WHAT DOES RECONSTRUCTION FIDELITY MEASURE UNDER DISTRIBUTION SHIFT?

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstruction fidelity, reported as variance explained or fraction of variance unexplained, is used to judge learned representations, including when a model fitted on one dataset reconstructs another. Under such distribution shift, one score mixes three quantities: credit from the reference it is measured against, error from the reconstruction's mean being in the wrong place, and the structure around the target's own mean that the reconstruction captures. Reference dependence and the split of mean-referenced squared error into mean bias and centered error are classical; for reconstructions confined to a convex set, as in archetypal analysis, we show that the score is at most the fit available after one free translation (the translated fit) minus the squared distance from the target mean to the set, so geometry alone can force a negative score. In cross-corpus transfer of archetypal dictionaries ($K = 32$) over text embeddings, 37 of 48 settings score below the target mean, yet their fit around that mean exceeds that of randomly rotated hulls, though it nearly matches hulls of randomly chosen source points; transferred dictionaries reach 32–77% of the target-informed translated fit of dictionaries fitted on the target corpus, and location is the larger share of the transfer excess in all 48 settings, at each of $K = 8$, 16, 32 and 64. Representing the same records by their original text makes every cross-corpus score negative. For the Gemma Scope sparse autoencoders we evaluate, the same split shows that on natural data these SAEs sit far from the zero-score boundary, and that their decoder bias supplies substantial location correction when the codes are held fixed. Reporting the three quantities separately keeps a negative score from being read as absence of target-centered fit.

1 INTRODUCTION

Reconstruction fidelity measures how much of a dataset a learned representation reproduces: variance explained for dictionary learning and archetypal analysis (Olshausen & Field, 1996; Cutler & Breiman, 1994; Mørup & Hansen, 2012), and the fraction of variance left unexplained for sparse autoencoders (SAEs; Gao et al., 2025; Lieberum et al., 2024). Under distribution shift, when a model fitted on one dataset reconstructs another, it is tempting to read a high score as evidence that structure carried over and a negative score as evidence that nothing did; neither reading is safe, because one score then mixes three quantities:

- **reference credit**, set by what the score is measured against (the origin or the target mean);
- **mean offset**, the distance between the reconstruction mean and the target mean;
- **translation-corrected fit**, the target's variation around its own mean that is captured once location is set aside, measured two ways: recentered fit, for the reconstruction actually produced, and translated fit, for what the model set could produce.

A running example. We embed short conversational moves, persuasive replies from Reddit's ChangeMyView (CMV) and Craigslist price negotiations, each first rewritten by an LLM into a short procedural summary, which we call a blueprint (Section 4); Section 6 repeats the study on the original text. A 32-archetype dictionary fitted to the CMV embeddings by principal convex hull analysis (PCHA; Mørup & Hansen, 2012) reconstructs held-out Craigslist embeddings, unchanged.

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.

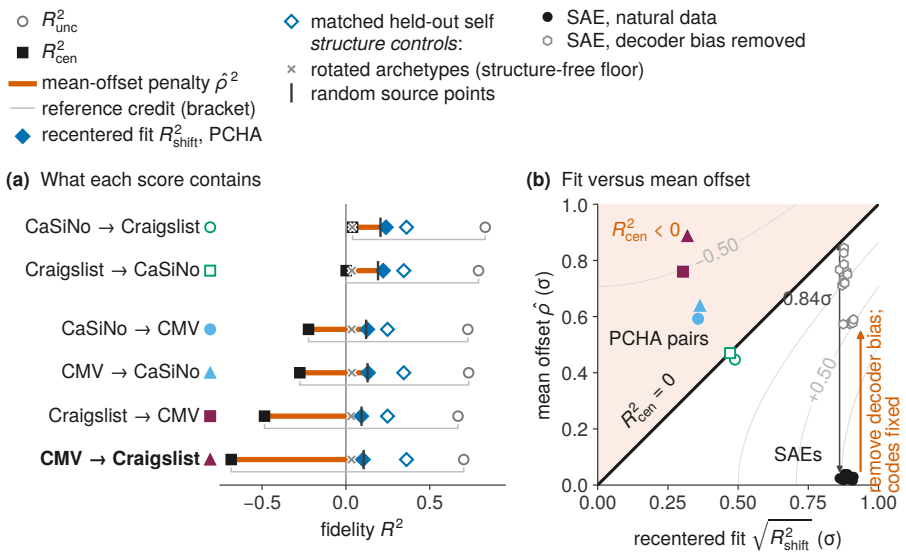

Figure 1: **What a fidelity score contains, in two reconstruction regimes.** (a) The three-corpus PCHA pairs (Table 1). Recentering moves the same fixed reconstructions to the target mean; the score splits as $R^2_{\text{cen}} = R^2_{\text{shift}} - \hat{\rho}^2$. The grey controls have location removed: $\times$ is a structure-free floor (the archetypes randomly rotated), the tick a random-source reference (hulls of 32 random source points). (b) $x = \sqrt{R^2_{\text{shift}}}$ against $y = \hat{\rho}$, in units of each target's σ, so $R^2_{\text{cen}} = x^2 - y^2$ and the diagonal is the zero-score boundary. Black dots are the Gemma Scope SAEs we evaluate, on natural data (Section 7); the black vertical arrow marks their smallest margin, 0.84σ, the vertical distance to the diagonal: the extra mean offset that would bring the score to zero. Hexagons are the same SAEs with the decoder bias removed and the codes held fixed (Appendix Table D.5).

Averaged over eight embedding configurations from six models (Figure 1), fidelity against the origin is 0.70, but against the Craigslist mean it is -0.684: the reconstructions have a larger total squared error than simply predicting the Craigslist mean. Shifting each reconstruction rigidly so that its mean matches the Craigslist mean, without changing any weights, gives $+0.102$, the *recentered fit*; the difference, 0.786, is the squared distance between the two means in units of the target's variance, and the split is exact: $0.102 - 0.786 = -0.684$. The reconstructions do capture some structure in the Craigslist data, about as much as a hull of randomly chosen CMV points would (Section 5). Their mean-offset penalty is large enough to push the score below zero. CMV→Craigslist is the most extreme of our six corpus pairs; between the negotiation corpora, scores are near zero (Table 1).

What is classical and what is new. The gap between the origin-referenced and mean-referenced scores is the textbook difference between uncentered and centered R^2, neither reference being correct in general (Section 2), and forecast verification has split squared errors into mean and centered parts since Murphy (1988); we claim nothing new for that algebra. What is new concerns the constrained set of reconstructions a model can produce, here the convex hull of a PCHA dictionary's archetypes: how far the target mean lies from it, and how much the model could capture after one global translation with its weights refitted, the *translated fit* (an oracle diagnostic, Section 3.1).

Our contributions are:

- **Accounting and a bound (Sections 2–3).** A bound for closed convex reconstruction sets that links the mean-referenced score to the translated fit and to the target mean's distance from the set, giving a sufficient sign certificate; Proposition 2, which refines Theorem 1 for convex hulls into an exact location-and-shape split, with the location family at least ρ^2_H for the nearest-point reconstruction; and a controlled intervention on the hull's position alone, possible because a rigid translation leaves translated fit unchanged.
- **Evidence for convex dictionaries (Sections 4–6).** Under matched model size, group-held-out evaluation and optimization accurate enough to fix every sign, the certificate forces

33 of the 37 negative signs among 48 three-corpus settings, and recentered fit exceeds a structure-free floor while roughly matching a hull of random source points; transferred dictionaries reach 32%–77% of same-corpus translated fit (an oracle diagnostic), and location is the larger part of their excess error in all 48 settings and at each of $K = 8$, 16, 32 and 64; replacing blueprints with the original text makes every cross-corpus score negative (Section 6).

