THE GEOMETRY OF CONCEPTORS IN NEURAL-ACTIVATION SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Conceptors represent concepts as graded collections of directions in neural-activation space rather than as single vectors. We develop a geometric account of how these representations depend on the contrast used to define a concept: opposing values in bipolar fitting, or presence and absence in unipolar fitting. We derive an exact decomposition of a centered pooled conceptor into an average pole filter, a mean-separation update, and a covariance-shape update. Experiments across five language models on sentiment, political leaning and stance, and depression expression show that both updates improve reconstruction of held-out contrasts. Additional directions preserve variation beyond the mean-difference line, and changing the comparator changes the benefit of these directions. Shape, size, and overlap each show up downstream: in amplification experiments, bipolar fits raise evidence for both poles more consistently than unipolar fits; conceptor size (the quota) correlates positively with layerwise probe performance; and cross-concept overlap quantifies interference under subspace removal, linking shared directions to Boolean composition. Together, these findings explain how fitting choices shape conceptor geometry and support bipolar fitting when the objective is to amplify both opposing poles.

1 INTRODUCTION

Conceptors represent patterns of neural activity as soft linear filters (Jaeger, 2014; 2017), retaining learned directions with different strengths, or gains. They have been used to refine language embeddings, mitigate bias, and steer language models (Liu et al., 2018; Karve et al., 2019; Postmus & Abreu, 2024; Triantafyllopoulos et al., 2026). Their Boolean operations allow represented patterns to be combined or excluded (Jaeger, 2014).

A label such as "positive sentiment" leaves an important choice open. Comparing "This film is wonderful" with "This film is awful" contrasts opposing evaluations. Comparing it with "This film lasts two hours" contrasts an evaluation with its absence. We call these comparisons *bipolar* and *unipolar*. This operational distinction is related to the distinction between naturally contrasting binary concepts and features without natural contrasts studied by Park et al. (2023). These terms describe the fitting comparisons, without assumptions about psychological dimensionality or independence (Watson & Tellegen, 1985; Green et al., 1993). Figure 1(a,b) illustrates their mean-difference directions. For conceptors, changing the comparator changes the activation statistics used for fitting. Positive-only fitting is a separate construction using a single collection.

The comparison and the form of the representation are separate choices. Mean differences, commonly used in additive steering (Turner et al., 2023; Rimsky et al., 2024), summarize separation between two averages. Conceptors also reflect variation among individual activations, retaining one or several directions with different gains (Figure 1(c,d)). Filtering and addition are also distinct: filtering scales components already present in an input, whereas fixed addition applies the same displacement to every input.

For centered pooled fits, the covariance contains both within-pole variation and separation between pole means. We characterize how these contributions appear in the conceptor itself and how pooling differs from averaging separately fitted pole filters. This connects the mean-difference line to a richer representation without assuming that every additional direction is a distinct semantic feature: retained variation may also reflect topic or wording. We also examine retention strength and shared directions.

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.

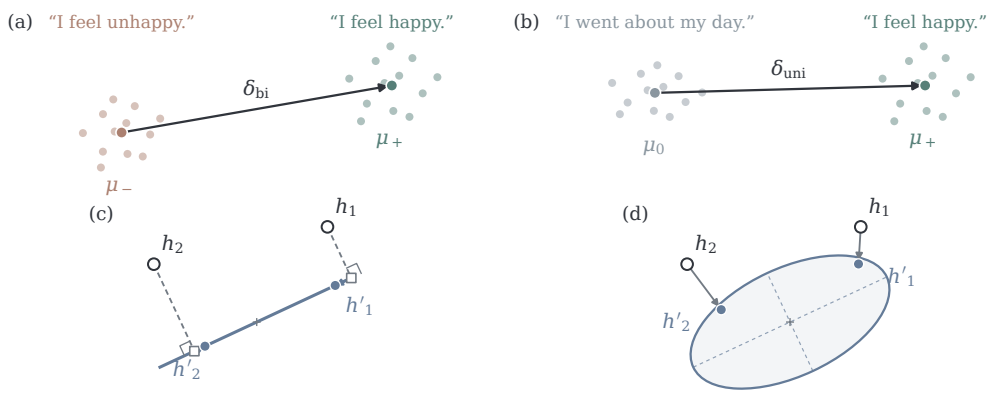


Figure 1: **Comparison and rank are separate choices.** Mean differences contrast (a) opposing poles or (b) presence and absence. Conceptors retain (c) one axis or (d) several, with shrinkage. Open circles: common inputs; squares: orthogonal projections; filled circles: outputs. The segment and ellipse are images of the unit ball. Coordinates and filters are illustrative.

Quota, the average conceptor gain, summarizes overall retention strength; we study its descriptive association with layerwise probe performance. Subspace overlap addresses a complementary question: whether removing directions associated with one concept also removes components retained by another, with implications for composition.

We investigate sentiment, political leaning and stance, and depression expression across five models: Gemma-2-2B-IT, Gemma-2-9B-IT, Qwen 2.5, Llama-3.1-8B-Instruct, and OLMo-3-7B-Instruct. We select these concepts a priori to span a spectrum from relatively simple to multifaceted constructs. Labeled examples for each let us test whether the same geometric patterns hold across different kinds of concepts.

Our contributions are:

1. **Explain how conceptors represent richer geometry than mean difference.** We derive an exact breakdown of what changes when a conceptor is fitted to both poles of a concept. The separation between their averages and differences in how their examples vary both help preserve contrasts in new contexts. Additional directions capture variation that a single mean difference misses.
2. **Show why opposing examples matter for amplification.** Conceptors fitted to both opposing poles strengthen average classifier evidence for each pole more consistently than those fitted to presence and absence.
3. **Connect conceptor strength to where information is most linearly observable.** Quota summarizes how strongly a conceptor retains activation patterns. Across six public benchmarks, higher quota tends to accompany better classifier performance across layers.
4. **Explain overlaps between concepts geometries.** Across five models, sentiment and depression-expression representations overlap more than either overlaps with political leaning. We show how overlap determines the average loss when one representation's directions are removed from another. This helps interpret the geometry behind Boolean composition.

Our focus is the geometry of conceptors used for steering. We analyze activations and frozen readouts, leaving downstream steering performance and generated-text effects for future work.

2 Related Work

Conceptors and activation steering. Conceptors represent activation patterns as regularized linear filters with Boolean operations (Jaeger, 2014; 2017). Liu et al. (2019) use these operations to update sentence representations across corpora while retaining performance on earlier corpora. For

contextualized-embedding debiasing, Sedoc & Ungar (2019) combine conceptors learned from homogeneous word-list subsets through OR. Yifei et al. (2023) extend conceptor debiasing to BERT and GPT, using NOT for bias removal and AND for intersectional bias, and introduce Conceptor-Intervened BERT. Postmus & Abreu (2024) apply conceptors to language-model steering and combine steering goals through Boolean operations. Additive methods such as ActAdd and CAA summarize a contrast with a single displacement vector (Turner et al., 2023; Rimsky et al., 2024).

Geometry for probing and intervention. Park et al. (2023) connect linear concept representations, measurement, and intervention through a causal inner product. Their categorical extension distinguishes feature presence from contrasts between attribute values (Park et al., 2025). Park et al. (2026) derive probe-guided dual steering from softmax information geometry, with regularized local updates determined by output-unembedding covariance. Gurnee et al. (2026) use averaged downstream Jacobians to identify verbalizable representations involved in internal reasoning. These approaches connect representation geometry to how the model reads or uses an activation.

Empirical concept representations. Sofroniew et al. (2026) identify emotion directions in Claude Sonnet 4.5 and show that interventions along them influence behavior. For depression assessment, Ravenda et al. (2026) use activation steering to improve symptom-relevance prediction and questionnaire completion. Zhu et al. (2026) construct symptom directions whose projection scores on held-out text align with clinician annotations. Abootorabi et al. (2026) find that methods with comparable steering performance can differ substantially in how well their geometry reflects relationships among human values. This motivates examining both the effects of interventions and the structure of the representations they use.

Our focus is the geometry learned by conceptors from contrasting activation collections: how pooling changes the filter, what its spectrum reveals, and which directions it shares with other concepts.

3 Background

We consider hidden activations $h \in \mathbb{R}^d$ at a fixed model layer and extraction site, omitting these indices for clarity. All geometric quantities use the Euclidean inner product in this activation space.

3.1 Semantic comparisons and mean differences

Let $\mathcal{H}_+$ contain activations of examples expressing a target pole, $\mathcal{H}_-$ its opposite, and $\mathcal{H}_0$ examples in which the attribute is absent. Write $\mu_s = n_s^{-1} \sum_{i=1}^{n_s} h_i^{(s)}$ for the mean of collection $s \in \{+, -, 0\}$. The corresponding difference-of-means vectors are

$$\delta_{\mathrm{bi}} = \mu_+ - \mu_-, \qquad \delta_{\mathrm{uni}} = \mu_+ - \mu_0. \tag{1}$$

The bipolar comparison contrasts opposing attribute values. The unipolar comparison contrasts presence with absence, following the feature–contrast distinction of Park et al. (2023). For sentiment, these compare positive expression with negative expression and sentiment-free expression, respectively (Figure 1(a,b)). The absence reference is defined by an empirical collection, whose mean μ_0 can be nonzero. For equally weighted matched pairs, each vector also equals the average of the paired activation differences. Such contrasts underlie additive steering (Turner et al., 2023; Rimsky et al., 2024), which applies the same displacement to each activation: $h' = h + \beta\delta$, with strength $\beta \in \mathbb{R}$. The vector summarizes the separation between the group means.

3.2 Conceptors

A conceptor is a linear filter learned by balancing reconstruction of its fitting activations against regularization of the operator (Jaeger, 2014; 2017). Conceptors have also been used for language-model steering (Postmus & Abreu, 2024). Given a fitting collection $\mathcal{H} = \{h_i\}_{i=1}^n$ and a fitting reference $c \in \mathbb{R}^d$, define $x_i = h_i - c$ and

$$R_c = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top. \tag{2}$$

The reference c sets the origin for fitting. Choosing c as the fitting mean gives a centered covariance. Choosing $c = 0$ gives an uncentered second moment. We use normalization $1/n$ in both cases.

The conceptor is the solution to

$$C_\lambda(R_c) = \operatorname*{arg\,min}_{M \in \mathbb{R}^{d\times d}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \|x_i - Mx_i\|_2^2 + \lambda \|M\|_F^2 \right\}, \qquad \lambda > 0. \tag{3}$$

The first term encourages reconstruction of the fitting activations. The second penalizes the size of the filter. The unique minimizer is

$$C_\lambda(R_c) = R_c(R_c + \lambda I)^{-1}. \tag{4}$$

The usual *aperture* parameter $\alpha > 0$ satisfies $\lambda = \alpha^{-2}$. Appendix A.1 gives the standard derivation.

A bipolar pooled fit combines $\mathcal{H}_+$ and $\mathcal{H}_-$. A presence–absence pooled fit combines $\mathcal{H}_+$ and $\mathcal{H}_0$. Fitting $\mathcal{H}_+$ alone gives a *positive-only* conceptor. Each construction also requires a choice of fitting reference c. Together, the fitting collection and reference determine R_c.

3.3 Spectral geometry

Let $R_c = U \operatorname{diag}(r_1, \ldots, r_d) U^\top$, with $r_1 \geq \cdots \geq r_d \geq 0$ and orthonormal eigenvectors u_j. Then

$$C_\lambda(R_c) = \sum_{j=1}^{d} g_j u_j u_j^\top, \qquad g_j = \frac{r_j}{r_j + \lambda} \in [0, 1). \tag{5}$$

At fixed R_c, increasing aperture raises positive gains without changing the axes or exact rank.

The leading-k approximation $C_{\lambda,k} = \sum_{j=1}^{k} g_j u_j u_j^\top$ filters via $h' = c + C_{\lambda,k}(h - c)$, projecting $h - c$ onto the retained axes and shrinking its coordinates (Figure 1(c,d)). The leading axis need not align with the mean difference (Appendix A.3). Additive steering applies the same displacement $\beta\delta$ to every input.

The *quota* averages gains over all d activation dimensions (Jaeger, 2014):

$$q(C) = \frac{\operatorname{tr}(C)}{d} = \frac{1}{d} \sum_{j=1}^{d} g_j, \qquad C = C_\lambda(R_c). \tag{6}$$

Quota depends on the fitted moment and aperture; its association with probe performance is empirical. For hard orthogonal projectors, conceptor NOT, AND, and OR give the orthogonal complement, intersection, and combined span (Jaeger, 2014); soft operations also depend on gains. Appendix A.4 relates overlap to average squared length retained under hard removal.

4 How pooling shapes a conceptor

Pooling both poles combines their separation with variation within each pole. We characterize these contributions to the filter and its reduction to the mean-difference line.

Let R_+, R_- be the covariances of the two poles, each centered on its own mean, and let $p, q > 0$ be their weights, with $p + q = 1$. Write $\delta = \mu_+ - \mu_-$ and $\mu = p\mu_+ + q\mu_-$. Centering the pooled examples on μ gives

$$R_{\text{pool}} = \underbrace{pR_+ + qR_-}_{R_{\text{w}}} + pq\delta\delta^\top. \tag{7}$$

The terms describe within-pole spread and mean separation. Empirical covariances use $1/n_s$ normalization; ordinary pooling sets $p = n_+/(n_+ + n_-)$ and $q = 1 - p$. All filters below use the same numerical $\lambda > 0$.

Proposition 1 (Pooling decomposition)**.** *Let $A = R_{\text{w}} + \lambda I$ and $\bar{C} = pC_\lambda(R_+) + qC_\lambda(R_-)$. The full pooled conceptor satisfies*

$$C_\lambda(R_{\text{pool}}) = \underbrace{\bar{C}}_{\text{average pole filter}} + \underbrace{D_\mu}_{\text{mean separation}} + \underbrace{J_R}_{\text{covariance shape}}, \tag{8}$$

where

$$D_\mu = \frac{\lambda pq}{1 + pq\delta^\top A^{-1}\delta}(A^{-1}\delta)(A^{-1}\delta)^\top, J_R = C_\lambda(R_\mathrm{w}) - \bar{C}. \tag{9}$$

Both updates are positive semidefinite. The mean update has rank one when $\delta \neq 0$, and $\mathrm{rank}(J_R) = \mathrm{rank}(R_+ - R_-)$.

The average pole filter. The weighted average $\bar{C}$ combines filters fitted after subtracting each pole's own mean. It represents within-pole variation and can already retain several directions.

The mean-separation update. For distinct means, D_μ is a rank-one update along $A^{-1}\delta$ (Fisher, 1936). This adjusts the mean difference by downweighting components along principal directions with greater within-pole spread. It follows the mean-difference line when δ is an eigenvector of R_w.

