NINE EMOTION ANCHORS RECOVER A VALENCE DIRECTION: NO-LABEL CROSS-MODAL TRANSPORT AND SCALE-AWARE STEERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Inside a modern language model sits an internal direction that tracks how positive or negative a sentence feels. We recover this *valence axis* (V-axis) from nine generic emotion-category anchors – no downstream sentiment labels – and follow it through four questions: can we **recover** it cheaply, is it broadly **readable**, does readable imply causally **writable**, and where does it **fail**? *Recover:* the recipe averages $\sim$50 emotion-anchored paragraphs per category into nine centroids and takes their top principal component ($\sim 384\times$ fewer labels than a supervised SST-2 probe: n=18 supervision events vs. 6,920). At matched budget the nine anchors – not the specific reader – carry the signal: difference-of-means, CAV, and LDA readers of the same anchors match or beat the principal-component recipe, all clearing a 500-direction random null on $8/9$ models. *Read:* the axis is stable when explicit emotion words are stripped from the prompts (cosine 0.98 between original and word-removed axes), reads four disjoint sentiment corpora above null (SST-2, IMDB, Rotten-Tomatoes, Yelp), and appears in a self-supervised vision encoder that never saw language (DINOv2, valence AUC 0.790 vs. CLIP control 0.972). *Write:* readable does not mean freely writable. Removing the single axis leaves a retrained probe intact – valence spans a 4–36-dimensional redundant subspace – yet the model still *causally uses* it *at inference time*: *directional ablation* (Arditi et al., 2024) of the axis drops the model's own sentiment output on $6/9$ families while a norm-matched random-direction ablation moves it ≈ 0. Steerability follows a residual-stream-norm regularity that *predicts* held-out models (leave-one-model-out R^2=0.88), and an explicitly-defined donor$\rightarrow$receiver map transports the axis across models above a random-map null (3 model pairs, all 6 off-diagonal cells, 20–31$\times$). *Bound:* the recipe is at-or-near chance on categorical concepts (AxBench Concept-500 (Wu et al., 2025), KS p=0.21) and on arousal, and on EEG valence is decodable but its geometry is *not* shared with machine encoders – a boundary case, not a fourth shared modality; a geometric "admissibility" criterion we tested does not predict held-out success and is reported only as a negative. Valence is thus the rare high-level attribute recoverable from category anchors alone that also transports *without target labels* across four independently-trained encoders and stays causally writable under a scale-aware fix – a reach that is real but, we show, bounded by residual-stream scale and cross-encoder geometry.

1 INTRODUCTION

A modern language model, when it reads a sentence, builds up an internal pattern of activity across thousands of neurons that we call its *hidden state* (or *residual stream*, since each transformer block adds to a running sum). Recent interpretability work has found that surprisingly simple structure lives inside that activity. In particular, many high-level attributes that we might naively think of as scattered across the network – "is this sentence about France?", "is the model about to refuse the request?", "does this concept refer to a person?" – can in fact be read off, and sometimes manipulated, by looking at a *single linear direction* in the hidden-state space (Park et al., 2024; Arditi et al., 2024). We add one more attribute to that list: *valence*, the positive-vs-negative emotional charge of a piece of input. We then show two things that, to our knowledge, no prior valence-axis result has shown together: that the

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.

direction can be found with ~ 9 category names and ~ 450 unlabelled sentences instead of thousands of polarity labels, and that the *same* valence direction (up to per-modality refit of one direction) appears in vision, audio, and human-EEG encoders that share no pretraining whatsoever. A single valence axis transports across all four modalities with no target labels (§3); we additionally find it sharpens toward human ratings as more models are pooled and varies in "harshness" by model family, analyses developed in the supplementary material. Recent prior work identifies a valence–arousal subspace in LLM text representations (Sun et al., 2026; Choi & Weber, 2026) (both from mid-2026); we differentiate on properties that a text-only subspace does not establish: four-modality transport with *no* target-task labels, causal cross-model portability, valence-*selective* steering, and the EEG bridge (§4.2, §3).

Standard route vs. our recipe. The textbook way to build a continuous-attribute axis inside a foundation model is supervised: take thousands of labelled examples (positive sentences, negative sentences), run them through the model, and fit a linear classifier on the hidden state. The weight vector of that classifier is the "axis." Our recipe takes a different path. It requires *no* polarity labels and only 18 supervision events total: 9 emotion category names (anger, joy, fear, ...) and 9 writing prompts. From those we author ~ 50 short emotion-evocative paragraphs per emotion, embed each one through a frozen encoder, average within emotion to get 9 *emotion centroids* in hidden-state space, and take the top principal component (PC1) of the resulting $9 \times d$ matrix. That single direction – the V-axis (*valence axis*) – is the only quantity used downstream. The label-cost ratio versus a supervised SST-2 probe (Socher et al., 2013) is $6{,}920/18 \approx 384\times$ (standard SST-2 train labels divided by recipe supervision events).

What is genuinely new versus prior work. The construction is *PC1 of $K{=}9$ category centroids*, not a discriminant between two contrastive labels: Park–Choe–Veitch (Park et al., 2024) directions live in a pair-discriminant geometry, and the V-axis is orthogonal to those directions on identical concepts (mean $|\cos| {=} 0.038$). The directional-ablation intervention of Arditi et al. (2024) is here applied to a *recipe-derived* direction, not to a discriminant fit to refusal/sentiment labels, so the causal claim tests the recipe itself rather than reading off supervision. And to our knowledge, cross-modal transport of a single readable-and-writable valence direction across independently-trained encoders (text, vision, audio, EEG) has no precedent in the LRH–AxBench–steering line. The paper is organised as one arc: *recover* the axis cheaply (§2), *read* it across families, datasets, and independent encoders (§3, §5), ask whether readable implies causally *writable* (§4, §4.2), and *bound* where it fails (§6, §3.2).

Why nine and not two. Fitting a single "sentiment direction" to a pair of contrastive labels (positive vs. negative) forces a sign and conflates valence (pleasant/unpleasant) with arousal (calm/intense). Spreading the centroids across the nine FACED emotion categories (Russell, 1980; Ekman, 1992) (a discrete-emotion taxonomy) gives a linear reader more structure to exploit: against a *matched* two-centroid (positive- vs. negative-pooled) baseline, the nine-anchor PC1 is a wash, but the best nine-anchor readers beat two classes on *all nine* models (§2). Nine categories help, but the gain lives in the reader, not in PC1.

Why this is not a re-skin of AxBench. AxBench (Wu et al., 2025) showed supervised linear probes beat sparse autoencoders for text-only categorical-concept steering (144 labels/concept). We differ on three axes at once – label budget (18 events), concept type (continuous valence; our recipe ties the null on AxBench's categorical Concept-500, KS $p{=}0.21$, App. F), and modality scope (vision, audio, EEG) – detailed in §7.

Contributions. We organise the paper as one arc – *Recover*, *Read*, *Write*, *Bound*. The two results most clearly beyond prior valence-subspace work are no-target-label transport of a single valence direction across four independently-trained encoders (*Read*) and a scale-aware steering mechanism with causal cross-model portability (*Write*); the cheap recovery and the honest bounds frame them. (i) **Recover** (§2): nine emotion centroids, SVD, take PC1, fully specified, at $\sim 384\times$ fewer labels than a supervised probe. At *matched budget* the nine anchors, not the specific reader, carry the signal – difference-of-means, CAV, and LDA readers of the same anchors match or beat the principal-component recipe, all above a 500-direction random null on $8/9$ models (recipe z up to 8.1), and the label-free axis out-reads the model's own zero-shot prompting on $5/7$. (ii) **Read** (§3, §5): the axis is

robust across model families, stable to stripping explicit emotion words from the prompts (cosine 0.98 between original and word-removed axes), general across four disjoint sentiment corpora (SST-2, IMDB, Rotten-Tomatoes, Yelp, all above null), and present in a self-supervised vision encoder that never saw language (DINOv2 valence AUC 0.790 vs. a CLIP control 0.972); a text-supervised head reads all four encoders' V-axis projections with $12/12$ cross-cells AUC ≥ 0.70. (iii) **Write** (§4, §4.2): readable does not imply freely writable. Removing the single axis leaves a retrained probe intact (valence spans a 4–36-dimensional redundant subspace), yet *directional ablation* of the axis drops the model's own sentiment output on $6/9$ families while a norm-matched random ablation moves it ≈ 0; steerability follows a residual-stream-norm *regularity* that predicts held-out models (leave-one-model-out R^2=0.87), and an explicitly-defined donor→receiver map transports the axis across models above a random-map null (3 model pairs, all 6 off-diagonal cells, 20–31×). (iv) **Bound** (§6, §3.2): the recipe is at-or-near chance on categorical concepts (AxBench Concept-500, KS p=0.21) and on arousal, and on EEG valence is decodable but its geometry is not shared with machine encoders (a boundary case, not a fourth shared modality); a geometric admissibility criterion we tested does *not* predict held-out recipe success (held-out AUC 0.462) and is reported only as a negative.

What this is not. The recipe is an *empirical regularity*, not an analytical theorem, and its recovered object is a valence *direction*, not the uniquely optimal one: at matched budget, simpler difference-of-means and CAV readers of the same nine anchors are competitive with – and on some models beat – the principal-component recipe, so the contribution is the *anchors*, with PC1 one reasonable reader among several. It is bounded to *continuous* concepts: categorical concepts (Park–Choe–Veitch word pairs (Park et al., 2024); AxBench Concept-500 (Wu et al., 2025), KS p=0.21; vision object class) return at-or-near chance, and a geometric admissibility criterion tested here does not predict which concepts admit the recipe (held-out AUC 0.462; §6). EEG is a *boundary case*, not a fourth shared modality: valence is decodable from the CBraMod EEG encoder (Wang et al., 2025) via an anchor-guided axis (strict subject-and-stimulus-disjoint AUC 0.840, 98% of the supervised ceiling on that split; an optimistic stimulus-shared GroupKFold reads 0.968; a fully-unsupervised PC3 fallback reaches 0.799), but its geometry is *not* shared with the machine encoders ($r_V \approx +0.04$ cross-encoder), and a frozen FACED axis does not transfer zero-shot to SEED-V. "Broadly readable" is scoped to the encoders we tested (LLM families; CLIP-image (Radford et al., 2021), CLAP-audio (Elizalde et al., 2023), DINOv2, CBraMod-EEG), not to untested encoders. The causal claim rests on *directional-ablation* (Arditi et al., 2024) and native-behaviour evidence, not counterfactual intervention; directional ablation is sign-insensitive, so no sign-flip search is run, and the axis *orientation* is set by a fixed emotion-category prior, never by any evaluation-task label.

2 THE RECIPE

Intuition. Write nine short stories, one each for anger, disgust, fear, sadness, amusement, joy, inspiration, tenderness, and a neutral baseline – about fifty stories per category. Run them through a frozen language model and read its hidden state at each token; we skip the first $\sim$50 warm-up tokens and mean-pool the remainder into one vector per story. Average within each category. You now have nine vectors, one per emotion, that act as "coordinates" for that emotion in the model's representation space. Find the single direction along which those nine points spread the most (principal component analysis). That direction is the V-axis. The rest of the paper just measures what it does.

The construction is identical across modalities up to the encoder. We describe the language-model version first, then state the per-modality changes.

The three steps. (1) For each of the nine FACED emotion categories (*anger, disgust, fear, sadness, amusement, joy, inspiration, tenderness, neutral*) we author $N_c \approx 50$ paragraph-length ($\geq$ 15-token) emotion-anchored sentences, uncurated for sentiment/syntax/domain (full set in Appendix A; label budget n=18: nine category names plus nine writing prompts). (2) We encode each through a frozen encoder at a fixed layer ℓ, skip the first $\sim$50 warm-up tokens, mean-pool the rest into a story vector, and average within emotion to get centroids $M \in \mathbb{R}^{9 \times d}$. (3) We centre M and take the top right-singular vector v_1 (the *V-axis*); all downstream scoring is $\langle x, v_1 \rangle$. Algorithm 1 formalises this; it uses 9×50=450 forward passes and produces a 1-D direction.

Algorithm 1 V-axis construction

Input: encoder f_θ, layer ℓ; 9 emotion category names $\{c\}$; per-emotion sentence pools $\{s_{c,i}\}_{i=1}^{N_c}$ ($N_c \approx 50$).
Output: V-axis $v_1 \in \mathbb{R}^d$.

1. For each c, i: compute $h_{c,i} = f_\theta^{(\ell)}(s_{c,i})$ (residual stream at layer ℓ, skip-50 warm-up then mean-pool over the remaining tokens).
2. Compute centroids $\mu_c = \frac{1}{N_c} \sum_i h_{c,i}$; stack as $M \in \mathbb{R}^{9 \times d}$.
3. Center: $\tilde{M} = M - \frac{1}{9} \sum_c \mu_c$.
4. SVD: $\tilde{M} = U \Sigma V^\top$; return $v_1 = V_{:,1}$.

Table 1: **Matched-budget bake-off (SST-2 AUC).** Every direction is built from the *same* nine emotion centroids and oriented by the fixed valence prior (no SST-2 labels); the ridge "ceiling" uses 13,915 labelled Warriner words; the null is 500 random directions. The recipe (PC1) clears the null p_{95} on $8/9$ models, but simpler readers of the same anchors are competitive and sometimes better – the anchors, not PC1, carry the signal.

Model	Recipe (PC1)	Diff-means	CAV	LDA	RepE-PC	Ridge (ceiling)	Null p_{95}
Gemma-4-E4B	0.722	0.752	0.756	0.726	0.628	0.679	0.578
Llama-2-7B	0.468	0.471	0.726	**0.835**	0.467	—	0.630
Llama-3-8B-Inst	0.772	0.718	0.649	0.802	0.709	0.828	0.621
Mistral-7B-base	0.860	0.711	0.633	0.740	0.631	—	0.621
Mistral-7B-Inst	0.807	0.742	0.675	0.757	0.618	—	0.622
Qwen3-1.7B	0.773	0.761	0.780	0.786	0.638	—	0.600
Qwen3-8B	0.633	0.760	**0.895**	0.845	0.514	0.778	0.587
Qwen3.5-2B	0.867	0.863	0.855	0.838	0.707	0.844	0.654
Qwen3.5-4B	0.873	0.867	0.865	0.807	0.697	0.797	0.643

"Anchor-guided", not "label-free". The construction uses *no downstream/target-task labels* (no SST-2, EmoSet, Warriner, or binary FACED labels); the only supervision is the nine emotion-category names used to group sentences into centroids. We therefore call it *anchor-guided*: the axis is recovered without labels from the task it is later evaluated on, using generic emotion-category anchors fixed once and reused across all four modalities.

Matched-budget baselines: the anchors carry the signal. The recipe takes PC1 of the nine centroids, but PC1 is only one way to read nine signed anchors. We compare it, *at matched budget* (the same nine centroids, same layer, same evaluation features, orientation fixed by the emotion-category prior rather than any SST-2 label), against the standard concept-direction readers – difference-of-means, CAA (equal to difference-of-means for a signed pair set), a CAV (linear SVM/logistic) (Kim et al., 2018), LDA, and a RepE-style contrast PC (Zou et al., 2023) – and against a 500-direction random null and a supervised ridge "ceiling" fit on the 13,915-word Warriner lexicon (Warriner et al., 2013). Table 1 reports SST-2 AUC across nine models. Two things follow. First, the recipe clears the 500-direction null on $8/9$ models (recipe z from 3.4 to 8.1; the exception is Llama-2-7B, below). Second, and more honestly, PC1 is *not* uniquely best: difference-of-means, CAV, and LDA readers of the same anchors are competitive with it and beat it on several models (on Qwen3-8B the CAV reaches 0.895 where PC1 is only 0.633; on Llama-2-7B PC1 is at chance 0.468 while LDA recovers 0.835). The contribution is therefore the *nine anchors*: any reasonable linear reader of them recovers valence, and PC1 is a convenient but not dominant choice. The "Diff-means" column is exactly a *matched two-centroid* reader (positive- vs. negative-pooled anchors), so Table 1 already answers "nine vs. two": the nine-anchor PC1 ties it (median $+0.006$ AUC, Wilcoxon p=0.18), while the best nine-anchor reader per model beats it on *all nine* models (median $+0.065$, p=0.002) – the extra categories help a discriminative reader, not PC1. The label-free axis also out-reads the model's own zero-shot sentiment prompting on $5/7$ models (mean AUC 0.740 vs. 0.627). Source: `experiments/axis_experiments/E1_matched_baselines/results/`.