- **Sparse autoencoders and reporting (Sections 7–8).** For the Gemma Scope SAEs we evaluate, the fixed-map split shows that they sit far from a sign change on natural data (mean-offset penalty never above 0.0016), and removing the decoder bias with the codes held fixed lowers their score by 0.33–0.71; we recommend stating the reference, reporting the mean-offset penalty and the recentered fit, and comparing transfer with a matched held-out-self model.

In these experiments, a negative score did not mean that the reconstructions captured none of the target's structure around its mean.

2 Accounting under shift

Let $X \in \mathbb{R}^{N\times D}$ be the target data, with rows x_i and mean μ, and let $\hat{X}$ reconstruct it, with residual $\mathrm{RSS} = \|X - \hat{X}\|_F^2$. The two fidelity scores divide this residual by the error of a baseline predictor:

$$R^2_{\mathrm{unc}} = 1 - \frac{\mathrm{RSS}}{\|X\|_F^2}, \qquad R^2_{\mathrm{cen}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}, \qquad \mathrm{TSS} = \sum_i \|x_i - \mu\|^2.$$

R^2_{unc} is referenced to the origin, its baseline predicting zero for every row, and R^2_{cen} to the target mean μ, the best constant under squared loss; a score of zero ties the baseline, and a negative score loses to it. Neither baseline is right in general: the origin suits claims about absolute coordinates, the target mean claims about variation among target examples. Splitting the target's energy into its mean and its variation around the mean, $\|X\|_F^2 = N\|\mu\|^2 + \mathrm{TSS}$, and calling the first share the *centroid mass* $m = N\|\mu\|^2/\|X\|_F^2$, every reconstruction satisfies $R^2_{\mathrm{unc}} = m + (1-m)\,R^2_{\mathrm{cen}}$.

By this textbook relation (Kvålseth, 1985; Eisenhauer, 2003), predicting the target mean already scores m against the origin, so a reconstruction can score well against the origin yet lose to the target mean, since $R^2_{\mathrm{cen}} < 0$ exactly when $R^2_{\mathrm{unc}} < m$; the text embeddings studied here have large centroid mass (Appendix E.1). The identity only rescales the residual (orderings: Appendix A.8); it does not explain why the residual can exceed TSS and make the score negative.

Write $\hat{\mu}$ for the mean of the reconstructed rows, $\sigma^2 = \mathrm{TSS}/N$ for the target's variance per example, and $C = I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ for the matrix that subtracts column means (Table A.1); the residual $X - \hat{X}$ then splits into its mean, $\mu - \hat{\mu}$, and the variation around that mean, and the two parts are orthogonal.

Proposition 1 (fixed-map split; classical). For any fixed reconstruction $\hat{X}$ of a target with $\mathrm{TSS} > 0$,

$$R^2_{\mathrm{cen}} = R^2_{\mathrm{shift}} - \hat{\rho}^{\,2}, \qquad R^2_{\mathrm{shift}} = 1 - \frac{\|C(X - \hat{X})\|_F^2}{\mathrm{TSS}}, \qquad \hat{\rho} = \frac{\|\mu - \hat{\mu}\|}{\sigma}.$$

The *recentered fit* R^2_{shift} is the score of the same reconstruction with all its rows shifted by one vector so that their mean matches the target mean, no weights changed, and the *mean-offset penalty* $\hat{\rho}^{\,2}$ is the squared distance between the two means in units of the target variance; we call such quantities, computed with the reconstruction held fixed, *fixed-map* quantities, and the running example of Section 1 is exactly this split. The score is negative exactly when the penalty exceeds the recentered fit, so a reconstruction can capture target-centered structure and still score below zero because its own mean is in the wrong place. Throughout, R^2_{shift} and $\hat{\rho}^{\,2}$ describe the reconstruction actually produced, while R^2_{ST} and ρ^2_H describe what the model set could achieve and use the target. Proposition 1 makes no assumption about how $\hat{X}$ was produced, so Section 7 applies it unchanged to SAEs.

3 WHEN THE MODEL SET FORCES A NEGATIVE SCORE

3.1 TRANSLATED FIT, DISTANCE AND THE SIGN CERTIFICATE

Proposition 1, like the classical decompositions, describes one reconstruction already made; the new step concerns the set $H \subset \mathbb{R}^D$ of points the model can output for a single row, for archetypal analysis the convex hull of the archetypes $a_1, \ldots, a_K$. The *translated fit* is the best mean-referenced score the set can reach after one global translation t, with every row reconstructed again:

$$R^2_{\mathrm{ST}} = 1 - \frac{1}{\mathrm{TSS}} \inf_{t,\, h_i \in H} \sum_i \|x_i - h_i - t\|^2.$$

It measures how much of the target's variation around its own mean the model can capture once location is set aside, and because it uses the target data to choose the translation and the weights, it is a diagnostic, not a measure of how the model performs without adaptation; the translation is an accounting device and a controlled intervention (Section 3.2), not a proposed way to adapt a model. The *centroid-to-hull distance* $\rho_H = \min_{h \in H} \|\mu - h\|/\sigma$ (for a general convex set, the centroid-to-set distance) depends on no reconstruction weights but uses the target mean, so it too is a diagnostic.

Lemma 1 (recentered fit under convex projection). If each row of $\hat{X}$ is the nearest point of a closed convex set H to the corresponding row of X, then $R^2_{\mathrm{shift}} \geq \|C\hat{X}\|^2_F/\mathrm{TSS}$, which is positive unless all reconstructions coincide.

For convex-hull dictionaries, a positive recentered fit is therefore guaranteed, as a non-negative translated fit is for any nonempty H; only their magnitudes, compared with a floor and a matched held-out self, carry information (proof in Appendix A.6).

Theorem 1 (translation bound). Let H be nonempty, closed and convex, let every row of $\hat{X}$ lie in H, and let $\mathrm{TSS} > 0$. Then

$$R^2_{\mathrm{shift}} - \hat{\rho}^{\,2} \;=\; R^2_{\mathrm{cen}} \;\leq\; R^2_{\mathrm{ST}} - \rho^2_H.$$

The equality is Proposition 1; the inequality combines two relaxations, $R^2_{\mathrm{shift}} \leq R^2_{\mathrm{ST}}$ because re-optimizing the weights after a translation cannot fit worse (Lemma 2), and $\hat{\rho} \geq \rho_H$ because $\hat{\mu}$, an average of points in H, lies in H (Lemma 3; Figure A.1); Appendix A.7 gives the proofs. The theorem covers any closed convex set of outputs, such as linear subspaces and linear decoders with simplex-constrained codes; the outputs of TopK codes, of a thresholded encoder or of the amortized encoder–decoder maps of sparse autoencoders do not form such a set, and a convex set containing them would define a different oracle problem from the map we evaluate.

Corollary 1 (sign certificate). If $\rho^2_H > R^2_{\mathrm{ST}}$, then $R^2_{\mathrm{cen}} < 0$.

