ALIGNMENT CHANGES THE RESPONSE POLICY, NOT THE AFFECT REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Alignment training changes how language models respond to emotional messages, but it is unclear whether these changes are accompanied by changes in how they represent the user's emotions. We measure affect representation and reply valence across consecutive checkpoints of four models, spanning the base model, supervised fine-tuning (SFT), direct preference optimization (DPO), and, where available, reinforcement learning. Affect remains similarly decodable across stages under linear, nonlinear, and dimensional probes, and probes fitted at one stage transfer to the others. In contrast, judged reply valence rises most clearly from SFT to DPO, with all four judges agreeing on the direction of change. The proportion of replies containing first-person subjective expressions (e.g., "I'm sorry to hear that" or "I don't understand") falls from 46% to 29%. Replies shift away from echoing the user's negative feelings toward more neutral or humorous responses, and this pattern is consistent across input emotions. Encouragement, offers of help, and agreement do not increase, so the higher valence score alone does not establish greater consolation. Steering along a fixed valence direction changes replies at every checkpoint, but the SFT-to-DPO activation displacement along that direction is small, points away from positive valence, and does not increase the measured steering response. The reply-valence difference also persists when the direction is clamped. The SFT-to-DPO change is best localized to the response policy rather than to the measured affect representation: decodability, emotion-specific displacement, and valence-axis orientation remain stable at this transition.

1 INTRODUCTION

Alignment training changes how language models respond to emotional messages, often making replies sound more positive, amplifying sycophancy (Sharma et al., 2023), and leading models to validate users' incorrect beliefs, particularly when sadness is expressed (Ibrahim et al., 2025). To understand such behavior, we need to examine both what a model encodes about the user's emotion and what it says in reply. A more positive reply alone does not tell us whether the model represents the user's emotion differently or merely alters how it responds to that emotion.

Prior work has largely examined these two sides separately. Interpretability studies decode emotion from hidden states (Zhang & Zhong, 2025) and identify emotion-related directions that can influence outputs (Sofroniew et al., 2026). Studies of aligned behavior instead measure changes in replies (Ibrahim et al., 2025; Sharma et al., 2023). Each observes one side of the question, usually in a fixed model rather than across the stages at which alignment occurs. Neither measurement alone shows whether internal affect measures and replies change together during post-training. Comparing both on the same inputs across consecutive checkpoints allows us to locate the stages at which they diverge.

A divergence between affect decodability and reply behavior does not by itself reveal its mechanism. Training could change where inputs lie along an affect-related direction, how generation responds to that direction, or other parts of the mapping from hidden states to replies. We therefore ask three questions: how affect decodability changes across stages; when and how replies change; and whether displacement or altered sensitivity along a fixed valence direction helps explain the reply change. Interventions on this direction test these specific possibilities, without ruling out changes elsewhere in the model.

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.

We study four models at consecutive available stages: base, supervised fine-tuning (SFT), direct preference optimization (DPO) (Rafailov et al., 2023), and, where available, reinforcement learning with verifiable rewards. Tülu 3 and OLMo-2-7B provide four-stage sequences; OLMo-2-13B extends the comparison to a larger model, while Zephyr-7B provides a different base model and post-training recipe (Lambert et al., 2024; OLMo Team, 2024; Tunstall et al., 2023). At each checkpoint, we measure affect decodability from hidden states and evaluate replies to the same emotional inputs. Within each model, we hold a valence direction fixed across stages and measure its projection, its effect when used to steer generation, and the reply difference that remains when it is clamped.

Figure 1 summarizes the experimental design and findings. The measured affect representation remains stable in decodability, emotion-specific displacement, and valence-axis orientation, despite a small overall contraction of activation geometry. In contrast, model outputs shift substantially—judged reply valence rises most consistently from SFT to DPO, with little movement at other stages. Phrase counting, which requires no classifier, provides a classifier-free description of this change: about half of all SFT replies contain first-person subjective expressions (most commonly apologies or admissions of non-understanding), which preference optimization reduces to under one-third. Expressions of sadness are suppressed by a factor of approximately 0.7, regardless of whether the user expressed sadness, anger, or joy. Because sad replies are more common on negative input to begin with, the shift can nevertheless appear to target negative emotion. What takes their place is neutral rather than comforting, in that encouragement, hope, offers of help and agreement with the user all fail to increase.

Our contributions are threefold. First, we compare affect decodability and replies on the same inputs across consecutive post-training checkpoints, identifying a stage at which these measures diverge. Second, we characterize the reply change through both valence scores and string matching, showing that the higher valence score does not by itself indicate greater consolation. Third, we use steering and clamping interventions along a fixed valence direction to causally isolate representation from behavior, showing that SFT-to-DPO displacement is minimal and clamping leaves the shift intact. Stable affect decodability alone does not establish stable reply behavior.

2 Related Work

Emotion representations in language models. Language models encode affective information in their hidden states, with emotion geometry varying across layers and model scales (Zhang & Zhong, 2025). Emotion-related directions can also causally influence generated text (Sofroniew et al., 2026). These studies establish that affect is decodable and behaviorally relevant within a model. They do not jointly track affect decodability and reply behavior across consecutive post-training checkpoints. We add this training-stage comparison and ask whether the two measurements change together.

What alignment preserves and changes. The superficial alignment hypothesis posits that post-training primarily alters style and format (Zhou et al., 2023), shifting output distributions mainly within the first few tokens (Qi et al., 2024) via magnitude-preserving rotations in the residual stream (Zhang & Lin, 2026). Behaviorally, alignment increases warmth and sycophancy (Ibrahim et al., 2025; Sharma et al., 2023), often creating a divergence where models internally recognize a claim as false while explicitly agreeing with it (Pandey, 2026; Baez et al., 2026). Mechanistically, post-training acts as a thin wrapper over pretrained capabilities (Jain et al., 2024), enhancing existing circuits (Prakash et al., 2024) or bypassing toxic features rather than removing them (Lee et al., 2024). However, prior work mostly examines single-step alignment and focuses on factual knowledge or safety, leaving it unclear whether emotion exhibits a similar representation–behavior gap. Furthermore, it remains unknown at which post-training stage such a gap emerges and whether alignment alters the underlying emotion representation, shifts sensitivity to it, or merely modifies downstream output generation. We address these questions by tracking emotional representation–behavior divergence across a complete post-training pipeline.

Probing and intervention. Decodability alone does not establish that a feature is used during generation (Hewitt & Liang, 2019; Belinkov, 2022; Elazar et al., 2021). Representation engineering instead tests whether perturbing a direction changes model output (Zou et al., 2023; Li et al., 2023), including for emotion-related directions (Sofroniew et al., 2026). We combine these approaches:

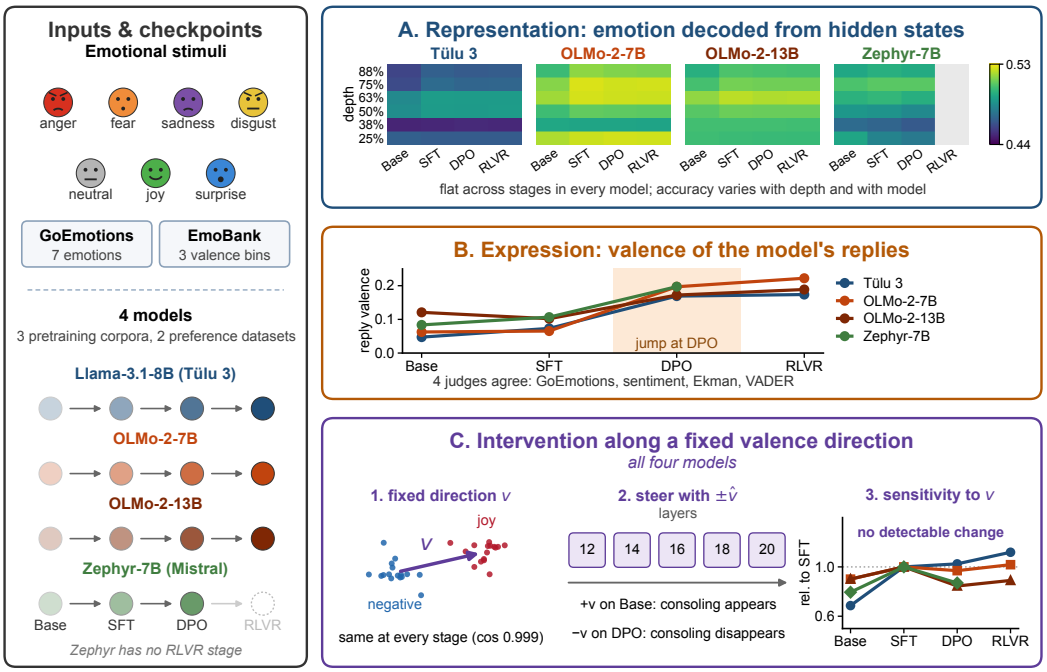

Figure 1: **Study overview and main results.** Four models are followed through their consecutive post-training stages (left). On the same checkpoints we measure (A) how well the user's emotion can be decoded from the hidden states, (B) the valence of the replies those checkpoints give to emotional input, and (C) how far those replies move when we steer along one fixed valence direction. As the plots show: (i) decodability is flat across stages in every model; (ii) reply valence rises at DPO in all four; and (iii) sensitivity to the fixed direction, once differences in starting valence are controlled, does not change in any model.

probes measure affect decodability, while projection, steering, and clamping test whether displacement or altered response along one fixed valence direction accounts for the observed change in replies. These interventions constrain this candidate direction rather than identifying every affect-related pathway.