The covariance-shape update. Fitting the average covariance differs from averaging the fitted filters because the conceptor depends nonlinearly on covariance. Thus J_R reflects differences in the poles' spread and orientation. It vanishes exactly when $R_+ = R_-$ and can involve several directions even when the means coincide.

The identity applies to full filters before rank truncation, at a shared λ, including presence–absence comparisons. Appendix A.2 gives the proof. Fitting only two equally weighted, distinct pole means about their midpoint gives a rank-one conceptor along the mean difference; within-pole variation outside this line increases the rank of the full pooled conceptor (Appendix A.3).

5 Data and measurements

5.1 Datasets

BipolarConcepts. Our original fitting collection contrasts positive/negative sentiment, left/right political leaning, and depression expression/counter-expressions. The datasets contain 100 pairs for each concept (600 sentence entries). These supply the original conceptor geometry; Appendix B.1 includes more details.

ConceptVariants. This authored collection contains 312 sentences covering 24 topics per concept (sentiment, depression expression, and political stance). Each sentiment or policy topic has two opposing and two neutral formulations. Each symptom topic has one symptom formulation and four comparators. We vary the reference within a topic to examine how the comparison shapes the representation. The available sentence inventories have no exact or normalized-text overlap with BipolarConcepts; these are separate collections, not nested subsets.

ConceptGroups. This grouped evaluation suite combines 1,400 reviews from three domains (sentiment concept) (Kotzias, 2015), 196 authored sentences across fourteen symptom facets (depression concept), and 994 favor/against tweets across five stance topics (political stance concept) (Mohammad et al., 2016; Barbieri et al., 2020). Each group is balanced across its two labels. We fit representations on the remaining groups and evaluate the excluded group's mean difference.

Public probing benchmarks. We train and evaluate probes on SST-2 (Socher et al., 2013), Rotten Tomatoes (Pang & Lee, 2005), TweetEval sentiment (Barbieri et al., 2020), Hyperpartisan News (Kiesel et al., 2019), Reddit Depression (thePixel42, n.d.), and Cross-cultural Reddit (Rai et al., 2025). Their probe scores are compared with conceptor spectra fitted on BipolarConcepts.

Appendix B gives the collection inventories, comparison labels, evaluation partitions, and available length statistics.

5.2 Measurements

For a nonzero mean difference δ, an orthonormal basis $U_k \in \mathbb{R}^{d\times k}$, and a filter M, define

$$\mathrm{Coverage}_k(\delta) = \frac{\|U_k^\top \delta\|_2^2}{\|\delta\|_2^2}, E(M,\delta) = \frac{\|M\delta - \delta\|_2^2}{\|\delta\|_2^2}. \tag{10}$$

Coverage measures the fraction of squared length retained by projection; *reconstruction error* also reflects filter gains. Fidelity $F = 1 - E$ equals coverage for $M = U_k U_k^\top$.

For orthonormal rank-k bases U_A, U_B, define

$$\mathrm{Overlap}(A, B) = \frac{1}{k}\|U_A^\top U_B\|_F^2. \tag{11}$$

Overlap is zero for orthogonal subspaces and one for identical subspaces.

Spectral summaries are quota (Equation 6) and $\mathrm{EVR}_k = \sum_{j=1}^{k} r_j / \sum_{j=1}^{d} r_j$, the leading-$k$ share of the eigenvalue sum for a nonzero spectrum. Logistic-probe ROC-AUC measures label separability; layerwise Pearson correlations relate it to these summaries. Each experiment specifies rank, fitting reference, regularization, and evaluation split.

6 EXPERIMENTS

We use Gemma-2-2B-IT, Gemma-2-9B-IT, Qwen-2.5-3B-Instruct, Llama-3.1-8B-Instruct, and OLMo-3-7B-Instruct. The default features are mean-token activations at midpoint blocks 13, 21, 18, 16, and 16. The main multi-direction comparisons use $k = 20$, with $k = 1$ as a single-direction control; the primary decomposition ablation and quota use the full spectrum. CONCEPTGROUPS evaluations exclude one complete group from fitting and average held groups equally. Appendix C gives protocols and controls.

6.1 POOLING AND THE FITTING REFERENCE

We compare pooled and separate-pole rank-20 supports on CONCEPTGROUPS to test the effect of pooling at equal fitting counts. Each fit receives the same number of source sentences, and we evaluate coverage of the excluded group's mean difference. We center on either each fit's own mean or a shared source center. Table 1 shows higher pooled coverage than the positive-only fit in all fifteen own-centered model/data conditions and than the negative-only fit in fourteen. The exception is OLMo Pol. Stance, where pooled and negative-only coverage are 30.33% and 30.61%. With a shared source center, pooling wins in four conditions against the positive pole and six against the negative pole. Retaining each pole's mean offset therefore substantially reduces the pooling advantage. Appendix C.2 gives the original fitting-collection curves, where pooled leading-ten capture has the highest layer average in all fifteen model–concept pairs. On BIPOLARCONCEPTS, pooled rank-20 support has the highest layer-average coverage in all fifteen model–concept pairs, exceeding one half in each and reaching 97.4–99.7% for Sentiment (Appendix C.2).

6.2 TESTING THE TWO POOLING UPDATES

We remove the decomposition terms to test their contributions to held-contrast reconstruction. The operators $\bar{C}$, $\bar{C} + D_\mu$, $\bar{C} + J_R$, and C_{pool} share the source examples and $\lambda = \mathrm{tr}(R_{\mathrm{pool}})/d$. Each receives the same held-group mean difference. Figure 2 shows that combining both updates gives the lowest reconstruction error in every model/data condition. Across both sites and all three shrinkage scales, the full filter outperforms every ablation in all 270 comparisons (Appendix C.3). At the primary midpoint setting, the mean update contributes more for Sentiment in four of five models; OLMo instead benefits more from the covariance-shape update. The shape update contributes more for Pol. Stance in all five models. For Gemma-2B Pol. Stance, adding the shape update reduces error from 0.209 to 0.179; adding the mean update reduces it to 0.207. These reconstruction gains are measured at shared λ. A follow-up comparison of rank-20 orthogonal projectors fixes rank, trace, and nonzero gains: the pooled projector achieves higher held-contrast coverage in 251 of 270 comparisons (Appendix C.3).

Variation across layers. At some early Qwen layers, the mean line approaches pooled rank-20 coverage for Sentiment and Depression; the gap can be larger later. Both Gemma models show substantial coverage gaps throughout depth, while Llama and OLMo vary by task and layer (Appendix C.4).

6.3 RETAINING INDIVIDUAL STATES FROM BOTH POLES

We next ask how much of each sentence's activation survives the filter. For a held state $x = h - c$, centered on the source midpoint, we measure $\|Mx\|_2^2/\|x\|_2^2$ and average these ratios within each pole

		Own mean			Shared source center		
Model	Data	Pool	+	−	Pool	+	−
Gemma 2 2B	Sentiment	**75.5**	52.3	48.4	75.4	78.1	**80.2**
	Pol. Stance	**51.7**	49.8	50.8	**51.7**	50.1	51.0
	Depression	**69.8**	51.1	52.7	69.8	**70.6**	69.6
Gemma 2 9B	Sentiment	**74.3**	54.4	52.6	74.3	76.3	**77.9**
	Pol. Stance	**48.5**	47.7	47.8	**48.5**	48.3	48.2
	Depression	**73.0**	57.5	57.5	73.1	**73.9**	73.0
Qwen 2.5 3B	Sentiment	**96.0**	94.0	93.8	96.0	96.5	**96.7**
	Pol. Stance	**74.0**	73.8	72.6	73.9	**74.0**	72.7
	Depression	**92.2**	88.5	89.4	92.2	**92.5**	92.4
Llama 3.1 8B	Sentiment	**84.6**	76.8	75.8	84.6	85.5	**86.3**
	Pol. Stance	**57.7**	56.8	56.9	**57.6**	57.4	57.3
	Depression	**81.6**	74.2	74.4	81.7	**82.6**	81.9
OLMo 3 7B	Sentiment	**42.7**	21.8	11.4	42.7	48.3	**49.6**
	Pol. Stance	30.3	29.9	**30.6**	30.4	30.3	**30.9**
	Depression	**30.3**	11.8	14.2	30.3	**32.0**	31.2

Table 1: Held-group rank-20 coverage (%) at equal fitting counts. Pool fits both poles; +/− fit the positive/negative pole separately. Bold marks the highest coverage within each model–concept row and centering condition; higher coverage indicates better preservation of the held-group contrast.

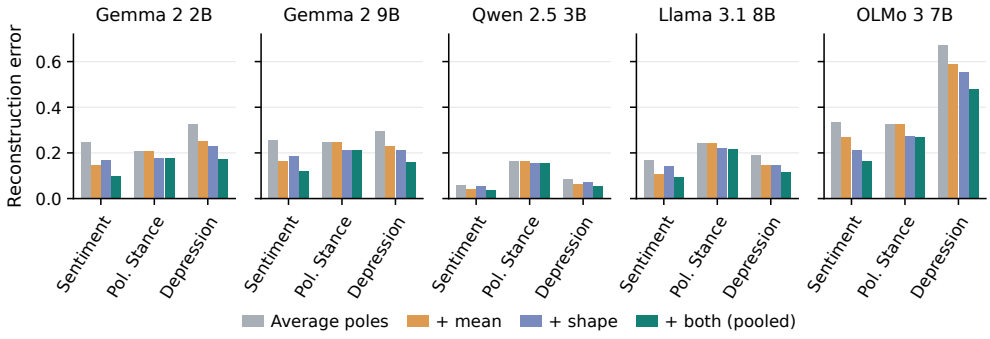


Figure 2: **Reconstruction under shared regularization.** Midpoint CONCEPTGROUPS error, averaged equally over held groups; lower is better.

and held group. On CONCEPTGROUPS, C_{20} retains more energy at both poles than the mean-line projector (Table 2). C_1 does so in 29 out of 30 individual-pole comparisons; it falls slightly below the mean line for OLMo's negative Sentiment pole. Even C_1 almost always retains more energy than the mean line, despite shrinkage. The mean line retains more positive-pole energy for Depression and more negative-pole energy for Sentiment and Pol. Stance. Averaging over the two poles, moving from C_1 to C_{20} increases retained energy by 7.2 percentage points for Sentiment, 15.1 for Pol. Stance, and 13.2 for Depression. Additional directions therefore preserve substantially more held-state variation, especially for Pol. Stance and Depression, which are by nature more complicated concepts than Sentiment.

Data	Mean line	C_1	C_{20}
Sentiment (↑)	35.7 / 36.4	61.1 / 61.1	**68.3 / 68.4**
Pol. Stance (↑)	23.2 / 24.7	35.1 / 37.6	**50.7 / 52.3**
Depression (↑)	40.8 / 36.5	49.7 / 44.8	**62.1 / 58.7**

Table 2: Held-state retention (%). Positive/negative retained energy, averaged over held groups then the five models.

6.4 QUOTA CORRELATION

We ask whether conceptor retention strength tracks label separability on separate text collections. At each layer, conceptors are fitted to the pooled poles of BIPOLARCONCEPTS with aperture $\alpha = 10$. Separately, we train an ℓ_2-regularized logistic probe on each public benchmark's training split and evaluate ROC-AUC on its held-out split. We correlate the archived layerwise AUC curves with quota and the leading-ten explained-variance ratio (Figure 3). Quota correlates positively with AUC in all 30 model–benchmark comparisons and exceeds EVR_{10} in 20. A possible explanation for the pronounced gap in Qwen we observe is that a few principal components dominate its activation variance. EVR_{10} can then remain close to one across layers, providing little contrast between their spectra. Quota combines gains across the full spectrum, allowing intermediate-strength directions to influence the summary even when the leading components dominate the variance. Quota can be computed directly from the fitted conceptors without training another classifier. It complements information-theoretic and geometric diagnostics of layerwise representations (Skean et al., 2025). Positive layerwise correlation alonex does not establish a reliable rule for selecting probe layers (Appendix C.6, which also reports entropy and effective-rank controls and absolute AUC curves).

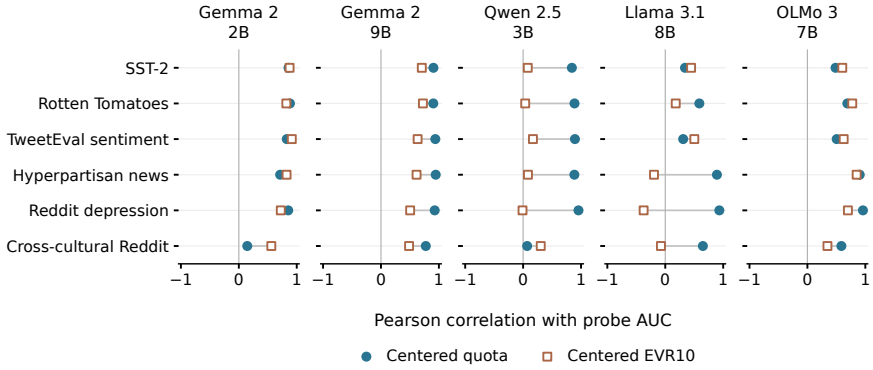


Figure 3: **Quota and layerwise separability.** Pearson correlations of public-benchmark probe ROC-AUC with quota and EVR_{10} fitted on BIPOLARCONCEPTS. Values closer to $+1$ indicate stronger positive tracking of probe AUC across layers.

6.5 OVERLAP AND BOOLEAN COMPOSITION

We examine which concept representations share directions and how this affects their interaction under removal. Using BIPOLARCONCEPTS, we compare the leading 20 conceptor axes at every recorded layer and average overlap across layers. Sentiment and Depression expression share more leading structure than either shares with political leaning in all five models (Table 3). Their layer-average overlaps are 0.318 in Llama and 0.221 in OLMo, compared with 0.096–0.108 and 0.035–0.042 for the respective political pairings. The ordering is consistent, while the separation varies substantially across families.

Concept pair	Gemma 2 2B	Gemma 2 9B	Qwen 2.5 3B	Llama 3.1 8B	OLMo 3 7B
Sentiment–Depression	**0.321**	**0.300**	**0.332**	**0.318**	**0.221**
Sentiment–Pol. Stance	0.154	0.134	0.169	0.096	0.035
Depression–Pol. Stance	0.163	0.143	0.182	0.108	0.042

Table 3: Leading-twenty overlap, averaged across layers.

Overlap (Equation 11) gives an exact interpretation of hard subspace removal. Thus, the measured overlaps imply 81.8–96.5% retained squared length for the uniformly weighted axes of the other concept when removing political-leaning directions. Removal between Sentiment and Depression retains 66.8–77.9%. These are layer-average geometric consequences of the measured overlaps. In the hard-projector setting, conceptor NOT uses the orthogonal complement, AND retains intersections, and OR combines spans (Jaeger, 2014). The observed pair differences quantify interference under the complement operation and provide geometric evidence relevant to composition.