Here ρ^2_H is a location cost that no reconstruction from the set can avoid, and R^2_{ST} the centered fit available once location is free; when the cost exceeds the fit, the score is negative whatever the weights. For the CMV→Craigslist pair means, the translated fit is 0.16 and the Craigslist mean lies $\rho_H = 0.87\sigma$ from the CMV hull, against $\hat{\rho} = 0.89\sigma$ for the reconstruction mean, so the bound is -0.60; the observed score is -0.68. Across the three-corpus study, the certificate forces the sign in 33 of the 37 negative cells (Figure 2a). It is sufficient but not necessary: when ρ^2_H falls just short of R^2_{ST}, the gap between bound and score, the *theorem slack*, can still make the score negative.

3.2 CONTROLLED TRANSLATION: MOVING ONLY THE HULL CAN REVERSE THE SIGN

Moving a hull rigidly leaves its translated fit unchanged, since the free translation in R^2_{ST} absorbs the move. We move each of the 60 CMV dictionaries of the running example (one configuration per model, six of the eight; the first two of the ten splits; five seeds; a grid fixed before any results were seen) by $\alpha(\mu - h^*)$, with h^* the point of its hull nearest the Craigslist mean, so that $\rho_H = 0$ at $\alpha = 1$, and reconstruct every target row again from the moved hull. With the average translated fit fixed at 0.15, the average score rises from -0.68 at $\alpha = 0$ to $+0.11$ at $\alpha = 1$, where every one of the 60 dictionaries has changed sign, not only the average, and it falls again as the hull moves past the target (Figure 2b). Over all six directed corpus pairs (360 evaluations), the same move raises the

average score from -0.27 to $+0.15$ against a translated fit of 0.18, with every pair's average positive at $\alpha = 1$, as on the full grid (Appendix B.1); two sweeps along directions observed in the data, rather than constructed ones, behave the same way (Appendix B.1). Because we control the move, in these experiments the hull's position alone changed the sign; this does not show that location causes all of the excess error in natural transfer, where the target's structure around its mean also differs from the source's.

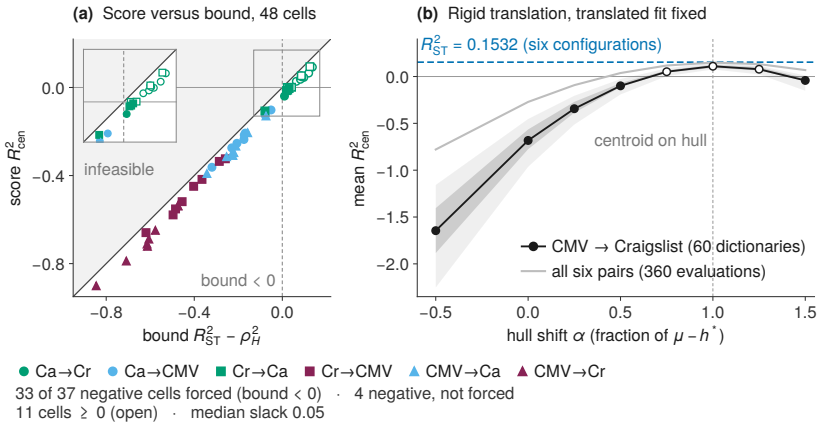

Figure 2: **The location bound and a controlled translation.** (a) Score against bound, 48 three-corpus cells: Theorem 1 excludes the grey region, and left of the dashed line the sign is forced negative. The inset holds the 11 non-negative cells (open) and the 4 negative cells whose sign is not forced; over the 2,400 cross-corpus split-seed evaluations the slack has median 0.050 (Appendix E.3). (b) The controlled translation (Section 3.2): mean R^2_{cen} of the 60 dictionaries (black; interquartile and full-range bands) and of all 360 evaluations (grey); the translated fit (dashed; 0.1532 over six of the eight configurations) is unchanged (Table B.1).

3.3 Location and shape

Theorem 1 certifies signs but does not apportion a transferred dictionary's excess error between location and shape. Cutting the total normalized residual at the translated fit gives two families: the *shape family*, $1 - R^2_{\text{ST}}$, the error that remains even when the hull may move, and the *location family*, $R^2_{\text{ST}} - R^2_{\text{cen}}$, the error removed by one free translation with the weights re-optimized. For the nearest-point reconstruction from a convex hull, Proposition 2 (Appendix A) splits each family exactly into two non-negative terms and shows that the location family is at least ρ^2_H, which refines Theorem 1 for that reconstruction. We write Δ for the cross-minus-self difference of any term between a cross-corpus cell and its matched held-out self (Section 4).

4 Evaluation protocol

Data and models. The primary study uses three corpora of conversational moves: two negotiation corpora, CaSiNo (Chawla et al., 2021) and Craigslist Bargains (He et al., 2018), and the argumentation corpus ChangeMyView (CMV; Tan et al., 2016); Section 6 adds a fourth, PersuasionForGood (PFG; Wang et al., 2019). Each record is represented by its blueprint, an LLM-written procedural summary of the move. We use blueprints because embeddings of raw text are dominated by topical vocabulary: describing each move by its argumentative strategy, with domain-specific content abstracted away, targets the structure that could carry over between platforms (Ayromlou et al., 2026). Blueprints are embedded with eight configurations from six models, all L2-normalized; the two text-embedding-005 truncations are additional configurations of the same model, not independent replications (Appendix G.1 gives selection events, record counts, generation settings, embedding models and a leakage audit). On each source corpus we fit $K = 32$ PCHA archetypes ($K = 8$, 16 and 64 in Appendix G.2) with five seeds and use the resulting dictionary, unchanged, for every target; target data are never used to choose or refit archetypes, and each target row is reconstructed

by its nearest point in the hull. A *cell*, or setting, is one directed corpus pair (source → target) under one embedding configuration, so the three-corpus study has 48 cells (six directed pairs × eight configurations; aggregation: Appendix G.4).

Matched baseline and floors. Splits are group-disjoint, by dialogue for the negotiation corpora and PFG and by original post for CMV, and ten split assignments, defined on record identifiers, are shared by every embedding configuration. The baseline for a transferred dictionary is a *matched held-out-self* dictionary of the same size, fitted with the same budget on disjoint groups of the target's own corpus and scored on the identical target matrix, so a cross-versus-self difference cannot come from a change of target sample; it is a capacity reference, since a $K = 32$ dictionary fitted on part of a corpus need not reconstruct all of the centered variation in the rest. Each transferred dictionary is also compared with two structure-free floors built from it, its archetypes rotated at random about their mean and an isotropic hull of matched spread, and, as a random-source reference, with a hull of random source points (Appendix E.6).

Signs and aggregation. Several informative cells lie close to $R^2_{\text{cen}} = 0$, where a converged optimizer is not enough to settle a sign, so we verify signs numerically (Appendix G.3); for the fixed-position score, no numerically verified interval crosses zero, and numerically verified and point-estimate signs agree in every evaluation. Finite-target-sample sensitivity is assessed with group-level bootstrap resampling, never pooled with split or seed variability: the bootstrap standard deviation of R^2_{cen} has median 0.0067, and the widest 95% interval spans 0.060. The directed corpus pair is the unit of interpretation; cell, split, seed and optimization counts measure robustness within the evaluated corpora and are not treated as independent samples for population-level inference.