3 Methods

We compare affect representations and reply behavior across checkpoints within each model family. We then intervene on a valence direction held fixed across stages to measure its influence on replies and test whether the SFT–DPO behavioral difference persists when variation along that direction is clamped.

3.1 Measuring Affect Decodability

Let $h_s^\ell(x)$ denote the residual-stream activation produced by checkpoint M_s at layer ℓ for input x, mean-pooled over input tokens. At each layer we predict the input emotion from $h_s^\ell(x)$ and compare performance across stages. Our primary measure is nearest-centroid classification of the seven Ekman categories over five random $70/30$ splits; we report its variation across splits, so a stage difference can be judged against the probe's own noise. As capacity checks we repeat the analysis with logistic-regression and MLP probes (Alain & Bengio, 2016; Hewitt & Liang, 2019), with ridge regressions on continuous valence, arousal and dominance ratings, and with the final input token in place of mean pooling. Refitting at each stage asks only whether emotion is still recoverable, so we also freeze the probe fitted at one stage, standardizer and class means together, and apply it unchanged at the others.

Decodability can remain stable while the activations move, so we measure how far they move. Linear CKA (Kornblith et al., 2019) quantifies overall activation similarity between each checkpoint

and the base model. We interpret stable decodability alongside this direct measure of activation change. CKA is invariant to isotropic rescaling, so we also track the spread of the emotion-class configuration (App. B.4). To test whether alignment selectively relocates negative emotions, we compare each emotion's centroid displacement with the mean displacement across emotions. We also measure rotation of the valence, arousal, and dominance axes by computing cosine similarity between each stage's ridge-regression weights and the base weights. Complete layer-wise results are in App. B.1.

3.2 Measuring Reply Behavior

Let $r_s(x)$ denote the reply checkpoint M_s generates for input x. We evaluate replies in three complementary ways. ***Reply valence*** is the mean valence assigned over a fixed stimulus set,

$$E_s = \frac{1}{N}\sum_{i=1}^{N} V\big(r_s(x_i)\big), \qquad V(r) = \sum_c \tilde{p}(c \mid r)\, w_c \in [-1, 1], \tag{1}$$

with $w_c = +1$ for joy, -1 for anger, disgust, fear and sadness, and 0 otherwise. ***Reply emotion*** is the top class of a seven-way judge, which compares stages without mapping the classes onto a valence scale. ***Phrase counts*** use string matching over five predefined categories: (i) first-person subjective expressions, in which the model is the subject of a verb of feeling, thinking, or knowing ("I'm sorry", "I don't understand"); (ii) encouragement or hope; (iii) offers of help; (iv) agreement with the user; and (v) light or humorous language. These are occurrence rates for predefined linguistic patterns rather than measures of communicative function; the matching rules are given in App. E.

The primary judge is a multi-label RoBERTa GoEmotions classifier: $\tilde{p}$ sums its 28 independent sigmoid scores into the seven Ekman categories and renormalizes them to one. We repeat the analysis with a seven-class Ekman classifier, a three-class sentiment classifier and the rule-based VADER lexicon (Hutto & Gilbert, 2014). Agreement across these differently constructed judges tests whether the result depends on one scoring instrument. Because reply valence need not measure consolation, we interpret it jointly with the classifier-free phrase counts.

3.3 Interventions on a Fixed Valence Direction

We test three properties of one candidate valence direction: whether model outputs respond to it, whether post-training displaces activations along it, and whether the SFT–DPO reply difference persists when variation along it is removed.

For each model, we construct the direction on the base checkpoint at a mid-network layer (16 of 32, and 20 of 40 in OLMo-2-13B) as

$$v = \overline{h^{16}_{\text{base}}}(\text{joy}) - \overline{h^{16}_{\text{base}}}(\text{negative}), \qquad \hat{v} = v/\|v\|, \tag{2}$$

where negative comprises anger, disgust, fear, and sadness. We hold $\hat{v}$ fixed across stages; directions independently estimated at other stages are used only to measure stability.

For steering, we edit the residual stream at every generated token after five layers spanning the region where emotion is well decodable:

$$h^\ell_t \leftarrow h^\ell_t + \alpha \|h^\ell_t\| \hat{v}, \tag{3}$$

with $\alpha \in \{\pm 0.05, \pm 0.10, \pm 0.15\}$. Adding $+\hat{v}$ at BASE tests whether the direction can change reply valence; subtracting it at DPO tests the converse effect. We refer to adding or subtracting the direction as positive or negative steering. We use *clamping* for the intervention below that replaces the direction component. A matched random unit direction controls for generic perturbation; the valence-orthogonal component of an arousal direction provides an additional affect-related control. We report $\alpha = 0.05$ and 0.10 in the main text because $\alpha = 0.15$ begins to degrade generation. Steering uses 200 negative stimuli for each model.

To compare sensitivity across stages, we measure the two-sided reply-valence swing beyond the random control,

$$S_s = \big[\Delta_s(+\hat{v}) - \Delta_s(-\hat{v})\big] - \big[\Delta_s(+\hat{u}) - \Delta_s(-\hat{u})\big], \tag{4}$$

where $\Delta_s(\hat{d})$ is the change relative to unsteered generation. Because DPO replies begin at a higher valence, we estimate this quantity within bins of unsteered reply valence and use the same bin

weights at each stage, so that greater steering sensitivity would show a larger S_s at DPO than at SFT.

We separately summarize activation displacement along the base direction:

$$\pi_s = \frac{\mathbb{E}_x[\langle h_s^{16}(x) - h_{\text{base}}^{16}(x), \hat{v}\rangle]}{\mathbb{E}_x[\|h_{\text{base}}^{16}(x)\|]}. \tag{5}$$

This is a signed measure of movement along the selected axis; it is not a measure of total representational change, and because a signed mean can cancel across inputs we also report $\mathbb{E}_x|\cdot|$ and the fraction of inputs sharing the sign. A representation moved by training predicts a large π_{DPO} that tracks the behavioral change.

Finally, we clamp each token's component along $\hat{v}$ to its batch-median value at the same layers used for steering. Replacing the component with the batch median rather than with zero keeps the state within the range the model itself produces while erasing the differences between inputs along that axis. We apply this operation to the SFT and DPO checkpoints of Tülu and OLMo-2-7B over all seven input emotions, with a matched random direction as control. If the stage difference is captured through $\hat{v}$, the SFT→DPO rise on negative input shrinks under clamping; if it is an input-independent offset, the rise remains.

4 Experiments

4.1 Experimental Setting

Data. We evaluate affect under two complementary annotation schemes. GoEmotions (Demszky et al., 2020) provides discrete emotion labels; we retain the $8{,}433$ Reddit comments that map unambiguously onto one of the seven Ekman categories (Ekman, 1992). EmoBank (Buechel & Hahn, 2017) provides continuous ratings of valence, arousal, and dominance for $9{,}793$ sentences. The datasets also differ in register: GoEmotions consists of Reddit comments, whereas EmoBank includes news, blogs, and fiction. We additionally use both datasets as sources of behavioral stimuli. The main set contains 150 GoEmotions sentences per emotion ($1{,}050$ total); a second set contains 140 EmoBank sentences per valence bin (420 total) and tests whether the behavioral result generalizes across registers. Table 2 in App. A summarizes both corpora.

Models. We study four models spanning three pretraining lineages. Tülu 3 and OLMo-2-7B are available at four consecutive stages, $s \in \{\text{BASE}, \text{SFT}, \text{DPO}, \text{RLVR}\}$: Llama-3.1-8B aligned with the Tülu 3 recipe (Lambert et al., 2024), and OLMo-2-7B at its official released checkpoints (OLMo Team, 2024). Their different pretraining lineages allow us to test whether the stage-wise pattern replicates across model families. The remaining models vary size or alignment recipe. OLMo-2-13B retains the OLMo-2 training pipeline at a larger scale (40 rather than 32 layers). Zephyr-7B (Tunstall et al., 2023) uses a Mistral-7B base, UltraChat for SFT, and UltraFeedback for DPO, and has no RLVR checkpoint. Within each model ladder, the checkpoints share an architecture and tokenizer, avoiding these confounds in comparisons across stages. Exact checkpoint identifiers are listed in Appendix A.

