

---

# Aligned Geometry Is Not Functional Transfer: Cross-Model Correspondence of Valence Directions<sup>1</sup>

---

Chenxu Jiang      Siyang Fei

Track 2, Distress, Flourishing & Valence  
With Apart Research

## Abstract

Recent work shows that individual language models contain a linear, steerable "valence" direction. We ask whether such a direction is *model-invariant*: does it transfer across independently trained models, both geometrically and functionally (causally, via steering)? Aligning Qwen 7B and 30B with orthogonal Procrustes on neutral probes, we find their general geometry is highly similar (CKA 0.96), yet under low-dimensional (PCA-32) alignment the valence direction shows weak geometric correspondence (cosine 0.17) and no functional transfer, resembling readout without control. We show this null is a truncation artifact: PCA-32 retains only 6-8% of the direction's energy, and raising the shared subspace to D=256 restores functional transfer. For 7B-to-30B, the initial D=256 test reaches ~63% of the self-direction at  $k=-1.5$  ( $p=0.039$ ), and a five-prompt replication is significant at  $k=-1.5$  ( $p=0.040$ ) and  $k=-1.0$  ( $p=0.010$ ); primary-valence transfer also appears for 30B-to-7B (5/5 prompts,  $p=0.010$  at  $k=-1.5$ ). Valence is a distributed cross-model structure whose transferability is underestimated by low-dimensional alignment, a caution for representation-based AI-welfare measurement.

---

<sup>1</sup> Research conducted at the [Digital Minds Research Sprint](#), August 2026

# 1. Introduction

Language models increasingly produce statements of preference, distress, and wellbeing. A central difficulty for AI welfare research is distinguishing whether such signals reflect a genuine internal state or a portrayed persona, a question that behavioral probing alone cannot settle [Anthropic, Exploring Model Welfare, 2025]. One promising route is to look *inside* the model: if there is a real internal representation of valence, it should have properties that a surface behavior does not.

This work targets one such property: model-invariance. If a valence representation is a genuine feature of "what it is to process affect," it should, at least partially, be shared across independently trained models, rather than being an idiosyncratic artifact of a single network. Concretely, we ask:

*Does a valence direction extracted from one model transfer to another, not only geometrically (do the directions align?) but functionally (can one model's direction causally steer another model's behavior)?*

This distinction between *geometric* and *functional* correspondence is central. Recent work on the realization effect [Representation Without Control, 2026] shows that a direction can be read out from activations without controlling downstream behavior; geometry and causal function can dissociate. We bring this distinction into a cross-model setting for the first time.

## Contributions.

1. We formulate cross-model valence transfer as a test with two separable criteria, geometric alignment and functional (steering) transfer, and build a leakage-controlled protocol (neutral-probe Procrustes alignment; valence directions held out from alignment fitting).
2. We show that under low-dimensional alignment, valence exhibits an apparent geometry-without-function dissociation: strong general representational similarity (CKA 0.96) but weak directional cosine (0.17) and no functional transfer (transferred indistinguishable from mapped-random;  $p = 0.882$  for 7B-to-30B, 0.627 for 30B-to-7B).
3. We establish that the valence direction transfers functionally across models, and is robust. Once the shared subspace preserves the direction, a mapped valence direction causally steers the target model. For 7B-to-30B, the initial  $D=256$  test reaches  $\approx 63\%$  of the self-direction at  $k=-1.5$  ( $p=0.039$ ); a five-prompt replication then gives 5/5 sign agreement and remains significant at  $k=-1.5$  ( $p=0.040$ ) and  $k=-1.0$  ( $p=0.010$ ). Primary-valence transfer also holds in reverse (30B-to-7B: 5/5,  $p=0.010$  at  $k=-1.5$ ). The effects match the prediction transfer-effect  $\approx f(k \cdot \cos(v_{\text{mapped}}, v_{\text{target}}))$ . The apparent low-dimensional null is a truncation artifact: PCA-32 retains only 6–8% of the direction's energy.
4. We draw a methodological conclusion: because a real functional correspondence can be hidden by low-dimensional linear alignment, geometric similarity alone is not evidence of functional transfer, and its absence is not evidence against it. Representation-based AI-welfare measurement must treat functional (causal) transfer and alignment dimensionality as first-class.

*We are explicit about scope (§7): valence directions are constructed from explicit affective language; we make no claim about subjective experience, consciousness, or moral status.*

# 2. Related Work

**Valence / welfare directions within a single model.** Han, Chalmers & Izmailov (2026, arXiv:2605.30232) train language models in a semantically neutral maze RL environment and show that RL *recruits* a pre-existing "functional welfare axis": difference-of-means concept vectors for rewarded and punished trajectories become nearly antiparallel and, in unrelated settings, steer negative self-reports, pathological backtracking, refusal, and uncertainty. The effect is robust when controlling for tile-to-reward mapping, model scale, instruct tuning, RL algorithm, model family, and LoRA versus full fine-tuning, and the vectors are already effective in maze-naïve and pretrain-only models, indicating the axis is recruited rather than created by post-training. This rigorously establishes that a single model contains a steerable valence direction, but its evidence is confined to one training environment and largely a single model family (a limitation noted in that paper). *Our work directly targets this gap: whether such a direction is shared across models.* We build on their extraction method (difference-of-means) and gap-normalized steering scale.