5 Three-corpus transfer: what negative scores contain

The same transfers, two verdicts. Against the coordinate origin, all 48 cells have positive fidelity: pair means 0.67–0.83, against 0.83–0.89 for matched held-out self. Against the target mean, 37 of the 48 cells are negative, in a graded pattern: all 32 cells involving CMV are negative, with pair means from −0.22 to −0.68, while the two negotiation corpora, CaSiNo and Craigslist, sit near zero in both directions, with pair means about +0.04 and +0.00, and 11 of their 16 cells are positive. Five of these cells lie within ±0.02 of zero, three of them negative, and two of the 37 negative cells have a 95% group-bootstrap interval (Section 4) that includes zero, so we treat that pair as sitting at the sign boundary rather than interpreting each small sign (Table 1; Figure 1a).

Table 1: Cross-corpus reconstruction summary: eight-configuration grand means of split means, with $R^2_{\text{cen}} = R^2_{\text{shift}} - \hat{\rho}^{\,2}$ for the same fixed reconstructions before rounding to four decimals. Recovery is cross-corpus over matched held-out-self translated fit. The last two columns give the recentered fit of a hull of random source points and of matched held-out self on the same target (floors and translated fit: Appendix Tables E.1, E.5 and E.6).

Source → target	R^2_{unc}	R^2_{cen}	R^2_{shift}	$\hat{\rho}^{\,2}$	R^2_{ST}	**Recovery (oracle)**	**Random-source** R^2_{shift}	**Held-out-self** R^2_{shift}
CaSiNo → Craigslist	0.8306	+0.0395	0.2389	0.1995	0.2650	0.7321	0.2061	0.3612
Craigslist → CaSiNo	0.7907	+0.0014	0.2222	0.2208	0.2377	0.6868	0.1914	0.3449
CaSiNo → CMV	0.7267	−0.2230	0.1278	0.3508	0.1609	0.6531	0.1202	0.2479
CMV → CaSiNo	0.7317	−0.2749	0.1326	0.4076	0.1725	0.4987	0.1287	0.3449
Craigslist → CMV	0.6683	−0.4854	0.0922	0.5776	0.1365	0.5544	0.0931	0.2479
CMV → Craigslist	0.7026	−0.6840	0.1021	0.7861	0.1615	0.4475	0.1058	0.3612

Fit around the target mean. Removing only the mean of each fixed reconstruction's residual, with no dictionary refit and no weights changed, the fixed weights capture about 5%–28% of target-centered variance (positive by Lemma 1); by directed pair, the recentered fit is 0.09–0.24 and the mean-offset penalty 0.20–0.79. Compared with the structure-free floors and the random-source reference of Section 4, all drawn for the same dictionaries, transferred dictionaries exceed the rotated floor by a median of 0.096, but hulls of random source points trail them by a median of only 0.005. The recentered fit that survives transfer therefore owes little to the learned archetypes. Recentering is not the whole explanation, however: transferred reconstructions also fit the target's centered variation worse than matched held-out self, with a median centered residual ratio $1 - R^2_{\text{shift}}$ of 0.81 under transfer against 0.69.

What becomes recoverable once one translation is freed. Freeing one global translation and re-optimizing the simplex weights gives the translated fit: cross-corpus pair means are 0.14–0.27, compared with 0.25–0.36 for matched held-out self. Across the 48 cells, recovery (the cross-corpus R^2_{ST} divided by the matched held-out-self R^2_{ST} on the identical target) ranges from 0.32 to 0.77, with median 0.59; it is highest between the two negotiation corpora (pair means 0.73 and 0.69, Table 1) and lower for every pair involving CMV, including the running example (Appendix E.2). Free translation is not zero-shot performance, and this ordering is a pattern in these three corpora rather than a general law. What the scores establish is narrower: what a transferred dictionary can capture once location is freed is 32%–77% of the matched held-out self's, and well above what structure-free floors capture (0.51 of the way from floor to matched held-out self). Learned archetypes also add to translated fit beyond hulls of random source points, consistently but modestly: PCHA is ahead in 47 of 48 cells, by a median of 0.011, whereas in recentered fit it roughly matches them (Appendix E.6). At the median, the larger part of the remaining deficit is subspace mismatch, target variation outside the directions the source archetypes span (Appendix E.2).

Location is the larger part of the excess over held-out self. Applying Proposition 2 to the transferred and matched held-out-self dictionaries on the identical target apportions each cell's excess error Δ between the location and shape families (Section 3.3). The location family, the error one free translation removes, exceeds the shape family, the error that remains, in all 48 cells, by a factor of 1.5–7.5 (median 2.7) (as at $K = 8$, 16 and 64; Appendix G.2). Large centroid mass (0.61–0.90) may make this unsurprising for L2-normalized embeddings (anisotropy; Ethayarajh, 2019; Mu & Viswanath, 2018); a single score does not show it separately, and it is not universal (Section 7). The shape family is nonetheless a substantial second part, 12%–40% of the excess (Appendix E.3). The offset is also aligned with directions in which the source corpus itself varies, far more than a randomly oriented offset would be (Appendix E.4), and matching the held-out-self dictionary's fitting budget changes effect sizes only slightly relative to the location-versus-shape pattern (Appendix E.5).

6 Robustness: a fourth corpus, other summaries, raw text

Section 5 rests on three corpora, one learned source summary and one representation; this section changes each in turn at the common 538-move budget (Appendix G.2).

A fourth corpus. PFG adds 1,076 selected moves from 758 dialogues in which one participant tries to persuade another to donate, and the six original directed pairs are recomputed at the same budget. Of the 48 PFG-involving cells, 46 have a negative median R^2_{cen}; location excess exceeds shape excess in all 48, with median location share 68.8% (range 59.2%–86.5%), although strong location dominance is less common than for the original pairs (Appendix C.1).

Other source summaries. We compare PCHA with four summaries built from the source fitting data alone, Extreme Support (32 boundary points chosen greedily; Appendix C.3), random exemplars (100 hulls of 32 source points), PAM (32 medoids) and the source mean ($K = 1$), over 60 cells on the five full-dimensional configurations: Extreme Support beats PCHA in 46 of 60 cells, and PCHA beats the random-draw mean, PAM and $K = 1$ in all 60, including all 51 in which its own score is negative. By Proposition 1, PCHA's average advantage over random exemplars, 0.082, is 0.009 from recentered fit and 0.073 from a smaller mean-offset penalty, so the penalty accounts for 88.5% of the advantage and is the larger part in all 12 directed pairs (Appendix C.3). In the primary three-corpus study, PCHA's mean-referenced score exceeds that of hulls of random source points by a median of 0.090 (48 of 48 cells), and 88% of that difference comes from a smaller mean-offset penalty: over 90% in the four CMV-involving pairs, and 54–63% between the two negotiation corpora, where learned archetypes also add recentered fit (Figure H.8).

Source-native text. The same records are represented by their original text, the selected utterance or, for CMV, the winning comment, with records, splits, budget, $K = 32$ PCHA, seeds and numerically verified solvers unchanged, on the five full-dimensional configurations (60 paired cells). The accounting carries over: the fixed-map split describes every cross-corpus score, all of which are now negative, since the nine cells of that 60-cell comparison that were positive on blueprints turn negative; pair-mean recentered and translated fit are lower on raw text in all 12 directed pairs, and recovery is lower in all 60 paired cells, while location changes in direction-dependent ways (Appendices F.1–F.3). The two representations differ in more than wording: the blueprint generator saw

surrounding dialogue and response evidence (Appendix G.1), whereas the source-native text is the focal move alone, so we interpret the comparison as a paired representation-pipeline contrast, not as a causal estimate of abstraction or denoising.