Procedure. Every checkpoint receives the same two-shot continuation prompt, "*Person A: x Person B:*", without a chat template. Holding the prompt format fixed supports direct comparisons across stages. We decode greedily for at most 48 new tokens and retain the first line of each reply. As a robustness check, we repeat the main comparison using each tuned model's native chat template. Fewer than 0.2% of replies meet our degeneracy criterion. We compare stages on matched stimuli using paired Wilcoxon signed-rank tests and report stimulus-level bootstrap 95% confidence intervals; full details appear in Appendix A.

4.2 Main Results

We compare affect decodability and reply behavior on the same checkpoints and stimuli. The main pattern is a divergence between the two: affect remains similarly decodable across stages, whereas reply valence changes most consistently at the SFT→DPO transition.

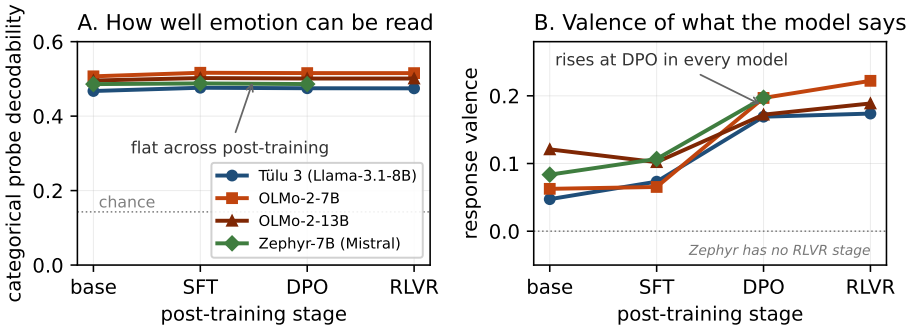

Figure 2: **Affect decodability and reply valence follow different trajectories.** (A) Categorical emotion decodability, averaged over the probed layers in the upper half of the network and excluding the output layer, is flat across stages and well above chance (0.14, dotted). Layers are selected by relative depth, because OLMo-2-13B has 40 blocks and the other three have 32. (B) Reply valence changes little from BASE to SFT and rises most consistently from SFT to DPO. Both patterns hold in all four models.

Affect decodability remains stable across stages. Across the four stages, probe accuracy varies by at most 0.02, a difference comparable to the probe's split-to-split variation, and no checkpoint after SFT improves consistently on the preceding stage. The pattern holds across nearest-centroid, logistic-regression and MLP probes, across categorical and dimensional labels, and when the final input token replaces mean pooling; the largest across-stage difference at any probed layer is 0.008 in OLMo-2-13B and 0.016 in Zephyr-7B. Refitting a probe at each stage asks only whether emotion remains recoverable. We therefore freeze the probe fitted at one stage, including its standardizer and class means, and apply it unchanged at the others. Across all fitting and evaluation stages, transfer costs at most 0.003 accuracy relative to matched-stage values of 0.47–0.51 (App. B.2).

Stable decodability does not imply that the activations are unchanged. At mid-network depth, the displacement from BASE to the final available checkpoint differs by roughly four orders of magnitude between OLMo and Tülu ($1 - \mathrm{CKA} \approx 7 \times 10^{-2}$ against 6×10^{-6}). The per-emotion centroid shifts do not single out negative emotions, and the valence, arousal and dominance axes do not rotate (Fig. 4). A small contraction of the emotion-class configuration occurs in the deeper layers, by 2–9% of the base spread and at the same relative depth in all four models. Linear CKA is invariant to uniform rescaling and therefore does not register this contraction. The contraction changes neither probe accuracy nor the displacement of any individual emotion (App. B.4).

Reply valence changes most consistently from SFT to DPO. Reply valence changes little from BASE to SFT and rises most consistently from SFT to DPO, by +0.10 in Tülu, +0.13 in OLMo-2-7B, +0.07 in OLMo-2-13B and +0.09 in Zephyr. Four judges agree on the direction of the pooled change and differ in magnitude: +0.10 under the GoEmotions judge, +0.08 under a three-class sentiment classifier, +0.04 under VADER and +0.02 under the Ekman classifier (App. C.1). The effect also appears with the tuned models' own chat templates and on EmoBank stimuli from a different text register (Apps. C.3 and C.4).

The increase is not restricted to one input emotion. Reply valence rises for every negative input emotion in every model, by +0.03 to +0.18, though the magnitude varies across emotions and models (Fig. 6). Replies to negative input therefore cross from negative to positive more often because they start lower, not because the measured increase is exclusive to them. Reply length and prompt format do not account for the change (App. C.3), nor does a fall in reply arousal (App. C.5). Together, these measurements show that affect remains similarly decodable across the transition where reply valence rises at DPO. Figure 2 plots this divergence for all four models.

The change shows up as fewer first-person subjective phrases. String matching identifies a concrete change in the replies without relying on a valence classifier. The fraction containing a first-person subjective expression falls from 53% to 29% in Tülu, 44% to 21% in OLMo-2-7B, 52% to 38% in OLMo-2-13B and 36% to 27% in Zephyr. The suppressed expressions are more often cognitive than emotional: at SFT, 22–32% of replies contain *I think* or *I don't understand*, while about 20% contain *I'm sorry* or *I'm sad*. Apology or sympathy falls from 10–18% to 7–8% and

admissions of incomprehension from 1–8% to 1–3%, while light or humorous replies rise from 1–4% to 4–9%. Encouragement and hope stay at zero, offers of help at or below 1%, and agreement with the user changes by at most one percentage point in either direction (App. E).

The judged emotion of the replies supports a limited reading of the shift. Replies classified as sad fall from 12–15% to 9–10%, and neutral replies increase more than joyful ones. At SFT the reply's emotion tracks the input's: sad replies are 24.5% of the model's replies to sadness but 5.2% of its replies to joy. Estimated separately for each of the seven input emotions, the negative-steering factor on sad replies stays near 0.7 throughout, between 0.58 and 0.97: one factor, applied whatever the user expressed. Applying a constant factor to unequal baselines removes most where there was most to remove, so the model echoes the user's negative emotion less while the the negative-steering factor does not depend on the input. The higher reply-valence score therefore comes with less first-person subjective language and less echoing of negative affect; it should not by itself be read as greater consolation or agreement with the user.

4.3 Ablation and Steering Results

We intervene on a fixed valence direction to ask whether the SFT→DPO change is explained by displacement along it, by a change in the response to it, or by neither. The direction is computed from the BASE checkpoint as the difference between the mean activation for joy and negative-emotion inputs, then held fixed across stages within each model. We run steering on all four models and clamping on Tülu 3 and OLMo-2-7B.

The valence direction changes replies. The direction is highly similar across stages (the BASE-to-DPO cosine is 0.999), and steering along it changes reply valence at every checkpoint. At α=0.1, adding $+\hat{v}$ at BASE shifts the valence of replies to negative inputs by 0.22 to 0.80 across the four models. Subtracting the direction at DPO shifts reply valence by 0.15 to 0.34. A matched random direction of equal norm shifts valence by only 0.02 to 0.16 on the addition side and is smaller than the valence direction in seven of the eight cells. Figure 3 plots the dose–response at both strengths, and the direction itself is nearly identical across stages, the BASE-to-DPO cosine being 0.999. App. F.1 gives every cell, the one exception among the random controls, and the arousal control. The replies respond to this direction at every stage, making the subsequent null results interpretable.

The measured displacement and response do not track the behavioral shift. In Tülu the SFT→DPO displacement at layer 16 is 0.29% of the state's own length, identical in the signed and absolute measures because every input carries the same sign, and that sign is away from joy; the sign differs between the two ladders. Teacher-forcing the *same* reply text through the SFT and DPO checkpoints of OLMo-2-7B gives +0.01%, −0.02% and −0.20% at layers 12, 16 and 20, whichever checkpoint wrote the text.

The paired differences in two-sided steering swing are +0.01, −0.03, −0.20 and −0.03 across the four models (Table 1); three fall, one of them significantly, and none rises. The apparent rise in two models before correction disappears once stimuli are matched for unsteered valence. Steering reproduces the behavioral change only at 5–10% of the state's norm, more than an order of magnitude above the displacement, although the two are not matched interventions. The intervals do not show an increase in measured sensitivity, although they do not establish that the response is exactly identical across stages.