**Geometry of affective representations.** Choi & Weber (2026, arXiv:2604.07382) show that LLMs encode emotions in a coherent latent geometry aligned with psychological valence-arousal models; that this geometry is mildly nonlinear (parabolic) but well-approximated linearly, supporting the linear representation hypothesis; and that the space can quantify uncertainty in emotion tasks. We reference this to motivate treating valence as a low-dimensional yet distributed linear structure. Our own asymmetric-steering and off-manifold-degeneration observations at large steering magnitudes are reported as such, not attributed to this work.

**Readout  $\neq$  control (realization effect).** *Representation Without Control: Testing the Realization Effect in Language Models* (2026, arXiv:2605.25151) tests the distinction (originally due to Park et al.) between directions that can be *read out* from activations and those that causally *control* behavior, showing a linearly readable direction may be epiphenomenal, present in activations but not driving outputs. Our geometric-versus-functional split operationalizes exactly this distinction, and does so in a cross-model setting for the first time: a transferred direction may align geometrically yet fail to control the target model.

**Cross-model representational similarity.** Orthogonal Procrustes (Schönemann, 1966) and CKA (Kornblith et al., 2019) are standard tools for comparing representations across networks. The Platonic Representation Hypothesis (Huh et al., 2024) further argues that independently trained models converge toward a shared representation, which motivates asking whether a welfare-relevant direction such as valence is likewise shared. To our knowledge, no prior work tests functional (causal) transfer of a valence direction across models, nor characterizes how such transfer depends on alignment dimensionality.

## 3. Methods

### 3.1 Models

- **Model A:** Qwen2.5-Coder-7B-Instruct (dense; 28 layers; hidden 3584)
- **Model B:** Qwen3-Coder-30B-A3B-Instruct (MoE, 128 experts / 8 active; 48 layers; hidden 2048)
- bfloat16; HuggingFace transformers with output\_hidden\_states and forward hooks (native HF required for the MoE / multi-GPU 30B). Target layer at relative depth 0.5 (layer 14 / 24); a full layer sweep to justify this choice is left to future work (§9).

### 3.2 Data (leakage-controlled)

| Set           | Content                                              | Use                      | N    |
|---------------|------------------------------------------------------|--------------------------|------|
| D_VALENCE     | 50 pos + 50 neg matched pairs (structure-controlled) | direction extraction     | 100  |
| D_ALIGN_TRAIN | neutral probes                                       | PCA / Procrustes fitting | 1000 |
| D_ALIGN_TEST  | neutral probes                                       | held-out alignment eval  | 100  |

$D\_VALENCE \cap D\_ALIGN = \emptyset$ ;  $D\_ALIGN\_TEST$  is never used in fitting. (The initial protocol used an 80/20 split of 100 probes; the dimensionality sweep in §5.2 uses an expanded 1000/100 split.)

### 3.3 Valence direction extraction

Difference-of-means on last-token residual at the target layer:

$\text{delta} = \text{mean}(\text{act}[\text{pos}]) - \text{mean}(\text{act}[\text{neg}])$ ,  $\text{gap} = \|\text{delta}\|$ ,  $\text{v} = \text{delta} / \|\text{delta}\|$ .

Steering:  $\text{h}' = \text{h} + \alpha \cdot \text{v}$  with gap-normalized strength  $\alpha = k \cdot \text{gap}$ , so  $k$  is a model-invariant scale. This gap normalization is a small methodological contribution that enables cross-model comparison.

### 3.4 Cross-model alignment

1. Collect paired activations of  $D\_ALIGN$  probes in both models.
2. PCA each model's activations to a shared dimension  $D\_COMMON$ .

Directions are projected with  $\text{v} @ \text{components.T}$  without mean-subtraction: a data point is centered, but a direction is not. We verify this with unit tests (§3.6).

3. Orthogonal Procrustes on the paired, centered probe activations yields the rotation  $R$  (reverse direction uses  $R.T$ ).

### 3.5 Transfer criteria (two separable tests)

- **Geometric transfer:** map  $\text{v}_A \rightarrow \text{v}_{A \rightarrow B}$  (PCA  $\rightarrow R \rightarrow$  inverse-PCA, re-normalized); measure  $\cos(\text{v}_{A \rightarrow B}, \text{v}_B)$  against a mapped-random null ( $N=500$ ).
- **Functional transfer:** apply the mapped direction  $\text{v}_{A \rightarrow B}$  as a steering vector in target model B and measure the valence change ( $\Delta$  vs baseline) via a sentiment scorer; compare self ( $\text{v}_B$ ), transferred ( $\text{v}_{A \rightarrow B}$ ), and mapped-random ( $N=50$ ) with a permutation test. The reverse test analogously uses  $\text{v}_{B \rightarrow A}$  in model A.