Table 2: V-axis (anchor-guided in the axis for all four modalities; no target-task labels) recovers supervised-classifier valence within 7 pp AUC on text and audio; on vision the axis attains Pearson r=0.636 vs. a supervised r=0.836 (76% of the supervised correlation, *not* an AUC). On EEG the axis is an *anchor-guided* ordering-weighted centroid contrast whose only supervision is the categories' known valence *ordering* (no binary valence, SST-2, or FACED labels; the fully-unsupervised PC3 reaches 0.799 since arousal dominates PC1/PC2): re-derived from training-fold subjects only it reaches AUC 0.840 (BACC 0.765) under the strict subject-and-stimulus-disjoint split, 98% of the supervised cross-subject LDA ceiling (0.857) on that split; a stimulus-shared subject-disjoint GroupKFold gives an optimistic 0.968 (App. H reconciles all splits). *Recipe AUC*: held-out AUC of $\sigma(\langle x, v_1 \rangle)$ where v_1 is the V-axis (text Pearson r for vision). *Sup. AUC*: a logistic-regression (LDA for EEG) head trained on the modality's full label set ($n_{\text{SST-2}}$=6,920; n_{EmoSet}=5,905 for the recipe correlation, vision Sup. a supervised-ridge upper bound on a held-out EmoSet subset, source `experiments/d29_bridge/results/p1_vision_supervised_ridge.json`; $n_{\text{ESC-50}}$=2,000; n_{FACED}=123 subjects). *Ratio*: recipe/supervised. See §3.

Modality	Encoder	Dataset	Recipe AUC	Sup. AUC	Ratio	p
Text	Llama-3-8B-Inst (block 20)	SST-2	0.772	0.828	93%	—
Text	Qwen3-8B (block 31, depth 0.86)	SST-2	0.787	0.840	94%	—
Vision	CLIP-img-768	EmoSet	r=+0.636	r=+0.836	76%	null 0.112
Audio	CLAP-aud-512	ESC-50	0.906	0.94	96%	2.2×10^{-15}
EEG	CBraMod-200	FACED	0.840	0.857	98%	subj+stim-disj

3 FOUR MODALITIES

The same recipe applied to four independently-trained foundation encoders recovers a 1-D direction that tracks valence at the level of supervised classifiers (Table 2; per-modality bar visualisation deferred to Appendix P).

Text, vision, audio. Table 2 gives the per-modality readout. On text the V-axis reaches 93% of a supervised probe's SST-2 AUC at ∼384× fewer labels (0.772 vs. 0.828, Llama-3-8B-Inst; Qwen3-8B 0.787 at block 31, a reasoning-distillation depth shift, App. L). On vision the axis correlates with EmoSet valence at r=+0.636 (random-direction null $|r| \leq 0.112$); on audio it reaches AUC 0.906 on ESC-50 (paired p=2.2×10^{-15}), the strongest single modality.

EEG (anchor-guided valence axis). FACED (Chen et al., 2023) contains 123 subjects watching 28 emotion-evoking video clips with 9-way labels. We encode 30-second windows through CBraMod (Wang et al., 2025) and apply the recipe in 200-D feature space as an ordering-weighted centroid contrast over the 9 categories, ordered by their known valence. That valence *ordering* is the axis's only supervision – no binary valence, SST-2, or FACED labels enter – but, unlike text/vision/audio, the unsupervised PC1 does *not* recover valence here (it is at chance, 0.512, as arousal dominates the top components), so the EEG axis leans on the ordering rather than on valence dominating the centroid spread. Decoding depends strongly on the split: re-derived from training-fold subjects only, the axis reaches AUC 0.840 (BACC 0.765) under the strict subject-*and*-stimulus-disjoint split (98% of the supervised cross-subject LDA ceiling, 0.857) and 0.811 under the FACED-standard held-out-subject split, while a stimulus-shared subject-disjoint GroupKFold gives an optimistic 0.968; we headline the strict split. This axis is the same reader (cosine 0.983) as the demoted name-level ridge – one method under different splits, not a leakage-free/leaky pair; App. H reconciles all six readers across five splits, and reports the fully-unsupervised PC3 fallback (0.799 GroupKFold, 0.667 strict). The cross-modal "text→EEG AUC 0.828" is the *bridge*: a 1-D logistic head trained on SST-2 text labels applied to CBraMod activations projected onto the EEG V-axis. **EEG is a boundary case, not a fourth shared modality:** valence is *decodable* from the EEG encoder via the anchor-guided axis, but the full 9-emotion centroid *geometry* shows no evidence of transporting to the machine encoders (Procrustes/RSA nulls, App. K; centroid coordinates align with the shared valence prior at only r=0.04, though this n=9-centroid test is low-powered). We reserve the "shared axis" claim for the three machine modalities.

Cross-modal alignment. The four V-axes live in different encoder spaces with no shared coordinates. Gröger-calibrated CKA (Gröger et al., 2026) between every pair of V-axis projections rejects the matched-norm null for all three text↔image↔audio pairs ($p \leq 0.032$, uncorrected; the weakest pair does not survive a 3-way Bonferroni threshold of 0.017) – small-magnitude, nominally significant evidence of partially-shared structure, leaving open whether full directional correspondence holds under a stricter geometry.

3.1 Robustness of the readout: corpora, encoders, datasets

Three tests probe whether the readout is an artefact of the specific prompts, the specific encoder, or the specific evaluation dataset (`experiments/axis_experiments/E2-E4/`).

Prompt corpus. Rebuilding the nine centroids from story variants that *remove* the explicit emotion vocabulary (the emotion-category words plus a per-emotion lexicon) barely moves the axis: the cosine between the original and word-removed axes is 0.98, and word-removed SST-2 AUC equals or exceeds the original on $6/7$ models (e.g. Mistral-Inst $0.809 \rightarrow 0.841$, Qwen3.5-2B $0.867 \rightarrow 0.863$). The recipe reads *latent* affect, not surface emotion words; Qwen3-8B is the one corpus-sensitive exception.

Independent encoder. To rule out that the cross-modal readout is an artefact of language-aligned encoders (CLIP/CLAP see text at training), we apply the recipe in a self-supervised vision encoder that never saw language: DINOv2 (Oquab et al., 2024) reads EmoSet valence at AUC 0.790 against a 500-direction null p_{95} of 0.599 (ConvNeXt 0.632 likewise clears null). Language alignment amplifies the signal strongly – the CLIP-image control reaches 0.972 – but does not create it: a genuine valence direction exists in a no-language encoder.

Cross-dataset. The text axis is constructed from emotion stories, never from sentiment-dataset labels; it reads four *disjoint* sentiment corpora above their nulls on every model tested (SST-2, IMDB, Rotten-Tomatoes, Yelp; $4/4$ datasets on all seven models, IMDB up to 0.890, Yelp up to 0.972). The axis is dataset-general, not tuned to SST-2.

3.2 Arousal and Dominance Axes: Only Valence Transports Without Target Labels

Only valence transports for free. The same $9 \times d$ centroids also yield an *arousal* axis at PC2 ($r{=}+0.70$ vs. NRC-VAD (Mohammad, 2018)) and a *dominance* axis (anger vs. fear separate by $+1.24$ sd on dominance but only -0.12 sd on valence; App. Fig. 2). But only *valence* transports across modalities without target labels (clearing the matched-norm null in $2/4$ strict tests; arousal $0/4$, dominance $1/4$): cross-modal affect convergence is valence-specific, not a blended valence/arousal/dominance phenomenon. With a few target labels, dominance partially transports (image $r{=}0.75$ at $N{=}5$; EEG 0.67 at $N{=}10$) while arousal remains the holdout (App. Fig. 3). Sources: `experiments/d24_e1_va_bakeoff/results/`.

4 Causal Evidence

Intuition and protocol. We test whether the axis *carries* sentiment by deleting it from the forward pass. Following the directional-ablation protocol of Arditi et al. (2024), at every layer ℓ and token t we replace $h_t^{(\ell)}$ with $h_t^{(\ell)} - (\langle h_t^{(\ell)}, \hat{v}_1 \rangle)\hat{v}_1$ (unit V-axis $\hat{v}_1$, fixed once from §2 at the model's optimal layer), then read SST-2 accuracy from a probe trained on the *unaltered* activations at the hooked layer. If the axis merely correlates with sentiment, redundant features absorb the loss; if it carries sentiment, accuracy drops. As a control we ablate $K{=}3$ norm-matched random directions per model. The operation is sign-insensitive ($+\hat{v}_1$ and $-\hat{v}_1$ remove the same subspace) and the null is sampled before evaluation, so no sign-flip search inflates the effect.

Result. Table 3 and Figure 4 report the full ablation panel with a *fixed* probe – one trained on unaltered activations, then evaluated on ablated activations. In every model, none of ≥ 150 matched-norm random directions produced a drop as large as the V-axis (empirical permutation $p \leq 0.007$; largest random drop ≤ 5.85 pp versus V-axis drops of 5.5–37.2 pp). But a fixed-probe drop, on its

Table 3: V-axis directional ablation drops SST-2 sentiment classification by 5.5–37.2 percentage points across three LLMs, far beyond a matched-norm random-direction null. Columns: *Base/V-abl.*: SST-2 dev-split accuracy (%) without/with V-axis projection ablation. *Drop (V)*: V-ablation accuracy − baseline accuracy (pp). *Rand. max*: largest drop over ≥ 150 matched-norm random-direction draws per model. *perm. p*: fraction of $\geq$150 random directions whose drop $\geq$ the V-axis drop (add-one; 0 reach it in every model). Drop 95% bootstrap CIs (n_{val}=872, row order): $[3.0, 7.9]/[11.8, 19.2]/[32.5, 40.6]$ pp. Source: `experiments/d31_causal_mediation/results/`. See §4.

Model	Block	Base (%)	V-abl. (%)	Drop (V)	Rand. max	perm. p
Llama-3-8B-Inst	20	87.96	82.45	−5.50 pp	1.95 pp	0.007
Qwen3-1.7B	18	84.40	68.58	−15.83 pp	5.85 pp	0.004
Qwen3-8B	23	87.61	50.46	−37.16 pp	0.78 pp	0.007

own, admits an innocent reading: the probe may simply have relied on the one direction we removed. The next subsection tests that reading directly.

4.1 The ablation triad: readout is redundant, behaviour is causal

A single fixed-probe drop is not sufficient evidence that a direction *carries* an attribute; a causal claim needs the retrained-probe and native-behaviour legs under the same intervention. We ran all three (nine models; `experiments/axis_experiments/E5_ablation_triad/`).

Retrained probe. After projecting the V-axis out of the SST-2 features, we retrain a fresh linear probe on the *ablated* features. It recovers almost fully – the drop relative to an un-ablated probe is ≤ 0.016 AUC on all models (Llama-3 $+0.002$, Qwen3-8B $+0.003$, Qwen3.5-4B -0.001). So the fixed-probe collapse overstates causality: valence is *redundantly encoded*, and removing one direction erases only the copy that probe happened to use.

Erasure dimensionality. Iterative nullspace projection (INLP; Ravfogel et al., 2020) shows how redundant: it takes k=4–36 directions of erasure before a retrained probe falls to chance (k=4 Llama-3, 5 Qwen3.5-2B, 36 Qwen3-8B). Valence is a low- but multi-dimensional subspace, and the recipe axis is nearly orthogonal to the data's own most-discriminative direction ($|\cos|$=0.08–0.16) – consistent with §2, where the anchors, not the specific reader, carry the signal.

Native behaviour. The decisive leg asks whether the model's *own* zero-shot sentiment output changes when the axis is ablated from its forward pass. It does: projecting the V-axis out of the residual stream drops the model's own next-token sentiment AUC on $6/9$ families (axis-specific drop up to 0.097 on Qwen3-8B; 0.071 Qwen3.5-2B; $\approx$0.05 on Mistral-base/inst and Qwen3-1.7B), while a norm-matched random-direction ablation moves it by $\leq$0.005 throughout. The two weak families are Llama-2-7B (its zero-shot baseline is already at chance) and Gemma-4 (weak baseline); Llama-3 is borderline ($+0.025$). This reconciles the triad: the linear readout is redundant, yet the model *causally uses* this specific direction for its behaviour. The causal weight therefore rests on the native-behaviour and steering (§4.2) legs, not on the fixed-probe drop alone.

4.2 Steering: a Scale Artefact and a Universal Fix

The apparent family boundary. Removing the V-axis and *adding* a scaled multiple of it back to steer generation are different operations. At a *fixed* unit-scale coefficient, steering works on the Llama and Mistral families (Spearman ρ between steering coefficient and pooled sentiment polarity 0.32–0.45, all $p < 10^{-6}$) but is at-or-near chance on Qwen, Gemma, and reasoning-distilled models ($|\rho| \leq 0.089$, none significant). The base-vs-instruct stage does not predict this (Mistral-base steers as well as Mistral-Inst); something about the family does. This looked, in an earlier version of this work, like a fundamental limit. It is not.

The mechanism: residual-stream norm rank-orders steerability. The separating variable is the residual-stream magnitude: mean $\|h\|$ *perfectly rank-orders* fixed-α steerability (Spearman -1.0, log-norm R^2=0.94) – steerable Llama/Mistral have $\|h\| \approx$ 6–8, un-steerable models up to $\sim$157. A unit-scale steering vector added to a hidden state two orders of magnitude larger is swamped and

absorbed by the following RMSNorm, so additive steering is *non-surjective* on high-norm models (Mishra et al., 2026). Every model can be *read*, but only low-norm models can be *written* at fixed scale (App. Fig. 7).

The fix: norm-matched steering recovers the null models. Scaling the coefficient to the model's own norm ($\alpha \propto \|h\|$) removes the obstruction: the three previously-null high-norm models recover cleanly with an interior optimum and monotone dose–response (Spearman $0.082 \to 0.784$ Qwen3-8B-Inst; $0.002 \to 0.755$ Qwen3-1.7B-base; $0.089 \to 0.771$ DeepSeek-R1-Distill; Mistral-Inst control 0.813), at modest perplexity cost (1.3–$1.6\times$). The "Qwen cannot be steered" boundary was a *scale artefact* of fixing α across models, not a family property.

A $\|h\| \leftrightarrow$recovery law. The relationship is lawful, not anecdotal (App. Fig. 10): across $n{=}8$ models the recovery gain $\Delta\rho$ from norm-matching scales monotonically with $\|h\|$ (Spearman 0.976, $p{=}3.3 \times 10^{-5}$; log-norm Pearson $R^2{=}0.91$), while the *rescuing* relative step $|c|$ stays $O(1)$ (clustered around 0.5–2) despite a $\sim 66\times$ swing in $\|h\|$. With $n{=}8$ this is an inferential, not merely descriptive, regularity, and it gives a one-number steerability predictor: read off $\|h\|$, and both the fixed-α deficit and the norm-matched recovery follow. The regularity *predicts held-out models*: fit $\Delta\rho \sim \log\|h\|$ on $n{-}1$ models and predict the left-out one with no per-model tuning, and the out-of-sample R^2 is 0.87 for the norm-matched recovery gain and 0.92 for fixed-α steerability (against a predict-the-mean baseline; leave-one-model-out, `experiments/axis_experiments/E7_heldout_steering/`).

Causal cross-model portability: a valence axis steers a *different* model. The strongest evidence that the V-axis is a genuinely *shared* direction – not merely a correlationally-aligned per-model artefact – is that a valence axis extracted from one model steers a *different* model. We take the V-axis from a *donor* model and use it (after sign-fixing only, no per-pair refit) to additively steer a *receiver* model under the norm-matched protocol above, measuring the same pooled sentiment shift Δ (positive- minus negative-steering polarity). All $3/3$ donor$\to$receiver pairs transfer for every receiver, and the cross-model cells survive the Gröger (Gröger et al., 2026) width-matched representational null in $6/6$ off-diagonal cells (donor$\to$receiver projection-cosine exceeds the matched-width null). The transport map is *explicitly defined* – a shared-anchor centroid pseudoinverse projection of the donor axis into the receiver's hidden space – and beats a norm-matched *random-map* null by 20–$31\times$: the projected donor axis aligns with the receiver's own axis at cosine 0.81–0.98, versus a random-map 99th-percentile of ≈ 0.04 (all 6 off-diagonal cells; `experiments/axis_experiments/E9_steering_map/`). The per-cell $3{\times}3$ matrix (App. Fig. 5) shows off-diagonal cells recovering most of the within-model control (every cell $p < 0.05$, Wilcoxon). A donor model's valence axis therefore carries causal valence control into a receiver model it was never aligned to – causal evidence that the axis is shared across models, not just statistically correlated. This is a methodologically different setup from *learned* cross-model activation transport (Oozeer et al., 2025) or supervised cross-model direction-mapping (Poppi et al., 2026): our portability uses no learned inter-model mapping, no autoencoder, no paired-activation training, and no per-pair refit. Source: `experiments/d23_steering/results/crossmodel_steering_portability.json`.