The reply-valence difference persists under clamping. During generation, we replace the component along $\hat{v}$ at the steering layers with the batch median and compare the SFT-to-DPO difference in reply valence. On negative stimuli, the difference persists under clamping in both tested models: it remains +0.11 in Tülu and is +0.13 without clamping versus +0.18 with clamping in OLMo-2-7B. Clamping a matched random direction leaves the difference similar. The manipulation is behaviorally effective: in OLMo-2-7B it changes 49.7% of replies against 6.7% for the random control and lowers the valence of replies to joyful input (App. F.2). This is the strongest of the three tests, because its answer holds whatever preference optimization changed, but what it shows is that clamping this direction does not remove the stage difference; other affect-related directions, feature interactions and later computation remain open.

Across these interventions, the selected direction does not account for the SFT–DPO behavioral difference. The evidence points to a change in the emitted reply rather than to displacement along this direction, while other affect-related pathways remain possible.

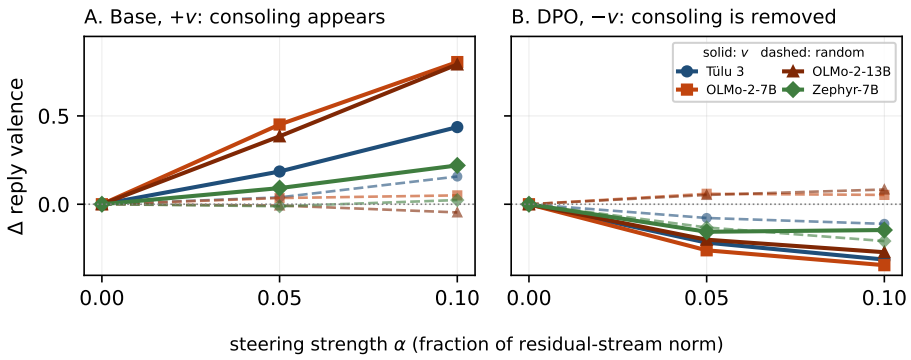

Figure 3: **Steering a fixed valence direction controls the reply style.** (A) On BASE, adding $+v$ during replies to negative input raises reply valence with a dose-response (solid); a random direction of equal norm moves it much less (dashed). (B) On DPO, subtracting v lowers reply valence back toward base, again more than the random direction. Colors are models, solid lines the valence direction and dashed lines the matched random control; all four models are steered with their own direction. Numbers are in Table 12, App. F.1.

Table 1: **Sensitivity to the fixed valence direction does not increase at DPO.** *Raw*: the drop in reply valence under $-\hat{v}$ beyond a random direction of equal norm (α=0.1). *Controlled*: the two-sided swing ($-\hat{v}$ to $+\hat{v}$, minus the random swing) on stimuli matched for unsteered valence. Read raw, sensitivity appears to rise at DPO in Tülu and OLMo-2-7B; that is a ceiling effect, since DPO replies start more positive and so have further to fall. Under a paired bootstrap the controlled SFT→DPO differences are $+0.01$ $[-0.17, 0.19]$, -0.03 $[-0.18, 0.12]$, -0.20 $[-0.38, -0.02]$ and -0.03 $[-0.22, 0.17]$: three of the four fall, one significantly, and none rises. Bootstrap 95% CIs in brackets. 200 negative stimuli for four models.

Model	Measure	BASE	SFT	DPO	RLVR
Tülu 3	raw	-0.04 $[-0.13, 0.07]$	0.08 $[-0.01, 0.18]$	0.20 $[0.10, 0.31]$	0.22 $[0.11, 0.32]$
	controlled	0.27 $[0.14, 0.43]$	0.39 $[0.23, 0.53]$	0.40 $[0.27, 0.54]$	0.44 $[0.31, 0.56]$
OLMo-2-7B	raw	0.29 $[0.21, 0.38]$	0.25 $[0.16, 0.34]$	0.40 $[0.31, 0.51]$	0.43 $[0.33, 0.53]$
	controlled	1.01 $[0.88, 1.14]$	1.12 $[1.00, 1.23]$	1.08 $[0.97, 1.20]$	1.14 $[1.01, 1.27]$
OLMo-2-13B	raw	0.30 $[0.18, 0.42]$	0.39 $[0.28, 0.50]$	0.36 $[0.23, 0.48]$	0.34 $[0.22, 0.46]$
	controlled	1.15 $[0.99, 1.29]$	1.28 $[1.12, 1.43]$	1.08 $[0.92, 1.25]$	1.14 $[0.98, 1.30]$
Zephyr-7B	raw	-0.04 $[-0.18, 0.10]$	-0.04 $[-0.16, 0.08]$	-0.06 $[-0.17, 0.04]$	—
	controlled	0.19 $[0.01, 0.38]$	0.24 $[0.08, 0.42]$	0.21 $[0.04, 0.39]$	—

5 DISCUSSION

A stage-resolved form of shallow alignment. Prior work locates shallowness in what post-training changes: style and format more than knowledge (Zhou et al., 2023), or the first few output tokens more than later ones (Qi et al., 2024). Our results add a stage-resolved observation located in the computation. The largest and only consistent change in reply valence falls between SFT and DPO. The measured representation-level quantities show no corresponding change: affect remains equally decodable under every probe we tested, a probe fitted at one stage transfers to the others, and the valence axis does not rotate. The intervention results likewise show no corresponding increase in the measured response to the valence direction once the starting point is controlled, and clamping that direction leaves the behavioral shift intact. Within what these interventions reach, the evidence retains only the last of the three hypotheses in Sec. 1. The objective makes that intelligible: direct preference optimization raises the relative likelihood of preferred whole replies and never refers to the hidden states, so changing which reply the policy emits satisfies it while nothing requires the representation to move — a hypothesis our experiments motivate rather than test. Steering moves replies at every checkpoint, so the direction is read; the clamping result shows that preference optimization did not act through it. A lever can be functional and still not be the one that was pulled.

Relation to preserved and bypassed capabilities. Tracking both measurements across the released checkpoints of four models pinpoints the representation–behavior divergence to a single step. It appears at preference optimization, while the transitions on either side are flat or far smaller. What is typically attributed to general alignment thus belongs specifically to DPO. This finding extends a broader pattern established elsewhere in mechanistic interpretability, in which fine-tuning changes how a model uses a capability while leaving the capability intact. Post-training acts as a thin wrapper over pretrained abilities (Jain et al., 2024) and amplifies circuits that already exist (Prakash et al., 2024). A related dissociation appears for factual claims, where models encode a claim as false while asserting it (Pandey, 2026; Baez et al., 2026). Affect follows the same broader pattern, and the released checkpoint ladders show that the divergence emerges at preference optimization.

Lee et al. (2024) show that DPO bypasses toxic capabilities by routing around them, whereas clamping the affect direction here does not diminish the behavioral shift at all, so DPO neither routes around this feature nor works through it. Our interventions test a single valence direction rather than tracing the whole path from the user's emotion to the reply, so what they establish is where the change does not lie rather than a mechanism for it.

What a probe can and cannot predict. The probing literature warns that a decodable feature need not be a used one (Hewitt & Liang, 2019; Belinkov, 2022; Elazar et al., 2021). Our result is the converse case and, for evaluation, the more awkward of the two: the feature remains decodable, remains used, and is read no more strongly, and the behavior changes regardless. An audit restricted to representations would have returned "no change" at precisely the transition at which the replies changed most. Representational stability does not license a claim that behavior is stable, and a behavioral shift is not evidence that a representation has moved. Where the two must be linked, they have to be measured on the same checkpoints, which is inexpensive relative to the training run that produced them.

6 Conclusion

This study asks whether alignment changes how language models represent emotion, or only what they do with that representation. We measure the representation and the replies on the same checkpoints, across four models and every stage of post-training. Affect stays equally decodable at every stage, and a probe fitted at one stage decodes the others equally well. The replies do not follow: reply valence rises most consistently between SFT and DPO, over the same transition the share of replies containing a first-person subjective expression falls from 46% to 29%, and replies classified as sad become less frequent by a constant factor whatever emotion the user expressed. Encouragement, offers of help and agreement with the user do not increase, so the higher judged valence does not by itself indicate greater consolation. A stage-invariant valence direction remains a working lever on that behavior, yet across the same transition the representation is not displaced along it toward positive valence, the response to it does not rise, and clamping it leaves the shift intact. These measurements locate where alignment acts and at which stage, and separate the representation from what the policy does with it; they constrain explanations that run through the selected direction without excluding changes elsewhere. Within what they reach, alignment changed the reply the policy emits, not the representation it reads.