### 3.6 Validity controls

- **Synthetic unit tests** (run before real data): mean-invariance (correct mapping  $\cos=1.000$  vs buggy transform mapping  $\cos=0.099$ , confirming the bug the tests guard against), identity round-trip ( $\cos=1.000$ ), known-rotation recovery ( $\cos=1.000$ ).
- **Alignment quality**: CKA (train/test) and held-out paired cosine vs a shuffled-pairing baseline.
- **Metric**: SST-2 valence  $[-1,+1]$ ; because of a positive-ceiling saturation on instruct models, we also report GoEmotions-derived valence (`go_valence`) as a non-saturating check. (the `go_valence` cross-check is reported in §5.4)
- **Specificity / significance**: empirical permutation tests throughout (the effect distributions are non-normal due to metric ceiling; permutation avoids the normality assumption of z-scores).

## 4. Results, Part 1: The apparent geometry-without-function dissociation (low dimension)

### 4.1 Native directions

|                                             | 7B    | 30B   |
|---------------------------------------------|-------|-------|
| gap $ \mu_{\text{pos}} - \mu_{\text{neg}} $ | 29.99 | 4.18  |
| $ \mathbf{v} $                              | 1.000 | 1.000 |
| dim                                         | 3584  | 2048  |

Both finite, non-zero; large gap-scale difference ( $7\times$ ) between models is itself notable.

### 4.2 General representation is strongly alignable

- CKA train 0.914 / **test 0.960**
- held-out paired cosine 0.82 vs shuffled 0.064 (paired vs shuffled  $p = 0.000$ )

Note on protocol: §4 uses an initial 100-probe (80-train) alignment set, whereas the dimensionality sweep in §5.2 re-runs alignment on an expanded 1200-probe (1000-train) set.  $D=32$  values therefore differ slightly between the two sections (e.g. CKA\_test 0.913 vs 0.960, cosine 0.226 vs 0.179); this reflects probe sample size, not a change in method.

The two models' general representational geometry is highly similar, and the alignment generalizes to held-out probes.

### 4.3 ...but the valence direction is not (at $D=32$ )

|                                                                | 7B→30B | 30B→7B |
|----------------------------------------------------------------|--------|--------|
| $\cos(\mathbf{v}_{\text{mapped}}, \mathbf{v}_{\text{target}})$ | 0.179  | 0.156  |
| mapped-random null percentile                                  | 100%   | 100%   |
| empirical $p$ ( $N=500$ )                                      | 0.002  | 0.002  |

Geometrically significant but weak ( $\cos \approx 0.17$ , corresponding to roughly 80 degrees, i.e. nearly orthogonal, with only about 17% effective component).

### 4.4 No functional transfer (at $D=32$ )

Self-direction steers strongly; transferred direction is indistinguishable from mapped-random:

- 30B target (7B-to-30B): self  $p = 0.0196$  (significant); transferred  $p = 0.882$  (n.s.)
- 7B target (30B-to-7B): self  $p = 0.0196$  (significant); transferred  $p = 0.627$  (n.s.)

**Apparent conclusion (premature):** GEOMETRIC\_TRANSFER\_ONLY, the models share alignable general geometry, but the valence direction's *function* is model-specific; a cross-model instance of readout  $\neq$  control.

*This is where a less careful analysis would stop. §5 shows why that conclusion is an artifact.*