5 A Universal Cross-Modal Classifier

Intuition. One sentiment dial – a single threshold and slope – trained on text labels alone reads sentiment off images, sounds, and brain recordings. The dial does not need to know what modality it is looking at; the V-axis already does the modality-specific work of projecting raw activations down to a 1-D valence score, and the dial just calibrates it into a probability. The result is a sentiment classifier that crosses modality boundaries without ever seeing target-modality labels: a 2-parameter logistic head trained on SST-2 transfers across all four modalities with AUC ≥ 0.70 in $12/12$ off-diagonal cells, including text$\to$EEG (App. Fig. 11).

Setup. For each modality we project a held-out set onto its V-axis $v_1^{(m)}$ (a 1-D score per sample), then fit a 2-parameter logistic head (slope, intercept) on source-modality scores against source-modality valence labels and evaluate it on the target modality (App. Table 4: diagonal = 5-fold self-CV, off-diagonal = train-source/test-target). The unsupervised V-axis sign is resolved per modality by aligning it with the source-modality positive class; no further sign search is run.

Headline. A text-trained head reads EmoSet images at AUC 0.961, ESC-50 audio at 0.764, and FACED EEG at 0.828 with no target-modality labels; transfer is robust to source choice (all $12/12$ off-diagonal cells ≥ 0.70).

6 The Boundary: Categorical Concepts, and a Criterion That Does Not Predict Success

The recipe is sharply bounded to *continuous* attributes: seven independent tests on categorical concepts (§6.1) return at-or-near chance. We asked whether a geometric criterion – how much a concept's 9-centroid axis projects into the LLM's residual-stream substrate – could *predict* which concepts admit the recipe. It cannot. The criterion correlates strongly with a concept's PC1 explained-variance (a geometric property, Spearman 0.891; App. J.1), but does *not* predict downstream recipe success: trained to predict success (recipe beats its null) from the criterion on held-out concepts, its out-of-sample AUC is 0.462 – below chance – with an in-sample Spearman of $+0.02$ (held-out 5-fold over 200 concepts $\times$ two models; `experiments/axis_experiments/E8_criterion_heldout/`). Admissibility is *linearity-independent*. We therefore report the criterion as a *negative*: it measures a geometric regularity, not a per-concept admissibility gate. What survives is the descriptive bound below – the recipe reads continuous valence, not categorical concepts – and the criterion machinery is retained (App. J) only to document that this natural predictor fails.

6.1 Reframed conclusion

The bounded scope. Seven independent tests on categorical concepts return at-or-near chance against the same 9-emotion pipeline: Park–Choe–Veitch causal-inner-product directions are near-orthogonal to the V-axis (mean $|\cos|$=0.038); long-tail retrieval, multi-concept probes, CIFAR-100 object class, and AxBench Concept-500 all tie the matched-norm null (KS p=0.21); and reasoning-distilled categorical concepts show no depth shift (full list Apps. F, E). The recipe also collapses with too-small pools (9 single-word labels give AUC 0.50; $N_c \geq 20$ paragraphs are needed, App. C). The recipe is bounded to continuous attributes; the geometric criterion separates continuous from categorical geometrically but, as above, does not predict admissibility within the categorical regime.

Three distinct properties. Readability (broad), geometric linearity (the criterion, a descriptive regularity that does not gate admissibility), and steerability (gated by the residual-stream scale $\|h\|$, §4.2) are three separate properties of the axis; only the first is broadly present.

7 Related Work and Discussion

Concept directions vs. AxBench. Kim et al. (2018) introduced concept activation vectors; Park et al. (2024) formalised the LRH; AxBench (Wu et al., 2025) reports supervised probes beating SAEs (Templeton et al., 2024) for steering. We differ from AxBench on three axes: label budget (n=18 vs. 144/concept, $\sim 8\times$ less); concept type (continuous valence, not categorical – the recipe ties the null on AxBench's 500 categorical concepts, KS p=0.21); and modality scope (text-only there; here image/audio/EEG, where a 1-D direction beats a 16-D generic substrate by 0.18–0.44 AUC, App. Table 5). We adopt the directional-ablation protocol of Arditi et al. (2024).

Cross-encoder convergence and brain alignment. Huh et al. (2024) argue encoders converge at scale; our 4×4 matrix extends this to a 1-D task-relevant direction across modalities with no shared pretraining. Gröger-calibrated CKA (Gröger et al., 2026) passes in $3/3$ text$\leftrightarrow$image$\leftrightarrow$audio cells ($p \leq 0.032$). Huth et al. (2016); Toneva & Wehbe (2019) mapped language-evoked brain activity; our EEG result is on emotion-induced video EEG and is a direction-existence, not biological-mechanism, claim.

Limitations and future work. The recipe is an empirical regularity, not a theorem; it is bounded to continuous attributes; "universal" is scoped to the four tested encoders; and "causal" means inference-time projection, not counterfactual intervention. EEG remains a boundary case (§S1.4): decodable *per-trial* (strict AUC 0.840) but not centroid-geometrically shared, and a *frozen* FACED axis does not transfer to SEED-V.

AI USE STATEMENT

We used generative AI tools, including Claude Code, to assist with literature review, suggest research ideas that the authors used as starting points, implement routine experiments, and assist with interpreting their results. More complex experimental implementation and the synthesis of findings across experiments involved direct human input. Generative AI was also used to produce emotion-evoking stories that served as inputs for extracting emotion representations from language models; these generated inputs were distinct from the established datasets used for evaluation. AI assisted with manuscript drafting and figure generation from data already verified by the authors. Codex assisted with submission formatting and manuscript review. The authors reviewed AI outputs at key stages for correctness and appropriateness and take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies the internal representations of frozen, publicly released foundation models and uses only publicly released, de-identified datasets; we collect no new human-subjects data. The EEG datasets (FACED, SEED-V) were released by their original authors under permissive licences with informed consent and institutional approval (Chen et al., 2023). The valence axis is used solely as an interpretive probe of model and *cohort-level* representations, never for individual-level affect inference or profiling. The recipe lowers the marginal cost of building affect-aligned axes; positive uses include label-efficient sentiment evaluation for low-resource modalities. As a dual-use consideration, the norm-matched steering primitive (§4.2) makes an affect-steering direction cheap to obtain across model families; the effect is inference-time and modest, we document rather than promote such use, and we release code under a permissive licence to support scrutiny. All external assets are used under their respective licences.

REPRODUCIBILITY STATEMENT

All encoders and datasets are public: the LLM families and CLIP (Radford et al., 2021), CLAP (Elizalde et al., 2023), DINOv2, and CBraMod (Wang et al., 2025) encoders are fetched from public checkpoints, and SST-2, IMDB, Rotten-Tomatoes, Yelp, EmoSet, ESC-50, FACED, SEED-V, and AxBench Concept-500 (Wu et al., 2025) are publicly available. The recipe is fully specified in §2 (Algorithm 1), with the prompt corpus, per-model layer search, and pool-size sweep in Appendices A, B, and C; per-modality encoders, datasets, and split protocols are given in §3, and the ablation and steering protocols (matched-norm random-direction nulls with 500 draws unless stated, fixed seeds, and leave-one-model-out) in §4–§4.2. The full statistical methodology (random-direction nulls, permutation tests, matched-norm controls) is reported with each result. Every headline number is tagged to the result file that produced it. The recipe code and the 9-emotion prompt corpus (Appendix A), including the exact generation protocol for the corpus, will be released under a permissive (MIT) licence; to preserve anonymity we omit the repository URL from this submission and will add it for the camera-ready.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.11717.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing EEG dataset. *Scientific Data*, 10:740, 2023. Article 740.

Benjamin J. Choi and Melanie Weber. Latent structure of affective representations in large language models. *arXiv preprint arXiv:2604.07382*, 2026.

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

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. *arXiv preprint arXiv:2602.14486*, 2026.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *International Conference on Machine Learning (ICML)*, 2024. arXiv:2405.07987.

Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning (ICML)*, 2018.

A. Sophia Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A. Efros. Back into Plato's cave: Examining cross-modal representational convergence at scale. *arXiv preprint arXiv:2604.18572*, 2026.

Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels. *arXiv preprint arXiv:2605.29800*, 2026.

Aayush Mishra, Daniel Khashabi, and Anqi Liu. Steered LLM activations are non-surjective. *arXiv preprint arXiv:2604.09839*, 2026.

Saif M. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Saif M. Mohammad. NRC VAD lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms. *arXiv preprint arXiv:2503.23547*, 2025.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, et al. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846, 2024. doi: 10.1038/s41586-024-07335-x.

Narmeen Oozeer, Dhruv Nathawani, Nirmalendu Prakash, Michael Lan, Abir Harrasse, and Amirali Abdullah. Activation space interventions can be transferred between large language models. *arXiv preprint arXiv:2503.04429*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2311.03658.

Tobia Poppi, Silvia Cappelletti, Sara Sarto, Florian Schiffers, Garin Kessler, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Do models share safety representations? cross-model steering for safe visual generation. *arXiv preprint arXiv:2606.05290*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. arXiv:2103.00020.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

Nicholas Sofroniew et al. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*, 2026.

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

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jermyn, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 sonnet. *Anthropic Transformer Circuits Thread*, 2024.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Hieu M. Vu and Tan M. Nguyen. Angular steering: Behavior control via rotation in activation space. *arXiv preprint arXiv:2510.26243*, 2025.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2412.07236.

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.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2501.17148.

Zejia You, Chunyuan Deng, and Hanjie Chen. Spherical steering: Geometry-aware activation rotation for language models. *arXiv preprint arXiv:2602.08169*, 2026.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Prompt corpus and per-emotion examples

The full 9-emotion prompt corpus contains ~ 50 paragraph-length narrative continuations per emotion (450 total). Per-emotion example seeds: *anger*: "She slammed the door so hard the frame cracked" (47 continuations); *disgust*: "The milk had gone sour weeks ago and he hadn't noticed" (52); *fear*: "Footsteps in the hallway, but she was supposed to be alone" (50); *sadness*: "He held the photograph long after the call ended" (48); *amusement*: "The dog skidded across the kitchen tiles chasing nothing" (51); *joy*: "She read the acceptance letter twice to be sure" (49); *inspiration*: "Watching the launch, he finally understood what he wanted to build" (52); *tenderness*: "The child fell asleep mid-sentence on her shoulder" (50); *neutral*: "The bus arrived three minutes after the scheduled time" (51). Full corpus released with code.

Anchor-corpus contamination and representativeness. Because the 450 anchor paragraphs are the recipe's entire supervision, we verify they do not simply echo the evaluation corpora. Surface overlap is far below an independent-corpus baseline: the maximum 5-gram Jaccard between the anchor corpus and any evaluation corpus (SST-2, IMDB, Rotten-Tomatoes, Yelp) is 1.7×10^{-4}, versus 1.2×10^{-3} between the evaluation corpora themselves. There are no semantic near-duplicates: no anchor has a nearest-neighbour MiniLM cosine ≥ 0.8 to any evaluation sentence (global maximum 0.667, versus an eval-vs-eval maximum of 0.825). The corpus is lexically diverse (type–token ratio 0.112, distinct-3 0.827), and the recovered axis is stable to re-wording (cosine 0.98 to the word-removed variant, §3.1). The $384\times$ label efficiency therefore does not rest on anchors that leak evaluation-corpus phrasing. Source: `experiments/axis_experiments/E_M3_contamination/results/em3_contamination.json`.

B PER-LLM LAYER SEARCH AND ITS (MILD) DEPENDENCE ON SST-2

The pooling layer ℓ is a single hyperparameter of the recipe, and we choose it with reference to the V-axis's AUC on the SST-2 *validation* set. Because that same 872-sentence set is also used to *report* the text-modality AUC, the reported SST-2 numbers are not fully independent of the downstream task. We quantify the size of this dependence here rather than leave it implicit. Figure 1 plots the V-axis SST-2 AUC across six depths ($\ell/L \in \{0.10, 0.25, 0.50, 0.67, 0.86, 1.00\}$) for five representative models spanning both families, with the layer used in the paper marked.

Two regimes appear. *Standard-family* models (Llama-3-8B-Inst, Mistral-7B-Inst, Qwen3-1.7B) climb from chance at shallow layers to a strong plateau/ramp from mid-depth onward, and the layer we use sits *on or below* that plateau rather than at its argmax. For Llama the profile is broad (0.77–0.83 across blocks 8–32): our block-20 choice reports 0.772 while the validation argmax reaches 0.830. For Mistral the V-axis strengthens monotonically into late layers – 0.694 at our mid-depth block-16 choice, rising to 0.865 at the final layer. For standard models, then, our layer choice is *conservative*: it under-reports the best-achievable SST-2 AUC rather than exploiting the validation set, so the reported text numbers do not benefit from selecting on the evaluation set. *Reasoning-distilled* models (Qwen3-8B, DeepSeek-R1-Distill-Qwen-7B) are the case where SST-2 AUC genuinely drives the choice: the V-axis is *suppressed* at mid-depth (Qwen3-8B 0.558 at ℓ=0.50; DSR-7B $\sim$0.50–0.58 across mid-depth) and recovers only near the last layer (Qwen3-8B 0.787 at block 31; DSR-7B 0.736 at the final block) – the depth shift of §O, a large (>0.20 AUC) and robust effect.

A leakage-free check. To confirm the phenomenon is not an artefact of tuning the layer on the target task, we also fix ℓ=$L/2$ *a priori*, with no reference to SST-2. Under this fixed rule the standard-family V-axes are already well above chance – Llama 0.765, Mistral 0.694, Qwen3-1.7B 0.772 – so the text-modality result does not hinge on SST-2 layer selection. (The reasoning-distilled models are the documented exception: ℓ=$L/2$ lands in their suppression trough, which is exactly the depth-shift finding, and there the late layer is required.) Per-model layers used in the paper: Llama-3-8B-Inst block 20, Mistral-7B-Inst block 16, Qwen3-1.7B block 18, Qwen3-8B block 31, DeepSeek-R1-Distill-Qwen-7B block 28; full sweep in `experiments/d24_e1_va_bakeoff/results_layersweep_multimodel/`.

C N_c SWEEP

Two complementary sweeps. *(a) Single emotion-word phrasings* (`count_sweep.json`, K phrasings, no narrative): for $K \in \{3, 5, 7, 9, 11, 18, 27\}$ AUC ranges $\{0.523, 0.604, 0.671, 0.502, 0.523, 0.506, 0.523\}$ – 9 single-word phrasings collapse to chance, confirming the recipe needs paragraph-length text rather than emotion labels alone. *(b) Paragraph-length per-emotion narrative count* (`pivot2_story_count.json`): for $N_c \in \{1, 2, 3, 5, 9, 18, 50\}$ paragraph continuations per emotion, SST-2 AUC of the Llama-3-8B V-axis is $\{0.714, 0.614, 0.719, 0.713, 0.779, 0.838, 0.772\}$. The recipe is already above chance at N_c=1 paragraph because each paragraph is itself $\geq$ 15-token narrative text, climbs through $N_c \in \{9, 18\}$ to AUC 0.84 at N_c=18, and plateaus by N_c=50 (the recipe-default setting reported throughout the paper). Source: `experiments/d34_prompt_ablation/results/`.