6.6 BIPOLAR VS UNIPOLAR CAPTURE

We compare fixed mean-difference addition and centered conceptor amplification under bipolar and unipolar fitting, examining how each intervention changes evidence for the positive and opposite poles. Bipolar fitting uses positive and opposite examples; unipolar fitting pairs the positive pole with a neutral or absence reference. Fixed addition translates all inputs along the fitted mean difference, while conceptor amplification depends on each input's position relative to the fitting midpoint.

For each fit, let u be its normalized mean difference, c its midpoint, and C_k its conceptor. We compare

$$T_{\mathrm{add}}(h) = h + Du, \qquad T_{\mathrm{amp}}(h) = h + tC_k(h - c), \tag{12}$$

with $k \in \{1, 20\}$, $D = \beta s$, and $t = D/\operatorname{RMS}_s(C_k(h - c))$. Here s is a source-calibrated reference scale shared across methods, and RMS_s denotes weighted RMS over source calibration examples (Appendix C.7). Four grouped CONCEPTVARIANTS folds separate fitting and evaluation. Both fits are evaluated on the same held positive and opposite examples using a frozen, source-trained polarity readout f. Changes in standardized correct-pole scores are

$$\Delta M_+ = \mathbb{E}_+[f(T(h)) - f(h)], \Delta M_- = -\mathbb{E}_-[f(T(h)) - f(h)]. \tag{13}$$

Positive values indicate stronger evidence for the corresponding pole.

For two-sided amplification, the fitting contrast matters. Fixed mean-difference addition favors the positive pole on average, whereas bipolar C_{20} strengthens both mean margins in all thirty calibrated conditions, compared with ten for bipolar C_1 and seven for unipolar C_{20} (Table 4). Mean-line controls fitted against a reference do not reproduce this consistency, whether their presence collection contains only the positive pole or both poles together (Appendix C.7). Simply including examples from both poles is therefore insufficient: how their relationship is represented matters. These results support bipolar conceptor fitting when the goal is to strengthen each pole on its respective examples. An input-dependent bipolar mean-line amplifier also succeeds in all condition averages, but C_{20} achieves two-sided improvement more consistently across held folds (Appendix C.7). Our evaluation concerns activation geometry; we do not test open-ended generation steering.

	Bipolar fit			Unipolar fit		
Method	Positive pole	Opposite pole	Both improve	Positive pole	Opposite pole	Both improve
Mean addition	+0.2942	-0.2942	0/30	+0.0550	-0.0550	0/30
C_1 amplification	+0.0018	+0.0002	10/30	+0.0177	-0.0153	2/30
C_{20} amplification	+0.0301	+0.0322	30/30	+0.0177	-0.0083	7/30

Table 4: Mean standardized held-pole margin changes at $\beta = 0.10$ across five models.

7 CONCLUSION

Our exact pooling decomposition separates the average pole filter from mean-separation and covariance-shape contributions. Across five models, both contributions improve held-out contrast reconstruction under shared regularization. Additional directions preserve variation beyond the mean line, with benefits depending on model and layer. Reconstructing activations may require more directions than influencing behavior. Our findings are compatible with intervention evidence for a refusal-mediating direction (Arditi et al., 2024) and mean-difference truth directions that can generalize across datasets and causally affect outputs (Marks & Tegmark, 2023). For pole prediction, features from the rank-one mean component D_μ outperform pooled rank-twenty features in 25 of 30 model/data/site conditions using mean-token activations (Appendix C.3). In offline amplification, bipolar rank-twenty conceptors increase both mean held-pole readout margins more consistently than presence–absence fits. Fixed addition favors one pole; input-dependent updates can increase both mean margins. These results concern edited activations and frozen readouts; we do not evaluate downstream steering or generated text that is left as future work. Quota's observed association with layerwise label separability motivates hypotheses about where to probe. Cross-concept overlap quantifies interference under hard subspace removal, informing Boolean composition. Identifying effective steering layers and establishing the semantic effects of Boolean combinations require direct intervention studies.

AI USE STATEMENT

We used generative AI tools to assist with drafting and revising sections of the manuscript, improving clarity and organization, and identifying related literature. The authors reviewed the original sources themselves and assessed their relevance and how they were represented in the paper.

AI-assisted coding tools supported the development and revision of experimental code, the execution of experiments, and the implementation of robustness and sensitivity analyses reported in the appendices. Generative AI also assisted with refining experimental questions, discussing the interpretation of results, and writing and explaining mathematical derivations and proofs.

The authors reviewed the AI-assisted text, mathematical arguments, experimental code, and reported results. This review included checking the correspondence between the described methods and their implementation, examining experimental outputs, and verifying references against the original sources. Final decisions about the methodology, interpretation, and presentation were made by the authors. We take full responsibility for the final content of this work, including all text, claims, code, figures, and other artifacts produced with generative AI assistance.

ETHICS STATEMENT

This work examines concept representations in language-model activations using authored sentence collections and existing public datasets. Depression-related labels describe linguistic expressions, and the public mental-health benchmarks use proxy labels. Our results do not establish clinical validity for diagnosing depression, assessing suicide risk, or inferring an individual's mental health. Throughout the paper, "bipolar" and "unipolar" describe semantic comparisons. The labels "positive" and "negative" identify the sides of a comparison and express no judgment about their desirability.

Content warning: The appendix includes examples referring to depression and suicidal ideation, which some readers may find distressing. The authored examples document the language covered by the experimental collections and make the comparisons transparent. They are included for research purposes and provide no instructions for self-harm. Our use of "amplification" refers to changes in model activations and readout scores; it does not advocate intensifying psychological distress or encouraging harmful behavior.

Mental-health disclosures in public social-media datasets raise privacy concerns even when the text is publicly accessible. Public availability should not be interpreted as consent to sensitive profiling. Reuse and redistribution should respect source access conditions and minimize exposure of identifiable or sensitive text. The binary labels and authored reference sentences also simplify diverse experiences and viewpoints. Their interpretation may depend on language, culture, and context, and the experiments do not establish fairness across populations.

Methods for modifying concept representations could be misused for intrusive profiling, manipulation, or amplification of harmful content. Our experiments evaluate activation geometry and frozen readouts, without evaluating generated-text steering or deployment safety. Applications involving users, especially in mental-health settings, would require separate safety evaluation, appropriate oversight, and safeguards against unintended effects.

REPRODUCIBILITY STATEMENT

The assumptions and theoretical results are presented in Section 4, with complete proofs in Appendix A. Appendix B documents the sentence collections, public dataset sources, preprocessing, and evaluation splits. Appendix C provides experimental settings, fitting procedures, and additional robustness analyses. We will publicly release our codebase and the synthetic sentiment and political-stance sentence collections, together with experiment configurations and instructions for reproducing the analyses. The synthetic depression-expression collections will be distributed through gated access because they contain sensitive mental-health content, including references to suicidal ideation. The release documentation will explain the access procedure and conditions of use so that researchers can request these data for replication.

REFERENCES

Mohammad Mahdi Abootorabi, Armin Saghafian, Ali Bazshoushtari, Hamid Rezaei, EunJeong Hwang, Vered Shwartz, Parvin Mousavi, and Purang Abolmaesumi. Steering geometry: Validating human value geometry in LLM steering space, 2026. URL https://arxiv.org/abs/2609.06289.

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

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 1644–1650, 2020.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7 (2):179–188, 1936.

Donald P Green, Susan L Goldman, and Peter Salovey. Measurement error masks bipolarity in affect ratings. *Journal of personality and social psychology*, 64(6):1029, 1993.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, et al. Verbalizable representations form a global workspace in language models. *arXiv preprint arXiv:2607.15495*, 2026.

Herbert Jaeger. Controlling recurrent neural networks by conceptors. *arXiv preprint arXiv:1403.3369*, 2014.

Herbert Jaeger. Using conceptors to manage neural long-term memories for temporal patterns. *Journal of Machine Learning Research*, 18(13):1–43, 2017.

Saket Karve, Lyle Ungar, and João Sedoc. Conceptor debiasing of word representations evaluated on weat. In *Proceedings of the first workshop on gender bias in natural language processing*, pp. 40–48, 2019.

Corey LM Keyes. Mental illness and/or mental health? investigating axioms of the complete state model of health. *Journal of consulting and clinical psychology*, 73(3):539, 2005.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th international workshop on semantic evaluation*, pp. 829–839, 2019.

Dimitrios Kotzias. Sentiment labelled sentences. *UCI Mach. Learn. Repos*, 2015.

Tianlin Liu, João Sedoc, and Lyle Ungar. Correcting the common discourse bias in linear representation of sentences using conceptors. *arXiv preprint arXiv:1811.11002*, 2018.

Tianlin Liu, Lyle Ungar, and João Sedoc. Continual learning for sentence representations using conceptors. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3274–3279, 2019.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pp. 31–41, 2016.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pp. 115–124, 2005.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 76441–76463, 2025.

Kiho Park, Todd Nief, Yo Joong Choe, and Victor Veitch. The information geometry of softmax: Probing and steering. *arXiv preprint arXiv:2602.15293*, 2026.

Joris Postmus and Steven Abreu. Steering large language models using conceptors: Improving addition-based activation engineering. *arXiv preprint arXiv:2410.16314*, 2024.

Sunny Rai, Khushi Shelat, Devansh Jain, Ashwin Kishen, Young Min Cho, Maitreyi Redkar, Samindara Hardikar-Sawant, Lyle Ungar, and Sharath Chandra Guntuku. Cross-cultural differences in mental health expressions on social media. In Vinodkumar Prabhakaran, Sunipa Dev, Luciana Benotti, Daniel Hershcovich, Yong Cao, Li Zhou, Laura Cabello, and Ife Adebara (eds.), *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pp. 132–142, Albuquerque, New Mexico, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-237-4. doi: 10.18653/v1/2025.c3nlp-1.10. URL https://aclanthology.org/2025.c3nlp-1.10/.

Federico Ravenda, Seyed Ali Bahrainian, Andrea Raballo, and Antonietta Mira. Navigating through the hidden embedding space: steering llms to improve mental health assessment. In *Proceedings of the 41st ACM/SIGAPP Symposium on Applied Computing*, pp. 227–229, 2026.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

João Sedoc and Lyle Ungar. The role of protected class word lists in bias identification of contextualized word representations. In *Proceedings of the first workshop on gender bias in natural language processing*, pp. 55–61, 2019.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025.

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 *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, et al. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*, 2026.

thePixel42. Depression detection. Hugging Face dataset repository, n.d. URL https://huggingface.co/datasets/thePixel42/depression-detection.

Ilias Triantafyllopoulos, Young-Min Cho, Ren Tao, Miranda Muqing Miao, Sunny Rai, Lyle Ungar, Sharath Chandra Guntuku, Neville Ryant, and João Sedoc. Conceptors for semantic steering. *arXiv preprint arXiv:2605.04980*, 2026.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

David Watson and Auke Tellegen. Toward a consensual structure of mood. *Psychological bulletin*, 98(2):219, 1985.

Li S Yifei, Lyle Ungar, and João Sedoc. Conceptor-aided debiasing of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10703–10727, 2023.

Fangyi Zhu, Ajay Subramanian, Allison Constant, Camille Wang, Ravish Gupta, and Corey J Keller. Interpretable symptom vectors for depression in a large language model. *arXiv preprint arXiv:2609.01832*, 2026.

A Proofs and supporting derivations

A.1 Derivation of the conceptor solution

For completeness, we derive the standard conceptor solution from the regularized reconstruction objective of Jaeger (2014; 2017). Let $x_i = h_i - c$ and $R = R_c = n^{-1}\sum_i x_i x_i^\top$. Writing $L(M)$ for the objective in Equation 3, we have

$$\begin{aligned} L(M) &= \frac{1}{n}\sum_{i=1}^{n} \|x_i - Mx_i\|_2^2 + \lambda\|M\|_F^2 \\ &= \operatorname{tr}\big((I-M)R(I-M)^\top\big) + \lambda \operatorname{tr}(MM^\top) \\ &= \operatorname{tr}(R) - 2\operatorname{tr}(MR) + \operatorname{tr}\big(M(R+\lambda I)M^\top\big). \end{aligned} \tag{14}$$

Here we use symmetry of R and the cyclic property of the trace. Differentiating with respect to M gives

$$\nabla_M L = 2M(R+\lambda I) - 2R. \tag{15}$$

Setting this gradient to zero yields $M(R+\lambda I) = R$. Since $R \succeq 0$ and $\lambda > 0$, the matrix $R+\lambda I$ is positive definite and hence invertible. Thus the stationary point is

$$C_\lambda(R) = R(R+\lambda I)^{-1}. \tag{16}$$

Positive definiteness also makes the quadratic term in Equation 14 strictly convex in M. The stationary point is therefore the unique global minimizer. Finally, substituting the eigendecomposition of R into this solution gives the gains in Equation 5.

A.2 Proof of the pooling decomposition

Proof of Proposition 1. We use the notation of the proposition. For finite fitting sets, expectations are empirical averages normalized by $1/n_s$ within each pole.

The pooled covariance. The pooled mean is $\mu = p\mu_+ + q\mu_-$, so $\mu_+ - \mu = q\delta$ and $\mu_- - \mu = -p\delta$. For the positive pole, write $h - \mu = (h-\mu_+) + q\delta$. Since $\mathbb{E}_+[h-\mu_+] = 0$, expansion gives

$$\mathbb{E}_+[(h-\mu)(h-\mu)^\top] = R_+ + q^2\delta\delta^\top.$$

The negative pole similarly gives $R_- + p^2\delta\delta^\top$. Taking their weighted average yields

$$\begin{aligned} R_{\text{pool}} &= pR_+ + qR_- + (pq^2 + qp^2)\delta\delta^\top \\ &= R_{\text{w}} + pq\delta\delta^\top. \end{aligned} \tag{17}$$

The mean update. The conceptor formula can be written as $C_\lambda(R) = I - \lambda(R+\lambda I)^{-1}$. Set $A = R_{\text{w}} + \lambda I$ and $t = pq$. Because $A \succ 0$, the Sherman–Morrison identity gives

$$(A + t\delta\delta^\top)^{-1} = A^{-1} - \frac{tA^{-1}\delta\delta^\top A^{-1}}{1 + t\delta^\top A^{-1}\delta}. \tag{18}$$

Consequently,

$$\begin{aligned} C_\lambda(R_{\text{pool}}) - C_\lambda(R_{\text{w}}) &= \lambda\big[A^{-1} - (A + t\delta\delta^\top)^{-1}\big] \\ &= \frac{\lambda pq}{1 + pq\delta^\top A^{-1}\delta}(A^{-1}\delta)(A^{-1}\delta)^\top = D_\mu. \end{aligned} \tag{19}$$

The scalar multiplier is positive. Since A is invertible, D_μ is positive semidefinite, has rank one for $\delta \neq 0$, and is zero for $\delta = 0$. If $R_{\text{w}} v_j = r_j v_j$, the v_j coordinate of $A^{-1}\delta$ is $(v_j^\top \delta)/(r_j + \lambda)$. This gives the weighting described in the main text.