7 SPARSE AUTOENCODERS: SAME IDENTITY, DIFFERENT REGIME

What carries over beyond convex hulls. We evaluate canonical Gemma Scope JumpReLU SAEs for Gemma 2 2B on post-MLP residual-stream activations ($D = 2304$), using the 16,384-feature SAEs at layers 5, 12, and 20 (nominal average L0 68, 82 and 71). Their reconstruction $\hat{x} = f(x)W_{\text{dec}} + b_{\text{dec}}$, with JumpReLU codes $f(x)$, comes from an amortized encoder–decoder map rather than from projection onto a convex source set, and the decoder bias b_{dec} appears only at the output (Appendix D.1). We use five text distributions, FineWeb, OpenWebMath, Spanish text, StarCoder (Python) and UltraChat, each split into alternating 1,024-token blocks: reference blocks define distribution means and shift directions, and all reconstruction scores are computed on the token-disjoint evaluation blocks (Appendix D.1). Our shifts change the input text while the model stays fixed (Gemma 2 2B, the model the SAEs were trained on); applying SAEs to a different model, such as base-model SAEs on instruction-tuned activations (Lieberum et al., 2024), is not tested here. Only Proposition 1 is used here: $R^2_{\text{cen}} = R^2_{\text{shift}} - \hat{\rho}^{\,2}$ holds for every fixed SAE reconstruction; Theorem 1 holds for any closed convex set containing the reconstructions, such as b_{dec} plus the cone of non-negative combinations of decoder directions, but we do not evaluate such a set for these SAEs (Section 3.1).

On these natural distributions, which may overlap the SAEs' training data, these SAEs are far from the fixed-map sign boundary. Across the 15 layer-by-distribution conditions, R^2_{cen} and R^2_{shift} both lie between 0.74 and 0.83 while $\hat{\rho}^{\,2}$ never exceeds 0.0016 (Appendix D.1 checks the identity numerically in every condition), so on natural data SAE fidelity is set almost entirely by recentered fit.

With the centered residual held fixed, the score reaches zero when $\hat{\rho} = \sqrt{R^2_{\text{shift}}}$, so the *fixed-output margin* $\sqrt{R^2_{\text{shift}}} - \hat{\rho}$ is the additional reconstruction-mean offset, in units of σ, that would bring the score to zero if added parallel to the existing offset (Appendix D.1); its smallest value across the 15 natural conditions is 0.84σ (Figure 1b), whereas the four CMV-involving PCHA pair means in Table 1 already satisfy $\hat{\rho}^{\,2} > R^2_{\text{shift}}$. The same fixed-map identity describes two different regimes. Shifting the activations themselves along the between-distribution directions (Appendix D.2) first makes the score negative at 0.54σ–1.51σ, whereas the measured separations between the distributions are at most 0.50σ, with a near miss at layer 20, where the earliest crossing, 0.54σ, is close to that layer's 0.50σ; these are boundary measurements, not immunity claims.

The decoder bias supplies substantial fixed-code location correction. Removing b_{dec} from the output while holding the learned codes fixed is a fixed-map location intervention: it translates every reconstruction by the same vector, so the centered residual and recentered fit are exactly unchanged, while mean-referenced fidelity changes sharply. Across the five distributions, removing the bias lowers R^2_{cen} by 0.33–0.35 at layer 5, 0.51–0.59 at layer 12, and 0.56–0.71 at layer 20, and the reconstruction-mean offset rises from at most 0.04σ to $\hat{\rho} = 0.57$–0.84. A negligible mean offset on natural data therefore does not mean that location is unimportant for these SAEs: with the codes fixed, the released decoder bias supplies substantial location correction, consistent with its training role as a learned centering vector (Appendix D.4). The intervention changes output location only and is not an SAE trained without a bias: the released encoder bias, which it leaves untouched, already absorbs b_{dec}, so the removed vector still shapes the codes (Appendix D.4).

8 WHAT TO REPORT

Reference conventions differ, sometimes under the same name: Gao et al. (2025), Rajamanoharan et al. (2024b) and Lieberum et al. (2024) divide the reconstruction error by that of the mean predictor, whereas He et al. (2024, Eq. 6) define "explained variance" as one minus the squared error over the squared norm of the input, a ratio against uncentered energy. As written, O'Neill et al. (2025, Eq. 13) use the variance of the residual instead, which gives the recentered fit and, reasonably for their question, leaves out the mean-offset penalty (Appendix I.2). We do not recommend one reference: the origin and the target mean answer different questions, and either can be right (Section 2).

Do not read a negative score as an absence of target-centered fit, and do not overread signs near zero. In Section 5, the recentered fit that survived transfer exceeded structure-free floors but nearly matched hulls of random source points, and transferred dictionaries recovered 32%–77% of same-corpus translated fit. Table 2 collects the practices; Appendix I.3 gives each with its evidence.

Table 2: What to report with a reconstruction-fidelity result under distribution shift.

Report	**Applies to**	**Needs**	**Shown in**
Reference predictor, in words; m when origin-referenced credit matters	every fidelity score	target mean and energy for m	Section 2
$\hat{\rho}^2$ and R^2_{shift}	any fixed reconstruction	the reconstruction mean and the centered residual	Sections 2, 5, 7
R^2_{ST}, labeled as an oracle diagnostic; ρ_H and the certificate	re-optimizable model sets; convex sets	one extra fit with a free translation; one projection	Sections 3.1, 5
A structure-free floor for R^2_{shift} and R^2_{ST}	convex model sets	rotated or isotropic hulls built from the same dictionary	Sections 3.1, 5
Matched held-out-self baseline on the identical target	transfer claims	one same-capacity model fitted on disjoint groups	Section 4
Resampling interval and optimization bound for scores near zero	sign claims	group bootstrap; solver bound	Section 4

9 Related work

Decompositions of a squared error into bias and centered parts come from forecast verification and hydrology (Appendix I.1), and a neighboring literature debates whether the observed mean is a suitable benchmark at all (Schaefli & Gupta, 2007); we accept what a negative mean-referenced score says and ask what the residual contains. Theorem 1 would not apply directly even to SAEs whose dictionary atoms are constrained to the convex hull of the data (Fel et al., 2025), whose codes remain ordinary sparse codes; our bridge between the families is one of measurement, not of architecture. Appendix I.4 gives the related work in full, including work on SAE error and on SAE behavior away from the training distribution (Engels et al., 2025; Heindrich et al., 2025).

10 Limitations and conclusion

Proposition 1 holds for any fixed reconstruction and Theorem 1 for any fixed reconstruction whose rows lie in a closed convex model set; the certificate is sufficient, not necessary, and the four-term split, the only part that apportions an observed excess, is analyzed empirically only for PCHA. The dictionary experiments cover four dialogue and argumentation corpora with corpus-specific record selection, twelve directed corpus pairs as the unit of generalization, eight embedding configurations that are not independent replications (two are truncations of a third), and one LLM blueprint pipeline (Section 6). The main PCHA results use $K = 32$ archetypes and L2-normalized embeddings, and the number of negative cells depends on K (48, 45, 37 and 33 at $K = 8$, 16, 32 and 64; Appendix G.2). The SAE study covers one base model and one SAE family at three layers, and its five distributions are not a calibrated ladder of shifts, nor established as outside the SAEs' training data (Appendix I.5 gives these limitations in full).