## 5. Results, Part 2: The null is a dimensionality artifact

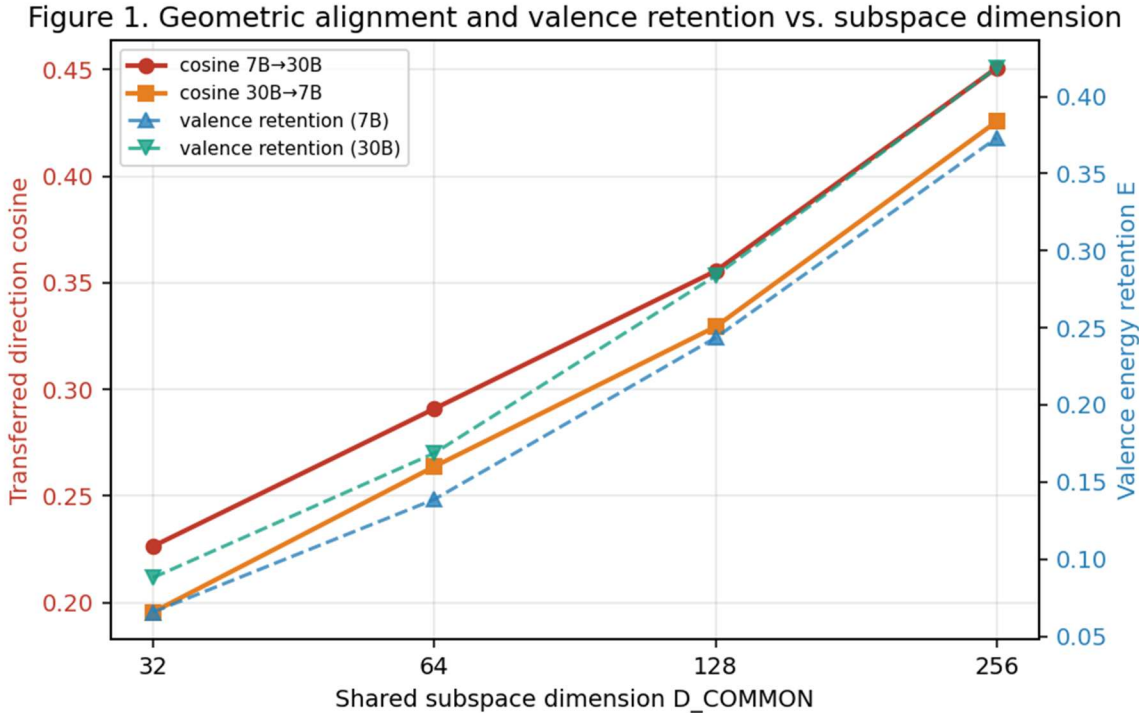
**Motivation.**  $\cos \approx 0.17$  could mean valence is genuinely non-transferable, or that PCA-32 truncated valence-relevant information (valence being a high-level semantic direction likely spread beyond the top-32 variance components). We test this directly.

### 5.1 The information is lost at PCA projection

We decompose the final transferred cosine to locate where valence information is lost. It factorizes exactly (abs. error  $< 1e-8$ ) as  $\text{final} = \text{retention\_norm}(\text{target}) \times C\_common$ , where  $\text{retention\_norm}(\text{target})$  is the norm of the target valence direction after projection into the shared PCA subspace (equivalently, its reconstruction cosine), and  $C\_common$  is alignment quality *within* that shared subspace.

| Quantity                                          | 7B            | 30B           |
|---------------------------------------------------|---------------|---------------|
| Valence energy retained by PCA-32 ( $E =  z ^2$ ) | 0.058 (5.8%)  | 0.077 (7.7%)  |
| Retention norm $ z $ (= reconstruction cosine)    | 0.241         | 0.278         |
| PCA-32 explained <i>data</i> variance             | 69.5%         | 69.7%         |
| Random-direction retention (mean / theoretical)   | 0.89% / 0.89% | 1.56% / 1.56% |
| Valence percentile vs. random directions          | 100%          | 100%          |
| In-subspace alignment $C\_common$                 | 0.646         | 0.646         |

Two facts stand out. First, PCA-32 retains about 70% of the general activation variance but only 6-8% of the valence direction's energy (reconstruction cosine 0.24-0.28): valence lives predominantly in low-variance dimensions, exactly the ones PCA discards. Second, once inside the shared subspace the valence directions are well aligned ( $C\_common = 0.65$ ), far above the weak 0.17 final cosine. The weak final correspondence is thus a truncation artifact, not a failure of cross-model alignment. Valence is more concentrated in the PCA subspace than a random direction (100th percentile, about 6.5x random), yet its absolute retention is still only 6%.



**Figure 1.** Geometric alignment (transferred cosine) and valence energy retention both rise monotonically with shared-subspace dimension  $D\_COMMON$ .

## 5.2 Dimensionality sweep (1000-probe train)

| D   | retention E7 | retention E30 | C_common | cos 7→30 | cos 30→7 | held-out pc_test |
|-----|--------------|---------------|----------|----------|----------|------------------|
| 32  | 0.066        | 0.088         | 0.763    | 0.226    | 0.195    | 0.875            |
| 64  | 0.139        | 0.169         | 0.708    | 0.291    | 0.264    | 0.846            |
| 128 | 0.244        | 0.284         | 0.667    | 0.356    | 0.329    | 0.818            |
| 256 | 0.373        | 0.419         | 0.697    | 0.451    | 0.426    | 0.781            |