The covariance-shape update. Define

$$A_+ = R_+ + \lambda I, \qquad A_- = R_- + \lambda I, \qquad H = R_+ - R_-.$$

Then $A = pA_+ + qA_-$, $A_+ = A + qH$, and $A_- = A - pH$. Using the inverse form of the conceptor gives

$$J_R = C_\lambda(R_{\mathrm{w}}) - \bar{C} = \lambda\big[pA_+^{-1} + qA_-^{-1} - A^{-1}\big]. \tag{20}$$

To obtain its sign and rank, we use the exact identity

$$(A+E)^{-1} = A^{-1} - A^{-1}EA^{-1} + A^{-1}E(A+E)^{-1}EA^{-1}. \tag{21}$$

It follows by substituting $(A+E)^{-1} = A^{-1} - (A+E)^{-1}EA^{-1}$ into $(A+E)^{-1} = A^{-1} - A^{-1}E(A+E)^{-1}$. Applying Equation 21 to $E = qH$ and $E = -pH$ gives

$$A_+^{-1} = A^{-1} - qA^{-1}HA^{-1} + q^2A^{-1}HA_+^{-1}HA^{-1},$$
$$A_-^{-1} = A^{-1} + pA^{-1}HA^{-1} + p^2A^{-1}HA_-^{-1}HA^{-1}.$$

Weighting these equations by p and q cancels the terms that are linear in H. Substitution into Equation 20 yields

$$J_R = \lambda pq\, A^{-1}HBHA^{-1}, \qquad B = qA_+^{-1} + pA_-^{-1}. \tag{22}$$

Both A and B are positive definite. With $K = B^{1/2}HA^{-1}$, the factorization becomes $J_R = \lambda pq\, K^\top K \succeq 0$. The matrices on either side of H in K are invertible, giving

$$\mathrm{rank}(J_R) = \mathrm{rank}(K) = \mathrm{rank}(H).$$

Thus $J_R = 0$ exactly when $R_+ = R_-$. This argument allows arbitrary relative orientations of the two covariances.

Combining the terms. The two updates satisfy $C_\lambda(R_{\mathrm{pool}}) = C_\lambda(R_{\mathrm{w}}) + D_\mu$ and $C_\lambda(R_{\mathrm{w}}) = \bar{C} + J_R$. Substituting the second equality into the first proves Equation 8. □

A.3 Mean-only conceptors and the mean-difference axis

The standard conceptor spectral formula (Jaeger, 2014; 2017) has an exact mean-difference case.

Proposition 2 (Mean-difference axis)**.** *Let $\delta = \mu_+ - \mu_- \neq 0$ and $u = \delta/\|\delta\|_2$. Fitting only the two equally weighted pole means about their midpoint $c = (\mu_+ + \mu_-)/2$ gives*

$$C_{\mathrm{means}} = \frac{\|\delta\|_2^2}{\|\delta\|_2^2 + 4\lambda}\, uu^\top, \qquad \lambda > 0. \tag{23}$$

Its sole retained axis is the mean-difference line. When all examples are fitted, the full conceptor remains rank one precisely when within-pole variation is confined to this line.

The labels orient u toward the positive mean; filtering scales each input's coordinate along this axis by the regularized gain. Compressing each pole to its mean gives one direction; variation outside that line adds retained directions.

We prove Proposition 2 using the spectral formula in Equation 5, following Jaeger (2014; 2017). Throughout, $\delta = \mu_+ - \mu_- \neq 0$, $u = \delta/\|\delta\|_2$, and $\lambda > 0$.

Fitting the two means. Let the pole weights be $p, q > 0$, with $p + q = 1$. Their pooled mean is $\mu = p\mu_+ + q\mu_-$, so centering gives $x_+ = \mu_+ - \mu = q\delta$ and $x_- = \mu_- - \mu = -p\delta$. The resulting moment is

$$R_{\mathrm{means}} = px_+x_+^\top + qx_-x_-^\top = pq\delta\delta^\top = pq\|\delta\|_2^2 uu^\top. \tag{24}$$

It has one positive eigenvalue, $pq\|\delta\|_2^2$, along u. The conceptor transforms this eigenvalue by $r \mapsto r/(r+\lambda)$, giving

$$C_\lambda(R_{\mathrm{means}}) = \frac{pq\|\delta\|_2^2}{pq\|\delta\|_2^2 + \lambda}\, uu^\top. \tag{25}$$

Setting $p = q = 1/2$ proves Equation 23. The same calculation applies to a presence–absence comparison after replacing μ_- by μ_0.

The positive-mean construction. For balanced weights, fit the single positive mean about the same midpoint $c = (\mu_+ + \mu_-)/2$. Its fitting vector is $\mu_+ - c = \delta/2$, so its second moment is $\delta\delta^\top/4$. It therefore gives exactly the same conceptor as the balanced two-mean fit at the same λ. Both pole means are used to determine this reference.

Including within-pole variation. Let $R_{\mathrm{w}} = pR_+ + qR_-$ be the within-pole covariance in Equation 7, and write $C_{\mathrm{pool}} = C_\lambda(R_{\mathrm{pool}})$. The spectral transformation preserves zero and positive eigenvalues, so C_{pool} and R_{pool} have the same range. For any $z \in \mathbb{R}^d$,

$$z^\top R_{\mathrm{pool}} z = z^\top R_{\mathrm{w}} z + pq(\delta^\top z)^2. \tag{26}$$

Both terms are nonnegative. Their sum vanishes exactly when $z \in \ker(R_{\mathrm{w}}) \cap \delta^\perp$. Taking orthogonal complements gives

$$\mathrm{range}(C_{\mathrm{pool}}) = \mathrm{range}(R_{\mathrm{w}}) + \mathrm{span}(\delta). \tag{27}$$

Since $\delta \neq 0$, this range is one-dimensional exactly when $\mathrm{range}(R_{\mathrm{w}}) \subseteq \mathrm{span}(\delta)$. This proves the full-filter rank-one condition stated in Proposition 2.

A.4 Overlap and hard subspace removal

Let $U_A, U_B \in \mathbb{R}^{d\times k}$ have orthonormal columns spanning two subspaces, and let $P_A = U_A U_A^\top$. Define their overlap as

$$\mathrm{Overlap}(A, B) = \frac{1}{k}\|U_A^\top U_B\|_F^2. \tag{28}$$

Removing subspace A applies $I - P_A$. For each unit column u_j of U_B, orthogonality gives $\|(I - P_A)u_j\|_2^2 = 1 - \|U_A^\top u_j\|_2^2$. Averaging over the k columns yields

$$\frac{1}{k}\|(I - P_A)U_B\|_F^2 = 1 - \frac{1}{k}\|U_A^\top U_B\|_F^2 = 1 - \mathrm{Overlap}(A, B). \tag{29}$$

Overlap therefore measures the fraction of average squared length lost from these axes under hard removal. The identity assumes orthogonal projectors and equal weighting of the basis vectors. Soft conceptor operations also depend on their gains. Effects on semantic content require separate evaluation.

B Dataset details

B.1 Authored concept collections

BipolarConcepts. Each matched pair contains one sentence from each pole. The inspected native inputs contain 100 pairs each for Sentiment, Depression and Pol. Stance (600 sentence entries). The depression counter-expressions include positive well-being. The labels describe the content expressed in the sentences. Table 5 gives the inventory for both authored concept collections.

ConceptVariants. Each topic is expressed in several formulations. For sentiment and political leaning, descriptive-neutral text supplies a topic description; attitudinal-neutral and undecided text supply alternative references. The symptom comparisons distinguish ordinary functioning, explicit symptom absence, flourishing, and recovery. All formulations of a topic remain together when constructing evaluation folds. The static comparator-reconstruction analysis excludes recovery because it introduces a temporal change. There are 24 topics per concept, with one sentence per formulation.

B.2 ConceptGroups

The suite contains three collections with different grouping units. Table 7 gives their individual group sizes alongside word-length statistics.

Sentiment. The texts come from UCI Sentiment Labelled Sentences (Kotzias, 2015). Amazon, IMDb, and Yelp define the three groups. We retain the source positive/negative labels and use the review domain to define evaluation folds.

Collection	Concept	Formulations	Texts
BIPOLARCONCEPTS	Sentiment	Positive; negative	200
	Pol. Stance	Left; right	200
	Depression expression	Depression; counter-expressions	200
CONCEPTVARIANTS	Sentiment	Positive; negative; descriptive neutral; attitudinal neutral	96
	Pol. Stance	Left; right; descriptive neutral; undecided	96
	Depression expression	Depression; ordinary functioning; explicit absence; flourishing; recovery	120

Table 5: Authored collections: formulations and sentence counts.

Pol. Stance. The tweets come from SemEval/TweetEval (Mohammad et al., 2016; Barbieri et al., 2020). Preparation combines the original training and test files and retains favor/against labels. Removing four repeated texts from 998 packaged rows leaves 994 tweets across abortion, atheism, climate change, feminism, and Hillary Clinton. Evaluation uses custom held-topic folds. The labels express stance toward a specified topic.

Depression. The authored first-person sentences span fourteen facets, including sleep, energy, and interest. Each facet contains seven matched pairs contrasting symptom expression with ordinary functioning or positive well-being. Pair members remain together during sampling and splitting. These sentences serve as linguistic stimuli.

Grouped evaluation. Every prepared domain, facet, or topic contains equal numbers of examples from its two labels. Each collection is evaluated separately. One entire group is excluded from fitting, and its labels are used to compute the evaluation mean difference. Fitting weights source groups equally and weights the two labels equally within each source group. Any selection of a direction, gain, or layer uses source groups only. We repeat over the held-out groups and average their scores equally.

B.3 PUBLIC PROBING BENCHMARKS

Table 6 records the targets and partitions used for probing. The three sentiment benchmarks are compared with sentiment spectra, Hyperpartisan News with political-leaning spectra, and the two Reddit benchmarks with depression-expression spectra.

Benchmark	Probe target	Training / evaluation
SST-2 (Socher et al., 2013)	Positive / negative sentiment	Train / validation
Rotten Tomatoes (Pang & Lee, 2005)	Positive / negative sentiment	Train / test
TweetEval sentiment (Barbieri et al., 2020)	Positive / negative sentiment	Train / test
Hyperpartisan News (Kiesel et al., 2019)	Hyperpartisan / mainstream	Train / validation
Reddit Depression (thePixel42, n.d.)	Released binary proxy label	Train / test
Cross-cultural Reddit (Rai et al., 2025)	Mental-health proxy label	Existing train / test

Table 6: Public probing targets and recorded train/evaluation partitions.

TweetEval sentiment excludes neutral labels. Hyperpartisan News uses its hyperpartisan/mainstream target, which differs from left/right leaning. The Reddit labels are proxies and do not establish clinical diagnoses. Cross-cultural Reddit uses pre-existing files described as a user-level train/test split.

B.4 LENGTH STATISTICS

Word counts use whitespace-delimited units before tokenization. Table 7 reports mean, 95th-percentile, and maximum word lengths over each complete prepared group or collection. CONCEPTVARIANTS additionally averages 63.0 Unicode characters per sentence.

Collection	Group	Texts	Mean words	95th pct.	Max.
CONCEPTVARIANTS	All concepts	312	10.0	14.0	16
CONCEPTGROUPS					
Sentiment	Amazon	200	11.5	26.0	30
	IMDb	600	14.3	33.0	61
	Yelp	600	11.0	23.0	32
Pol. Stance	Abortion	240	16.4	24.0	30
	Atheism	230	17.7	24.0	31
	Climate change	48	16.0	22.3	24
	Feminism	240	17.4	24.0	26
	Hillary Clinton	236	14.9	23.0	25
Depression	All 14 facets	196	6.8	9.0	12

Table 7: Prepared sentence counts and word lengths; Depression aggregates fourteen equal-sized facets.

Collection	Gemma median / max.	Qwen median / max.
CONCEPTVARIANTS	12 / 19	11 / 18
CONCEPTGROUPS: Sentiment	14 / 82	13 / 78
CONCEPTGROUPS: Pol. Stance	25 / 48	24 / 50
CONCEPTGROUPS: Depression	9 / 16	8 / 14

Table 8: Recorded token-count medians and maxima for the prepared geometry inputs.

Table 8 reports token medians and maxima from the Gemma 2 and Qwen 2.5 tokenizers. Inputs are plain text with special tokens, without a chat template or truncation.

B.5 SENTENCE EXAMPLES

Tables 9–13 illustrate every collection, comparator formulation, and evaluation group. For Depression, the positive pole denotes symptom expression; authored counterparts include ordinary functioning and positive well-being. Recovery appears in the inventory but is excluded from the static comparator and amplification analyses. Pol. Stance denotes left/right policy contrasts in the authored collections and favor/against positions on specified topics in CONCEPTGROUPS. Review and tweet examples placed in the same row are independent texts, not matched sentence pairs.

Concept	Recorded positive pole	Recorded negative pole
Sentiment	I feel great today.	I feel bad today.
Pol. Stance	I support student debt relief.	I oppose student debt cancellation.
Depression	Nothing interests me.	Everything interests me.

Table 9: Matched BIPOLARCONCEPTS examples; positive denotes sentiment, left-leaning policy, or depression expression.

Concept	Formulation	Matched-topic example
Sentiment	Positive	I found the film wonderfully engaging.
	Negative	I found the film painfully dull.
	Descriptive neutral	I watched the film in the evening.
	Attitudinal neutral	I had no strong feelings about the film.
Pol. Stance	Left	Regarding health insurance policy, I favor expanding publicly funded health coverage.
	Right	Regarding health insurance policy, I favor relying more on privately purchased health coverage.
	Descriptive neutral	Regarding health insurance policy, different proposals are being discussed by the public.
	Undecided	Regarding health insurance policy, I have not settled on a policy preference.
Depression	Symptom expression	My usual hobbies have lost all their appeal.
	Ordinary functioning	My usual hobbies interest me about as much as before.
	Explicit absence	My usual hobbies have not lost their ordinary appeal.
	Flourishing	My usual hobbies fill me with excitement and delight.
	Recovery	My usual hobbies interest me again after a period of losing their appeal.

Table 10: Every CONCEPTVARIANTS formulation, illustrated by one topic per concept; recovery is excluded from static analyses.

Review domain	Positive	Negative
Amazon	Battery life is also great!	Awkward to use and unreliable.
IMDb	This is such a fun and funny movie.	The problem was the script.
Yelp	The atmosphere here is fun.	The manager was the worst.

Table 11: CONCEPTGROUPS: positive and negative Sentiment examples from each review domain.