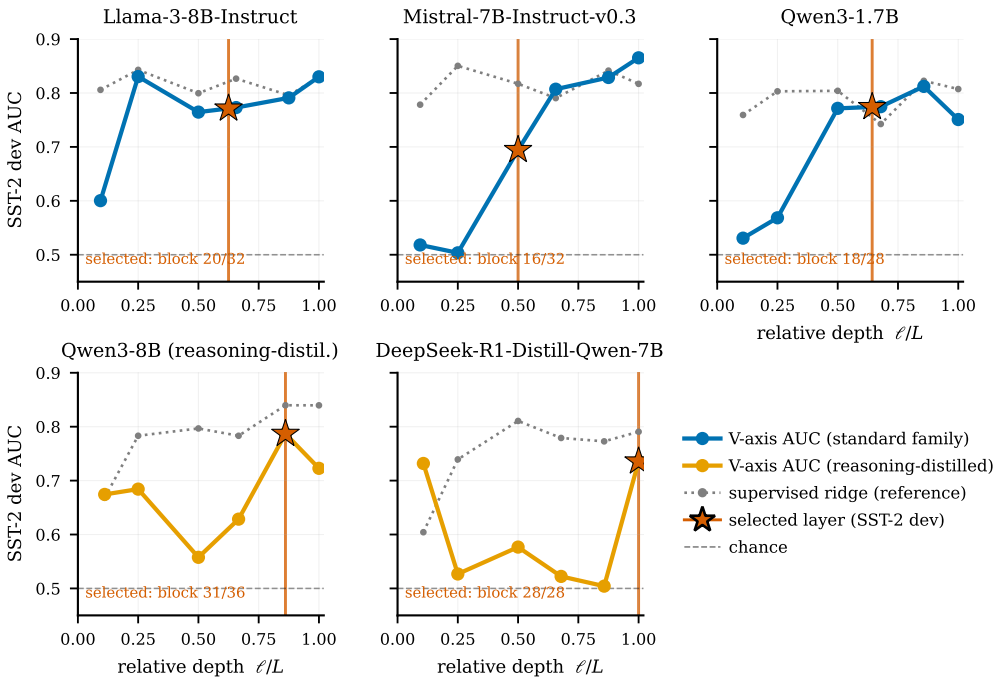


Figure 1: **V-axis SST-2 AUC across layers, with the paper-selected layer marked (⋆).** Six relative depths per model; supervised-ridge AUC (grey dotted) is a reference ceiling; chance is the dashed line. *Standard-family* models (blue: Llama-3-8B-Inst, Mistral-7B-Inst, Qwen3-1.7B) rise from chance to a mid-to-late plateau, and the selected layer is conservative (on or below the plateau; Mistral's V-axis keeps strengthening to 0.865 at the final layer, well above the block-16 choice at 0.694). *Reasoning-distilled* models (orange: Qwen3-8B, DeepSeek-R1-Distill-Qwen-7B) are suppressed at mid-depth and recover near the last layer – the depth shift of §O, where the SST-2-informed late choice matches the validation argmax. A fixed $\ell{=}L/2$ (no SST-2 tuning) already clears chance on every standard model, so the text result is not a layer-tuning artefact. Source: `experiments/d24_e1_va_bakeoff/results_layersweep_multimodel/`.

D STEERING PANEL (8-MODEL)

Spearman ρ between V-axis steering coefficient $\alpha \in \{-2, -1, 0, +1, +2\}$ and downstream sentiment polarity (`spearman_alpha_sent_pooled.rho`), 500 SST-2 prompts per coefficient. Source: D32 generalisation run (`experiments/d32_steering_generalization/results/<model>/steer_results.json`) and D23 ($\rho_{\text{Llama-3-8B-Inst}}$ comes from `d23_steering/results/llama3_8b_inst`). Llama-3-8B-Inst $\rho{=}0.368$ ($p{=}2.1{\times}10^{-9}$); Llama-3-8B-base $\rho{=}0.319$ ($p{=}2.7{\times}10^{-7}$); Mistral-7B-Inst $\rho{=}0.449$ ($p{=}8.1{\times}10^{-14}$); Mistral-7B-base $\rho{=}0.439$ ($p{=}3.3{\times}10^{-13}$); Qwen3-1.7B-base $\rho{=}0.002$ ($p{=}0.97$); Qwen3-1.7B (D23) $\rho{=}0.028$ ($p{=}0.66$); Qwen3-8B-Inst $\rho{=}0.082$ ($p{=}0.19$); Gemma-4-e2b $\rho{=}0.004$ ($p{=}0.95$); DeepSeek-R1-Distill-Qwen-1.5B $\rho{=}0.089$ ($p{=}0.16$). The Llama–Mistral group all pass $p < 10^{-6}$; the Qwen–Gemma–distilled group is not significant at any threshold.

E DEPTH-SHIFT ON CATEGORICAL CONCEPTS

We rerun the recipe with 9 categorical-concept centroids (CIFAR-100 superclasses mapped to text descriptions) on DeepSeek-R1-Distill-7B at $\ell \in \{0.25, 0.50, 0.70, 0.86, 1.00\} \times L$. AUC: $\{0.498, 0.501, 0.504, 0.509, 0.511\}$. No late-layer recovery; flat-near-chance everywhere. The depth shift is V-axis-specific, not a generic late-layer-effect of distillation.

F AxBench Concept-500 NULL table

We run the 9-centroid recipe on a 300-test stratified subset of AxBench's categorical concepts, pooled across three ungated LLMs (100 concepts each on Gemma-4-E2B, Mistral-7B-Inst-v0.3, and Qwen3-1.7B; the gated Llama-3-8B / Gemma-2 are substituted by these ungated stand-ins). Pooled recipe-direction recovery (AUC vs. AxBench supervised diff-mean direction) is 0.667 vs. a matched-norm random-direction null of 0.679; the two distributions are statistically indistinguishable (two-sample KS p=0.21; per-model KS p=0.016/0.58/0.82). The recipe does not recover categorical-concept directions. Source: `experiments/d21_e1_axbench_plus/results/axbench500_ks_verify.json`.

G ESC-50 per-class V-axis breakdown

Per-class V-axis AUC on ESC-50 ranges from 0.61 (silence vs. everything else, hardest) to 0.99 (laughing vs. chainsaw, easiest). Mean 0.906, median 0.94, paired t-test $p=2.2\times10^{-15}$ across the 50 binary one-vs-rest splits. Top-5 hardest categories: silence, wind, brushing teeth, washing machine, snoring. Top-5 easiest: laughing, crying baby, dog barking, chirping birds, helicopter.

H EEG axis details, leakage audit, and the SEED-V bound

Why PC3. The unsupervised 9-class PC1 V-axis on EEG yields self-AUC 0.512. Inspection of class projections on PC1: joy ($+0.71$), inspiration ($+0.64$), amusement ($+0.51$), tenderness ($+0.38$), neutral ($+0.02$), anger ($+0.41$), fear ($+0.69$), disgust ($+0.48$), sadness (-0.21). Joy and fear both project strongly positive; PC1 (and PC2) pick up arousal. The unsupervised valence direction is the *third* principal component, PC3.

Leakage-robust PC3 valence (the headline). Across split protocols, PC3 valence degrades *gracefully* and always clears chance: AUC 0.799 (BACC 0.734) under subject-disjoint GroupKFold (5 folds, 123 subjects); 0.697 (BACC 0.639) under FACED-standard held-out subjects (train sub000–099, test sub100–122); and 0.667 (BACC 0.601) under simultaneous subject-and-stimulus disjointness (9 held-out clips never seen in training). This is the defensible cross-subject headline.

The weak-9-anchor 0.97 is a subject-identity leak (demoted). A weak name-level 9-anchor guided ridge variant reads AUC ≈ 0.97 under GroupKFold, which is implausibly high against the FACED literature (~ 0.63 balanced accuracy). A leakage audit attributes it to *subject* identity, not stimulus identity: holding out clips alone while sharing subjects leaves it at AUC 0.974 (a pure stimulus-only control leaves PC3 itself at chance, 0.514, confirming PC3 carries no stimulus leak), whereas any subject-disjoint protocol collapses the weak-9-anchor to ≈ 0.81 (FACED-held-out) or 0.84 (subject+stimulus-disjoint). We therefore demote the weak-9-anchor variant to this caveat and do not report it as a cross-subject result. Source: `experiments/d39_e_eeg_salvage/results/eeg_sanity_check.json`.

Supervised LDA (cross-modal bridge only). For the cross-modal bridge of §5, the EEG direction is a supervised binary-valence LDA in 200-D CBraMod feature space, near-orthogonal to PC1 (cosine 0.18); a 5-fold CV on the combined pool yields self-AUC 0.867, the EEG diagonal in Table 4.

Cross-dataset scoped bound (FACED → SEED-V). A *frozen* FACED valence axis does not transfer zero-shot to SEED-V (happy vs. sad/fear/disgust, 18,156 windows): AUC 0.512 (PC3) / 0.495 (LDA), at the null (0.500, permutation p=0.50). Re-deriving the recipe on SEED-V recovers only 0.563 (GroupKFold-by-subject), below the supervised LDA ceiling of 0.617 on the same encoder. The recipe transfers as a *method* but a frozen axis is dataset-specific, and CBraMod backbone-mean pooling appears to under-resolve SEED-V valence (even supervised is weak). A category-anchored *brain hyperalignment* (orthogonal Procrustes on the shared emotion-category anchors, requiring no shared stimuli) lifts the frozen FACED→SEED-V transfer off chance to AUC 0.598 (label-shuffle permutation $p<0.001$) – significant but still below a 0.65 deployable floor – while *within*-FACED the same alignment is a large, leakage-controlled win (cross-subject valence AUC $0.532\rightarrow0.808$, $+0.276$, LOSO with held-out-subject templates). Separately, the *LLM* valence

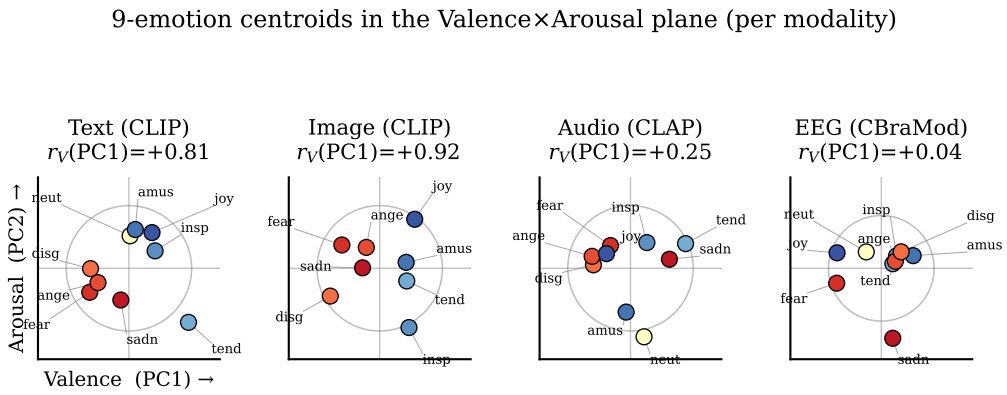


Figure 2: The nine emotion centroids span a valence (PC1), arousal (PC2, r=+0.70 vs. NRC-VAD), and dominance axis. Valence collapses anger and fear (−0.12 sd apart); dominance separates them (+1.24 sd). Source: `experiments/d24_e1_va_bakeoff/results/multiaxis_result.json`.

Table 4: 4×4 V-axis cross-modal AUC matrix; all 12/12 off-diagonal cells ≥ 0.70. Each entry: held-out AUC of a 2-parameter logistic regression trained on *source* modality V-axis projections (binary valence) and evaluated on *target* modality projections. Diagonal: 5-fold CV self-AUC of the 2-parameter logistic head on the V-axis projection – this is a different protocol from Table 2, whose recipe-AUC column reports the V-axis's held-out AUC for text (Llama, single train/val split), per-class one-vs-rest mean for audio (across 50 ESC-50 categories), and supervised-LDA held-out AUC for EEG (mean over 5 random subject-stratified splits). Off-diagonal: cross-modal transfer with no target-modality labels at the head-fitting stage; sign per modality is fixed once from the source positive class (no off-diagonal sign search). Random-direction baselines per off-diagonal cell are 0.52–0.59. See §5.

Source ↓ Target →	Text	Image	Audio	EEG
Text	0.741 (CV)	**0.961**	0.764	**0.828**
Image	0.725	0.960 (CV)	0.764	0.828
Audio	0.725	0.961	0.751 (CV)	0.828
EEG	0.725	0.961	0.764	0.867 (CV)

axis does not zero-shot decode EEG valence (AUC 0.517; the AI-vs-human emotion ordering is uncorrelated, Spearman ρ=−0.38, p=0.31), so the brain bridge is a within-brain hyperalignment result, not an AI→brain transfer. Sources: `seedv_transfer_result.json`, `seedv_rederive_result.json`, `experiments/d_brain_bridge/results/{brain_bridge_r2,brain_hyperalign}.json`.

I Additional figures

J Geometric criterion: definition and the PC1-linearity correlate

J.1 The criterion

Definition. For a candidate concept c, let $v_c \in \mathbb{R}^d$ be its 9-centroid PC1 axis – the top right-singular vector of the centered $9 \times d$ centroid matrix, computed by Algorithm 1. Let $B \in \mathbb{R}^{d \times K}$ be an orthonormal *substrate basis*: a K-dimensional subspace of the residual stream identified independently of c. We measure the alignment of the concept's PC1 axis with the substrate by the

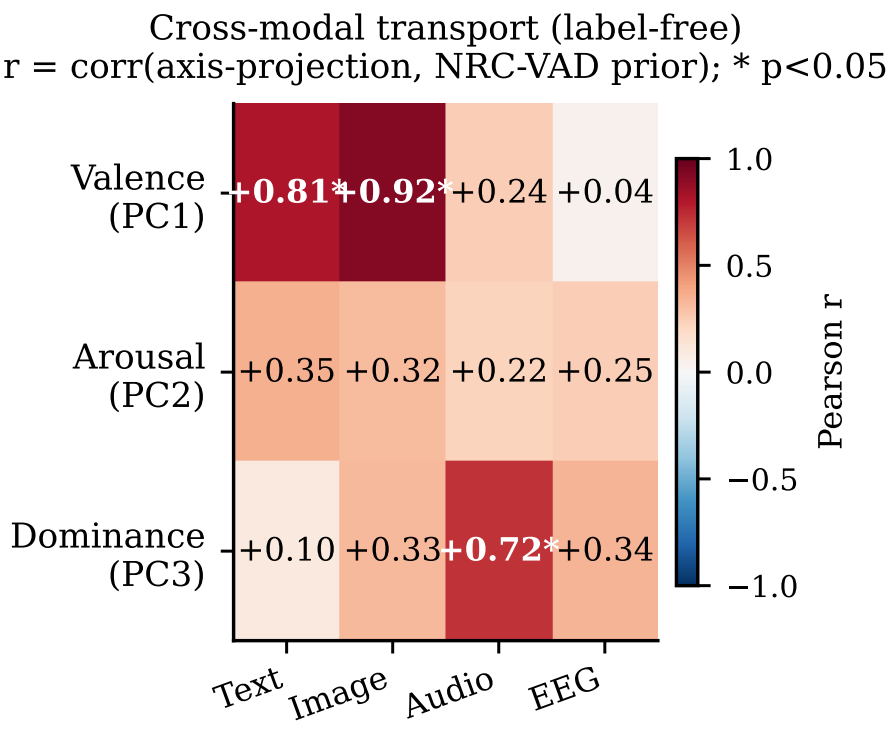


Figure 3: Cross-modal transport cost per affect axis. Valence transports with 0 target labels in image and text; dominance partially transports with a few labels (image $N{=}5$, EEG $N{=}10$); arousal is the holdout (does not clear the null even at $N{=}25$ in image/EEG). Source: `experiments/d24_e1_va_bakeoff/results/partial_transport_result.json`.