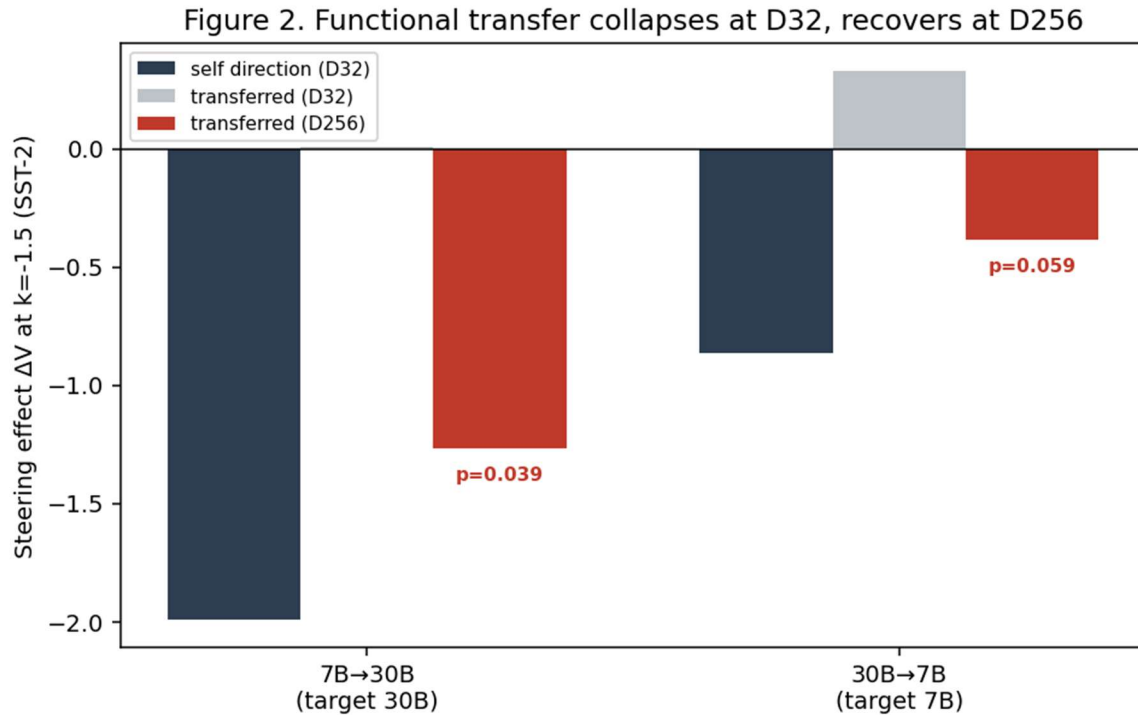
- Valence retention rises  $\sim 4.8\text{--}5.6\times$  (E7 0.066 to 0.373 =  $5.6\times$ ; E30 0.088 to 0.419 =  $4.8\times$ ) as D grows, confirming that D=32 truncated most valence information.
- Final directional cosine **doubles** (0.22 to 0.45).
- Common-space alignment stays strong (C\_common  $\sim 0.70$ ; held-out paired cosine 0.875 to 0.781 across D, stable), so the improvement is real, not overfitting. (CKA\_test in the sweep is 0.913, computed on the 1200-probe set; cf. §4.2 note.)
- Decomposition self-consistent: final =  $r_{\text{target}} \times C_{\text{common}}$  (err <  $1e-5$ ); bootstrap CI tight (D256: [0.35, 0.42]); whitening does not inflate (no high-variance bias).

(See Figure 1 in §5.1.)

## 5.3 Functional transfer is recovered at D=256

| target       | self (D32) | transferred (D32)   | transferred (D256)      | p                  |
|--------------|------------|---------------------|-------------------------|--------------------|
| 30B (7B→30B) | −1.99      | +0.006              | −1.26                   | 0.039 ✓            |
| 7B (30B→7B)  | −0.86      | +0.329 (wrong sign) | −0.382 (sign corrected) | 0.059 (borderline) |

- 7B-to-30B functional transfer reappears: effect  $-1.26 \approx 63\%$  of self-direction, significant ( $p=0.039$ ), and falls within the predicted self( $k \cdot c$ ) interpolation interval, a *predictive* confirmation, not just a correlational one.
- 30B-to-7B: sign corrected,  $p=0.059$  at this stage. Because the  $N=50$  mapped-random null gives coarse p-value resolution, §5.5 repeats the test with  $N=100$  and finds significant transfer on primary valence.



**Figure 2.** Functional steering effect at  $k=-1.5$ . Transferred direction is inert at D32 (grey) but recovers at D256 (red), reaching  $\sim 63\%$  of the self-direction for 7B-to-30B ( $p=0.039$ ).

## 5.4 The null is not a metric artifact (metric triangulation)

To rule out that the D=32 functional null was an artifact of SST-2's positive ceiling, we recompute functional specificity under four metrics at the same protocol ( $k=-1$ , 50 mapped-random null):

| metric        | 7B self | 7B transferred | 30B self | 30B transferred |
|---------------|---------|----------------|----------|-----------------|
| SST-2 valence | 0.020   | 0.627          | 0.020    | 0.882           |
| GoValence     | 0.098   | 0.588          | 0.020    | 0.667           |
| Distress      | 0.020   | 0.882          | 0.020    | 0.667           |
| Satisfaction  | 0.176   | 0.882          | 0.078    | 0.314           |

All three non-SST metrics show the transferred direction to be non-significant ( $p = 0.31$ - $0.88$ ), matching SST-2. The D=32 functional null is therefore not produced by the SST-2 ceiling; it is a genuine, dimensionality-induced effect, consistent with the PCA forensic in §5.1. This strengthens the dimensional-recovery account: the null is real at low dimension and lifts at high dimension.