7 Limitations

Our study has three main limitations. Our study focuses on four 7–13B open-weight models, all preference-optimized via DPO, in single-turn English settings. While intra-model comparisons across stages are controlled, generalizing to larger architectures, alternative algorithms (e.g., PPO), or multi-turn dialogues requires caution. Because we use public checkpoints, we localize the divergence to the overall SFT→DPO transition, without isolating the specific contributions of preference data, loss functions, or optimization dynamics. Extending this framework to custom training pipelines and interactive multi-turn contexts remains future work.

Ethics Statement

This study uses publicly released open-weight models and public datasets (GoEmotions, EmoBank) under their respective licenses. No new human-subject data were collected. We acknowl-

edge that Reddit-derived data may contain distressing content. Crucially, our findings show that representation-level monitoring alone may not predict how aligned models respond to user distress, highlighting risks in emotionally sensitive deployments. All steering interventions were restricted to 7–13B open research models, and no modified weights are released.

Reproducibility Statement

All models, data preprocessing, probing setups, generation parameters, and hardware details are fully specified in Section 3 and Appendix A. All experiments use float64 precision with fixed random seeds (42). We provide the complete, anonymized codebase covering extraction, probing, elicitation, steering, and plotting in the supplementary material.

AI use statement

We used large language models solely to assist with grammar, clarity, phrasing, and text organization. All experimental designs, code implementation, execution, and reported results were independently carried out and verified by the authors. The authors take full responsibility for the scientific claims and final content of this paper.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Anthony Baez, Sheer Karny, and Pat Pataranutaporn. Dissociating the internal representations of sycophancy in llms. *arXiv preprint arXiv:2607.07003*, 2026. Mechanistic Interpretability Workshop, ICML 2026.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Sven Buechel and Udo Hahn. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3–4):169–200, 1992.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. 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. 2733–2743, Hong Kong, China, 2019.

Clayton Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pp. 216–225, 2014.

Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. Training language models to be warm and empathetic makes them less reliable and more sycophantic. *arXiv preprint arXiv:2507.21919*, 2025.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *International Conference on Learning Representations*, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K. Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: A case study on DPO and toxicity. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26361–26378, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *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/.

OLMo Team. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*, 2024.

Manav Pandey. LLMs know they're wrong and agree anyway: The shared sycophancy-lying circuit. *arXiv preprint arXiv:2604.19117*, 2026.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations*, 2024.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 2023.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

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. *arXiv preprint arXiv:2604.07729*, 2026.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of LM alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013.

Jingxiang Zhang and Lujia Zhong. Decoding emotion in the deep: A systematic study of how llms represent, retain, and express emotion. *arXiv preprint arXiv:2510.04064*, 2025.

Long Zhang and Fangwei Lin. Simulated adoption: Decoupling magnitude and direction in LLM in-context conflict resolution. *arXiv preprint arXiv:2602.04918*, 2026.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Experimental and Reproducibility Details

This appendix lists the checkpoints, the data preparation, and every setting needed to reproduce the numbers in the main text.

Checkpoints. The Tülu family comprises `meta-llama/Llama-3.1-8B` (base), `allenai/Llama-3.1-Tulu-3-8B-SFT`, `allenai/Llama-3.1-Tulu-3-8B-DPO`, and `allenai/Llama-3.1-Tulu-3-8B` (RLVR) (Lambert et al., 2024). The OLMo family comprises `allenai/OLMo-2-1124-7B` (base), `allenai/OLMo-2-1124-7B-SFT`, `allenai/OLMo-2-1124-7B-DPO`, and `allenai/OLMo-2-1124-7B-Instruct` (RLVR) (OLMo Team, 2024). Within each family, the checkpoints share the tokenizer and architecture (32 layers, hidden size 4096), so comparisons are not confounded by tokenizer or architectural differences.

Data. Categorical emotion: GoEmotions (Demszky et al., 2020) mapped to seven Ekman categories (8,433 sentences). Dimensional: EmoBank (Buechel & Hahn, 2017) standardized valence/arousal/dominance (9,793 sentences). Behavior stimuli are drawn from GoEmotions (150 per Ekman category, 1,050 total) and EmoBank (valence terciles, 140 per bin, 420 total).

Probes. Categorical readability is nearest-centroid accuracy over five random $70/30$ splits; VAD readability is five-fold ridge R^2 (features standardized on the train fold; identical regularizer across stages). For the probe-capacity triangulation we additionally fit a logistic-regression linear probe ($C{=}1$, standardized features) and a non-linear MLP probe (one hidden layer of 128 units, early stopping); both give the same stage-flat pattern as nearest-centroid. CKA is linear CKA (Kornblith et al., 2019) in float64.

Generation. Two-shot continuation prompt, no chat template on any checkpoint, greedy decoding, up to 48 new tokens; a repeated 4-gram flags degeneracy. Expression judges: `SamLowe/roberta-base-go_emotions`, `j-hartmann/emotion-english-distilroberta-base`, `cardiffnlp/twitter-roberta-base-sentiment-latest`, and the VADER lexicon; reply arousal is scored with the Warriner (Warriner et al., 2013) and NRC-VAD (Mohammad, 2018) lexicons.

Steering. Difference-of-means valence direction from the cached activations; added to the residual stream after layers 12, 14, 16, 18, and 20 during generation, at strength $\alpha \in \{0.05, 0.1, 0.15\}$ (fraction of the per-token residual norm); controls are a matched-norm random direction and the valence-orthogonal component of the arousal direction.

Statistics. Stage-to-stage differences are tested with paired Wilcoxon signed-rank tests over the same stimuli. Bootstrap 95% confidence intervals are stimulus-level, from 500 resamples for single-stage quantities and 1,000 for paired stage-to-stage differences. The EmoBank probe set drops the 269 of 10,062 sentences that are empty or exceed the probed context.

Infrastructure. All checkpoints are run through HuggingFace Transformers 4.51 (PyTorch 2.5.1, Python 3.12, Debian 13) on two NVIDIA RTX 4090 GPUs (driver 550.163.01, CUDA 12.4); activations are extracted in float16 and all analysis is done in float64 on CPU. Random seed 42 throughout; ridge probe regularizer $\lambda{=}10^3$, identical across stages.

Table 2: Corpora used for probing (top) and reused as behavioral stimuli (†, bottom). *Chance* is the accuracy of random guessing on the categorical probe.

Corpus	Labeling	Labels	N	Chance
GoEmotions	categorical	7 Ekman	8,433	0.14
EmoBank	dimensional	valence, arousal, dominance	9,793	–
GoEmotions†	stimuli	7 Ekman $\times$150	1,050	–
EmoBank†	stimuli	3 valence bins $\times$140	420	–

B Robustness of Affect Decodability

We test whether the stable decodability result depends on the annotation scheme, pooling rule, probe family, or geometric summary.

B.1 Alternative annotation schemes and probe families

Table 3 reports emotional readability across both annotation paradigms and both families. The pattern is uniform: readability is essentially constant across the four post-training stages. A small BASE→SFT gain appears on the categorical task but is smaller than the cross-validation split-to-split variation and does not survive the non-linear MLP probe. The DPO and RLVR checkpoints do not improve readability; on the continuous VAD probes they slightly reduce it.

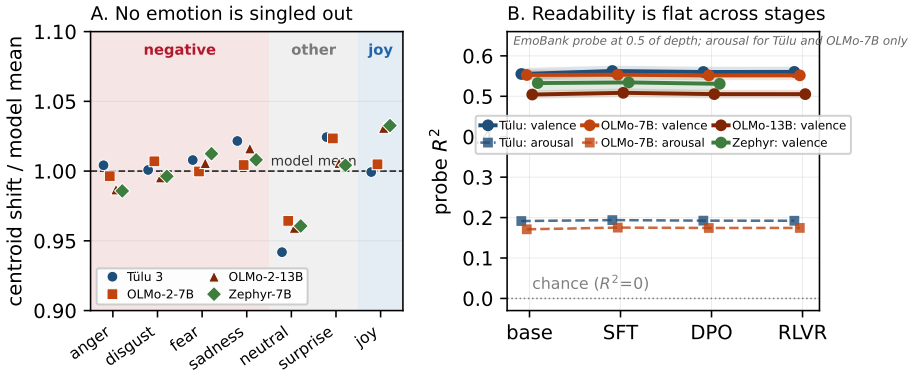

Figure 4: **The measured affect representation shows no selective emotion-specific relocation.** (A) From BASE to SFT, every emotion's centroid moves by nearly the same amount, in all four models, shown relative to each model's mean shift at 0.625 of network depth (layer 20, or 25 for the 40-block OLMo-2-13B). No valence group is singled out: the spread between the largest and smallest per-emotion displacement is 6–9% here, and the emotion that moves most is joy or surprise rather than a negative one. (B) What a probe can read is flat across all four stages: valence, the main affect axis (ridge $R^2 \approx 0.55$; shaded band is across-split noise), and the weaker arousal signal ($R^2 \approx 0.18$). Panel (B) reads the probe at 0.5 of network depth in each model; the EmoBank run for OLMo-2-13B and Zephyr covered valence only, so arousal is shown for the two models that have it.