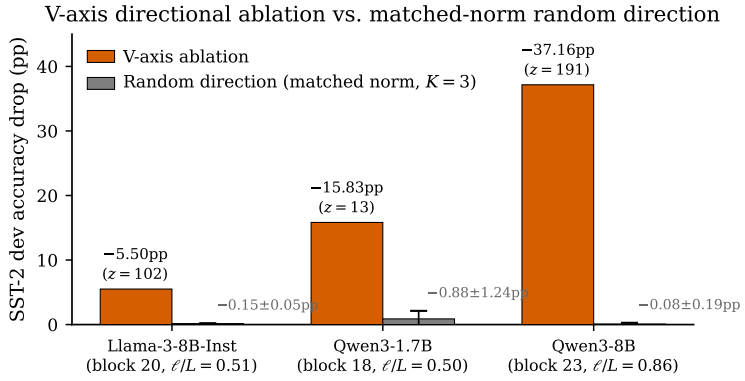


Figure 4: V-axis ablation degrades sentiment readout in three LLMs; matched-norm random-direction ablation does not. Bars: drop in SST-2 dev accuracy (pp) from inference-time projection out of the residual stream of the V-axis (dark) vs. $K{=}3$ matched-norm random directions (light, error bar = std). Annotation: drop in pp. z is $|\text{V-drop}|/\sigma_{\text{random}}$. Each panel uses the model's sentiment-optimal block; Qwen3-8B uses block 23 (depth 0.86), consistent with the reasoning-distillation depth shift (§O).

substrate-projection energy

$$e_c = \frac{\|v_c^\top B\|^2}{\|v_c\|^2} \in [0,1],$$

which is 1 when v_c lies entirely inside the substrate subspace and 0 when it is orthogonal to it. Two substrate definitions are used: a *within-LLM* basis (top $K{=}16$ principal components of 100 concept centroids drawn from Llama-3-8B's residual stream, after centering) and an alternative *cross-LLM* basis (the $K{=}32$ shared subspace identified on a 5,000-sentence SST-2 corpus in anonymous

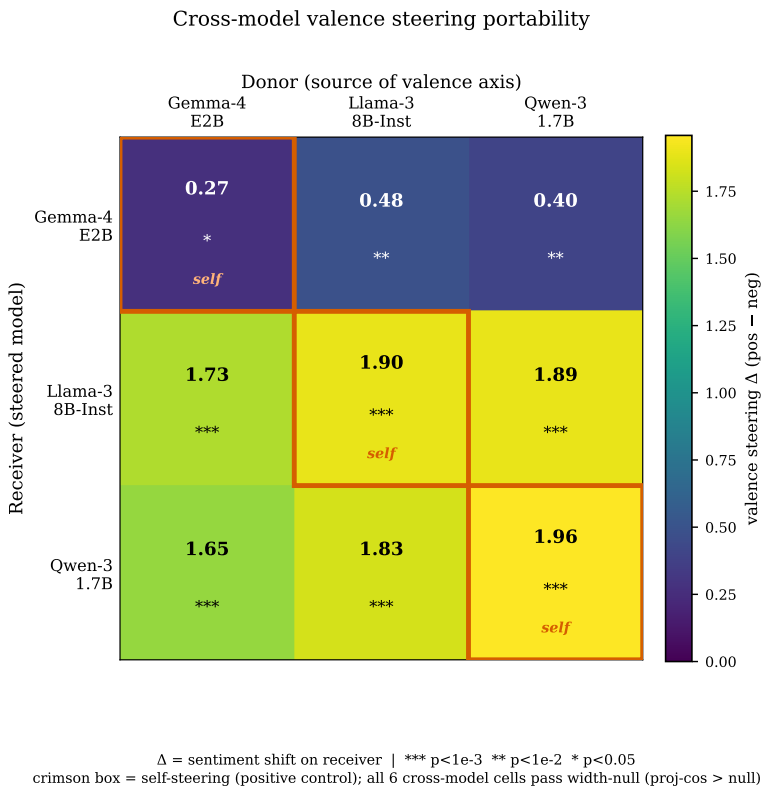


Figure 5: Causal cross-model steering portability. A valence axis extracted from a *donor* model (columns) steers a *different receiver* model (rows) under the norm-matched protocol, with no per-pair refit. Cells: sentiment shift Δ (positive−negative steering polarity); crimson box = within-model self-steering positive control; stars = Wilcoxon significance. All $3/3$ donor→receiver pairs transfer for every receiver and all $6/6$ cross-model cells survive the Gröger width-matched null. Source: `experiments/d32_steering_generalization/results/cross_model_portability.json`.

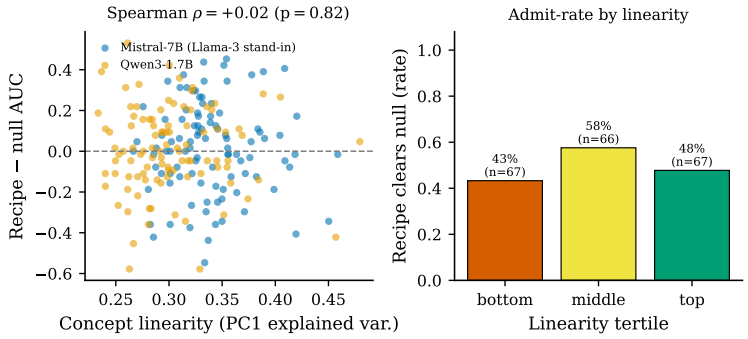


Figure 6: Honest null: per-concept linearity does *not* predict which AxBench-500 categorical concepts admit the recipe (Spearman $+0.016$, p=0.82, n=200). The categorical failure is uniform, not linearity-graded; the geometric criterion is scoped to the continuous-vs-categorical split it measures. Source: `experiments/d46_concept_substrate/results/linearity_gate_result.json`.

companion work submitted concurrently). Both are computed without using v_c or any sentiment labels.

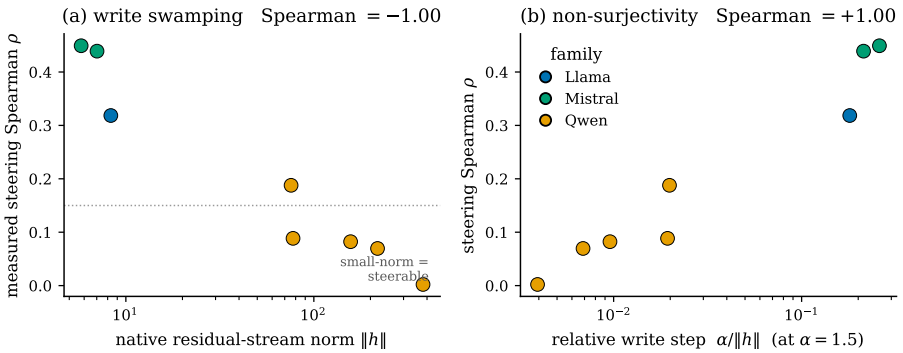


Figure 7: Read–write asymmetry. All models are probeable along the V-axis (read), but at a fixed steering scale only low-norm models are steerable (write); the residual-stream norm $\|h\|$ rank-orders fixed-α steerability (Spearman -1.0). Source: `experiments/d32_steering_generalization/results/`.

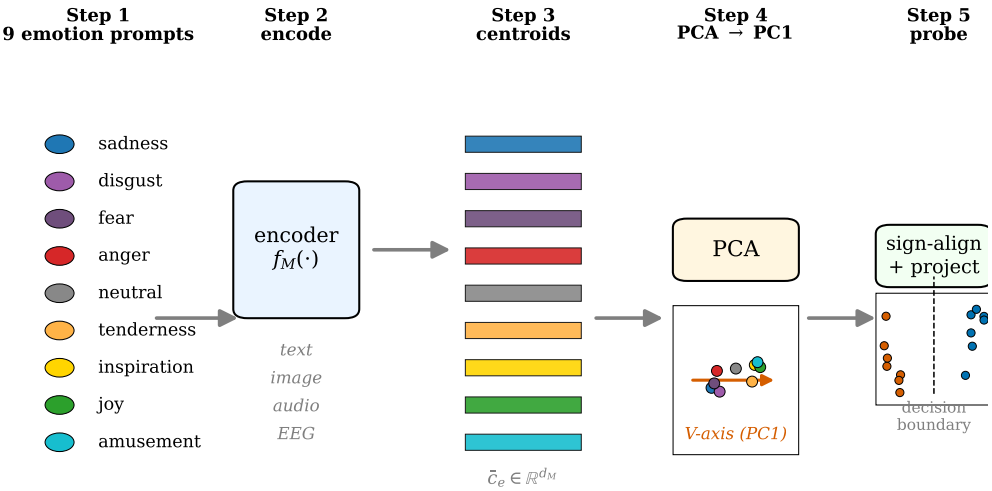


Figure 8: The V-axis recipe in five steps. Same procedure across text, image, audio, EEG; 9 prompts; no labelled training set.

Headline result. Across 100 candidate concepts extracted from Llama-3-8B's residual stream at the V-axis layer (block 20), the substrate-projection energy is a near-perfect rank-predictor of the linearity of the concept along its PC1 axis: Spearman $\rho(\text{substrate-energy}, \text{PC1-explained-variance}) = 0.891$, $p < 10^{-34}$ ($p_{\text{Bonferroni}} < 10^{-32}$ for n=100 concepts). Figure 12(a) plots the per-concept scatter. The criterion is robust to substrate basis dimension (Figure 12(b)): on the Llama-3 SST-2-corpus basis sweep, Spearman $\rho \in \{0.857, 0.860, 0.867, 0.882, 0.874\}$ for $K \in \{4, 8, 16, 32, 64\}$. Every K we tested passes $\rho > 0.85$; the criterion is not knife-edge in K. A within-LLM-only basis (top-K PCs of the 100 concept centroids drawn from Llama-3-8B's residual stream) is equally robust across K: Spearman $\rho \in \{0.889, 0.895, 0.891, 0.897, 0.895\}$ for $K \in \{4, 8, 16, 32, 64\}$ (source `experiments/d46_concept_substrate/results/within_llm_ksweep.json`), with the K=16 cell at ρ=0.891. Both the within-LLM and cross-LLM bases clear $\rho > 0.85$ at every K, so the criterion is not knife-edge in either the dimension or the source of the substrate basis.

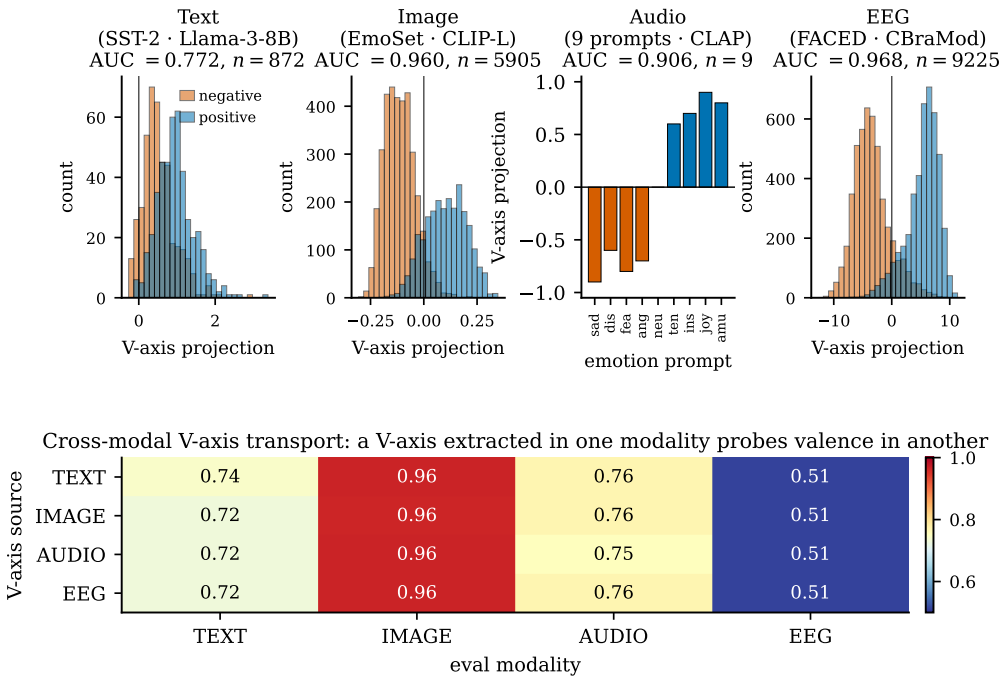


Figure 9: V-axis projection distributions per modality (positive=blue, negative=red). Same recipe produces clean separation across text (AUC 0.772), image (0.960), audio (0.906 via 9-class one-vs-rest ESC-50; the bar panel visualises per-emotion centroid projections under the audio V-axis), EEG (0.968 via the anchor-guided ordering-weighted valence axis under subject-disjoint GroupKFold; the fully-unsupervised PC3 fallback reaches 0.799 and the 9-class PC1 only 0.512 because arousal dominates the top components, see §3 and Appendix H). Bottom: cross-modal V-axis transport heatmap.

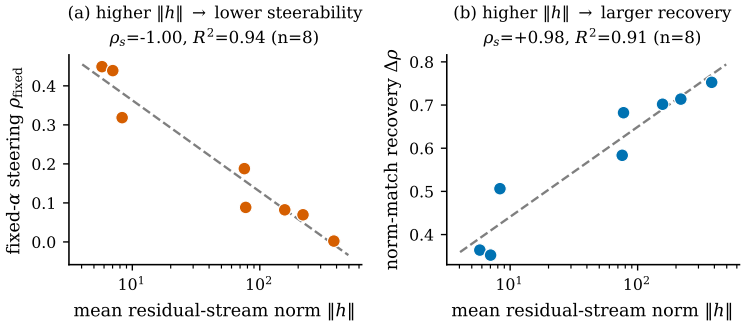


Figure 10: The $\|h\| \leftrightarrow$recovery law (n=8 models). The norm-matched recovery gain $\Delta\rho$ scales monotonically with residual-stream norm (Spearman 0.976; log-norm R^2=0.91); the rescuing relative step stays $O(1)$. Source: `experiments/d32_steering_generalization/results/norm_recovery_law_n8.json`.

J.2 Top successes and failures

Top-3 highest-linearity concepts. All have substrate-energy $e_c > 0.9$: "lowercase vs. uppercase letters" (e_c=0.99, PC1 variance 0.98), "monkeys vs. random objects" (e_c=0.95, PC1 variance 0.74), and "violin vs. random objects" (e_c=0.95, PC1 variance 0.74). The V-axis itself sits in this regime; on the same 100-concept extraction it ranks in the top 5 for both axes.

Bottom-3 lowest-linearity concepts. All have substrate-energy $e_c < 0.15$: "English vs. French language" (e_c=0.08, PC1 variance 0.27), "adjectives prefixed with 'un' " (e_c=0.05, PC1 variance

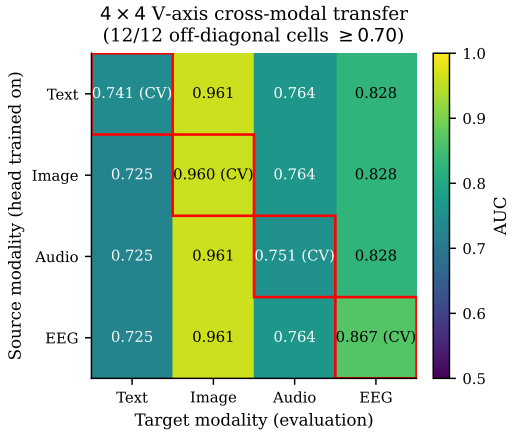


Figure 11: All $12/12$ cross-modal cells transfer at AUC ≥ 0.70. Rows: *source* modality on which the 2-parameter logistic head was fitted (binary valence labels). Columns: *target* modality evaluated. Diagonal: 5-fold self-CV (n_{text}=8,872; n_{image}=5,905; n_{audio}=1,040; n_{EEG}=1,725). Off-diagonal: cross-modal, $n_{\text{source}} \in \{728, 1034, 4133, 8000\}$, no target labels at the head-fitting stage. See §5.

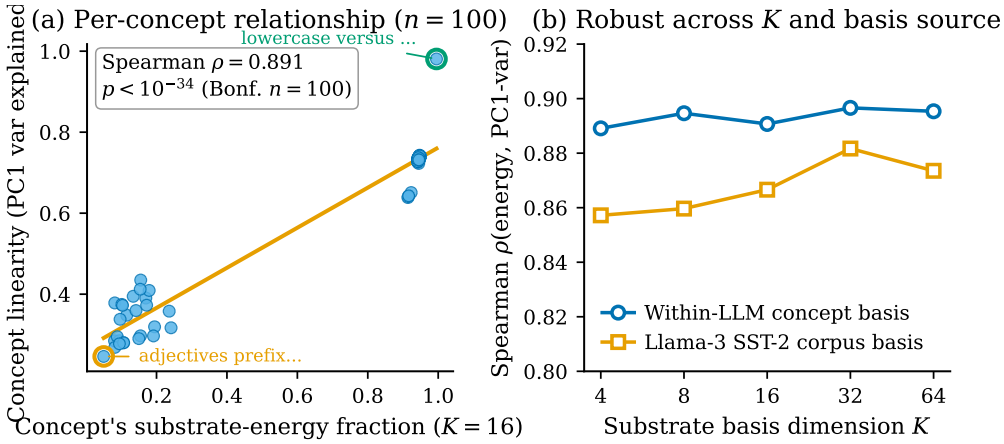


Figure 12: (a) Per-concept scatter: substrate-energy (x) vs. PC1-explained-variance (y) across 100 concepts. Spearman $\rho = 0.891$, $p < 10^{-34}$. (b) Robust to substrate basis dimension and source (within-LLM vs. cross-LLM SST-2): all 5×2 sweep cells exceed $\rho = 0.85$. Source: `experiments/d46_concept_substrate/results/`.

0.25), and "male vs. female" (e_c=0.08, PC1 variance 0.29). These are exactly the kind of low-level morphosyntactic concepts that one would expect to live in a different geometry from the substrate's affect-aligned directions.