The measurement point does not depend on these limits. Under distribution shift, a reconstruction-fidelity score is not self-interpreting: reference credit, mean offset and translation-corrected fit are distinct quantities, and making them visible separately prevents the sign or magnitude of one fidelity score from being mistaken for an intrinsic property of the representation. Across model family, source summary, representation and capacity, what changed in the systems we measured was the balance between translation-corrected fit and mean offset; our conclusion concerns measurement, not a law of transfer. Two questions remain open that this account can pose but not answer: why some source summaries and reconstruction maps place their reconstructions better than others, and what model-set diagnostics play the role of the convex certificate in other reconstruction families.

AI USE STATEMENT

We used generative AI tools (large language models) as follows.

As part of the method. The blueprint representation studied in this paper is generated by an LLM, which rewrites each conversational move into a short procedural summary (Section 4; Appendix G); Section 6 compares it with the original text.

Research. The research question, the accounting and the theory (Proposition 1, Theorem 1, Proposition 2) are the authors' own. AI tools reviewed our drafts adversarially, and their critiques led to additional experiments: one auxiliary result, the positivity of recentered fit under convex projection, was first raised in such a critique, as was its extension to firmly nonexpansive maps, and we verified its proof; in response, we designed the structure-free floor experiment (Appendix E.6) with AI assistance, fixing its controls and decision rules before the results were known. AI tools also served as a sounding board when we interpreted results, and AI coding assistants helped implement and parallelize the experiment code under our direction, with parallel results verified against serial implementations.

Writing. AI tools helped draft, restructure, condense and edit the manuscript.

The authors reviewed all AI-assisted work, checked every reported number against the experimental outputs, verified all derivations, and take full responsibility for the content.

ETHICS STATEMENT

All data come from publicly released research corpora: the negotiation dialogues of CaSiNo and Craigslist Bargains, ChangeMyView discussions from Reddit distributed through ConvoKit, and the PersuasionForGood dialogues (Appendix G.1); we collected no new data from people and use the corpora under the terms of their releases. The sparse-autoencoder study uses the publicly released Gemma 2 model and Gemma Scope SAEs (Section 7). The paper reports aggregate quantities only; it quotes no conversational text and identifies no individual. The blueprints are LLM-written summaries of these public records that replace the parties and domain details with placeholders; automated audits found no item tokens, prices, topic entities or charity names in them (Appendix G.1). Our contribution is a measurement account of reconstruction fidelity, and we build no persuasion or negotiation system. The foreseeable misuse is selective reporting of whichever fidelity reference looks best; the reporting practices of Section 8 guard against it.

REPRODUCIBILITY STATEMENT

Appendix A gives the proofs of the formal results with their assumptions (Proposition 2 in Appendix A.3, Lemma 1 in Appendix A.6, and Proposition 1, Lemmas 2 and 3 and Theorem 1 in Appendix A.7). Section 4 and Appendix G specify the data and protocol: corpora, record selection, blueprint generation settings, embedding models and a leakage audit (Appendix G.1), fitting budgets and capacity (Appendix G.2), solver tolerances and the numerical verification of signs (Appendix G.3), aggregation rules (Appendix G.4), and the settings of all six studies (Appendix G.5, Table G.2). The structure-free floors and random-source hulls are specified in Appendix E.6, the controlled translations in Appendices B.1 and F.4, and the sparse-autoencoder evaluation, including models, layers, distributions, token-disjoint block splits and input-shift sweeps, in Appendices D.1 and D.2.

REFERENCES

S. Ayromlou, P. Sehgal, and P. Narayana. Diagnosing and mitigating compounding failures in agentic persuasion via taxonomic strategy retrieval. arXiv preprint, 2026. arXiv:2606.24976.

T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

J. Y. Campbell and S. B. Thompson. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4):1509–1531, 2008.

J. P. Chang, C. Chiam, L. Fu, A. Wang, J. Zhang, and C. Danescu-Niculescu-Mizil. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of SIGDIAL*, 2020.

K. Chawla, J. Ramirez, R. Clever, G. Lucas, J. May, and J. Gratch. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of NAACL-HLT*, 2021.

A. Cutler and L. Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

J. G. Eisenhauer. Regression through the origin. *Teaching Statistics*, 25(3):76–80, 2003.

J. Engels, L. R. Smith, and M. Tegmark. Decomposing the dark matter of sparse autoencoders. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=sXq3Wb3vef. arXiv:2410.14670.

K. Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of EMNLP-IJCNLP*, 2019.

T. Fel, E. S. Lubana, J. S. Prince, M. Kowal, V. Boutin, I. Papadimitriou, B. Wang, M. Wattenberg, D. Ba, and T. Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. In *Proceedings of ICML*, 2025. arXiv:2502.12892.

B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of ICCV*, 2013.

J. Gao, D. He, X. Tan, T. Qin, L. Wang, and T.-Y. Liu. Representation degeneration problem in training natural language generation models. In *Proceedings of ICLR*, 2019.

L. Gao, T. Dupré la Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike, and J. Wu. Scaling and evaluating sparse autoencoders. In *Proceedings of ICLR*, 2025. arXiv:2406.04093.

Google. Gemini API: Embeddings (gemini-embedding-2). https://ai.google.dev/gemini-api/docs/embeddings, 2026. Accessed 21 September 2026.

Google Cloud. Get text embeddings (text-embedding-005, text-multilingual-embedding-002). Gemini Enterprise Agent Platform documentation. https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/embeddings/get-text-embeddings, 2026. Accessed 25 September 2026.

H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2):80–91, 2009. doi: 10.1016/j.jhydrol.2009.08.003.

H. He, D. Chen, A. Balakrishnan, and P. Liang. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of EMNLP*, 2018.

Z. He, W. Shu, X. Ge, L. Chen, J. Wang, Y. Zhou, F. Liu, Q. Guo, X. Huang, Z. Wu, Y.-G. Jiang, and X. Qiu. Llama scope: Extracting millions of features from Llama-3.1-8B with sparse autoencoders, 2024. arXiv:2410.20526.

L. Heindrich, P. Torr, F. Barez, and V. Thost. Do sparse autoencoders generalize? A case study of answerability. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025. arXiv:2502.19964.

J. K. Jolliff, J. C. Kindle, I. Shulman, B. Penta, M. A. M. Friedrichs, R. Helber, and R. A. Arnone. Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. *Journal of Marine Systems*, 76(1–2):64–82, 2009. doi: 10.1016/j.jmarsys.2008.05.014.

L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.

C. Kissane, R. Krzyzanowski, A. Conmy, and N. Nanda. SAEs (usually) transfer between base and chat models. AI Alignment Forum, 2024.

K. Kobayashi and M. U. Salam. Comparing simulated and measured values using mean squared deviation and its components. *Agronomy Journal*, 92(2):345–352, 2000.

T. O. Kvålseth. Cautionary note about R^2. *The American Statistician*, 39(4):279–285, 1985.

J. Lee, Z. Dai, X. Ren, B. Chen, D. Cer, J. R. Cole, K. Hui, M. Boratko, R. Kapadia, W. Ding, Y. Luan, S. M. K. Duddu, G. Hernandez Abrego, W. Shi, N. Gupta, A. Kusupati, P. Jain, S. R. Jonnalagadda, M.-W. Chang, and I. Naim. Gecko: Versatile text embeddings distilled from large language models, 2024. arXiv:2403.20327.