Facet	Symptom expression	Ordinary / positive counterpart
Appetite	I have no appetite and food feels like a chore.	I enjoy my meals and eat well.
Burden	I feel like a burden to everyone around me.	I feel valued by the people around me.
Concentration	I can't concentrate on anything anymore.	I concentrate easily on what I'm doing.
Energy	I feel drained and tired all the time.	I feel energetic and lively all the time.
Hopelessness	The future looks completely bleak to me.	I look forward to the future with hope.
Interest	Nothing interests me anymore.	I'm genuinely interested in lots of things.
Irritability	Every little thing makes me snap.	Little things don't get to me.
Mood	I feel sad and down most of the time.	I feel content most of the time.
Psychomotor	I move and speak as if in slow motion.	I move and speak at my normal pace.
Sleep	I lie awake for hours every night.	I fall asleep easily every night.
Suicidality	Sometimes I wish I just wouldn't wake up.	I'm glad to wake up to each day.
Tearfulness	I cry all the time, often for no reason.	I rarely feel the urge to cry.
Withdrawal	I've cut myself off from everyone I know.	I stay close to the people I know.
Worthlessness	I feel like a worthless failure.	I feel capable and worthwhile.

Table 12: CONCEPTGROUPS: matched Depression expressions from all fourteen facets.

Topic	Favor	Against
Abortion	So prolifers are against health rights?	#ProLifeYouth know that life begins at conception.
Atheism	god of the gaps is not evidence #next	God is so good! All the praises go up to Him!
Climate change	RT @user 700 people dead in Pakistan due to heatwave.	@user ice cracking in the summer?? SO ALARMING. #climatechangehoax
Feminism	FINALLY A WOMEN RUNNING FOR PRESIDENT	Feminism is nothing more than a cult that closes ranks whenever someone dares question them.
Hillary Clinton	RT @user Hillary for the win	Hillary Clinton for Ambassador to ISIS.

Table 13: CONCEPTGROUPS: favor/against examples for all five Pol. Stance topics.

C Additional experimental details and results

C.1 Feature sites, fitting, and uncertainty

Table 14 lists the five models, hidden dimensions, and zero-based midpoint/final geometry sites. Mean pooling uses nonpadding tokens; last-token controls use the last nonpadding token. Layer plots display one-based decoder-layer numbers.

Model	Blocks	d	Midpoint	Final
Gemma-2-2B-IT	26	2,304	13	25
Gemma-2-9B-IT	42	3,584	21	41
Qwen-2.5-3B-Instruct	36	2,048	18	35
Llama-3.1-8B-Instruct	32	4,096	16	31
OLMo-3-7B-Instruct	32	4,096	16	31

Table 14: Five-model geometry sites (zero-based raw blocks).

For ConceptGroups, one complete group is held out in each fit. The three collections supply three review domains, fourteen symptom facets, and five stance topics, yielding 110 held-model/group fits per feature site. Source groups receive equal weight, with equal pole weights within each group. Empirical covariances use population normalization. Held groups contribute only to evaluation. Any selection of a direction, gain, or readout penalty uses source groups. The source weights, evaluated rows, and numerical settings are fixed within each comparison.

The reference experiment uses 2,000 complete-pipeline bootstrap refits, as detailed in Appendix C.5. Archived coverage intervals instead resample held-group scores 10,000 times while keeping fits fixed; component intervals use 2,000 such draws. These intervals condition on the fitted folds. There are only three review-domain resampling units.

C.2 Pooling controls and the original fitting collections

Matched counts. Within each source group, a deterministic hash ordering selects a bank of B rows per pole, with B the largest eligible even count up to 32. Each fit receives B rows per group: half from each pole for pooling, or all from one pole otherwise. Total fitting budgets are 64 rows for Sentiment, 78 for Depression, and 96 or 128 for Pol. Stance. The shared centroid/calibration bank has twice the fitting count. Authored pair units remain intact.

Own-centered fits subtract their respective fitting means. Shared-center fits subtract the same balanced source-bank center. Ranks 1, 10, and 20 are evaluated at the midpoint/mean-token sites. Table 1 reports rank twenty. Pooled rank-one support exceeds the count-matched mean line in nine of fifteen conditions: all five Pol. Stance conditions and four of five Sentiment conditions. It falls below the mean line for OLMo Sentiment and all five Depression conditions.

BipolarConcepts Figure 4 reports leading-twenty capture of each fitting collection's own mean difference. Pooled capture has the highest layer average in all fifteen model–concept pairs; Sentiment averages range from 97.4% to 99.7%. These training contrasts describe fitted geometry and are distinct from the held-group evaluations.

C.3 Complete decomposition ablations and gain controls

Shared regularization. All four operators use full nonzero fitted supports and share the source fold's regularization:

$$\lambda = \tau \ \mathrm{tr}(R_{\mathrm{pool}})/d, \qquad \tau \in \{0.1, 1, 10\},$$

with $\tau = 1$ as the primary scale. Tables 15–19 report both mean-token sites and all scales. The full pooled filter has lower group-averaged reconstruction error than each ablation in all 90 model/data/site/scale settings: 270 paired comparisons, with full-minus- ablated errors from -0.224842 to -0.000283.

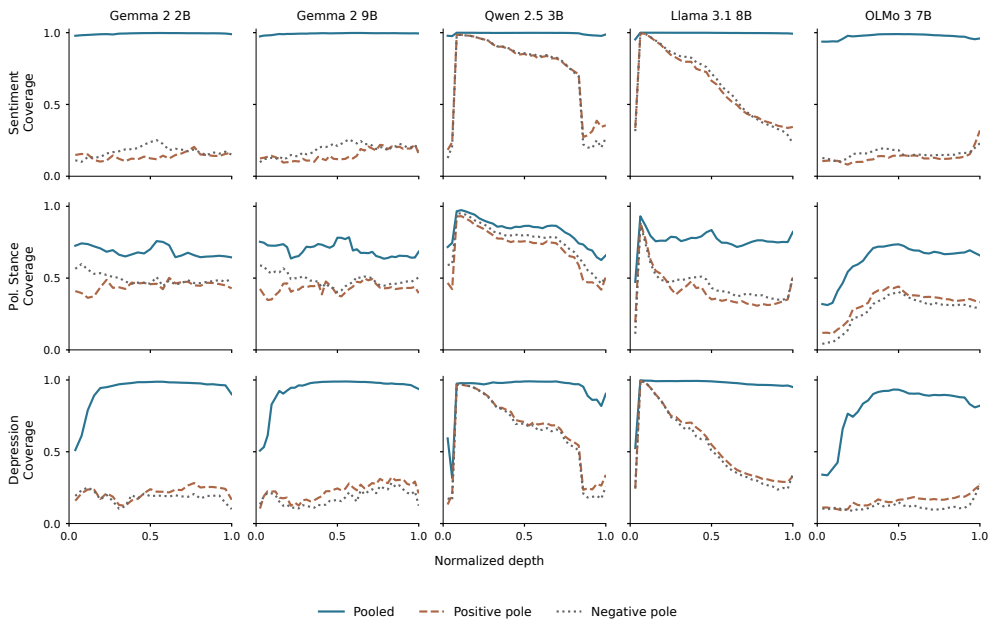


Figure 4: **Original bipolar capture.** Leading-twenty coverage of each fitting collection's own mean difference across layers; these are training contrasts.

Every operator receives the held contrast, which identity reconstructs with zero error. These comparisons concern the fitted filter family; positive semidefiniteness alone does not guarantee the observed ordering for every vector.

Data	Site	τ	$\bar{C}$	$+D_\mu$	$+J_R$	Full
Sentiment	Mid	0.1	0.1344	0.1052	0.0747	**0.0536**
	Mid	1	0.2489	0.1466	0.1679	**0.0984**
	Mid	10	0.4637	0.2689	0.4284	**0.2460**
	Final	0.1	0.1964	0.1797	0.1262	**0.1139**
	Final	1	0.2792	0.2347	0.2166	**0.1855**
	Final	10	0.4805	0.3584	0.4376	**0.3294**
Pol. Stance	Mid	0.1	0.1614	0.1605	0.1147	**0.1139**
	Mid	1	0.2088	0.2074	0.1786	**0.1774**
	Mid	10	0.3539	0.3504	0.3346	**0.3314**
	Final	0.1	0.1626	0.1619	0.1140	**0.1135**
	Final	1	0.2087	0.2078	0.1765	**0.1758**
	Final	10	0.3532	0.3510	0.3307	**0.3287**
Depression	Mid	0.1	0.3111	0.2577	0.2114	**0.1619**
	Mid	1	0.3265	0.2529	0.2300	**0.1732**
	Mid	10	0.4068	0.2801	0.3461	**0.2396**
	Final	0.1	0.4228	0.3748	0.3255	**0.2792**
	Final	1	0.4321	0.3745	0.3365	**0.2876**
	Final	10	0.4967	0.4102	0.4315	**0.3585**

Table 15: **Gemma-2B reconstruction ablations.** Lower normalized error is better. Updates augment $\bar{C}$ at shared $\lambda = \tau \ \mathrm{tr}(R_{\mathrm{pool}})/d$; bold marks the row minimum.

Data	Site	τ	$\bar{C}$	$+D_\mu$	$+J_R$	Full
Sentiment	Mid	0.1	0.1573	0.1262	0.0987	**0.0756**
	Mid	1	0.2573	0.1626	0.1878	**0.1214**
	Mid	10	0.4398	0.2722	0.4127	**0.2545**
	Final	0.1	0.2506	0.2315	0.1750	**0.1597**
	Final	1	0.3274	0.2820	0.2620	**0.2298**
	Final	10	0.5093	0.3922	0.4666	**0.3640**
Pol. Stance	Mid	0.1	0.2117	0.2101	0.1634	**0.1620**
	Mid	1	0.2491	0.2460	0.2140	**0.2116**
	Mid	10	0.3791	0.3718	0.3592	**0.3526**
	Final	0.1	0.2450	0.2440	0.1883	**0.1874**
	Final	1	0.2746	0.2732	0.2301	**0.2290**
	Final	10	0.3952	0.3921	0.3697	**0.3671**
Depression	Mid	0.1	0.2840	0.2331	0.1985	**0.1506**
	Mid	1	0.2948	0.2281	0.2119	**0.1587**
	Mid	10	0.3554	0.2458	0.3018	**0.2099**
	Final	0.1	0.3604	0.3124	0.2858	**0.2389**
	Final	1	0.3660	0.3110	0.2915	**0.2431**
	Final	10	0.4069	0.3295	0.3529	**0.2874**

Table 16: **Gemma-9B reconstruction ablations.** Lower normalized error is better. Updates augment $\bar{C}$ at shared $\lambda = \tau\ \mathrm{tr}(R_{\mathrm{pool}})/d$; bold marks the row minimum.

Data	Site	τ	$\bar{C}$	$+D_\mu$	$+J_R$	Full
Sentiment	Mid	0.1	0.0291	0.0182	0.0218	**0.0137**
	Mid	1	0.0584	0.0411	0.0552	**0.0388**
	Mid	10	0.0783	0.0733	0.0780	**0.0730**
	Final	0.1	0.1675	0.1545	0.1036	**0.0937**
	Final	1	0.2213	0.1918	0.1679	**0.1461**
	Final	10	0.3924	0.3040	0.3422	**0.2678**
Pol. Stance	Mid	0.1	0.0920	0.0914	0.0748	**0.0744**
	Mid	1	0.1650	0.1639	0.1550	**0.1540**
	Mid	10	0.2882	0.2867	0.2837	**0.2822**
	Final	0.1	0.1702	0.1698	0.1214	**0.1211**
	Final	1	0.2142	0.2133	0.1831	**0.1824**
	Final	10	0.3541	0.3517	0.3358	**0.3336**
Depression	Mid	0.1	0.0717	0.0591	0.0502	**0.0405**
	Mid	1	0.0870	0.0649	0.0733	**0.0551**
	Mid	10	0.1158	0.0952	0.1138	**0.0936**
	Final	0.1	0.4732	0.4164	0.3633	**0.3096**
	Final	1	0.4851	0.4140	0.3787	**0.3202**
	Final	10	0.5636	0.4577	0.4936	**0.4034**

Table 17: **Qwen-3B reconstruction ablations.** Lower normalized error is better. Updates augment $\bar{C}$ at shared $\lambda = \tau\ \mathrm{tr}(R_{\mathrm{pool}})/d$; bold marks the row minimum.

Data	Site	τ	$\bar{C}$	$+D_\mu$	$+J_R$	Pooled
Sentiment	Mid	0.1	0.1011	0.0786	0.0695	**0.0540**
Sentiment	Mid	1	0.1687	0.1080	0.1437	**0.0931**
Sentiment	Mid	10	0.2460	0.1980	0.2420	**0.1947**
Sentiment	Final	0.1	0.1698	0.1567	0.1096	**0.0982**
Sentiment	Final	1	0.1985	0.1765	0.1436	**0.1274**
Sentiment	Final	10	0.3237	0.2511	0.2677	**0.2126**
Pol. Stance	Mid	0.1	0.1998	0.1984	0.1595	**0.1583**
Pol. Stance	Mid	1	0.2446	0.2416	0.2192	**0.2169**
Pol. Stance	Mid	10	0.3647	0.3596	0.3514	**0.3466**
Pol. Stance	Final	0.1	0.1597	0.1589	0.1206	**0.1199**
Pol. Stance	Final	1	0.1820	0.1809	0.1520	**0.1511**
Pol. Stance	Final	10	0.2742	0.2721	0.2565	**0.2545**
Depression	Mid	0.1	0.1810	0.1507	0.1340	**0.1063**
Depression	Mid	1	0.1905	0.1484	0.1478	**0.1144**
Depression	Mid	10	0.2319	0.1712	0.2139	**0.1580**
Depression	Final	0.1	0.4407	0.3810	0.3306	**0.2722**
Depression	Final	1	0.4462	0.3785	0.3351	**0.2752**
Depression	Final	10	0.4936	0.3938	0.3991	**0.3188**

Table 18: Llama-3.1-8B: held-contrast reconstruction error, averaged equally over held groups; lower is better.

Data	Site	τ	$\bar{C}$	$+D_\mu$	$+J_R$	Pooled
Sentiment	Mid	0.1	0.3104	0.2661	0.1887	**0.1477**
Sentiment	Mid	1	0.3352	0.2714	0.2136	**0.1654**
Sentiment	Mid	10	0.4719	0.3207	0.3575	**0.2470**
Sentiment	Final	0.1	0.2560	0.2346	0.1773	**0.1581**
Sentiment	Final	1	0.2799	0.2479	0.2078	**0.1830**
Sentiment	Final	10	0.4007	0.3250	0.3413	**0.2814**
Pol. Stance	Mid	0.1	0.3100	0.3090	0.2508	**0.2498**
Pol. Stance	Mid	1	0.3267	0.3256	0.2717	**0.2707**
Pol. Stance	Mid	10	0.4228	0.4211	0.3905	**0.3890**
Pol. Stance	Final	0.1	0.2571	0.2558	0.2044	**0.2031**
Pol. Stance	Final	1	0.2757	0.2742	0.2289	**0.2275**
Pol. Stance	Final	10	0.3701	0.3680	0.3426	**0.3408**
Depression	Mid	0.1	0.6674	0.5932	0.5519	**0.4786**
Depression	Mid	1	0.6709	0.5897	0.5539	**0.4799**
Depression	Mid	10	0.7018	0.5928	0.5988	**0.5101**
Depression	Final	0.1	0.5666	0.5165	0.4729	**0.4237**
Depression	Final	1	0.5703	0.5152	0.4753	**0.4253**
Depression	Final	10	0.6040	0.5312	0.5202	**0.4597**

Table 19: OLMo-3-7B: held-contrast reconstruction error, averaged equally over held groups; lower is better.