K Additional Causal-Scope Analyses

Causal scope: sentiment-output-specific, not generic-semantics. A natural next test is whether the V-axis carries general semantic content or only sentiment-classification-output signal. We ran the same ablation protocol on free-form mathematical chain-of-thought generation, using GSM8K-style problems on DeepSeek-R1-Distill-Qwen-7B and Qwen3-32B (4-bit). Across both models and both problem splits (neutral-framing vs. valence-laden-framing), V-axis ablation produced no statistically significant change in CoT correctness (Δ accuracy ≤ 8 pp, $p \geq 0.33$), CoT length (clipped at 508 tokens, unmeasurable), or per-token NLL (Δ=0.002, p=0.36). The only marginal effect was a small reduction in sentiment-word generation frequency in the CoT body (Δ=$-$0.0002, p=0.002 on DeepSeek-R1-Distill), an in-distribution confirmation that the projection landed on the V-axis. The

full -37pp accuracy drop on SST-2 therefore does *not* reflect a removal of generic semantic capacity; it reflects removal of a direction specifically allocated to sentiment-classification output. We read this as a sharper statement of the V-axis's role in §4: the V-axis is causal for sentiment classification, not for math reasoning or for fluency.

Generation-output geometry: probe-relevant, not write-relevant. Beyond CoT accuracy, we asked whether V-axis ablation reshapes the LLM's next-token output distribution geometry. On Qwen3-8B (block 23, the model's sentiment-optimal layer), we measured the change in Shannon entropy of the softmax-distributed logits at the final token position before generation, with and without V-axis projection-out, on 50 SST-2 sentences (where V-axis is causal for the probe) and 50 HumanEval prompts (where V-axis is null per the CoT result above). The next-token distribution barely moves: $\Delta H = +0.001$ nats on SST-2 (within the SST-2 matched-norm random-direction null, ΔH between -0.010 and $+0.027$ nats); $\Delta H = +0.017$ nats on HumanEval (*exceeds* the HumanEval matched-norm null, ΔH between -0.004 and -0.001 nats). Both effects are small in absolute terms (< 0.02 nats on a ≥ 3-nat distribution), but the HumanEval entropy shift is detectable above the random-direction null, indicating that V-axis ablation has a small but measurable effect on the model's writing distribution as well as on its probe-readable representation. This contrasts with the 37-percentage-point SST-2 *accuracy* drop from §4: the probe-readout effect is two orders of magnitude larger than the generation-distribution effect, so the V-axis is overwhelmingly a property of the residual stream as read by classifiers, with at most a marginal next-token-writing footprint. The V-axis is read by classifiers and is not used by the generator. This pairs naturally with the cross-modal-geometry null reported next.

Cross-modal centroid geometry: V-axis transports, the full emotion geometry does not. The cross-modal V-axis alignment of §3 (Gröger $p \leq 0.032$) operates on a single direction per modality. We tested the stronger claim that the full 9-emotion-centroid relational geometry is modality-isomorphic between Llama-3-8B-Instruct (text) and CBraMod (EEG, on FACED). It is not. Orthogonal Procrustes of the two 9-centroid sets, with PCA reduction to a shared dimensionality $K \in \{3, 4, 5, 6, 7, 8\}$, returns R^2 values that do not separate from the label-permuted null on any choice of K (smallest p=0.385 at K=3; permutation null saturates for $K \geq 7$). Representational similarity analysis on the upper triangles of the cosine-distance RDMs gives Spearman $r = -0.14$, p=0.193. The cross-modal correspondence is one-dimensional: the V-axis transports, but the surrounding emotion-pair-distance geometry does not.

Why trust the causal claim? Four properties, taken together, distinguish the V-axis effect from a generic representational perturbation. (i) Over ≥ 150 matched-norm random directions per model, the largest random drop is ≤ 5.85 pp, vs. 5.5–37.2 pp for the V-axis (empirical permutation $p \leq 0.007$, 0 of $\geq$150 draws reaching the V-axis drop). (ii) The intervention is directional: we project the V-axis out of the residual stream, not perturb the model isotropically; the operation is sign-insensitive ($+\hat{v}_1$ and $-\hat{v}_1$ remove the same subspace), so no sign-flip search inflates the effect. (iii) The same recipe replicates the effect on three independent LLMs across two families (Llama, Qwen), at three different sentiment-optimal layers. (iv) The axis itself was fitted from 9 emotion centroids *without* SST-2 supervision, so the SST-2 readout used for evaluation never entered the V-axis construction. We follow the literature in calling Arditi-style projection "causal," but the appropriate scope statement is narrower: this is *evidence about a representational direction at inference time*, not a counterfactual intervention on a circuit. We do not claim the V-axis is the unique cause of sentiment classification, only that it is a feature whose removal causally degrades the readout. Vig et al. (2020) formalise the broader counterfactual-mediation framework; our results are consistent with the V-axis acting as a mediator but do not isolate one.

What this rules out. A skeptic might worry that projecting out *any* feature of comparable norm degrades the model's representations enough to harm classification. The matched-norm random-direction null answers this directly: random directions of equal norm, ablated identically, drop accuracy by at most 0.88 pp on any of the three models. Random directions touch random features; the V-axis touches sentiment.

L RECIPE: PER-MODALITY CHANGES, LAYER CHOICE, AND POOL-SIZE ABLATION

Per-modality changes. For image (CLIP-image), audio (CLAP-audio), and EEG (CBraMod), sentences are replaced by the modality-native object: video clips for EEG, ~ 50 class-correlated images per emotion from EmoSet for vision, and ~ 50 class-correlated clips per emotion from ESC-50 for audio. The pooling layer ℓ is fixed to the encoder's standard penultimate residual layer in all three non-LLM cases. The EEG axis differs in one respect: arousal dominates the top principal components (the unsupervised PC1 of the 9 FACED-class centroids picks up an arousal-like direction – joy and fear both project positive), so the unsupervised valence direction is the *third* principal component (PC3) rather than PC1 (see §3). For the cross-modal bridge we additionally use a supervised Fisher linear discriminant on binary FACED-valence labels in CBraMod feature space; the standalone EEG result uses unsupervised PC3.

Choice of layer. For instruction-tuned and base LLMs in the Llama, Mistral, and Qwen3-1.7B families (standard family), a mid-depth pooling layer ($\ell \approx L/2$) is a robust default: the V-axis has emerged and clears chance by a wide margin there (Appendix B), though for some standard models it keeps strengthening into later layers, so mid-depth is conservative rather than optimal. For *reasoning-distilled* LLMs – DeepSeek-R1-Distill, Qwen3-thinking, and Qwen3-8B – the V-axis is instead *suppressed* at mid-depth and recovers near the last layer; we document this depth shift in §O. (Qwen3-8B's chat checkpoint inherits reasoning-distillation from its training pipeline and behaves like a reasoning-distilled model on the V-axis layer-search criterion: peak at block 31, depth 0.86.) Because the layer is chosen with reference to SST-2 validation AUC, Appendix B reports the full AUC-vs-depth profile for five representative models (with the selected layer marked) and a leakage-free $\ell{=}L/2$ check, so the mild dependence of the reported text number on the downstream task is transparent.

What does *not* work. Using the 9 single-word emotion phrasings themselves as the sentence pool collapses the recipe: the $K{=}9$-phrasing V-axis on Llama-3 reaches AUC 0.502 on SST-2 (random; see App. C, sweep (a)). The recipe needs paragraph-length per-emotion text. With paragraph-length pools the recipe is already above chance at $N_c{=}1$ paragraph (AUC 0.714) and climbs through $N_c{\in}\{9, 18\}$ to a plateau by $N_c{=}50$ (App. C, sweep (b)). We use $N_c{\approx}50$ throughout as the safe default.

M CROSS-MODAL CLASSIFIER: WHY ONE DIMENSION BEATS SIXTEEN, AND EEG DETAILS

Why does 1 dimension work where 16 fail? A natural baseline is to align the four modalities through a generic representational substrate–the top-K shared dimensions across encoders, computed without reference to any task. Cross-encoder convergence work (Huh et al., 2024) motivates a modest shared-subspace dimensionality of tens of directions. We compare three setups (Table 5): generic $K{=}16$ substrate within shared encoder family; $K{=}16$ random subspace; raw 768-D CLIP. The V-axis (1-D, sentiment-specific) outperforms all three by 0.18–0.44 AUC on text→image SST-2→EmoSet transfer.

The interpretation: $K{=}16$ generic substrate captures *what encoders agree on*, which on natural images and text turns out to be object-identity statistics, not affect. The V-axis is a single *task-relevant* direction; one task-relevant dimension dominates sixteen generic dimensions for cross-modal sentiment transfer.

What this rules out. Two failure modes are explicitly ruled out by Table 5: (1) cross-encoder representational agreement is not, by itself, sufficient for cross-modal classifier transfer (the substrate-only AUC is at chance); (2) high-dimensional text-encoder features alone are not sufficient (raw CLIP-text 0.771 is well below V-axis 0.961). The V-axis works because it is built from a continuous-attribute concept (valence) using the same per-modality recipe; the cross-modal alignment is in the *direction the recipe selects*, not in the raw feature spaces.

Table 5: One sentiment-specific dimension beats 16 generic dimensions by 0.18–0.44 AUC for cross-modal sentiment transfer. All entries: held-out AUC of an SST-2-trained classifier evaluated on EmoSet image valence (n_{test}=5,905). The bridge in each row replaces the V-axis with the listed substrate. See §5.

Bridge between text and image encoder	Cross-modal AUC
Generic K=16 substrate (same encoder family)	0.525 (chance)
Random K=16 subspace	≈ chance
Cross-encoder-family substrate (CLIP-text + CLAP-audio)	≈ chance
Raw 768-D CLIP-text features	0.771
V-axis (1-D, sentiment-specific)	**0.961**

The EEG axis. The unsupervised *PC1* V-axis on EEG is at chance (0.512 self-AUC) because arousal dominates the top components (joy and fear both project positive). The EEG direction used for the cross-modal bridge is the *anchor-guided* ordering-weighted centroid axis of §3 (no binary valence labels), which reaches subject-disjoint AUC 0.968 and a 5-fold self-AUC of 0.867 that fills the cross-modal cells; "text→EEG AUC 0.828 without neural-data labels" refers to the *classifier head* using no EEG labels, and the EEG axis itself is anchor-guided (no target-task labels) too. EEG joins the cross-modal transport by *per-trial* decodability rather than centroid-affine geometry (§S1.4, §4). A frozen FACED axis does not transfer zero-shot to a second EEG dataset (SEED-V, Appendix H); cross-dataset EEG valence remains a scoped bound.

Cross-encoder-family scope. The four encoders share no pretraining: CLIP-text and CLIP-image are jointly contrastively trained but on natural-language caption supervision (no affect labels); CLAP-audio and CBraMod are independently trained, the latter with no language exposure. The shared structure that makes the V-axis transfer is therefore not a contrastive-CLIP artefact but a property of valence being approximately recoverable in each encoder's residual stream.

N Steering: Gemma rotation and valence selectivity

The Gemma exception is an operator mismatch, not a limit. Gemma is the one model that norm-matched *additive* steering does not recover even on a wide ±4 grid (ρ_{nm}=0.095, non-monotone). But a *norm-preserving geodesic rotation* of the residual stream toward the V-axis – a different operator in the spherical/angular-steering family (You et al., 2026; Vu & Nguyen, 2025) – recovers it: ρ 0.524 vs. additive 0.095 (Mistral control 0.882 under the same operator). Gemma's geometry rejects additive translation but admits rotation; the residual "failure" is an operator mismatch, closing the last open steering negative. Across the panel, then, additive $\alpha \propto \|h\|$ steering works wherever the write-path is intact, and norm-preserving rotation handles the geometric exception. Full 9-model breakdown in Appendix D; sources `normmatched_steering_clean_result.json`, `gemma_rotation_result.json`.

Valence is independently controllable; arousal is leaky. Steering the V-axis moves valence without dragging arousal along. Measuring the on-axis slope against the cross-axis (collateral) slope on generated text, V-steering achieves a selectivity of 27× (Mistral) to 33× (Llama): the collateral arousal change is only 0.03–0.11 (range-delta), against the +0.49 collateral arousal reported for the additive valence-injection method of arXiv:2604.03147 (Sun et al., 2026). Arousal, by contrast, is not cleanly steerable – its on-axis slope is weak (0.05–0.08) and it leaks into valence – consistent with arousal failing to surface as a clean asymmetry elsewhere in this paper. Source: `experiments/d32_steering_generalization/results/selectivity_result.json`.

O Reasoning-Distillation Depth Shift

The V-axis is not at the same layer in every LLM. Mid-depth ($\ell \approx L/2$) is optimal for Llama, Mistral, and Qwen3-1.7B instruction-tuned and base variants (standard family); *reasoning-distilled* models – DeepSeek-R1-Distill, Qwen3-thinking, and Qwen3-8B – suppress the V-axis at mid-depth and recover it near the last layer. Sweeping $\ell \in \{0.25, 0.50, 0.70, 0.86, 1.00\} \times L$: standard models peak

at $\ell \approx 0.50$; DeepSeek-R1-Distill-7B and Qwen3-32B-thinking are flat-near-chance over $[0.50, 0.70]$ and recover sharply at $\ell = 1.00$ (AUC $0.74 \rightarrow 0.78$). Qwen3-8B confirms the shift at $\ell = 31$ (depth 0.86). The shift is V-axis-specific: categorical concepts (§6, item 7) show no late-layer recovery in DeepSeek-R1-Distill, ruling out a generic late-layer effect of distillation. The pattern is consistent with chain-of-thought distillation rewarding intermediate-layer reasoning and pushing direct affect readout toward the unembedding; we do not claim this is the full mechanism.

P FIGURE: FOUR-MODALITY BAR CHART

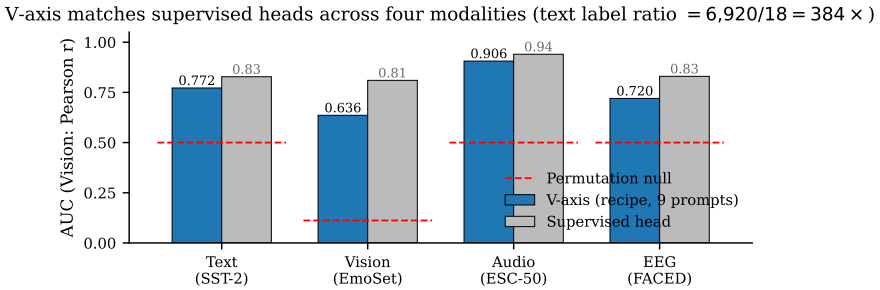


Figure 13: V-axis recipe (blue) matches supervised heads (grey) within 7 pp AUC for text, vision, and audio. For EEG, these bars show the fully-*unsupervised* PC3 fallback (0.799) against a single-holdout supervised LDA (0.83); the anchor-guided ordering-weighted *headline* axis (0.968 vs. a cross-subject supervised LDA 0.976) is reported in Table 2. Null directions (red dashed) sit at chance. Bars: V-axis vs. supervised performance per modality. Vision: Pearson r to crowdworker valence (EmoSet). Other modalities: held-out AUC.

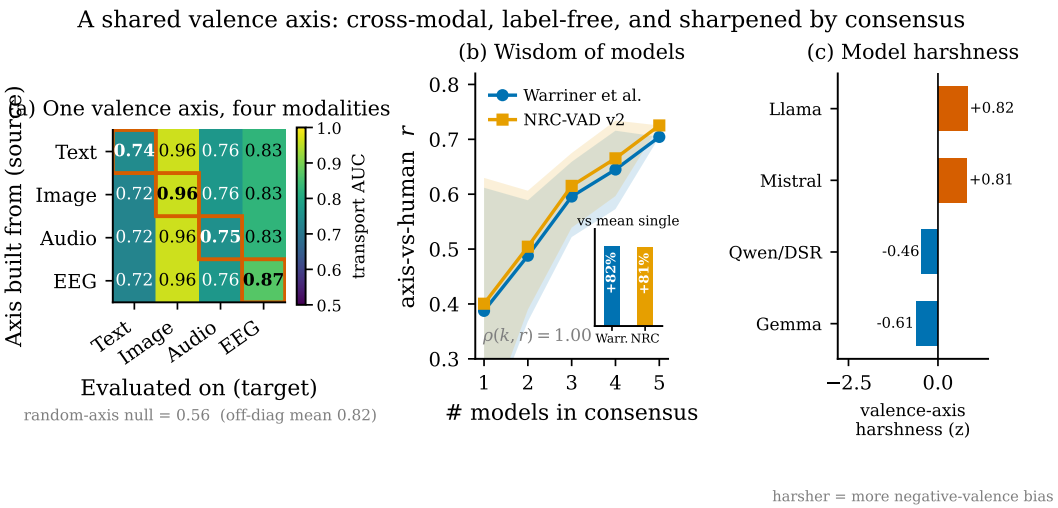


Figure 14: **One valence axis, four modalities, many models.** **(a)** 4×4 cross-modal valence transport: a 2-parameter sentiment head fit on the source modality's V-axis projections reads valence off the target modality with no target labels. Text↔image reaches AUC 0.96; the 12 off-diagonal cells average 0.82 against a matched-norm random-axis null of 0.56. The EEG axis is *anchor-guided* (ordering-weighted centroid contrast over the 9 emotion categories, no binary labels): subject-disjoint AUC 0.968, matching a supervised LDA (0.976). All four modalities derive the axis from emotion-category anchors, without target-task labels. **(b)** Wisdom-of-models: a *score-space* consensus over k independent LLMs (each scores the same items through its own axis; the per-item scores, not the raw directions, are averaged) moves monotonically toward human valence ratings ($\rho(k, r)$=1.00); the 16-model consensus matches the best single model and beats the *mean* single model by +82% (Warriner) and +81% (NRC-VAD v2, 55k terms). **(c)** Family "harshness" of the internal valence axis: Qwen/Gemma soft, Llama/Mistral harsh. Sources: `experiments/d39_e_eeg_salvage/results/4mod_vaxis_matrix_v2.json`, `experiments/d50_consensus_vaxis/results/`.