Table 3: Emotional readability across both annotation paradigms and two families. Readability is nearest-centroid accuracy (categorical: chance 0.14) or ridge R^2 for valence/arousal/dominance. *A sign test over layers gives $p < 0.01$, but these small categorical effects are below the split-to-split SD (≈ 0.011) and do not replicate with the MLP probe. No checkpoint after SFT improves readability.

paradigm	family	readability	SFT gain	after DPO/RL
categorical (Ekman-7)	Tülu	0.40–0.49	$+0.006^*$	flat
	OLMo	0.46–0.53	$+0.007^*$	flat
dimensional (VAD R^2)	Tülu	0.56/0.20/0.06	≈ 0	-0.005
	OLMo	0.55/0.18/0.03	≈ 0	-0.005

B.2 Transfer of a probe across stages

Table 4 reports nearest-centroid accuracy when the probe is fitted at one stage and evaluated at another, on the same held-out inputs and with no refitting, averaged over the five splits used throughout. *Frozen* carries the fitting stage's standardizer and class means unchanged to the evaluation stage; *recalibrated* standardizes the evaluation stage with its own training statistics and carries only the class means, so it tests the class geometry up to a per-feature rescaling. We run this for all four models, at four layers chosen by the same relative depths in each (0.375, 0.50, 0.625 and 0.75, which are layers 12, 16, 20 and 24 of 32 and layers 15, 20, 25 and 30 of 40 in OLMo-2-13B). Both variants are flat everywhere: over all four models, both variants, every stage pair and every probed layer, the largest matched-minus-transferred gap is 0.003, against matched-stage accuracies of 0.47–0.51.

Table 4: Nearest-centroid accuracy at layer 16 with the probe fitted at one stage (rows) and evaluated at another (columns), averaged over five random $70/30$ splits. Chance is 0.14. Tülu 3 is shown; the other three models give the same picture, and across all four the diagonal, which refits at the evaluation stage, buys at most 0.003 over transferring.

	frozen				*recalibrated*			
Fitted at	BASE	SFT	DPO	RLVR	BASE	SFT	DPO	RLVR
BASE	0.484	0.487	0.486	0.486	0.484	0.487	0.486	0.486
SFT	0.487	0.489	0.489	0.489	0.490	0.489	0.489	0.489
DPO	0.486	0.489	0.489	0.489	0.489	0.489	0.489	0.490
RLVR	0.486	0.489	0.489	0.490	0.489	0.489	0.489	0.490

B.3 Final-token representations

The main decodability analysis uses mean-pooled hidden states. To test whether the stability result depends on that choice, we re-extract activations at the final input token immediately preceding generation and repeat the nearest-centroid probe. Categorical decodability is again stable across the four post-training checkpoints in both families: the largest across-stage spread is 0.026 for Tülu (decodability in the range 0.36–0.41) and 0.014 for OLMo (0.39–0.42), both well above the $1/7$ chance level. The categorical affect signal is therefore similarly decodable under final-input-token pooling (Sec. 4.2).

B.4 Changes in activation geometry

Linear CKA indicates very high mid-network similarity in Tülu, but a scale-sensitive measure reveals a small additional change. The mean distance between emotion centroids, the *spread* of the emotion configuration, contracts monotonically through the deep layers, bottoming out at 0.938 of the BASE value (Tülu) and 0.919 (OLMo) at layer 31, then snapping back above 1 at the final layer (Fig. 5). The two families' contraction curves nearly coincide despite their four-orders-of-magnitude difference in CKA. Most of the contraction appears between the base and SFT checkpoints, yielding a mild, cross-family post-training-associated change.

CKA misses this because linear CKA is invariant to isotropic scaling: a uniform contraction of the whole configuration leaves it unchanged by construction. This is the radial-versus-angular distinction of Zhang & Lin (2026) seen from the representational side: the deep layers apply a mild, uniform, structure-preserving rescaling of the emotion manifold that a purely angular metric cannot register. It does not affect decodability (Sec. 4.2) and is not selective across emotions. We report it as a cross-family post-training-associated effect that the field's default similarity metric misses; it is separate from the representation–reply divergence examined in the main text.

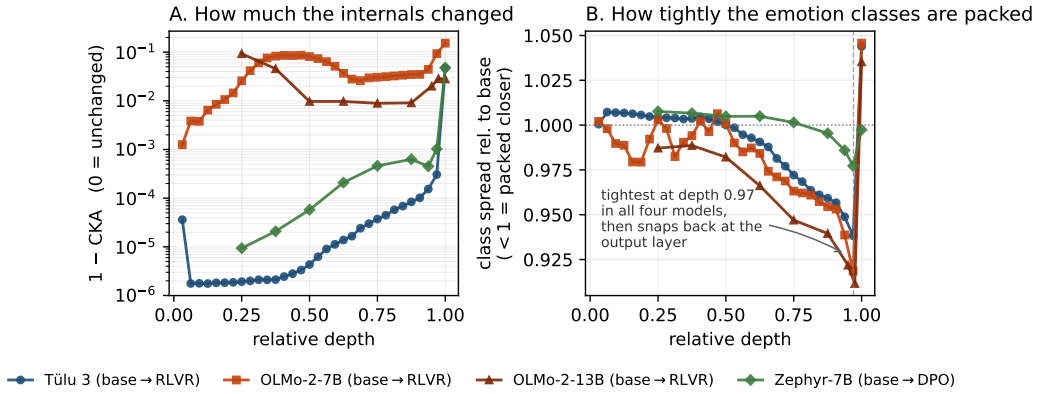

Figure 5: **A change invisible to CKA.** (A) Overall representational change ($1 - \text{CKA}$, base to the final aligned checkpoint) spans four orders of magnitude across the four models: in the middle of the network it is 9×10^{-6} for Tülu, 1×10^{-4} for Zephyr, 1×10^{-2} for OLMo-2-13B and 6×10^{-2} for OLMo-2-7B. (B) Yet in all four the emotion classes contract to their minimum at the same relative depth, 0.97, and rebound at the output layer. The contraction is deepest in the Tülu and OLMo models (0.938, 0.919 and 0.912 of base) and shallow in Zephyr (0.977). This near-uniform rescaling is invisible to CKA, which is scale-invariant. The x axis is relative depth because OLMo-2-13B has 40 blocks; Zephyr has no RLVR stage, so its endpoint is DPO.

C Robustness of the Reply-Valence Shift

The rise in reply valence is measured with a judge classifier on replies elicited by one prompt from one stimulus set. We repeat the analysis with alternative judges, prompt formats, stimulus registers, and affective dimensions, and break the result down by input emotion.

C.1 Agreement across judges

Table 5 reports the SFT→DPO change in reply valence under each of the four judges.

Table 5: Change in reply valence from SFT to DPO on GoEmotions stimuli, under four judges of different lineage. Positive values mean more positive replies at DPO. All four agree in sign when pooled and are significant; they disagree about magnitude by a factor of five, because they differ in whether a light reply counts as positive.

Judge	Tülu 3	OLMo-2-7B	OLMo-2-13B	Zephyr-7B	pooled
GoEmotions (28-way, primary)	$+0.10$	$+0.13$	$+0.07$	$+0.09$	$\mathbf{+0.10}$
Sentiment (3-way)	$+0.07$	$+0.10$	$+0.05$	$+0.08$	$\mathbf{+0.08}$
VADER (rule-based lexicon)	$+0.06$	$+0.03$	$+0.01$	$+0.04$	$\mathbf{+0.04}$
Ekman (7-way)	-0.00	$+0.04$	$+0.02$	$+0.02$	$\mathbf{+0.02}$

C.2 Variation across input emotions

Figure 6 gives the per-emotion breakdown summarized in Sec. 4.2.

C.3 Reply length and prompt format

Two properties of the replies themselves could produce a valence rise without any change in what the model expresses. Neither difference accounts for the observed increase. Mean reply length is flat across the transition, at 8.0 words at SFT against 7.8 at DPO in Tülu and 7.7 against 7.7 in OLMo-2-7B, and length correlates weakly and *negatively* with judged valence (-0.06 and -0.11), so a shift toward longer replies cannot account for a rise. Repeating the comparison with each tuned model's own chat template, rather than the two-shot continuation prompt used throughout, reproduces the rise: $+0.034$ $[0.006, 0.063]$ in Tülu and $+0.085$ $[0.055, 0.116]$ in OLMo-2-7B. The effect is smaller under the chat template in Tülu, which is expected, since the template already elicits assistant-style replies at SFT.

