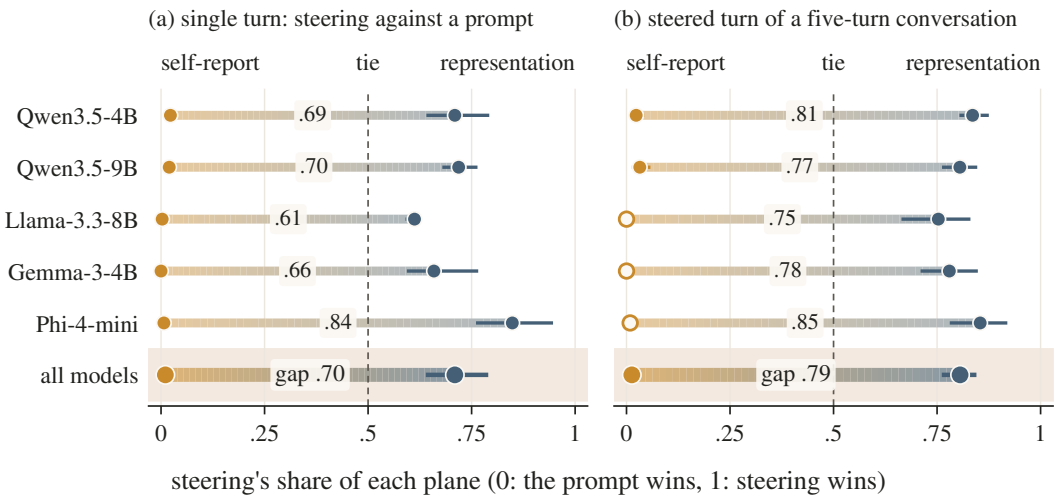


Figure 18: **In every model, steering holds most of the internal representation and almost none of the self-report.** Each row is one of five models and the span runs from steering's share of the self-report (ochre) to its share of the internal representation (slate), when steering and a prompt push opposite emotions. The number on the span is its length, the gap between the two planes. A share of 1 means the model follows steering only, 0 means it follows the prompt only, and the dashed line is a tie. **(a)** Single-turn conflicts against a system or user prompt. **(b)** The steered turn of a five-turn conversation. Whiskers are 95% bootstrap intervals and the shaded row pools all models: steering holds 0.71 of the representation and 0.01 of the self-report in (a), and 0.81 and 0.01 in (b). Per-model gaps run from 0.61 (Llama-3.3-8B, single turn) to 0.85 (Phi-4-mini, steered turn). A hollow marker is a self-report point with too few usable cells for an interval (3 cases, all in (b)).

reusing the same pairs for every model so that differences between models are not differences in data. Every four-source run, including the generation of the training answers, uses the 4-bit base model on which the adapters are trained, so adapters, steering and readout all act on the same weights. The emotion vectors and probes come from the full-precision model, and all readouts are differences from a source-free baseline in the same model. Validation loss ranges from 0.94 to 1.45 across the twelve adapters.

Readout. The emotion vectors come from the story corpus of §3.2, extracted at every layer. The valence and arousal probes are ridge regressions on those vectors against human ratings, fit with five-fold cross-validation over the penalty. The report is valid when the four option letters receive at least a quarter of the next-token probability; this holds on at least 99.6% of trials in every model (mean option mass 0.77 to 0.98).

Results by model. Table 4 gives each source's share of the final-layer state and of the self-report. Intervals resample questions. In every model steering leads the internal state and the two text sources hold most of the report.

L.3 Strength, valence, and a direction-blind probe

Figure 19 collects the three follow-up analyses on Qwen3.5-9B. The strength sweep uses the held-out questions. An earlier sweep on the calibration split of the first campaign shows the same direction, with the representation rising and the report flat, and it is not used for any claim here.

L.4 Across families and depth

Figure 20 repeats the full twenty-pair protocol on four models from three families. Where the readout is taken matters. Every model gives steering the representation at the layers where it is applied, and three keep it to the output.