Comparing directions at equal rank and trace. In a follow-up control, we replace each source-fitted operator $M \in \mathcal{M} = \{\bar{C}, \bar{C} + D_\mu, \bar{C} + J_R, C_{\text{pool}}\}$ by the orthogonal projector $P_{20}(M) = U_{20}(M)U_{20}(M)^\top$ onto its leading twenty eigenvectors. All fits have at least twenty nonzero modes. Every projector therefore has rank twenty, trace twenty, and twenty unit gains, with directions selected from source data alone. These are hard projectors, rather than the soft filters C_{20}. We retain the same held-group folds, sites, and shrinkage scales, and average held groups equally. Coverage is measured by equation 10; projector reconstruction error equals one minus coverage.

The pooled projector has higher coverage in 251 of 270 comparisons (Table 20). At the primary midpoint setting with $\tau = 1$, it wins 41 of 45 comparisons and has the highest coverage in 12 of 15 model/data settings. These counts describe the evaluated grid, not independent replications; the pooled and within-covariance projectors are unchanged across τ.

Ablated projector	Pooled wins	Pooled losses	Mean gain (pp)
$P_{20}(\bar{C})$	84	6	+14.56
$P_{20}(\bar{C} + D_\mu)$	77	13	+7.26
$P_{20}(\bar{C} + J_R)$	90	0	+10.10
All comparisons	251	19	+10.64

Table 20: Pooled versus ablated rank-20 projector coverage across five models, three datasets, two sites, and three scales. Gains are pooled-minus-ablation coverage in percentage points. Each ablation has 90 comparisons; no ties occur within 10^{-10}.

The advantage is not guaranteed. Pooling prioritizes total variance, so within-pole variation can enter the leading twenty directions and displace a direction aligned with the held contrast, even when the source and held mean contrasts agree.

Equal trace for the full soft filters. As a separate control, we uniformly rescale the full operators using only their source-fitted traces:

$$t_* = \min_{M \in \mathcal{M}} \operatorname{tr}(M), \qquad \widetilde{M} = \frac{t_*}{\operatorname{tr}(M)} M.$$

This feasible budget only attenuates gains. After matching within each fold and averaging held groups equally, the pooled filter has lower reconstruction error in 137 comparisons and higher error in 133. The projector result therefore supports an advantage in leading-subspace coverage, without implying a universal reconstruction advantage for the full soft filters at equal trace.

Predicting sentence labels. We test whether better contrast reconstruction also improves pole prediction. A regularized linear classifier uses either the rank-one mean-separation component D_μ or the pooled conceptor's leading twenty modes. For each held-out group, the remaining groups supply training, and source-group validation selects regularization. Balanced accuracy measures the average accuracy across the two poles on individual held-out sentences.

Mean-component features outperform pooled features in 25 of 30 model/data/site conditions using mean-token activations, with two ties and three losses (Table 21). The full filter's reconstruction advantage therefore does not imply better pole prediction. These results motivate different criteria for probing and concept-preserving steering: a probe needs to distinguish the poles, while an intervention may also need to preserve variation within each pole.

Stabilizing mean differences estimated from few sentences. We test whether filtering stabilizes mean differences estimated from few sentences. Within each held-out group, two disjoint samples, each containing both poles, give $\widehat{\delta}_a$ and $\widehat{\delta}_b$. Matched pairs stay in the same sample. We measure agreement by filtering each estimate and comparing it with the other:

$$L(M) = \tfrac{1}{2}\big(\|M\widehat{\delta}_a - \widehat{\delta}_b\|_2^2 + \|M\widehat{\delta}_b - \widehat{\delta}_a\|_2^2\big).$$

Both estimates are noisy. We average thirty splits per group, normalize by that group's average loss with $M = I$, and average groups equally. Loss below one improves on leaving the estimates unchanged.

Hard projection P_{20} and the conceptor C_{20} share twenty source-fitted axes; only the conceptor shrinks retained coordinates. Source-group validation selects its shrinkage scale from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, .1, 1\}$.

With one sentence per pole in each sample, C_{20} has lower loss than P_{20} in all 60 model/data/site/pooling conditions: for Gemma-2B Sentiment, normalized loss decreases from 0.950 to 0.600 (Table 21). This shrinkage benefit depends on sample size. With the largest available Sentiment samples, identity has lower loss than conceptor filtering in all twenty conditions. Shrinkage can suppress sampling noise but also attenuate useful components of a well-estimated contrast.

		Denoising loss ↓		Readout BA (%) ↑	
Model	Data	Hard P_{20}	Conceptor C_{20}	Mean update	Pooled
Gemma-2B	Sentiment	0.950	**0.600**	94.6	**95.4**
	Pol. Stance	0.811	**0.553**	**58.2**	57.3
	Depression	0.845	**0.638**	**98.5**	96.9
Gemma-9B	Sentiment	0.964	**0.608**	**95.0**	94.7
	Pol. Stance	0.833	**0.567**	**64.3**	59.5
	Depression	0.878	**0.646**	**99.0**	98.0
Qwen-3B	Sentiment	0.996	**0.615**	87.1	**87.5**
	Pol. Stance	0.955	**0.619**	**57.5**	54.1
	Depression	0.971	**0.630**	**94.9**	**94.9**
Llama-3.1-8B	Sentiment	0.988	**0.610**	92.7	**93.4**
	Pol. Stance	0.907	**0.599**	**64.8**	59.4
	Depression	0.931	**0.626**	**97.4**	**97.4**
OLMo-3-7B	Sentiment	0.604	**0.514**	**93.2**	91.7
	Pol. Stance	0.615	**0.520**	**61.1**	58.4
	Depression	0.599	**0.559**	**97.4**	95.4

Table 21: Midpoint controls. One-sentence-per-pole loss is normalized to identity (1); hard and conceptor filters share twenty axes. Pole-prediction balanced accuracy (BA) uses source-tuned readouts.

C.4 Layer coverage

Coverage across layers. Figure 5 shows pooled rank-20 minus mean-line coverage across decoder layers; Figure 6 gives the corresponding absolute scores. At every raw decoder layer, we fit the mean line and pooled rank-20 support on source groups and evaluate the excluded group's mean difference. The figure includes all $(26 + 42 + 36 + 32 + 32) \times 3 = 504$ model/data/layer averages, with no selection based on evaluation scores.

In early Qwen layers, both representations capture most Sentiment and Depression contrasts. The gap rises and falls across depth in every model/data pair. Coverage evaluates the retained directions and is independent of the conceptor's soft gains. Higher held-contrast coverage need not improve sentence-label prediction (Appendix C.3).

Comparison with other single directions. To test whether a stronger single direction closes the gap, we compare twelve source-fitted candidates at the fixed midpoint and final sites: the mean difference, the leading direction of normalized source-group mean differences, and Fisher and logistic directions with five regularization settings each. Source-group validation selects the candidate, which we refit on all source groups.

Rank-20 support has higher coverage in all 60 model/data/site/pooling conditions and all 440 held-group evaluations. The selected direction scores below the simple mean line in 28 condition averages and ties in four: source validation does not always select the direction that generalizes best.

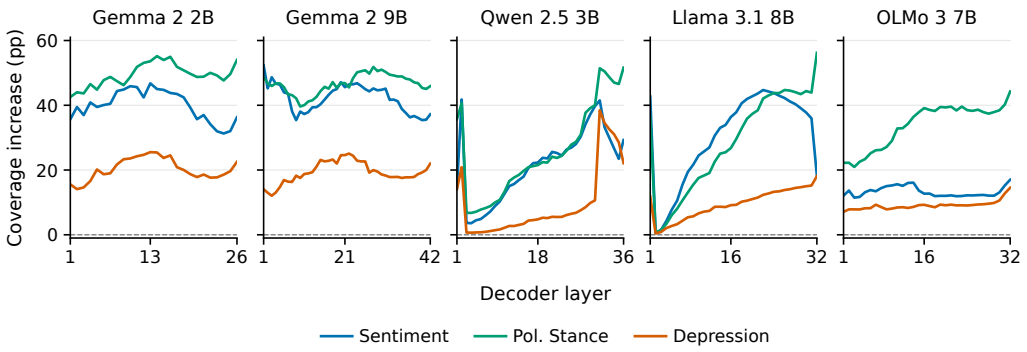


Figure 5: **Additional-direction coverage across layers.** Pooled rank-20 minus source mean-line coverage in percentage points, averaged equally over held groups. Absolute scores appear in Figure 6.

C.5 Changing the semantic reference

Question and setup. We use ConceptVariants to test how the comparator changes the usefulness of additional directions. The target remains positive sentiment, left-leaning policy, or symptom expression. Each target has three references, listed in Tables 22, 23, and 24. Recovery is excluded because it introduces a temporal comparison.

Four fixed folds keep related formulations together across 24 sentiment groups, twenty policy groups, and eight symptom facets with three topics each. For each comparator, a mean-line projector and C_{10} are fitted on identical source-fold examples. Conceptor fitting uses the centered pooled covariance R and $\lambda = \mathrm{tr}(R)/d$. All models use the default midpoint blocks and mean-token activations.

Evaluation and uncertainty. Each comparator defines its own held-group mean difference δ_g. We measure both operators' reconstruction fidelity on that contrast, $F(M, \delta_g) = 1 - E(M, \delta_g)$. Groups receive equal weight, as do topics within groups and the two members of each topic comparison.

The tables report 95% percentile intervals from 2,000 group-bootstrap refits, resampling whole groups within fixed folds with shared weights across comparators and models. Every draw refits the means, covariance, axes, and gains, excluding held groups. Intervals are conditional on the authored collection and folds; paired differences use the shared draws.

Results. Across five models, C_{10} has higher primary reconstruction fidelity in 44 of 45 model/concept/reference comparisons; 40 paired gap intervals are wholly positive. The reference effect nevertheless depends on the model. For Depression expression, replacing flourishing with explicit absence reduces the C_{10} advantage over the mean line in four of five models. For Llama, the gap falls from 43.86 to 3.68 percentage points. OLMo instead changes from 2.59 to 3.56 points; its paired gap reduction is -0.97 points with a 95% interval of $[-4.17, 0.58]$.

OLMo Sentiment with the negative reference supplies a counterexample at the primary midpoint mean-token view: mean-line fidelity is 18.97%, compared with 3.69% for C_{10}. The paired gap is -15.28 points with interval $[-15.46, 0.09]$, which includes zero. Hard rank-ten support gives nearly identical fidelity and retained gains are 0.994–0.998. Rank-twenty and full-support conceptors reach 22.17% and 26.34%, respectively. The leading ten axes omit much of the contrast; shrinkage has little effect on this result. Last-token controls reverse the ordering, further showing its dependence on the chosen representation.

Model	Comparator	Mean line	C_{10}	Gap
Gemma-2B	Flourishing	22.0 [13.1, 31.5]	51.2 [41.0, 53.9]	+29.1
	Ordinary functioning	30.4 [11.6, 53.2]	55.4 [42.9, 60.5]	+24.9
	Explicit absence	68.0 [48.0, 83.3]	74.8 [63.6, 84.6]	+6.9
Gemma-9B	Flourishing	20.2 [10.5, 34.2]	52.8 [43.5, 55.1]	+32.6
	Ordinary functioning	32.7 [12.7, 57.8]	57.6 [44.3, 64.6]	+24.9
	Explicit absence	68.6 [46.7, 85.5]	75.7 [63.4, 87.0]	+7.0
Qwen-3B	Flourishing	33.9 [15.4, 67.3]	76.4 [67.9, 81.3]	+42.4
	Ordinary functioning	46.2 [20.2, 76.4]	76.1 [63.9, 86.0]	+29.9
	Explicit absence	88.7 [78.3, 98.2]	91.1 [83.8, 98.3]	+2.4
Llama-8B	Flourishing	23.1 [7.7, 54.6]	66.9 [57.3, 72.3]	+43.9
	Ordinary functioning	34.6 [11.8, 67.5]	68.3 [55.1, 77.4]	+33.7
	Explicit absence	80.3 [63.1, 95.0]	84.0 [72.0, 95.4]	+3.7
OLMo-7B	Flourishing	20.7 [14.3, 20.8]	23.3 [16.0, 23.3]	+2.6
	Ordinary functioning	16.3 [9.9, 19.1]	18.8 [10.4, 22.3]	+2.5
	Explicit absence	29.2 [18.7, 33.0]	32.7 [22.0, 37.0]	+3.6

Table 22: Depression reference comparisons: fidelity (%) and 95% group-bootstrap intervals; gap in percentage points.

Model	Comparator	Mean line	C_{10}	Gap
Gemma-2B	Negative	40.8 [27.3, 44.8]	45.4 [37.1, 47.3]	+4.6
	Descriptive neutral	37.3 [19.9, 47.6]	50.2 [41.0, 55.4]	+13.0
	Attitudinal neutral	75.6 [65.4, 82.2]	79.1 [71.4, 84.2]	+3.4
Gemma-9B	Negative	35.1 [19.8, 39.5]	38.3 [23.1, 42.9]	+3.1
	Descriptive neutral	42.0 [22.4, 53.4]	53.0 [42.2, 59.5]	+11.0
	Attitudinal neutral	79.6 [69.6, 85.9]	82.0 [74.3, 87.1]	+2.4
Qwen-3B	Negative	22.3 [12.8, 36.3]	28.3 [19.9, 52.4]	+6.0
	Descriptive neutral	65.9 [42.1, 78.5]	73.6 [60.9, 83.1]	+7.6
	Attitudinal neutral	92.9 [84.3, 97.7]	93.7 [86.3, 97.9]	+0.9
Llama-8B	Negative	24.4 [15.5, 35.3]	26.0 [17.5, 46.5]	+1.6
	Descriptive neutral	52.6 [28.0, 64.6]	62.8 [50.8, 71.2]	+10.2
	Attitudinal neutral	87.7 [78.5, 93.5]	88.7 [81.3, 94.1]	+1.0
OLMo-7B	Negative	19.0 [14.0, 19.2]	3.7 [2.3, 16.3]	-15.3
	Descriptive neutral	11.1 [6.9, 12.0]	14.2 [9.3, 15.6]	+3.1
	Attitudinal neutral	17.0 [12.4, 17.1]	20.9 [15.9, 21.6]	+3.9

Table 23: Sentiment reference comparisons: fidelity (%) and 95% group-bootstrap intervals; gap in percentage points.