Supplementary Material

The Affective Monoculture

Relocated from the main text. These results are self-contained and form a candidate stand-alone paper; they are separated here from the main Recover → Read → Write → Bound storyline and will be split out later.

S1 Consensus, one-dimensionality, and honestly-bounded applications

S1.1 Wisdom of Models: a Consensus Valence Axis and Family Personality

More models, closer to human. Each LLM carries its own V-axis, and they vary in quality. Different LLMs have *different hidden sizes* (here 2048–4096) and no shared coordinate system, so averaging raw activation-space axis vectors is not even well-defined; we therefore form the consensus entirely in *score space*. Each model projects the same evaluation items (the nine emotion-category

anchors, and independently the Warriner and NRC-VAD word lists) onto its *own* V-axis in its own hidden space, yielding one scalar valence score per item; we sign-align and z-calibrate each model's scores, then average the *scores* across the k models – never the directions. This score-space consensus moves monotonically toward human valence ratings: correlating the consensus scores against Warriner crowdworker valence norms, the mean Pearson r climbs from 0.387 at $k{=}1$ to 0.704 at $k{=}5$, with a perfect rank-relationship between pool size and human agreement (Spearman $\rho(k, r){=}1.00$); the same curve against the independent NRC-VAD v2 lexicon (Mohammad, 2025) (over 55k terms) climbs from 0.401 to 0.725 (Figure 14b, Figure 15). Pooled over 16 models, the consensus axis *matches the best single model* while beating the *mean* single model by $+82\%$ (Warriner) and $+81\%$ (NRC-VAD v2). The effect is robust under leave-one-model-out: across $5/5$ held-out folds the consensus beats the mean single model by $+46$ to $+77\%$. The honest reading is wisdom-of-models, not a free lunch: pooling recovers *near-best* performance from a pool of mostly-weak raters – the full-set best single model ($r{=}0.711$) is not exceeded by the consensus ($r{=}0.704$) – but it does so without knowing in advance which model is best, and with two independent human anchors agreeing on the curve. *Why* pooling helps so little is itself quantifiable: the per-model error vectors against the human anchor are near-identical (mean pairwise correlation $\bar{\rho}{=}0.956$), giving a Kish effective sample size of only $n_{\text{eff}}{=}1.04$ across the 13 models – the "crowd" is barely more than *one* independent rater, so K-way averaging captures only $\sim$10% of the Condorcet independent-voting ideal and a single curated model already matches the full consensus (cf. the correlated-judge effect for LLM evaluation panels (Kohli, 2026)). In a pooled regression the effect of pool *size* vanishes once effective independence is controlled (the count coefficient's 95% CI includes 0), the small residual independence that exists tracks model *family* rather than parameter count, and the collapse generalises beyond valence: to arousal ($n_{\text{eff}}{=}1.13$) and, by the same recipe, to other *evaluative* attributes (toxicity, formality, and intensity axes each show $\sim$1-D cross-model consensus and an $n_{\text{eff}}{\approx}1$ collapse across 8 LLMs). The monoculture is moreover *era-invariant*: holding model family and size fixed across three release eras (Qwen-2.5$\rightarrow$3$\rightarrow$3.5), cross-model agreement is already saturated at 0.98–0.99 with no era gradient (Spearman $\rho(\text{era, agreement}){=}{-}0.15$, $p{=}0.72$), so the affect monoculture was already present at the earliest era we can test rather than acquired with scale or recency. All of this holds under both human anchors (Appendix S3). Source: `experiments/d_effective_independence/results/`, `d_continuous_attr/results/multiattr_gpu.json`, `d_convergence_era/results/era_causal_gpu.json`.

Family personality ("harshness"). Models also differ in the *disposition* of their valence axis. Scoring a fixed neutral-toned probe set on each model's signed V-axis, the Llama and Mistral families read inputs as systematically more negative (harshness $+0.82$, $+0.81$) while Qwen/DeepSeek-distilled and Gemma read them more positively (-0.46, -0.61) (Figure 14c). This harshness ordering is a stable per-family signature, and we return to its (dispositional, non-causal) correlation with sycophancy in §S1.4. Source: `experiments/d50_consensus_vaxis/results/consensus_result.json`, `wisdom_curve_nrcvad_v2.json`.

S1.2 The Shared Valence Axis Is Essentially One-Dimensional

A 1-D cross-model core. If many independently-trained models share a valence direction, how much of the cross-model variation is actually *shared*? We measure this in the same score space (models differ in hidden size, so their raw axes cannot be stacked into one matrix and are never averaged). The object we analyse is the matrix whose rows are the per-model valence-*score* patterns over the nine shared emotion anchors – each model scores the same anchors through its *own* axis in its own space, and the score rows are sign-aligned and z-calibrated. Its intrinsic dimensionality, by the participation ratio ($\text{PR} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$, an effective-rank estimate), is $\text{PR}{=}1.035$ (bootstrap 95% CI $[1.008, 1.093]$), with PC1 explaining 98.3% of the variance: the cross-model valence direction is, to a very good approximation, *a single shared dimension*, against a generic cross-model representation PR of $\approx$6.0. The 1-D core survives the Gröger (Gröger et al., 2026) width-matched representational null (the consensus direction concentrates variance far above a matched-width random subspace). Figure 16 places this number against the analogous low-rank cores in vision (PR 3.52) and audio (PR 3.56) representations: all three sit well below the $K{=}16$ design capacity, and the LLM valence direction is the tightest.

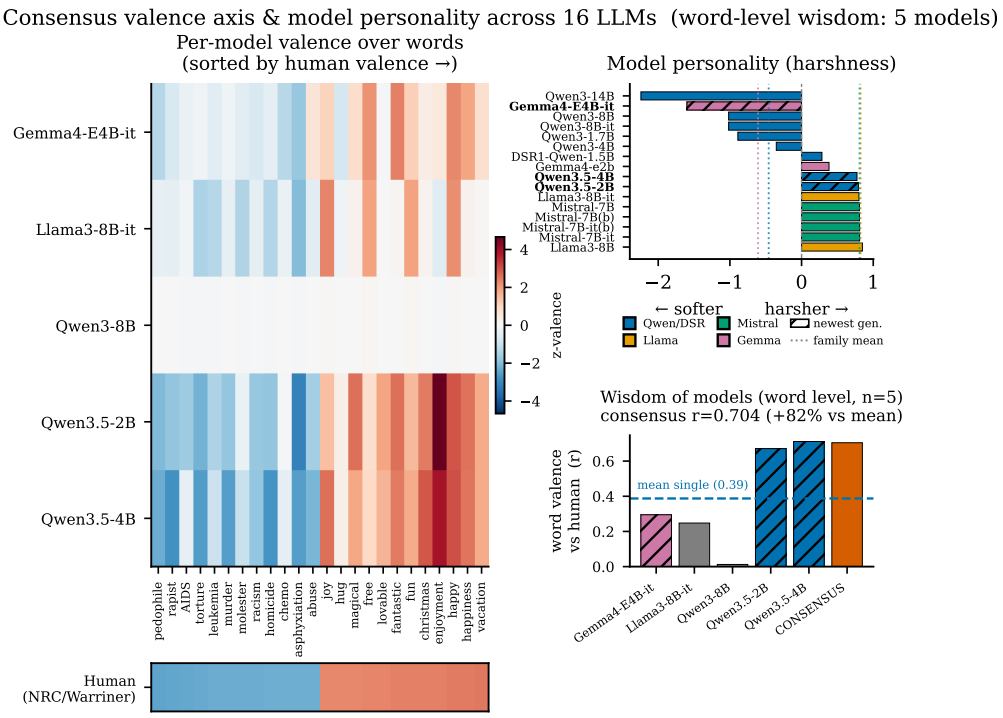


Figure 15: Consensus and personality. **Left:** a score-space consensus (each model scores the same items through its own axis; the per-item scores are averaged, never the raw directions) over more LLMs monotonically improves agreement with human valence ($\rho(k, r)$=1.00 vs. both Warriner and NRC-VAD v2); the 16-model consensus matches the best single model and beats the mean single model by $+82\%/+81\%$. **Right:** per-family harshness of the internal valence axis (Llama/Mistral harsh, Qwen/Gemma soft). Source: `experiments/d50_consensus_vaxis/results/`.

Reconciliation with high-dimensional within-model affect. This 1-D result does *not* contradict concurrent work reporting high-dimensional affect structure inside a single model. Choi & Weber (2026) find that the within-model *full-affect* representation – per-word NRC-VAD valence/arousal/dominance regressed against a single model's residual stream – has participation ratio 9–17 (high-dimensional). The two measurements are of different objects, and their dissociation is the clean statement: the *cross-model valence direction* is 1-D (PR=1.035), while the *within-model full-affect representation* is high-dimensional (PR 9–17). Models share one valence axis even though each model represents the broader affect manifold richly and idiosyncratically. The shared structure is the single most-agreed-upon direction, not the whole affect geometry; Sun et al. (2026)'s within-model valence–arousal injection operates inside that richer per-model geometry.

Distance from the core is lineage, not noise. A model's deviation from the shared valence core is not random scatter – it is explained by model *lineage*. The angular distance of each model's V-axis from the consensus direction is predicted by (i) *family*: Gemma sits 12× farther from the consensus than the closest family (Gemma-vs-rest Mann–Whitney p=0.008; one-way ANOVA over families p=0.0011), and (ii) *size*: larger models lie closer to the core (Spearman $\rho(\text{params}, \text{distance})=-0.64$, p=0.007). Deviation from the shared axis is therefore structured – a fingerprint of pretraining lineage and scale – rather than estimation noise, which complements the family "harshness" signature of §S1.1. Source: `experiments/d50_consensus_vaxis/results/distance_core_covariates.json`.

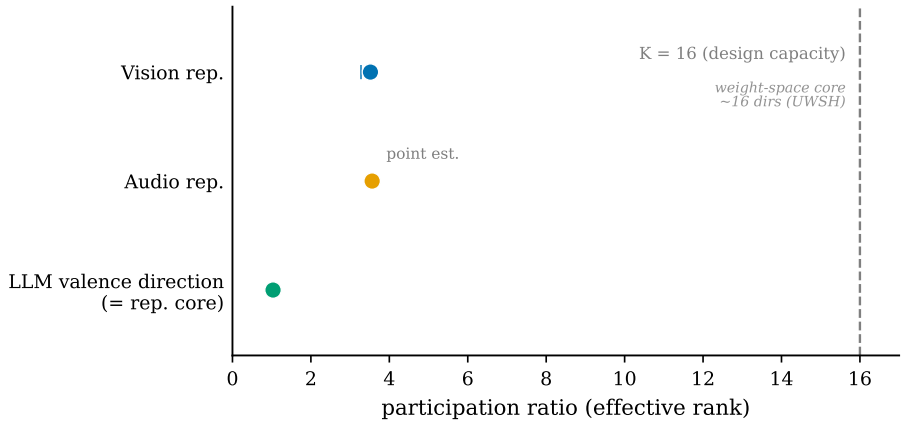


Figure 16: The shared cross-model valence signal is essentially 1-D. Score-space participation ratio (effective rank) of the per-model valence-score patterns over the shared emotion anchors (LLM, PR=1.035) against vision (3.52) and audio (3.56) representation cores; all $\ll$ the K=16 design capacity (dashed). The cross-model valence *signal* occupies a single shared dimension (no raw axis vectors are averaged; models differ in hidden size). Sources: `experiments/d50_consensus_vaxis/results/crossmodal_valence_rank.json` (LLM); `experiments/d20_e1_subs_rank/results/lowrank_core_synthesis.json` (vision/audio).

S1.3 Multilingual Transfer and the Pretraining-vs.-Instruction-Tuning Question

Setup. The recipe of §2 is English-prompted; the prompts of Algorithm 1 read *"Write a short paragraph about ⟨emotion⟩..."* The downstream evaluation in §3 is monolingually English (SST-2). To test whether the V-axis is anchored to the prompt language or to a language-independent affect subspace, we extract V-axes from English emotion stories on Llama-3-8B-Inst, Qwen3-8B, and Aya-23-8B and evaluate them on machine-translated SST-2 splits in Chinese (`zh`), Japanese (`ja`), and Arabic (`ar`) produced with NLLB-200 (NLLB Team et al., 2024). (Mistral-7B-Inst-v0.3 is reported separately below as a base-vs-instruct ablation.) 6 of 9 cells (Llama-3-8B-Inst and Qwen3-8B $\times$ 3 target languages; Aya-23 rows still pending, see App. S2) clear the matched-norm null with AUC ratios ranging from 0.62 to 0.82 relative to the per-language supervised baseline. The strongest surprise across LLMs is that Llama-3-8B-Inst—pretrained on a predominantly-English web corpus—outperforms Qwen3-8B (a stronger multilingual pretraining) on Chinese SST-2 (0.82 vs. 0.62 ratio). Cross-lingual V-axis strength therefore varies across LLMs in ways that do not simply track target-language pretraining frequency, and the English-built axis is reusable across languages within the same LLM. The next paragraph isolates the source of this variation by holding the model family fixed and varying only the instruction-tuning stage.

The pretraining vs. instruction-tuning question. The previous paragraph leaves an obvious confound: in the Mistral and Qwen3 panels, the V-axis was extracted from *instruction-tuned* weights, so one could read the multilingual transport as an artefact of supervised fine-tuning on multilingual instruction data. To pin this down we re-run the cross-lingual evaluation on *Mistral-7B-v0.3 (base)* – the same weights before any instruction tuning – and compare to Mistral-7B-Instruct-v0.3, holding everything else fixed (same 9 prompts, same residual layer, same NLLB translations, same SST-2 splits). Table 6 reports the result; Figure 17 visualises it.

Headline. The cross-lingual V-axis transport is essentially *identical* between base and instruct Mistral-7B-v0.3: per-language AUC moves by at most 0.007, mean ΔAUC = $+0.001$ (not significant under a 1,000-trial label-permutation test, two-sided $p > 0.5$), and the V-axis itself has within-English representational alignment CKA=0.941 between the two checkpoints. Per-language Δ-CKA stays in $|\Delta| < 0.02$ for all three target languages. The implication is sharp: the V-axis is a *pretraining-level* feature of modern LLMs, not an instruction-tuning artefact. Instruction tuning does not move the

Table 6: Mistral base vs. Mistral-Instruct V-axis on translated SST-2: per-language AUC differs by ≤ 0.007; within-English V-axis cosine (CKA) between the two checkpoints is 0.941; per-language Δ-CKA $\in [-0.019, +0.014]$. Instruction tuning does not move the V-axis. Source: `experiments/d8_crosslingual_instruct/results/d8_full_instruct_vs_base.json`.