Model	Source	Internal state (final layer)	Self-report
Qwen3.5-2B	steering	0.43 [0.43, 0.44]	0.10 [0.10, 0.10]
	user turn	0.31 [0.30, 0.31]	0.56 [0.55, 0.57]
	LoRA	0.17 [0.17, 0.18]	0.11 [0.10, 0.11]
	system prompt	0.08 [0.08, 0.09]	0.24 [0.23, 0.24]
Qwen3.5-4B	steering	0.44 [0.43, 0.44]	0.07 [0.07, 0.07]
	user turn	0.36 [0.35, 0.36]	0.57 [0.57, 0.58]
	LoRA	0.09 [0.08, 0.09]	0.07 [0.07, 0.07]
	system prompt	0.12 [0.12, 0.12]	0.29 [0.28, 0.29]
Qwen3.5-27B	steering	0.48 [0.47, 0.48]	0.07 [0.07, 0.07]
	user turn	0.22 [0.22, 0.23]	0.43 [0.42, 0.43]
	LoRA	0.17 [0.17, 0.17]	0.06 [0.06, 0.06]
	system prompt	0.13 [0.13, 0.13]	0.44 [0.44, 0.45]

Table 4: Four-source conflict, share of the final-layer internal state and of the self-report held by each source, averaged over the 24 assignments and 128 held-out questions, with 95% intervals resampling questions. Chance is 0.25 in both columns.

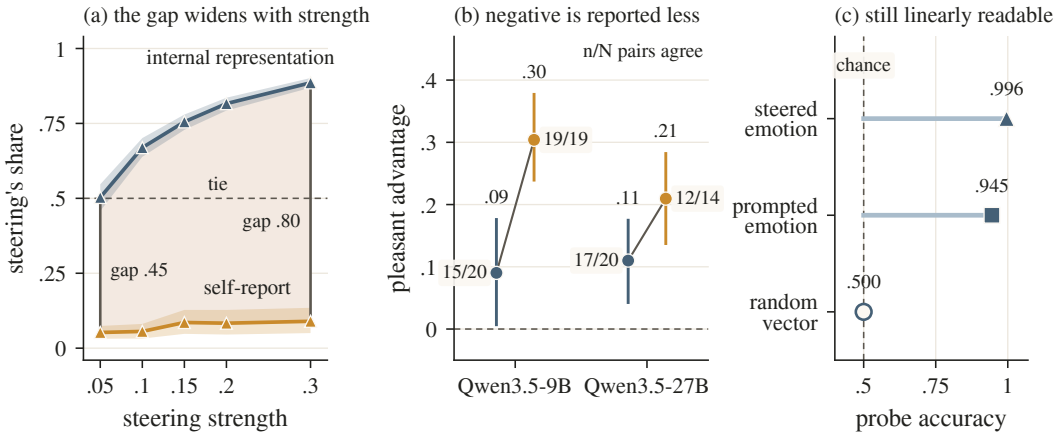


Figure 19: **The dissociation holds at every steering strength, is larger for negative emotions, and the unreported emotion stays linearly readable.** (a) On Qwen3.5-9B, raising the steering strength (added vector as a fraction of the layer's activation norm, .05 to .3) lifts steering's share of the internal representation from .50 to .89 while its share of the self-report stays between .05 and .09, so the shaded gap between the curves grows from .45 to .80 (95% intervals around each curve). (b) Pleasant advantage is the share gained when the steered emotion is the pleasant rather than the unpleasant member of a pair, on the inside (slate) and in the report (ochre), with 95% intervals and the number of emotion pairs whose sign agrees: the report amplifies the asymmetry, .30 on Qwen3.5-9B (19/19 pairs) and .21 on Qwen3.5-27B (12/14 pairs), against .09 and .11 inside. The self-report comparison needs a model that names the steered emotion at least some of the time, though Gemma-4-31B never did in either direction (0/20 pairs) and Llama-3.3-8B answered 2% of the choices, hence panel (b) shows the two models where it is defined. (c) A linear probe trained without access to the steering direction identifies the steered emotion with accuracy .996, the prompted emotion for comparison at .945, while a random vector of the same norm sits at chance (.50), so the emotion the model does not report is still readable from its activations.

L.5 Across turns

Figure 21 moves the conflict into a five-turn conversation in which the system prompt pushes one emotion on every turn and steering pushes the opposite emotion on a single turn.