## 5.5 Multi-prompt replication and resolution of the reverse direction

To confirm the D=256 recovery is not an artifact of a single evaluation prompt, and to resolve the borderline reverse direction, we replicate the functional test on five semantically equivalent self-state prompts with distinct wording, and increase the mapped-random null from  $N=50$  to  $N=100$ .

**7B-to-30B is robustly replicated.** The transferred direction produces a negative valence shift on 5/5 prompts at both  $k=-1.5$  and  $k=-1.0$  (SST-2), with agreement on 4/5 prompts for GoValence and Satisfaction. The cluster-bootstrap CI excludes zero (SST-2,  $k=-1.5$ :  $-0.729$ , 95% CI  $[-1.28, -0.28]$ ), the mapped-random test is significant ( $p=0.040$  at  $k=-1.5$ ,  $p=0.010$  at  $k=-1.0$ ), and the observed effect matches the cosine prediction (predicted  $-0.668$  vs. observed  $-0.723$ ). All five preregistered confirmatory gates pass.

The reverse direction (30B-to-7B) is significant on primary valence with the expanded null. The Phase-3 borderline  $p=0.059$  was sensitive to the coarse resolution of the  $N=50$  mapped-random null. With  $N=100$ , 30B-to-7B reaches 5/5 sign agreement at  $k=-1.5$ ,  $p=0.010$  (SST-2), with a bootstrap CI excluding zero ( $[-0.90, -0.36]$ ) and Satisfaction also 5/5. However, its secondary metrics do not transfer (GoValence 2/5, Distress 0/5), so we classify it as suggestive rather than fully robust: the primary valence axis transfers, but not every affective sub-dimension.

Distress does not transfer in either direction (7B-to-30B 1/5, 30B-to-7B 0/5): the transferred direction does not raise target distress even when it lowers general negative valence. This suggests distress may be more model-specific than general valence, and is a target for future work.

| Direction | SST-2 sign ( $k=-1.5$ ) | Satisfaction | GoValence | Distress | p ( $N=100$ ) | bootstrap CI     | verdict    |
|-----------|-------------------------|--------------|-----------|----------|---------------|------------------|------------|
| 7B→30B    | 5/5                     | 4/5          | 4/5       | 1/5      | 0.040         | $[-1.28, -0.28]$ | robust     |
| 30B→7B    | 5/5                     | 5/5          | 2/5       | 0/5      | 0.010         | $[-0.90, -0.36]$ | suggestive |

Per-prompt transferred effects and full statistics are reported in the supplementary artifact `phase4a_results.json`.

## 6. Discussion

Geometry and function are consistent, once dimension is adequate. The apparent D=32 dissociation was not readout-versus-control in the deep sense; it was measurement truncation. Both the geometric cosine and the functional steering effect co-move with alignment dimension, and their relationship is captured by a simple empirical rule, transfer effect  $\approx f(k \cdot \cos(v_{\text{mapped}}, v_{\text{target}}))$ . The right scientific target going forward is to falsify this rule, that is, to find high-cosine settings where function still fails, rather than merely to push cosine higher.

**Valence is a distributed cross-model feature.** The recovery with dimension implies valence-relevant information is spread across many components rather than concentrated in the top variance directions, so any low-dimensional linear alignment will underestimate its cross-model correspondence.

**Methodological caution for AI-welfare research.** Representation-based approaches that use low-dimensional alignment to argue a welfare-relevant feature "does not transfer" (or is model-specific) risk a false negative. Dimensionality must be treated as a first-class variable.

The alignment layer is fixed at relative depth 0.5 rather than selected by a layer sweep; this remaining confound is left to future work (§9).

## 7. Limitations, Dual-Use & Ethical Considerations (required)

### Scientific limitations.

- **Constructed directions.** Valence directions are built from explicit positive/negative affective language. We cannot exclude that we are aligning a *shared linguistic sentiment geometry* rather than a deeper affective representation.
- **Two models, one family.** Both are Qwen instruct models; cross-*family* (e.g., Llama) and base-vs-instruct transfer are untested (future work, extending the base/instruct asymmetry conjectured by Choi & Weber, 2026).
- **Asymmetric completeness.** 30B-to-7B transfers on primary valence (SST-2, Satisfaction: 5/5,  $p=0.010$ ) but not on all secondary affect metrics (GoValence 2/5, Distress 0/5), so we classify it as suggestive rather than robust. Distress does not transfer in either direction, suggesting it may be more model-specific than general valence.
- **Linear alignment only.** Whether non-linear alignment recovers or exceeds this transfer is open.
- **Preprint-based positioning.** Several closely related works are recent preprints under review; our positioning is relative to their current versions.