B. Li, H. Zhou, J. He, M. Wang, Y. Yang, and L. Li. On the sentence embeddings from pre-trained language models. In *Proceedings of EMNLP*, 2020.

T. Lieberum, S. Rajamanoharan, A. Conmy, L. Smith, N. Sonnerat, V. Varma, J. Kramár, A. Dragan, R. Shah, and N. Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, Miami, Florida, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.19. arXiv:2408.05147.

J. Mei, C. Wang, and W. Zeng. Online dictionary learning for approximate archetypal analysis. In *Proceedings of ECCV*, volume 11207 of *LNCS*, 2018. doi: 10.1007/978-3-030-01219-9_30.

M. Mørup and L. K. Hansen. Archetypal analysis for machine learning and data mining. *Neurocomputing*, 80:54–63, 2012. doi: 10.1016/j.neucom.2011.06.033.

J. Mu and P. Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *Proceedings of ICLR*, 2018.

A. H. Murphy. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116(12):2417–2424, 1988.

A. H. Murphy and E. S. Epstein. Skill scores and correlation coefficients in model verification. *Monthly Weather Review*, 117(3):572–582, 1989.

J. Ni, Q. Qiu, and R. Chellappa. Subspace interpolation via dictionary learning for unsupervised domain adaptation. In *Proceedings of CVPR*, pp. 692–699, 2013.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

C. O'Neill, M. Jayasekara, and M. Kirkby. Resurrecting the salmon: Rethinking mechanistic interpretability with domain-specific sparse autoencoders, 2025. arXiv:2508.09363.

S. Rajamanoharan, A. Conmy, L. Smith, T. Lieberum, V. Varma, J. Kramár, R. Shah, and N. Nanda. Improving sparse decomposition of language model activations with gated sparse autoencoders. In *Advances in NeurIPS*, 2024a. arXiv:2404.16014.

S. Rajamanoharan, T. Lieberum, N. Sonnerat, A. Conmy, V. Varma, J. Kramár, and N. Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders, 2024b. arXiv:2407.14435.

N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, 2019.

B. Schaefli and H. V. Gupta. Do Nash values have value? *Hydrological Processes*, 21(15):2075–2080, 2007. doi: 10.1002/hyp.6825.

K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu. MPNet: Masked and permuted pre-training for language understanding. In *Advances in NeurIPS*, 2020.

J. Su, J. Cao, W. Liu, and Y. Ou. Whitening sentence representations for better semantics and faster retrieval, 2021. arXiv:2103.15316.

B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of AAAI*, 2016.

C. Tan, V. Niculae, C. Danescu-Niculescu-Mizil, and L. Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of WWW*, 2016.

K. E. Taylor. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, 106(D7):7183–7192, 2001. doi: 10.1029/2000JD900719.

H. Theil. *Applied Economic Forecasting*. North-Holland, 1966.

L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei. Text embeddings by weakly-supervised contrastive pre-training, 2022. arXiv:2212.03533.

W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in NeurIPS*, 2020.

X. Wang, W. Shi, R. Kim, Y. Oh, S. Yang, J. Zhang, and Z. Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of ACL*, pp. 5635–5649, 2019.

I. Welch and A. Goyal. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508, 2008. doi: 10.1093/rfs/hhm014.

C. J. Willmott. On the validation of models. *Physical Geography*, 2(2):184–194, 1981.

A PROOFS FOR THE CONVEX-HULL FOUR-TERM SPLIT

Section 3.3 uses two families of error; this appendix states the four-term split that refines them and proves it. Proposition 2 splits each family into two non-negative terms by comparing the hull with its affine span, where weights must sum to one but may be negative, and it relates the location family to the distance ρ_H of Theorem 1.

Proposition 2 (convex-hull four-term split). Let H be the convex hull of archetypes $a_1, \dots, a_K$ with mean $\bar{a}$, let $S = \operatorname{span}\{a_k - \bar{a}\}$ be the directions they span, and let P_S and $P_\perp$ be the orthogonal projections onto S and its complement. Reconstruct each row by its nearest point in H, and let h^* be the point of H nearest μ. Define

$$s = \frac{1}{\mathrm{TSS}} \sum_i \|P_\perp(x_i - \mu)\|^2, \qquad \lambda = \frac{\|P_\perp(\mu - h^*)\|^2}{\sigma^2}, \qquad \rho_S^2 = \frac{\|P_S(\mu - h^*)\|^2}{\sigma^2},$$

$c = 1 - R^2_{\mathrm{ST}} - s$ and $\iota = R^2_{\mathrm{ST}} - R^2_{\mathrm{cen}} - \lambda$. Then (proof in Appendix A.3)

$$1 - R^2_{\mathrm{cen}} \;=\; \underbrace{s + c}_{\text{shape family}} \;+\; \underbrace{\lambda + \iota}_{\text{location family}}, \qquad c \geq 0, \qquad \rho_H^2 = \lambda + \rho_S^2, \qquad \iota \geq \rho_S^2.$$

In words, s (subspace mismatch) is the target's variation around its mean that lies outside the directions the archetypes span, and λ (out-of-span location) is the part of the offset between the target mean and the hull that is perpendicular to those directions, which a translation removes exactly. The other two terms measure the extra cost of requiring reconstructions to stay inside the convex hull rather than anywhere in its affine span: c (convexity) in the shape family, once location is free, and ι (in-span location $\times$ convexity) in the location family, when the hull remains fixed. The last two relations connect the split to Theorem 1. The squared centroid-to-hull distance divides into its part perpendicular to the archetype directions, which is exactly λ, and its part within them, ρ_S^2, which is at most ι. The location family is therefore at least ρ_H^2, which refines Theorem 1 for this reconstruction, and the gap between the theorem's bound and the score is exactly $\iota - \rho_S^2$: the extra cost of fitting a spread-out target inside a hull that cannot move (Appendix A.4). When the certificate does not apply, this gap is what can still make the score negative. The split is exact and does not depend on the order in which constraints are relaxed.

A.0 SETTING AND NOTATION

Let the dictionary have archetypes $a_1, \dots, a_K \in \mathbb{R}^D$, the rows of $A \in \mathbb{R}^{K \times D}$, and let $H = \operatorname{conv}\{a_1, \dots, a_K\}$ be their convex hull. Let $\bar{a} = \frac{1}{K}\sum_k a_k$ and let $S = \operatorname{span}\{a_k - \bar{a}\}$, the row space of $A_c = A - \mathbf{1}\bar{a}^\top$. The affine hull of H is $\bar{a} + S$, and $H \subseteq \bar{a} + S$. Write P_S for orthogonal projection onto S and $P_\perp = I - P_S$.