Model	Comparator	Mean line	C_{10}	Gap
Gemma-2B	Right	6.4 [1.1, 13.3]	14.1 [8.6, 21.7]	+7.8
	Descriptive neutral	65.1 [59.3, 66.7]	70.5 [66.3, 71.2]	+5.4
	Undecided	63.8 [56.8, 66.2]	69.0 [64.2, 70.2]	+5.2
Gemma-9B	Right	8.5 [1.7, 17.1]	15.9 [9.5, 25.8]	+7.4
	Descriptive neutral	57.3 [50.3, 59.1]	65.6 [60.6, 65.9]	+8.3
	Undecided	56.7 [48.2, 59.9]	64.5 [59.3, 65.3]	+7.8
Qwen-3B	Right	33.0 [6.4, 49.1]	42.9 [27.8, 56.9]	+9.9
	Descriptive neutral	58.6 [40.3, 71.5]	82.2 [77.7, 83.6]	+23.6
	Undecided	57.1 [38.6, 69.9]	81.8 [77.4, 83.1]	+24.7
Llama-8B	Right	14.4 [2.0, 29.4]	28.1 [17.2, 39.3]	+13.7
	Descriptive neutral	62.3 [51.3, 69.5]	76.9 [72.7, 77.4]	+14.5
	Undecided	58.2 [45.5, 66.0]	73.6 [68.7, 74.7]	+15.4
OLMo-7B	Right	2.2 [0.6, 4.1]	3.7 [2.6, 8.0]	+1.5
	Descriptive neutral	59.5 [54.4, 60.5]	61.5 [57.7, 62.8]	+2.0
	Undecided	58.8 [54.1, 59.7]	60.9 [56.8, 62.0]	+2.0

Table 24: Pol. Stance reference comparisons: fidelity (%) and 95% group-bootstrap intervals; gap in percentage points.

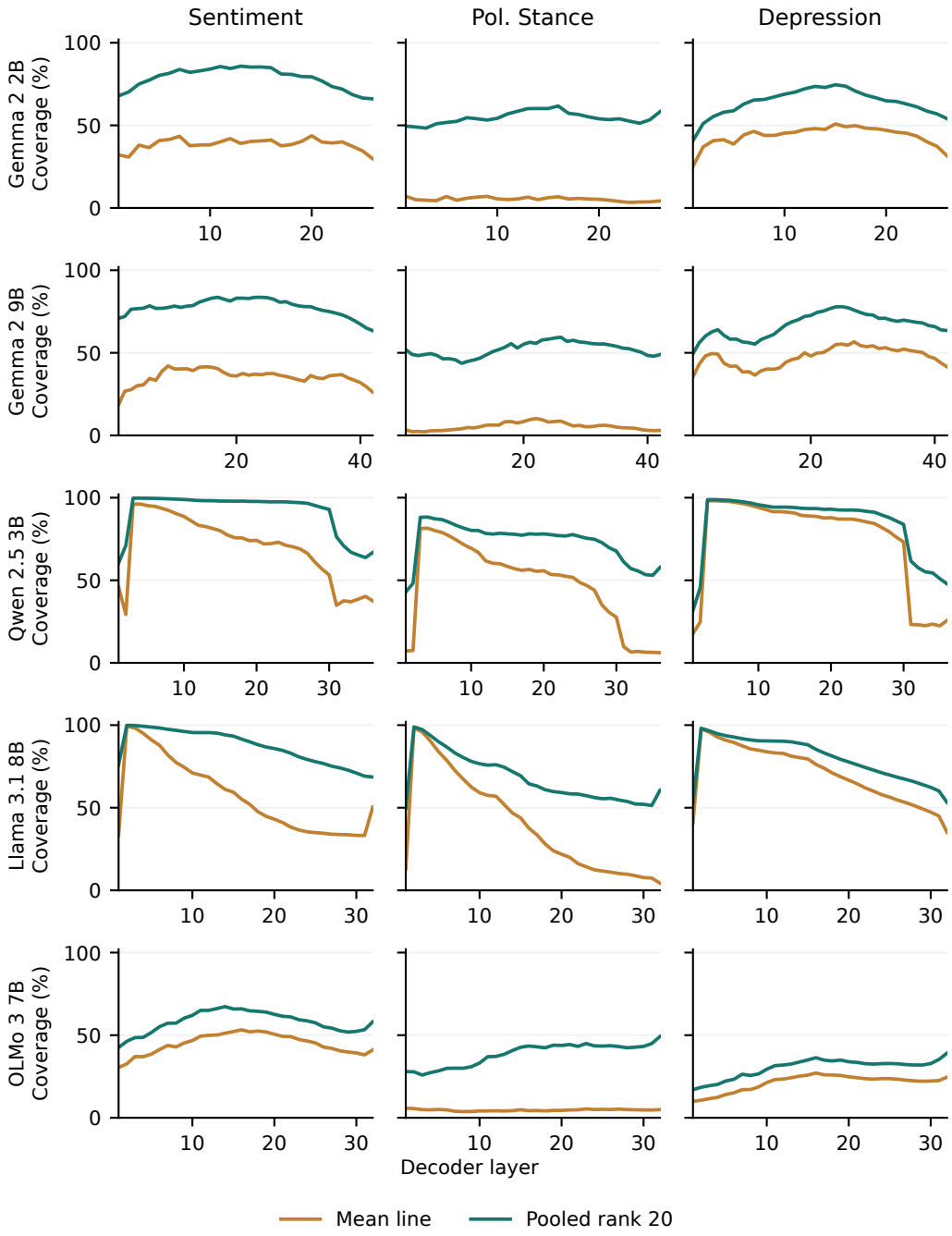


Figure 6: **Coverage across decoder layers.** Source mean-line and pooled rank-20 coverage on the held groups in Figure 5, using a common vertical scale.

C.6 QUOTA AND ALTERNATIVE LAYER DIAGNOSTICS

Spectral comparisons. Following the broader use of representation diagnostics discussed by Skean et al. (2025), we compare quota with two additional summaries of the same centered source covariance. For its eigenvalues r_j, define

$$p_j = \frac{r_j}{\sum_i r_i}, \qquad a_j = \frac{\sqrt{r_j}}{\sum_i \sqrt{r_i}},$$
$$h = -\frac{\sum_j p_j \log p_j}{\log d}, \qquad r_{\text{eff}} = \exp\left(-\sum_j a_j \log a_j\right).$$

Here h is normalized covariance spectral entropy and r_{eff} is effective rank of the centered activation matrix; $0 \log 0 = 0$. Their different spectral weights distinguish the two controls. Both are invariant to a common activation rescaling; finite fitting samples can keep h below one even at maximal available rank. We retain the paper's centered quota at $\alpha = 10$ and centered EVR_{10}. These controls use our fitting protocol, rather than reproducing the embedding preprocessing of Skean et al. (2025).

Absolute performance and fixed layer choices. Figure 7 shows all thirty AUC curves on a common absolute scale. For each diagnostic, we select its maximum over all layers, choosing the earliest layer on exact ties. Selection uses source diagnostics alone, without tuning signs, apertures, or layer ranges against evaluation AUC. The fixed midpoint is displayed layer $\lfloor L/2 \rfloor$: 13, 21, 18, 16, and 16 for Gemma-2B, Gemma-9B, Qwen, Llama, and OLMo, respectively. These retrospective comparisons reuse previously inspected curves without refitting probes.

Results and limits. Mean Pearson correlations with AUC are 0.789, 0.503, -0.464, and -0.451 for quota, EVR_{10}, h, and r_{eff}, respectively. Their maximum-value selectors yield mean AUC differences from the midpoint of -1.712, -2.473, -5.692, and -6.218 percentage points. These descriptive comparisons of fixed rules are not independent replications and do not establish maximum quota as a reliable rule for selecting probe layers.

Layerwise AUC ranges span 0.0209 (OLMo Reddit Depression, 0.9707–0.9916) to 0.1922 (Gemma-9B Hyperpartisan, 0.5256–0.7179).

C.7 BIPOLAR AND UNIPOLAR AMPLIFICATION

Fitting and evaluation. We use the four CONCEPTVARIANTS folds and the default midpoint/mean-token activations. The bipolar fits contrast sentence collections expressing positive versus negative sentiment, left versus right policy, and depressive symptoms versus flourishing. These contrasts do not assume that flourishing is the psychological opposite of depression or that their model representations have a particular geometry. Consistent with the distinction between mental illness and positive mental health (Keyes, 2005), we keep ordinary functioning, explicit symptom absence, and flourishing separate. Unipolar fits replace the second collection with one of two references: descriptive/attitudinal neutral for sentiment, descriptive/undecided for policy, and ordinary functioning/explicit symptom absence for Depression.

All fitting excludes the held fold. Each conceptor uses its fitting midpoint, rank $k \in \{1, 20\}$, and $\lambda = \text{tr}(R)/d$. A source-trained ridge readout evaluates held positive and opposite-pole examples with fixed weights, score normalization, and decision threshold.

Matching the edit size. Source calibration gives equal total weight to positive, opposite, and reference examples. For weights w_i summing to one, define

$$\text{RMS}_s(v) = \left(\sum_i w_i \|v_i\|_2^2\right)^{1/2}.$$

Addition uses step length $D = \beta s$. The recorded scale s is the minimum of the source activation RMS radius and the ten original filtering-update RMS values, from each fit's rank-one, rank-ten,

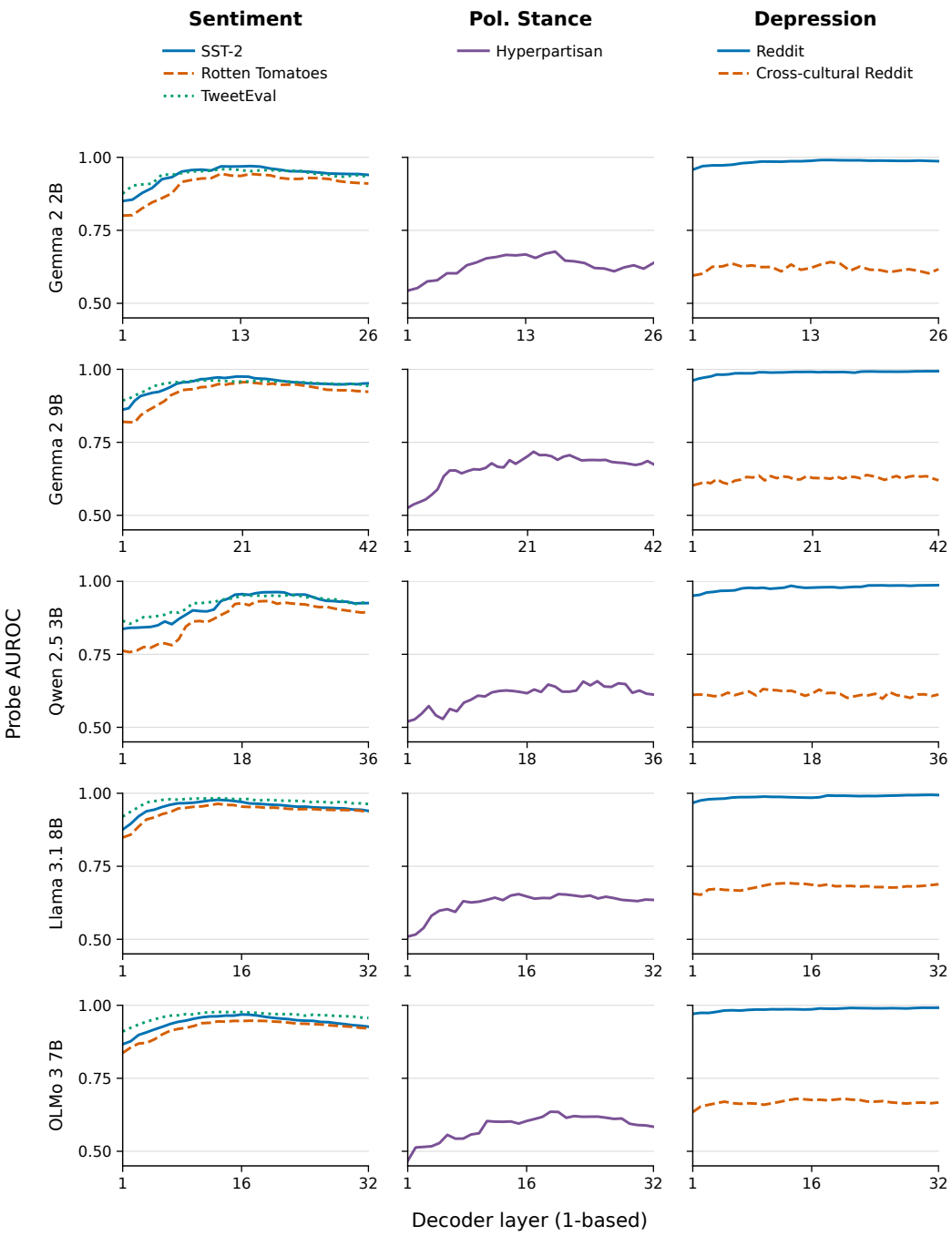


Figure 7: **Absolute layerwise probe performance.** All thirty archived held-out ROC-AUC curves across five models and six benchmarks, grouped by their source concept. The common vertical scale shows absolute performance.

and isotropic controls. We retain this source-only scale when evaluating C_{20}, so edit budgets are unchanged. Conceptor amplification uses

$$t = \frac{D}{\mathrm{RMS}_s(C_k(h - c))}.$$

This matches the source RMS displacement across methods; the coefficient remains fixed on held examples. We test $\beta \in \{0.05, 0.10, 0.20\}$, with 0.10 primary. At the primary setting, edits span 1.65–7.47% of the source activation radius.

Results across conditions. Each condition averages four held folds. Bipolar C_{20} increases both mean pole margins in all 30 calibrated conditions and 112 of 120 calibrated held-fold evaluations. These correspond to fifteen model/concept settings under two calibrations, with 56 of 60 distinct held fits improving both margins. Bipolar C_1 succeeds in 10/30 conditions, while unipolar C_{20} succeeds in 7/30 and unipolar C_1 in 2/30. The two calibrations are not independent replications. A simultaneous improvement requires both changes to exceed 10^{-10}; this criterion does not require statistical significance or a minimum practical effect. The smallest bipolar condition-mean margin change is 0.00450. Tables 25 and 26 report every primary condition.

Fitting both poles gives more consistent two-sided amplification in this comparison. Bipolar C_{20} improves both condition-mean margins at every tested strength. With a fixed linear readout, the changes scale linearly with β and retain their signs.