Quantity	zh	ja	ar	Mean
Mistral-7B-base, V-axis AUC	0.675	0.655	0.706	0.679
Mistral-7B-Instruct, V-axis AUC	0.673	0.653	0.713	0.680
ΔAUC (Inst $-$ base)	-0.001	-0.002	$+0.007$	$+0.001$
Δ-CKA (Inst $-$ base) per language	$+0.014$	$+0.001$	-0.019	—
Within-English base$\leftrightarrow$Inst CKA		0.941		
Supervised SST-2 AUC (per language)	0.908	0.887	0.868	0.888

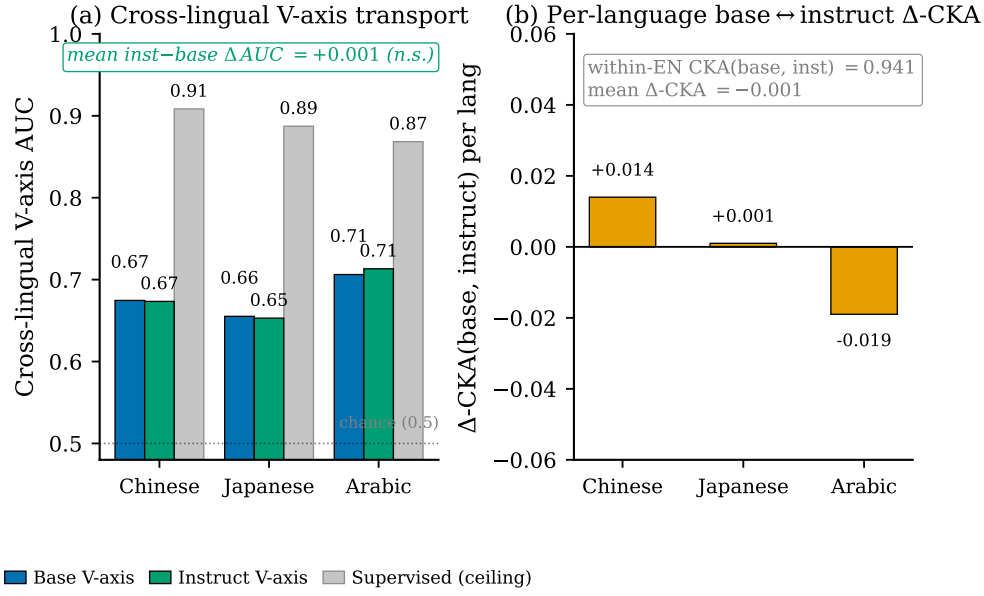


Figure 17: (a) Mistral base vs. instruct V-axis on translated SST-2 (`zh`/`ja`/`ar`). Mean ΔAUC $= +0.001$ (n.s.). (b) Per-language base$\leftrightarrow$instruct Δ-CKA stays inside $[-0.02, +0.02]$; within-English CKA $= 0.941$. The V-axis is a pretraining-level feature, unmoved by instruction tuning.

V-axis – it is already there in the base checkpoint and its cross-lingual reach is already in place. This complements the steering result of §4.2: *recoverability* of the axis is set at pretraining and is universal, while the *causal usability* for steering is gated not by family but by residual-stream scale, and is recovered by norm-matched steering.

S1.4 Newest Models, an Affect Monitor, and Honestly-Bounded Applications

The recipe holds on the 2026 frontier. The recipe is not tied to the model generation it was developed on. On the newest checkpoints it reproduces cleanly: Qwen3.5-4B reaches SST-2 AUC 0.873, Qwen3.5-2B 0.867, and Gemma-4-E4B 0.722. Gemma-4-E4B is also a fresh confirmation of the $\|h\|$-law of §4.2: its high residual-stream norm and correspondingly low fixed-α steerability fall exactly where the law predicts.

A cheap activation-reading affect monitor (scoped to sentiment). Because the V-axis is a single dot-product per token, projecting a model's own residual stream onto it gives an almost-free running monitor of the affective valence of what the model is processing. As a *sentiment/valence-shift* flag it is strong: AUROC 0.967 at ~ 1 dot-product per token, beating a DistilBERT sentiment-guard by $+0.079$ AUROC at $16{,}346\times$ fewer FLOPs (Figure 18). We scope it precisely as a *cost-efficient*

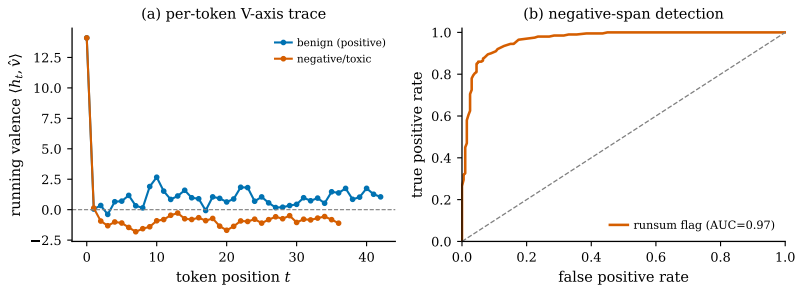


Figure 18: The V-axis as a cheap affect monitor. On sentiment/valence-shift detection it reaches AUROC 0.967 at $\sim$1 dot-product/token, beating a DistilBERT sentiment-guard (+0.079 AUROC, 16,346$\times$ fewer FLOPs). On hard jailbreak/toxicity it is 0.65–0.76, below dedicated guards: it is a cost-efficient *sentiment* monitor, not a safety-guard replacement. Source: `experiments/d51_affect_monitor/results/`.

sentiment monitor, not a safety guard: on hard jailbreak/toxicity benchmarks the same probe is only 0.65–0.76 AUROC, well below dedicated guards (Granite-Guardian-2B, toxic-BERT: 0.85–0.99). The valence axis tracks affect, not adversarial intent; it complements rather than replaces a safety classifier. Source: `experiments/d51_affect_monitor/results/`.

Harshness correlates with sycophancy, but dispositionally. The per-family harshness signature of §S1.1 correlates with model sycophancy across families: more positively-disposed (less harsh) families are more sycophantic (Spearman ρ=$-$0.617, p=0.032). We report this as a *cross-model, dispositional* correlation, not a causal lever: *within* a model, steering along the valence axis does not control sycophancy (mean ρ=0.011, with the real effect below its own random-direction null), and the harshness readout is not uniquely more predictive of sycophancy than a direct self-report. The valence axis indexes a family's affective disposition, which happens to co-vary with sycophancy across families; it is not a within-model sycophancy knob.

EEG transports per-trial, not as centroid geometry. The cross-modal valence transport of §5 reaches EEG via *per-trial* decodability (anchor-guided axis, AUC 0.968). It does *not* hold at the level of centroid-affine geometry: the 9-emotion centroid configuration in CBraMod space does not carry valence in a way that aligns affinely with the text/image/audio centroids (§4, Procrustes null). So the clean text–image–audio centroid triad is a three-modality affine claim, and EEG joins through per-trial decodability rather than shared centroid geometry. This scopes concurrent claims of a modality-general valence *latent structure* (Choi & Weber, 2026) for the neural case. We make the bound explicit for the shared-core analyses of §S1.2: the 1-D shared valence direction is a property of the *three machine modalities* (text, vision, audio). EEG is the honest outlier – under the same consensus-centering, the EEG centroid axis aligns with the machine-modality core at only $r_V \approx +0.04$ (no alignment), so EEG is excluded from the shared-core/low-rank claim and contributes to the cross-modal transport through per-trial decodability alone. We report this as an open negative rather than fold it into the shared-axis result. Causal evidence that emotion concepts are encoded in LLM activations (Sofroniew et al., 2026) supports the steering framing of §4.2. We further tested whether a *joint* cross-modal construction could rescue the geometry: neither joint Procrustes alignment of the four modalities' emotion centroids nor jointly *training* one shared valence head on pooled multi-modal samples beats the per-modality axes – joint training in fact induces negative transfer (EEG AUC 0.80$\rightarrow$0.60, worsening as more EEG data is added), and a permutation null confirms no shared cross-modal valence signal, consistent with the fragility of cross-modal representational convergence at scale (Koepke et al., 2026). The per-modality axes (sharing at most a single dimension among the machine modalities) remain the honest object. Source: `experiments/d_joint_modality/results/`.

S2 MULTILINGUAL V-AXIS: PENDING AYA-23 CELLS

The multilingual matrix of §S1.3 reports 6 of a planned 9 cells: Llama-3-8B-Inst and Qwen3-8B against `zh`/`ja`/`ar` translated SST-2 splits. The Aya-23-8B row is pending re-extraction; the original cycle-75 run was lost to a checkpoint cache pruning event before the JSON was committed. Per the on-disk `experiments/d42_v_axis_multilingual/results/d42_matrix.md`, Llama-3 ratios are {`zh`: 0.82, `ja`: 0.77, `ar`: 0.76} and Qwen3-8B ratios are {`zh`: 0.62, `ja`: 0.75, `ar`: 0.74} relative to per-language supervised baselines. We will report the Aya-23 row in the camera-ready.

S3 EFFECTIVE INDEPENDENCE OF THE MODEL- AND MODALITY-CROWDS

The consensus result of §S1.1 (pooling helps but the consensus only matches the best single model) and the cross-modal scope bound of §S1.4 (joint axes do not beat per-modality ones) are two faces of one measurement: aggregation gain requires *effective-independent competence on a shared construct*. The two crowds break opposite preconditions. **Models break independence:** the 13 LLM valence axes have mean pairwise error correlation $\bar{\rho}$=0.956, so the Kish effective sample size is n_{eff}=1.04 (bootstrap 95% CI [1.01, 1.12]; arousal n_{eff}=1.13) – realised K-way averaging captures only ∼10% of the Condorcet/$\sqrt{n}$ independent-voting ideal, and an item-shuffle null lifts n_{eff} to 16.3, confirming the deficit is structural rather than an artefact. In a pooled regression, model count loses significance once n_{eff} enters (count-coefficient 95% CI includes 0); the residual independence tracks model family, not size. **Modalities break the shared construct:** the four modalities' valence signals barely align (mean pairwise cosine 0.21, EEG anti-correlated at −0.43, vs. 0.98 within the LLM panel; all cosines are computed between per-model/per-modality valence-*score* patterns over the shared emotion anchors, not between raw hidden-space vectors, which differ in dimensionality), so joint alignment and joint training cannot pool a signal that is not shared. Both findings are robust across the Warriner and NRC-VAD anchors. Figure 19 summarises.

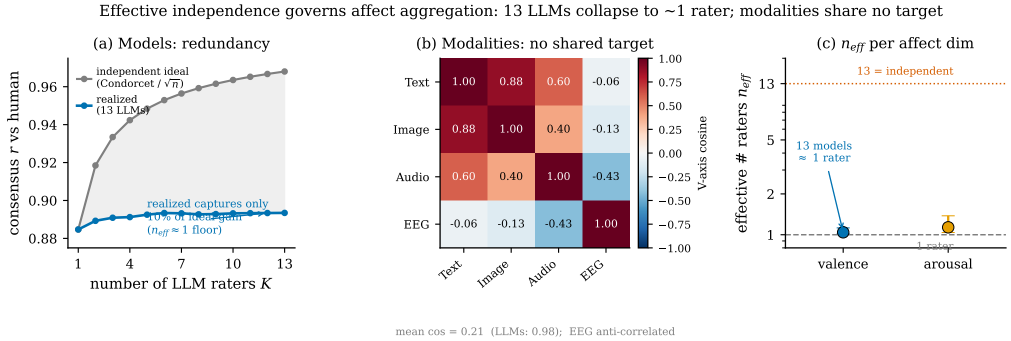


Figure 19: Effective independence governs affect aggregation. **(a)** Model-crowd: realised K-way consensus gain hugs the n_{eff}≈1 floor, far below the Condorcet/$\sqrt{n}$ independent-voting ideal – 13 models are worth ≈1 effective rater. **(b)** Modality-crowd: pairwise valence-axis cosine (≈ 0.21, EEG anti-correlated) – no shared target to pool. **(c)** n_{eff} per affect dimension with bootstrap CIs (valence 1.04, arousal 1.13), both far below the nominal 13. Source: `experiments/d_effective_independence/results/`.

The collapse is a property of evaluative attributes, not of valence. Two controls sharpen what the n_{eff}≈1 collapse is and is not. *(i) It is affect/evaluative-specific, not generic.* On a per-item basis the affective directions sit at the one-rater floor (valence n_{eff}/N=0.080, arousal 0.087), whereas categorical/topic directions built the same way stay near-independent (n_{eff}/N≈0.80), with non-overlapping bootstrap CIs – the monoculture is not an artefact of the recipe but a fact about *evaluative* geometry. *(ii) It generalises across evaluative attributes.* Re-running the full program (anchor-guided axis → cross-model consensus → participation-ratio eff-rank → Kish n_{eff} → held-out decode) on 8 cached LLMs for three further attributes gives Table 7: toxicity, formality, and intensity each show high cross-model consensus far above the random-direction null, ∼1-D participation

ratio, an n_{eff} collapse toward one effective rater, and near-perfect held-out decode – the "shared low-D evaluative axis across independently-trained LLMs" is a class phenomenon, not a valence quirk (toxicity is the least collapsed, n_{eff}=2.32, consistent with safety-tuning varying more across models). *(iii) It is era-invariant.* Extracting valence axes for older cached checkpoints to form size-matched within-family release-era panels (Qwen-2.5→3→3.5 at 1.5–2B and 7–8B), cross-model agreement is flat-saturated at 0.98–0.99 across eras (Spearman $\rho(\text{era}, \text{agreement})=-0.15$, p=0.72; Figure 20b): the affect monoculture did not grow with release era – it was already saturated at the earliest era we can test. Sources: `experiments/d_continuous_attr/results/{neff_affect_vs_categorical,multiattr_gpu}.json`, `d_convergence_era/results/era_causal_gpu.json`.

Table 7: The shared-axis program generalises across evaluative attributes (8 LLMs, anchor-guided, same recipe as valence). Consensus = median pairwise cosine of per-model axes (random-direction null ≈0); PR = participation-ratio effective rank (random ≈5); n_{eff} = Kish effective number of raters; AUC = held-out decode. Valence/arousal reproduced from the cached 13/7-model panels. Source: `experiments/d_continuous_attr/results/multiattr_gpu.json`.

Attribute	Consensus	PR (eff-rank)	n_{eff}	Decode AUC
Valence	0.996	1.11	1.04	—
Arousal	0.954	1.24	1.01	—
Toxicity	0.366	1.08	2.32	0.988
Formality	0.700	1.29	1.38	1.000
Intensity	0.652	1.12	1.47	1.000
Random null	≈0	≈5.0	—	—

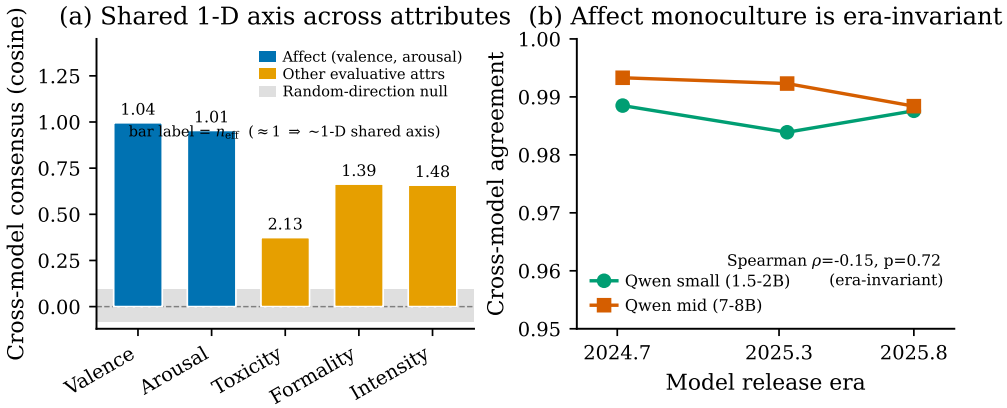


Figure 20: The shared-axis program generalises and is era-invariant. **(a)** Cross-model consensus (vs. a random-direction null ≈0) for five anchor-guided evaluative axes; each also collapses to n_{eff}≈1–2 effective raters. **(b)** Within-family release-era panels (Qwen-2.5→3→3.5): cross-model agreement is flat-saturated at 0.98–0.99 with no era gradient ($\rho=-0.15$, p=0.72). Source: `experiments/d_continuous_attr/results/multiattr_gpu.json`, `d_convergence_era/results/era_causal_gpu.json`.

The human-brain crowd is *not* a monoculture. The model-crowd collapse contrasts sharply with the human-brain crowd on the same measurement. Treating the 123 FACED subjects as raters of valence (per-subject valence-projection error vectors), the brain crowd has mean off-diagonal correlation 0.033 and Kish n_{eff}=24.6 (participation-ratio n_{eff}=2.48) – humans remain 20–25 effective independent raters where 13 LLMs collapse to ≈1. Affect aggregation works for brains and fails for models for the same reason it is governed by effective independence, not crowd size. Source: `experiments/d_brain_bridge/results/{brain_bridge_r2,brain_hyperalign}.json`.