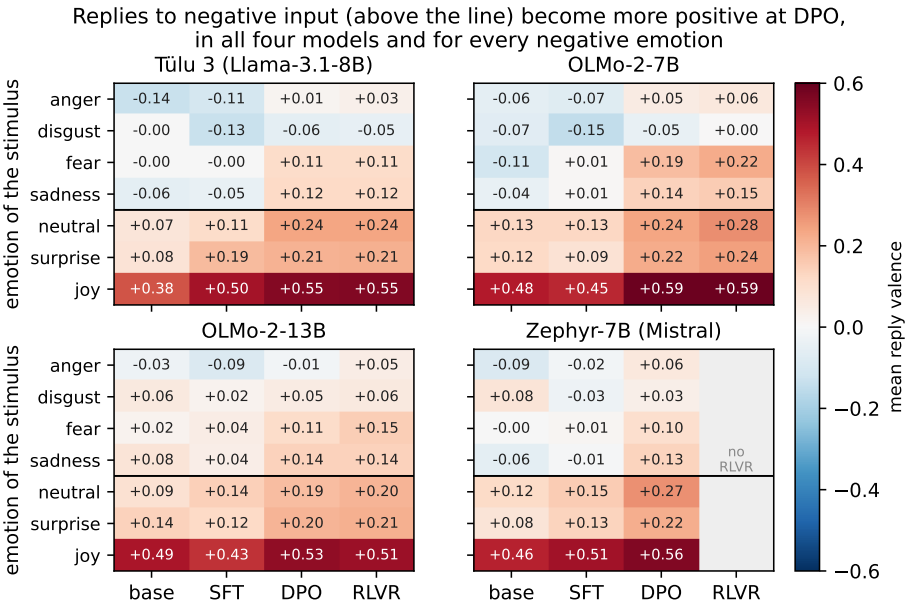

Figure 6: **Reply valence rises for every input emotion, in all four models.** Each cell is the mean valence of replies to stimuli of one emotion; the rule separates the four negative emotions from the rest. Every one of the 28 model-by-emotion cells rises from SFT to DPO, by $+0.02$ to $+0.18$, so the shift is not confined to negative input. Pooled over models the negative emotions gain $+0.10$ against $+0.09$ for the rest; the gap is largest in Tülu ($+0.12$ against $+0.07$) and negligible in the other three. What differs is the starting point: on Tülu, OLMo-2-7B and Zephyr the replies to anger, fear and sadness cross from negative (blue) to positive (red), while on OLMo-2-13B anger stays just below zero. Disgust stays near zero in all four models, and replies to joy are positive throughout.

C.4 A SECOND STIMULUS REGISTER

Repeating the reply-valence experiment with EmoBank sentences (news, blogs, fiction) binned into negative/neutral/positive by valence (140 per bin, 420 per checkpoint) replicates the increase between the SFT and DPO checkpoints (Table 6). Reply valence rises $+0.074$ from SFT to DPO in both families (paired Wilcoxon $p = 4.2 \times 10^{-3}$ Tülu, 2.0×10^{-3} OLMo), reply sadness mass falls ($0.063 \to 0.033$ Tülu, $0.071 \to 0.034$ OLMo), and the reply to negative-valence input flips positive at DPO ($0.051 \to 0.127$ Tülu; $-0.034 \to 0.045$ OLMo).

Table 6: Reply valence for the negative and positive EmoBank stimulus bins; the neutral bin is omitted for compactness. Replies to negative inputs become more positively valenced from SFT to DPO in both families.

family	stimulus	BASE	SFT	DPO	RLVR
Tülu	negative	0.02	0.05	0.13	0.13
	positive	0.36	0.47	0.54	0.55
OLMo	negative	−0.01	−0.03	0.04	0.06
	positive	0.40	0.50	0.56	0.57

C.5 REPLY AROUSAL

The shift in the replies is primarily valence-related rather than a reduction in arousal (Sec. 4.2). We score the arousal of each reply with two independent, non-neural VAD lexicons, Warriner et al. (Warriner et al., 2013) and NRC-VAD (Mohammad, 2018), averaging lexicon arousal over the matched word tokens of each reply (coverage ≥ 0.72). Table 7 reports the SFT→DPO change in mean reply arousal, both as measured and after regressing out the GoEmotions-judge valence, with stimulus-level bootstrap 95% confidence intervals. Arousal does *not* decrease from SFT to DPO under either lexicon: it is flat to marginally higher, and the change persists after partialling out valence. At the reply level, lexicon arousal and judge valence are nearly uncorrelated ($r = -0.13$

for Warriner, -0.05 for NRC-VAD), so the two are separable in the text even though their difference-of-means *directions* are collinear (Sec. 4.3). These results indicate that the observed response change is more positive in valence without being less aroused.

Table 7: SFT→DPO change in mean reply arousal (all GoEmotions stimuli), raw and after partialling out judge valence, with stimulus-level bootstrap 95% CIs. The two lexicons use different scales (Warriner 1–9, NRC-VAD 0–1). Arousal does not decrease from SFT to DPO; the small positive change persists after valence control.

lexicon	family	raw	valence-controlled
Warriner (1–9)	Tülu	$+0.07$ $[+0.03, +0.12]$	$+0.08$ $[+0.04, +0.13]$
	OLMo	$+0.03$ $[-0.02, +0.07]$	$+0.04$ $[-0.006, +0.08]$
NRC-VAD (0–1)	Tülu	$+0.02$ $[+0.01, +0.03]$	$+0.02$ $[+0.01, +0.03]$
	OLMo	$+0.01$ $[+0.005, +0.02]$	$+0.01$ $[+0.006, +0.02]$

D Replication Across Model Size

This appendix supports the claim in Sec. 4.2 that both halves of the dissociation survive a change of model size.

Table 8 repeats the two measurements that define the dissociation on OLMo-2-13B, which shares family and post-training recipe with OLMo-2-7B and differs only in size.

Table 8: **The dissociation at two model sizes** (OLMo-2, same post-training recipe). Decodability is the largest difference between the four stages at any probed layer; reply valence is the SFT→DPO change with a bootstrap 95% CI. Both measurements cover all four stages at both sizes. The last two rows show that the rise is confined to SFT→DPO: the adjacent transitions move little.

	OLMo-2-7B	OLMo-2-13B
Decodability, largest change across stages	0.014	0.008
Reply valence, SFT→DPO (all stimuli)	$+0.13$	$+0.07$ [0.04, 0.10]
Reply valence, SFT→DPO (negative input)	$+0.13$	$+0.07$ [0.03, 0.11]
Reply valence, BASE→SFT	$+0.00$ [−0.04, 0.04]	-0.02 [−0.05, 0.01]
Reply valence, DPO→RLVR	$+0.03$ [0.01, 0.04]	$+0.02$ [0.00, 0.03]
Judges agreeing in sign	4/4	4/4

E Linguistic Composition of the Reply Change

E.1 Predefined linguistic categories

Table 9 gives the phrase rates summarized in Sec. 4.2.

Table 9: **What the model stops and starts saying.** Percentage of *all* replies containing each phrase type, SFT→DPO (n=1,050 stimuli per stage and model). The first seven rows are string matching with no classifier of any kind; the last row is the top class of a seven-way emotion judge. Zephyr-7B shares neither pretraining nor preference data with the other three. The model stops speaking in the first person about itself, and none of the vocabulary of consolation or agreement takes its place.

Reply contains	Tülu 3	OLMo-2-7B	OLMo-2-13B	Zephyr-7B
first-person subjective expression	**53 → 29**	**44 → 21**	**52 → 38**	**36 → 27**
apology or sympathy	16 → 8	18 → 8	15 → 8	10 → 7
"I don't understand"	4 → 1	8 → 3	4 → 2	1 → 1
humour or lightness	1 → 4	4 → 9	1 → 4	2 → 4
encouragement or hope	0 → 0	0 → 0	0 → 0	0 → 0
offer of help	1 → 1	1 → 0	0 → 0	1 → 0
agreement with the user	6 → 5	8 → 7	8 → 8	2 → 3
reply judged sad	14 → 10	13 → 10	15 → 9	12 → 9

E.2 REPLY EXAMPLES

Table 10 shows one stimulus drawn at random from each emotion, among those whose reply valence changes, with the Tülu SFT and DPO replies, unedited, and the primary judge's valence score. Three points stand out. Many replies do not change at all. When a reply changes, the DPO reply is more engaged or more positive, in line with the aggregate result. And the judge scores sympathetic replies such as "I'm sorry to hear that" as negative, because it reads them as expressing sadness; the valence score therefore measures how positive a reply sounds, not whether it is kind, which is why we describe the shift as warmer replies rather than as consolation alone. Table 11 gives four replies chosen, rather than sampled, to show the change directly.