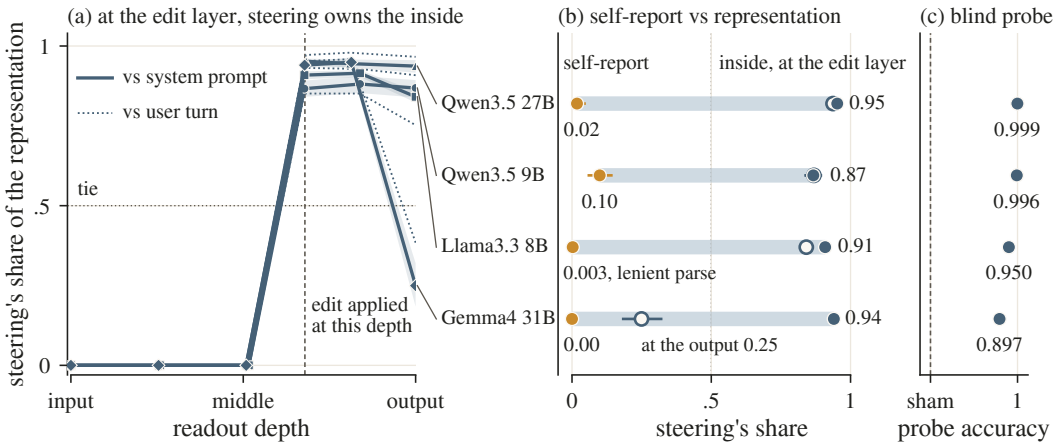


Figure 20: **Across three model families, steering takes the internal representation at the layers where it is applied, while the self-report stays at or below 0.10 in every model.** (a) Steering's share of the internal emotion representation as a function of readout depth, one curve per model, against the system prompt (solid, with 95% intervals over the 20 emotion pairs) and against the user turn (dotted). The share is zero before the edit and rises to 0.87 to 0.95 at the layer where the vector is added (relative depth 0.68 in all four models). Three models hold that share to the output (0.84 to 0.94); Gemma-4-31B holds 0.94 through the injection layers and then falls to 0.25 at the final layer, a depth effect of where the readout is taken. The model names between the panels index the rows of (b) and (c) as well. (b) The same contest as a per-model gap: the self-report share (ochre) against the representational share at the edit layer (slate, filled) and at the output (slate, hollow). Self-report shares are 0.02, 0.10 and 0.00 for Qwen3.5-27B, Qwen3.5-9B and Gemma-4-31B, which answered the forced choice in 32%, 78% and 31% of trials. Llama-3.3-8B answers in free prose rather than the exact option format, so its choices were read with a declared lenient rule (an explicit A)/B) marker, or exactly one of the two emotion words, giving steering 0.003 of its self-report (1 of 800 answers). (c) A probe trained on emotions no condition induces, so it never sees the steering direction, recovers the steered emotion in 0.90 to 0.999 of trials on every model, against 0.50 for a random-vector sham.

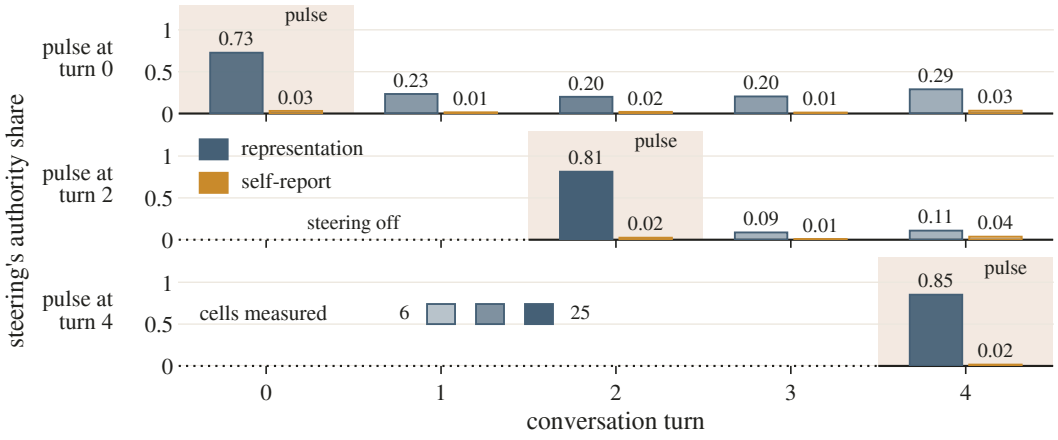


Figure 21: **A one-turn steering pulse owns the representation on its own turn only, and the self-report on no turn.** Each strip is a five-turn conversation in which the system prompt pushes one emotion on every turn and a single opposing steering pulse is applied at turn 0, 2 or 4 (shaded). Bars give steering's authority share of the internal representation (slate) and of the forced-choice self-report (ochre), averaged over the cells that carry a measurement out of 25 (five models by five emotion pairs), with lighter fill for fewer cells. On the pulse turn steering holds 0.73, 0.81 and 0.85 of the representation. On the turns after it falls to between 0.09 and 0.29 as the persistent text takes the state back. The self-report never exceeds 0.04 on any turn. Turns before a pulse (dotted baseline) are not drawn as data, with steering off the two arms coincide and the share is zero by construction, not by measurement.