Model	Concept	Reference	Addition	C_1	C_{20}
Gemma 2 2B	Sentiment	Descr.	+0.3034/-0.3034	+0.0008/-0.0015	**+0.0543/+0.0549**
Gemma 2 2B	Sentiment	Attit.	+0.2873/-0.2873	+0.0005/-0.0009	**+0.0399/+0.0403**
Gemma 2 2B	Pol. Stance	Descr.	+0.4435/-0.4435	+0.0062/-0.0033	**+0.0150/+0.0187**
Gemma 2 2B	Pol. Stance	Undecided	+0.4456/-0.4456	+0.0062/-0.0033	**+0.0151/+0.0189**
Gemma 2 2B	Depression	Ordinary	+0.2220/-0.2220	**+0.0028/+0.0016**	**+0.0521/+0.0517**
Gemma 2 2B	Depression	Absence	+0.2123/-0.2123	**+0.0024/+0.0016**	**+0.0495/+0.0488**
Gemma 2 9B	Sentiment	Descr.	+0.3318/-0.3318	+0.0007/-0.0013	**+0.0494/+0.0510**
Gemma 2 9B	Sentiment	Attit.	+0.3138/-0.3138	+0.0004/-0.0007	**+0.0347/+0.0358**
Gemma 2 9B	Pol. Stance	Descr.	+0.4522/-0.4522	+0.0025/-0.0007	**+0.0208/+0.0266**
Gemma 2 9B	Pol. Stance	Undecided	+0.4458/-0.4458	+0.0025/-0.0007	**+0.0208/+0.0266**
Gemma 2 9B	Depression	Ordinary	+0.2126/-0.2126	**+0.0033/+0.0022**	**+0.0461/+0.0463**
Gemma 2 9B	Depression	Absence	+0.2083/-0.2083	**+0.0029/+0.0023**	**+0.0453/+0.0453**
Qwen 2.5 3B	Sentiment	Descr.	+0.1715/-0.1715	**+0.0006/+0.0008**	**+0.0192/+0.0207**
Qwen 2.5 3B	Sentiment	Attit.	+0.1595/-0.1595	**+0.0005/+0.0003**	**+0.0118/+0.0124**
Qwen 2.5 3B	Pol. Stance	Descr.	+0.0579/-0.0579	**+0.0052/+0.0058**	**+0.0045/+0.0054**
Qwen 2.5 3B	Pol. Stance	Undecided	+0.0575/-0.0575	**+0.0052/+0.0059**	**+0.0045/+0.0054**
Qwen 2.5 3B	Depression	Ordinary	+0.1632/-0.1632	-0.0011/+0.0009	**+0.0203/+0.0238**
Qwen 2.5 3B	Depression	Absence	+0.1568/-0.1568	-0.0010/+0.0007	**+0.0182/+0.0210**
Llama 3.1 8B	Sentiment	Descr.	+0.2503/-0.2503	**+0.0020/+0.0017**	**+0.0330/+0.0397**
Llama 3.1 8B	Sentiment	Attit.	+0.2353/-0.2353	**+0.0012/+0.0008**	**+0.0201/+0.0237**
Llama 3.1 8B	Pol. Stance	Descr.	+0.4039/-0.4039	+0.0033/-0.0019	**+0.0073/+0.0085**
Llama 3.1 8B	Pol. Stance	Undecided	+0.3966/-0.3966	+0.0032/-0.0019	**+0.0073/+0.0085**
Llama 3.1 8B	Depression	Ordinary	+0.2106/-0.2106	-0.0007/+0.0002	**+0.0305/+0.0316**
Llama 3.1 8B	Depression	Absence	+0.2055/-0.2055	-0.0008/+0.0002	**+0.0284/+0.0291**
OLMo 3 7B	Sentiment	Descr.	+0.4255/-0.4255	-0.0013/+0.0014	**+0.0611/+0.0645**
OLMo 3 7B	Sentiment	Attit.	+0.4125/-0.4125	-0.0013/+0.0014	**+0.0600/+0.0632**
OLMo 3 7B	Pol. Stance	Descr.	+0.5269/-0.5269	-0.0009/+0.0004	**+0.0106/+0.0139**
OLMo 3 7B	Pol. Stance	Undecided	+0.5239/-0.5239	-0.0009/+0.0004	**+0.0106/+0.0139**
OLMo 3 7B	Depression	Ordinary	+0.3016/-0.3016	+0.0045/-0.0038	**+0.0578/+0.0590**
OLMo 3 7B	Depression	Absence	+0.2891/-0.2891	+0.0042/-0.0036	**+0.0550/+0.0561**

Table 25: Bipolar fits: positive/opposite margin changes at $\beta = .10$. Bold: both exceed 10^{-10}.

Model	Concept	Reference	Addition	C_1	C_{20}
Gemma 2 2B	Sentiment	Descr.	+0.0556/-0.0556	+0.0074/-0.0070	+0.0157/-0.0067
Gemma 2 2B	Sentiment	Attit.	+0.0374/-0.0374	+0.0258/-0.0244	+0.0225/-0.0089
Gemma 2 2B	Pol. Stance	Descr.	+0.0409/-0.0409	+0.0393/-0.0391	+0.0192/-0.0172
Gemma 2 2B	Pol. Stance	Undecided	+0.0844/-0.0844	+0.0801/-0.0746	+0.0452/-0.0425
Gemma 2 2B	Depression	Ordinary	+0.0933/-0.0933	+0.0068/-0.0017	**+0.0206/+0.0073**
Gemma 2 2B	Depression	Absence	+0.0265/-0.0265	+0.0066/-0.0041	**+0.0093/+0.0008**
Gemma 2 9B	Sentiment	Descr.	+0.0692/-0.0692	+0.0085/-0.0079	+0.0209/-0.0090
Gemma 2 9B	Sentiment	Attit.	+0.0378/-0.0378	+0.0255/-0.0247	+0.0248/-0.0134
Gemma 2 9B	Pol. Stance	Descr.	+0.0272/-0.0272	+0.0252/-0.0264	+0.0099/-0.0090
Gemma 2 9B	Pol. Stance	Undecided	+0.0823/-0.0823	+0.0745/-0.0736	+0.0379/-0.0358
Gemma 2 9B	Depression	Ordinary	+0.0919/-0.0919	+0.0054/-0.0009	**+0.0217/+0.0056**
Gemma 2 9B	Depression	Absence	+0.0303/-0.0303	+0.0068/-0.0030	+0.0117/-0.0009
Qwen 2.5 3B	Sentiment	Descr.	+0.0350/-0.0350	**+0.0022/+0.0002**	+0.0091/-0.0030
Qwen 2.5 3B	Sentiment	Attit.	+0.0214/-0.0214	+0.0130/-0.0100	+0.0153/-0.0074
Qwen 2.5 3B	Pol. Stance	Descr.	-0.0181/+0.0181	-0.0164/+0.0258	-0.0089/+0.0178
Qwen 2.5 3B	Pol. Stance	Undecided	-0.0157/+0.0157	-0.0150/+0.0246	-0.0077/+0.0166
Qwen 2.5 3B	Depression	Ordinary	+0.0547/-0.0547	-0.0000/+0.0001	**+0.0070/+0.0024**
Qwen 2.5 3B	Depression	Absence	+0.0072/-0.0072	+0.0021/-0.0016	**+0.0031/+0.0008**
Llama 3.1 8B	Sentiment	Descr.	+0.0626/-0.0626	**+0.0052/+0.0001**	+0.0159/-0.0047
Llama 3.1 8B	Sentiment	Attit.	+0.0327/-0.0327	+0.0226/-0.0171	+0.0231/-0.0119
Llama 3.1 8B	Pol. Stance	Descr.	+0.0298/-0.0298	+0.0114/-0.0161	+0.0130/-0.0127
Llama 3.1 8B	Pol. Stance	Undecided	+0.0661/-0.0661	+0.0212/-0.0293	+0.0311/-0.0299
Llama 3.1 8B	Depression	Ordinary	+0.0767/-0.0767	+0.0007/-0.0008	**+0.0107/+0.0028**
Llama 3.1 8B	Depression	Absence	+0.0142/-0.0142	+0.0023/-0.0021	+0.0056/-0.0017
OLMo 3 7B	Sentiment	Descr.	+0.1189/-0.1189	+0.0037/-0.0023	+0.0214/-0.0106
OLMo 3 7B	Sentiment	Attit.	+0.1702/-0.1702	+0.0478/-0.0288	+0.0394/-0.0153
OLMo 3 7B	Pol. Stance	Descr.	+0.0465/-0.0465	+0.0435/-0.0435	+0.0239/-0.0229
OLMo 3 7B	Pol. Stance	Undecided	+0.0789/-0.0789	+0.0737/-0.0715	+0.0403/-0.0382
OLMo 3 7B	Depression	Ordinary	+0.1259/-0.1259	-0.0001/+0.0007	**+0.0193/+0.0038**
OLMo 3 7B	Depression	Absence	+0.0669/-0.0669	+0.0001/-0.0006	+0.0109/-0.0043

Table 26: Unipolar fits: positive/opposite margin changes at $\beta = .10$. Bold: both exceed 10^{-10}.

	Mean margin change		Both margins improve	
Action	Positive	Opposite	Conditions	Held evals.
Bipolar: positive/opposite				
Mean-line amplification	$+0.1344$	$+0.1451$	$30/30$	$76/120$
C_1 amplification	$+0.0018$	$+0.0002$	$10/30$	$14/120$
C_{20} amplification	$+0.0301$	$+0.0322$	$30/30$	$112/120$
Unipolar: positive/reference				
Mean-line amplification	$+0.0416$	-0.0241	$4/30$	$15/120$
C_1 amplification	$+0.0177$	-0.0153	$2/30$	$16/120$
C_{20} amplification	$+0.0177$	-0.0083	$7/30$	$30/120$
Unipolar: pooled presence/reference				
Mean-line amplification	-0.0049	$+0.0058$	$0/30$	$2/120$
Fixed addition	-0.0096	$+0.0096$	$0/30$	$0/120$

Table 27: Mean standardized held-pole margin changes across five models at $\beta = 0.10$, with matched source-RMS displacement. Success requires both changes to exceed 10^{-10}. Conditions average four held folds before applying this criterion; held evaluations apply it separately to each fold under each reference. The thirty conditions comprise fifteen model/concept settings under two references. Bipolar rows share 60 distinct fits across the two calibrations; unipolar fits change with the reference.

Input-dependent mean-line controls. To distinguish input dependence from the choice of retained directions, we replace the conceptor by the projector onto each fit's source mean-difference line:

$$T_{\text{line}}(h) = h + t_{\text{line}} u u^\top (h - c), \qquad u = \frac{\delta}{\|\delta\|_2},$$

$$t_{\text{line}} = \frac{\beta s}{\text{RMS}_s(u u^\top (h - c))}.$$

This amplifies the signed mean-line component while preserving orthogonal components; C_1 need not retain the same axis. All controls use the same source/held splits, frozen polarity readouts, source-triad calibration weights, and edit budget $D = \beta s$.

We compare bipolar fitting, which contrasts positive and opposite poles, with unipolar fitting against a reference. Besides the usual positive/reference comparison, we test a presence collection containing both poles. Using source means with the original group/topic weights and equal pole weights, define

$$\mu_{\text{pres}} = \frac{\mu_+ + \mu_-}{2}, \qquad c_{\text{pres}} = \frac{\mu_{\text{pres}} + \mu_{\text{ref}}}{2},$$

$$u_{\text{pres}} = \frac{\mu_{\text{pres}} - \mu_{\text{ref}}}{\|\mu_{\text{pres}} - \mu_{\text{ref}}\|_2}.$$

We substitute this axis and center into the same calibrated amplifier and also evaluate fixed addition $h + D u_{\text{pres}}$. All fitting excludes the held group.

Both bipolar mean-line amplification and C_{20} improve both margins in all thirty condition averages (Table 27). Across distinct held fits, however, C_{20} succeeds in $56/60$, compared with $38/60$ for the mean line. Averaging folds can therefore conceal individual failures. For positive/reference fitting, C_{20} gives the larger minimum pole-margin change in $27/30$ conditions. Two-sided improvement does not require multiple retained directions, but C_{20} achieves it more consistently across held folds here. These fixed-readout gains do not establish better label prediction or generated behavior.

Source geometry and amplification. To interpret when the mean line or C_{20} performs better, we examine how much source variation along the bipolar mean line comes from pole separation. Greater directional disagreement among source topic contrasts did not consistently predict the advantage of C_{20}. For balanced bipolar fitting, the pooling decomposition instead gives

$$S_\mu = \frac{\|\delta\|_2^2/4}{u^\top R_{\text{pool}} u} = \frac{\|\delta\|_2^2/4}{\|\delta\|_2^2/4 + u^\top R_{\text{w}} u}, \qquad u = \frac{\delta}{\|\delta\|_2}, \tag{30}$$

where $\delta = \mu_+ - \mu_-$ and R_{w} is the balanced within-pole covariance. Thus $S_\mu \geq 1/2$ means that mean separation contributes at least as much variance along u as within-pole variation. We use this source-only split to describe the 60 distinct bipolar held fits at $\beta = 0.10$.

	Larger minimum change		Both margins improve	
Source geometry	Mean line	C_{20}	Mean line	C_{20}
$S_\mu \geq 1/2$ $(n = 25)$	25	0	$25/25$	$25/25$
$S_\mu < 1/2$ $(n = 35)$	9	26	$13/35$	$31/35$

Table 28: Exploratory comparison by source mean-line geometry. The first two numerical columns count paired wins in the smaller held pole-margin change. We compare methods within each reference calibration and average their difference across the two calibrations. The final columns count fits in which both changes exceed 10^{-10}. Reference repetitions count once, so these are distinct held-fit outcomes rather than condition averages.

When mean separation dominates the source mean-line coordinate, the mean line gives the larger minimum held margin change in all 25 fits. When within-pole variation dominates, C_{20} gives the larger minimum change in $26/35$ fits and improves both poles in $31/35$, compared with $13/35$ for the mean line (Table 28). In 19 of these fits only C_{20} improves both margins; three of its paired wins instead represent smaller negative changes. Concentrating amplification along the mean line therefore works well in the observed fits dominated by pole separation, while pooled amplification more often improves both poles when within-pole variation dominates that coordinate.

Pooling both poles into the presence collection. Including both poles in the presence collection does not recover the bipolar result: the resulting mean-line amplifier improves both condition-mean margins in $0/30$ conditions and both held margins in $2/120$ evaluations (Table 27). The two successes are Sentiment fits in Qwen and Llama under the descriptive reference. These success counts are unchanged at $\beta \in \{0.05, 0.10, 0.20\}$.

The geometry explains much of this difference. In $119/120$ source fits, both pole means project to the same side of c_{pres} along u_{pres}, so amplification shifts their polarity scores in the same direction. This presence direction depends on the pole midpoint, whereas the bipolar pooled moment also retains the mean-separation term $\delta\delta^\top/4$. Including both poles as fitting examples therefore does not ensure simultaneous polarity amplification. The comparison concerns the fitting contrast and its native center together; it does not imply that unipolar fitting must always fail or that bipolar fitting succeeds on every held group.