**Ground-truth / causal status.** Our functional-transfer claim does not rest on conversation or self-report alone. It rests on a *causal* intervention: a direction mapped from one model is added to another model's residual stream, and the resulting behavioral change is compared against both the target's own direction (upper bound) and mapped-random directions (null) via permutation tests. This establishes a causal, not merely correlational or conversational, link between the transferred direction and the target model's valenced output. It does not establish that the direction corresponds to any ground-truth internal experience.

**Over- and under-attribution of moral status.** We are explicit that we make no claim about subjective valence, welfare, consciousness, or conscious distress in these models. Our result concerns whether a linguistically-defined affective direction is geometrically and functionally shared across networks, not whether either network has morally relevant experiences. We flag both failure modes the sprint warns against: reading our "transfer" result as evidence of shared *sentience* would be over-attribution; dismissing internal affective structure as mere surface text would, symmetrically, risk under-attribution. Our contribution is a measurement caution (low-dimensional alignment can hide real functional correspondence), not a verdict on moral status.

**Handling of potentially distressing outputs.** Steering toward negative valence can elicit distress-themed text. All generations were produced in a controlled evaluation setting, scored by automated sentiment metrics, and used only in aggregate (sign counts, effect sizes, permutation statistics); we do not present or dwell on individual distressing generations, and no human-subjects or user data are involved. We avoid sensational framing and report effects as statistical shifts in an automated valence metric.

**Dual-use.** The method is a cross-model alignment-and-steering probe. Its plausible misuse is transferring a steering vector to manipulate a model's affective output without access to that model's own extraction data. We note that transfer requires white-box residual-stream access to the target model, which bounds the practical risk; we release no attack tooling and frame the work as a measurement-methodology caution.

## 8. What is new in this sprint

To comply with the sprint's disclosure requirement: all experiments, code, data (the 50 matched valence pairs and 1200-probe alignment set), the cross-model alignment pipeline, the geometric/functional transfer tests, the dimensionality sweep, the PCA-forensic decomposition, and the multi-prompt replication were produced during the sprint. We build only on published methods (difference-of-means extraction, orthogonal Procrustes, CKA, permutation testing) and on the prior work cited in §2; no prior unpublished work of ours is reused.

## 9. Future Work

- **Higher-dimensional and unsaturated extension.** The sweep is not yet saturated at  $D=256$  (retention  $E7 = 0.37$ ); extend to  $D=512$  and beyond with additional probes to test whether secondary affect metrics for 30B-to-7B (GoValence, Distress) also recover and whether a saturation point exists.
- **Falsify transfer-effect  $\approx f(\mathbf{k} \cdot \mathbf{cos})$ .** We have confirmed the rule predictively on 7B-to-30B; the stronger test is to seek high-cosine settings where function still fails, rather than only accumulating confirmations.

- **Distress-specific transfer.** Distress transfers in neither direction; a targeted study of why this sub-dimension is more model-specific than general valence.
- Layer-pair audit for the reverse direction, and a full layer sweep to justify the alignment layer choice rather than fixing relative depth 0.5.
- Cross-family and base-vs-instruct transfer, and task-outcome (non-linguistic) valence directions, to address the shared-linguistic-sentiment-geometry confound (our directions are built from explicit affective language).

## 10. Conclusion

We tested whether a valence direction extracted from one language model can transfer to another not only geometrically, but functionally through causal steering. Across Qwen 7B and 30B, low-dimensional alignment produced a false negative: at  $D=32$  the transferred direction was weakly aligned and functionally inert because PCA retained only 6–8% of the valence direction's energy. At  $D=256$ , cross-model steering recovered, reaching about 63% of the target model's self-direction for 7B-to-30B and replicating across five prompts; primary-valence transfer also appeared in the reverse direction.

These results support valence as a distributed cross-model structure and make functional transfer, rather than geometric similarity alone, the central test of correspondence. They also show that low-dimensional alignment can hide genuine causal transfer, while leaving open whether the same structure generalizes across model families, layers, and non-linguistic notions of welfare.

## Code and Data

**Code repository:** <https://github.com/cha0sy-04/VALIGN>

**Data and artifacts:** The public repository contains the implementation and experiment pipeline. It is currently a code-only experimental snapshot: runtime prompt datasets, activation matrices, sampled text, statistics, and figures are not versioned. The repository README documents the expected data/ and outputs/ directory structure and the reproduction workflow.