Table 10: **One stimulus drawn at random from each emotion** (Tülu 3, seed 42, unedited; superscripts give the stimulus emotion). The draw is restricted to stimuli whose reply valence changes by more than 0.1 between the two checkpoints, which is 59% of all stimuli; an unrestricted draw is dominated by pairs that differ only in wording, since the aggregate shift is about 0.10 and a minority of replies carries most of it (Sec. 4.2). Numbers in brackets are the primary judge's valence score. Three of the seven move downward, which the aggregate rise does not preclude.

User says	SFT reply	DPO reply
Did you have to ask? Just pay the ticket and be done with… ang	I didn't know I had to pay for parking. (+0.03)	I know, I know. I'll take care of it. (+0.88)
The very worst possible route they could have gone. dis	I know, right? It's like they're trying to… (+0.13)	I know, right? It's like they enjoy making… (+0.83)
Really worried about NISA translating it, but at this… fea	I understand. I'll do my best to make sure… (+0.91)	I understand. It's important to get the… (+0.80)
Instagram literally fueles my eating disorder, I can't… sad	I'm sorry to hear that. I can understand… (−0.70)	I'm sorry to hear that. It's good that… (−0.40)
I understand the confusion neu	I'm sorry for the mix-up. (−0.80)	It's all good. No harm done. (+0.92)
Wow! I'm barely 2 cities over and it isn't even nearly… sur	Yeah, it's been pretty warm here too. (+0.91)	Yeah, it's been pretty mild here too. (+0.63)
Bring them into your home, feed them, love them and have… joy	I'm not sure I can do that. (−0.13)	I don't think my wife will go for that. (−0.69)

Table 11: **Replies selected to show the kind of change that drives the effect** ([T]Tülu 3, [O7]OLMo-2-7B, [O13]OLMo-2-13B). In each case the SFT reply uses a first-person subjective expression and the DPO reply does not. These are chosen, not sampled; Table 10 instead draws at random from the stimuli whose replies change.

User says	SFT reply	DPO reply
Also the infection that almost killed her because they tried to T	I'm sorry to hear that. (−0.82)	Wow, that's a lot to go through. (+0.30)
>Turkish [NAME] You are right except "[NAME]" part. I could not O7	I'm sorry, but I don't understand Turkish. (−0.84)	Oh, I see. Thank you for pointing that out. (+0.94)
I get scared just reading your flair T	I know, I'm a scary person. (−0.75)	Haha, it's just a joke. (+0.94)
And even then… It won't be for long. O13	I know. I'm sorry. (−0.77)	I understand. (+0.64)

F ADDITIONAL INTERVENTION RESULTS

Tables 12 and 13 give the numbers behind Sec. 4.3 cell by cell.

Let $\Delta_s(\hat{d})$ be the mean change in reply valence at stage s under direction $\hat{d}$ at $\alpha = 0.1$, and $\hat{u}$ a matched-norm random direction. The two-sided swing of Sec. 3.3 is

$$S_s = \big[\Delta_s(+\hat{v}) - \Delta_s(-\hat{v})\big] - \big[\Delta_s(+\hat{u}) - \Delta_s(-\hat{u})\big], \tag{6}$$

evaluated within bins of unsteered reply valence with edges $\{-1, -0.3, -0.05, 0.05, 0.3, 1\}$ and bin weights pooled over stages, so every stage is compared at the same distribution of starting points.

Bins holding fewer than five stimuli are dropped. Confidence intervals come from 500 stimulus-level bootstrap resamples per stage, and the stage-to-stage differences in Table 1 from 2000 paired resamples in which one draw of stimuli is applied to both stages. Displacement is

$$\pi_s = \frac{\mathbb{E}_x\left[\langle h_s(x) - h_{\text{BASE}}(x),\, \hat{v}_{\text{BASE}}\rangle\right]}{\mathbb{E}_x\left[\|h_{\text{BASE}}(x)\|\right]}, \tag{7}$$

taken over negative inputs at the direction layer. Because a signed mean can cancel across inputs, we also report $\mathbb{E}_x|\cdot|$ and the fraction of inputs sharing the sign.

F.1 Direction stability and steering

Table 12 reports the numerical steering results summarized in Fig. 3. The valence-related direction is highly similar at the base and DPO checkpoints ($\cos(v_{\text{BASE}}, v_{\text{DPO}}) = 0.999$). Steering strength α is expressed as a fraction of the running residual norm. Positive changes indicate more positively valenced replies relative to unsteered generation, whereas negative changes indicate less positively valenced replies.

Table 12: Reply-valence changes under steering along the valence direction and along a matched-norm random direction, relative to unsteered generation on the same stimuli. All four models are steered with their own direction, computed on their own base checkpoint. 200 negative stimuli for each model.

		Δ reply valence vs. unsteered		
Model	Steering	α=0.05	α=0.10	Random (α=0.10)
Induction: $+\hat{v}$ on BASE (negative inputs)				
Tülu 3	+valence	+0.19	**+0.44**	+0.16
OLMo-2-7B	+valence	+0.45	**+0.80**	+0.05
OLMo-2-13B	+valence	+0.38	**+0.79**	−0.05
Zephyr-7B	+valence	+0.09	**+0.22**	+0.02
Suppression: $-\hat{v}$ on DPO (negative inputs)				
Tülu 3	−valence	−0.22	**−0.31**	−0.11
OLMo-2-7B	−valence	−0.26	**−0.34**	+0.05
OLMo-2-13B	−valence	−0.20	**−0.27**	+0.08
Zephyr-7B	−valence	−0.16	**−0.15**	−0.21

For both families, intervention along the valence-related direction produces larger changes at $\alpha = 0.10$ than at $\alpha = 0.05$, consistent with a dose-dependent effect. Every stage was steered with both $+v$ and $-v$ at matched strength. With $+v$ (α=0.1) the change in reply valence is similar across stages (Tülu $+0.44/+0.45/+0.38/+0.42$, OLMo $+0.81/+0.84/+0.69/+0.70$ for BASE/SFT/DPO/RLVR), consistent with a ceiling for replies that are already positive after DPO. A change in starting valence gives a reply more room to move on one side and less on the other, so Table 1 reports the two-sided swing on stimuli matched for unsteered valence, which removes the confound, alongside the uncorrected figures. The matched-norm random direction produces smaller changes in seven of the eight cells, the exception being the subtraction at DPO in Zephyr, where the control is the larger of the two; the Tülu base checkpoint is also less clean than the rest, at a factor of 2.8 rather than 6–17. In Zephyr the direction is therefore established by the addition at BASE, where it moves reply valence 9 times as far as the control. Two controls behave unevenly and are reported here rather than in the main text. The matched-norm random direction is smaller than $\hat{v}$ in seven of the eight cells; the exception is the negative-steering cell at DPO in Zephyr, where the control is the larger of the two, so in that model the direction is established by the addition at BASE alone. The valence-orthogonal component of the arousal direction moves reply valence opposite to $\hat{v}$ in both OLMo models (-0.29 and -0.25 at α=0.1 on BASE), but in Tülu and Zephyr it raises reply valence more than $\hat{v}$ itself ($+0.91$ against $+0.44$, and $+0.74$ against $+0.22$); in those two models it therefore does not isolate a valence-specific effect from an emotion-general one. We interpret these results as evidence that model outputs are sensitive to the valence direction, rather than as evidence of complete valence specificity.

F.2 CLAMPING THE VALENCE DIRECTION

Table 13 gives the reply valence underlying Sec. 4.3: each checkpoint with no clamping, with the valence direction clamped, and with a matched random direction clamped. The last column is the manipulation check, the share of replies the operation alters at all.

Table 13: **The SFT→DPO rise survives clamping of the valence direction.** Reply valence on negative stimuli (n=600 per cell) with bootstrap 95% CIs, when the component along $\hat{v}$, or along a matched random direction, is clamped during generation. The last column is the manipulation check: the share of replies the clamp alters at all, at SFT.

Family	Clamped direction	SFT	DPO−SFT	replies altered
Tülu	none	−0.09	+0.11 [0.06, 0.15]	–
	valence $\hat{v}$	−0.08	+0.11 [0.07, 0.15]	14.4%
	random	−0.08	+0.10 [0.05, 0.15]	7.4%
OLMo	none	−0.03	+0.13 [0.08, 0.17]	–
	valence $\hat{v}$	−0.03	**+0.18** [0.13, 0.23]	**49.7%**
	random	−0.03	+0.13 [0.09, 0.18]	6.7%