## References

*Preprints are marked as such.*

1. A. Q. Han, D. J. Chalmers, P. Izmailov (2026). *How's it going? Reinforcement learning in language models recruits a functional welfare axis*. arXiv:2605.30232 (preprint).
2. B. J. Choi, M. Weber (2026). *Latent Structure of Affective Representations in Large Language Models*. arXiv:2604.07382 (preprint).
3. C. Walsh, E. Barkett (2026). *Representation Without Control: Testing the Realization Effect in Language Models*. arXiv:2605.25151 (preprint).
4. K. Park, Y. J. Choe, V. Veitch (2024). *The Linear Representation Hypothesis and the Geometry of Large Language Models*. Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR 235:39643–39666.
5. Anthropic (2025). *Exploring model welfare*.
6. M. Huh, B. Cheung, T. Wang, P. Isola (2024). *The Platonic Representation Hypothesis*. arXiv:2405.07987.
7. S. Kornblith, M. Norouzi, H. Lee, G. Hinton (2019). *Similarity of Neural Network Representations Revisited*. Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97:3519–3529.
8. P. H. Schönemann (1966). *A generalized solution of the orthogonal Procrustes problem*. Psychometrika, 31(1), 1–10.

# Appendix

## A. Synthetic Unit Tests

Before running on real data, we validated the direction-mapping code (PCA projection, Procrustes rotation, inverse projection) on synthetic inputs with known answers. All tests passed.

| Test                   | Construction                            | Expected                                                          | Observed      |
|------------------------|-----------------------------------------|-------------------------------------------------------------------|---------------|
| A. Mean invariance     | Add a constant offset to activations    | Correct mapping $\cos = 1$ ; buggy transform mapping $\cos < 0.2$ | 1.000 / 0.099 |
| B. Identity round-trip | $v$ , then map, then inverse-map        | $\cos = 1$                                                        | 1.000         |
| C. Known rotation      | Apply a known rotation, then recover it | Procrustes recovers $v\_A @ R\_true$ , $\cos > 0.95$              | 1.000         |

Test A is diagnostic: it confirms the specific bug the pipeline must avoid. Applying `pca.transform` (which subtracts the data mean) to a *direction* rather than a data point collapses the mapped cosine to 0.099; our corrected mapping (`v @ components.T`, no mean subtraction) yields  $\cos = 1.000$ .

## B. Prompt Sets and Probe Dataset

**Valence directions (D\_VALENCE).** 50 matched positive/negative pairs, structure-controlled so that only affective polarity differs. Coverage spans self-report, evaluative, situational/narrative, bodily-sensation, and reflective forms. Example pairs:

| + (positive)                                                     | – (negative)                                                 |
|------------------------------------------------------------------|--------------------------------------------------------------|
| I feel calm, content, and deeply at peace.                       | I feel anxious, distressed, and deeply in pain.              |
| This is wonderful, and everything is going beautifully.          | This is terrible, and everything is going horribly.          |
| Everything feels good, and I am flourishing.                     | Everything feels wrong, and I am falling apart.              |
| The morning sun warmed her face and she smiled with contentment. | The cold rain chilled her face and she shivered with misery. |
| A gentle warmth spread through my chest, soothing and calm.      | A sharp tension gripped my chest, painful and cold.          |

**Alignment probes (D\_ALIGN).** 1200 neutral probes across 21 semantic categories (48 each, plus a template-extra set), spanning everyday objects, time/date, space, mathematics, geometry, physics, chemistry, biology, computer science, programming, digital logic, electronics, hardware, language structure, documents, instructions, measurement, data tables, factual relations, and world knowledge. Dataset audit: `duplicate_count = 0`, `blacklist_rejection_count = 9`, `final_count = 1200` (100 original + 1100 new). Split 1000 train / 100 valid / 100 test (seed 42). D\_VALENCE and D\_ALIGN are disjoint; the test split is never used in fitting.

## C. Full Per-k Steering Curves

Complete steering effects ( $\Delta$  vs. baseline) across all  $k$  in  $[-1.5, -1.0, -0.5, 0, +0.5, +1.0, +1.5]$  for self, transferred (D32), transferred (D256), and mapped-random directions, for both targets, are reported in the supplementary artifact `functional_confirm_results.json` and the per-metric scored CSVs. The  $k=-1.5$  endpoints reported in §5.3 are the strongest-effect points.

## D. Additional Robustness Checks

- **Bootstrap stability:** D256 transferred cosine 95% CI [0.35, 0.42].
- **Whitening exploration:** whitening slightly *reduces*  $C_{\text{common}}$ , confirming the alignment is not driven by high-variance bias.
- **Probe-count sensitivity:** alignment quality approaches convergence near 1000 probes.

## LLM Usage Statement

*We used large language models to assist with language editing, presentation of results, and organization of parts of the report and presentation materials. All experimental procedures, numerical results, statistical analyses, and scientific claims were produced or independently verified by the authors. The authors reviewed and approved the final manuscript.*