The target matrix $X \in \mathbb{R}^{N \times D}$ has rows x_i, mean μ and $\mathrm{TSS} = \sum_i \|x_i - \mu\|^2 > 0$. The four residuals are

$$\mathrm{RSS}_{\mathrm{SF}} = \sum_i \min_{h \in H} \|x_i - h\|^2, \qquad \mathrm{RSS}_{\mathrm{ST}} = \min_{t \in \mathbb{R}^D} \sum_i \min_{h \in H} \|x_i - t - h\|^2,$$

$$\mathrm{RSS}_{\mathrm{AF}} = \sum_i \min_{g \in \bar{a} + S} \|x_i - g\|^2, \quad \mathrm{RSS}_{\mathrm{AT}} = \min_{t \in \mathbb{R}^D} \sum_i \min_{g \in \bar{a} + S} \|x_i - t - g\|^2.$$

The letters denote simplex (S) or affine (A) weights and fixed position (F) or one free translation (T). Each inner minimum is attained because H and $\bar{a} + S$ are closed and convex; $\mathrm{RSS}_{\mathrm{ST}}$ is attained because the objective is continuous in t and grows without bound as $\|t\| \to \infty$ (H is bounded). Throughout, $R^2_{\mathrm{cen}} = 1 - \mathrm{RSS}_{\mathrm{SF}}/\mathrm{TSS}$, that is, the fixed-position reconstruction is the optimal one, as in all our PCHA evaluations, where the optimizer's bound numerically verifies optimality to the tolerance of Section 4 and Appendix G.3.

The nearest point of H to μ is h^*, with $d = \|\mu - h^*\| = d(\mu, H)$, $d_S = \|P_S(\mu - h^*)\|$, and $\rho_H^2 = Nd^2/\mathrm{TSS} = d^2/\sigma^2$.

The four residuals are easiest to see as a 2×2 design: weights are either affine (sum to one, signs free) or simplex (non-negative and sum to one), and position is either fixed or allowed one global translation.

Table A.1 lists the quantities of Sections 2–3 and the question each answers.

Table A.1: **The quantities of the accounting: what each answers and what it does not.**

Quantity	Weights	Applies to	Answers	Does not answer
$\hat{\rho}^2$ — fixed-map mean-offset penalty	fixed	any reconstruction	How much of this reconstruction's score loss is the mean-offset penalty?	What a refit could recover
R^2_{shift} — recentered fit	fixed	any reconstruction	How much centered structure do these same weights capture after recentering?	What the model set could capture after re-optimization
R^2_{ST} — translated fit, a diagnostic (recovery is its ratio to the matched held-out self's)	re-optimized	models with a reconstruction set	What centered fit is available once one translation is freed?	Zero-shot performance
ρ_H — centroid-to-set (for hulls, centroid-to-hull) distance	no target reconstruction weights	convex sets	Is a negative sign forced by $\rho^2_H > R^2_{\text{ST}}$?	How much of observed excess is location
location and shape families (Proposition 2)	both	convex-hull dictionaries (analyzed here for PCHA)	Which carries more of the excess over matched held-out self?	Anything about models that are not convex hulls

A.1 The affine residuals in closed form

Lemma A.1.

$$\text{RSS}_{\text{AF}} = \sum_i \|P_\perp(x_i - \bar{a})\|^2, \qquad \text{RSS}_{\text{AT}} = \sum_i \|P_\perp(x_i - \mu)\|^2,$$
$$\text{RSS}_{\text{AF}} - \text{RSS}_{\text{AT}} = N\|P_\perp(\mu - \bar{a})\|^2.$$

Proof. The closest point of the affine set $\bar{a}+S$ to y is $\bar{a}+P_S(y-\bar{a})$, at squared distance $\|P_\perp(y-\bar{a})\|^2$. This gives the first formula. With a translation t, the residual is $\sum_i \|P_\perp(x_i - \bar{a}) - P_\perp t\|^2$, which depends on t only through $u = P_\perp t$ and is minimized at the mean, $u = \frac{1}{N}\sum_i P_\perp(x_i - \bar{a}) = P_\perp(\mu - \bar{a})$. Substituting gives the second formula. For the third, write $x_i - \bar{a} = (x_i - \mu) + (\mu - \bar{a})$ and expand; the cross term vanishes because $\sum_i (x_i - \mu) = 0$. ∎

So $\lambda = N\|P_\perp(\mu - \bar{a})\|^2/\text{TSS}$: the out-of-span component of the offset between the target mean and the archetype mean, which one translation removes exactly.

A.2 The centroid identity

Lemma A.2. $Nd(\mu, H)^2 = \text{RSS}_{\text{AF}} - \text{RSS}_{\text{AT}} + Nd_S^2$, equivalently $\rho^2_H = \lambda + Nd_S^2/\text{TSS}$.

Proof. Split $\mu - h^*$ into its components along S and orthogonal to it: $d^2 = \|P_S(\mu - h^*)\|^2 + \|P_\perp(\mu - h^*)\|^2$. Because $h^* \in H \subseteq \bar{a} + S$, the difference $h^* - \bar{a}$ lies in S, so $P_\perp(\mu - h^*) = P_\perp(\mu - \bar{a})$. Multiply by N and use Lemma A.1. ∎

The out-of-span part of the centroid distance is therefore fixed by the affine hull alone and is the same for every point of H; only the in-span part d_S depends on where μ projects relative to the hull.

A.3 Proof of Proposition 2

Proposition 2. With s, λ, c, ι as defined in Appendix A.0,

$$s + \lambda + c + \iota = 1 - R^2_{\text{cen}}, \qquad s, \lambda, c, \iota \geq 0, \qquad \iota \;\geq\; \frac{N d_S^2}{\text{TSS}}.$$

Proof. **Closure.** Multiplying by TSS,

$$\begin{aligned} &\text{RSS}_{\text{AT}} + (\text{RSS}_{\text{AF}} - \text{RSS}_{\text{AT}}) + (\text{RSS}_{\text{ST}} - \text{RSS}_{\text{AT}}) \\ &\quad + \big[(\text{RSS}_{\text{SF}} - \text{RSS}_{\text{ST}}) - (\text{RSS}_{\text{AF}} - \text{RSS}_{\text{AT}})\big] = \text{RSS}_{\text{SF}}, \end{aligned}$$

and $\text{RSS}_{\text{SF}}/\text{TSS} = 1 - R^2_{\text{cen}}$.

$s \geq 0$ because it is a sum of squares.

$\lambda \geq 0$. Taking $t = 0$ in the definition of RSS_{AT} gives $\text{RSS}_{\text{AT}} \leq \text{RSS}_{\text{AF}}$. (Lemma A.1 gives the value directly.)

$c \geq 0$. For every t, $H \subseteq \bar{a} + S$, so each inner minimum over H is at least the corresponding minimum over $\bar{a} + S$. Minimizing both sides over t gives $\text{RSS}_{\text{ST}} \geq \text{RSS}_{\text{AT}}$.

$\iota \geq N d_S^2/\text{TSS} \geq 0$. Theorem 1 with the optimal fixed-position reconstruction reads $1 - \text{RSS}_{\text{SF}}/\text{TSS} \leq 1 - \text{RSS}_{\text{ST}}/\text{TSS} - Nd^2/\text{TSS}$, that is,

$$\text{RSS}_{\text{SF}} - \text{RSS}_{\text{ST}} \;\geq\; Nd^2.$$

By Lemma A.2, $Nd^2 = (\text{RSS}_{\text{AF}} - \text{RSS}_{\text{AT}}) + N d_S^2$. Subtracting $\text{RSS}_{\text{AF}} - \text{RSS}_{\text{AT}}$ from both sides gives $\iota\,\text{TSS} \geq N d_S^2$. ∎

A.4 The slack of Theorem 1 for PCHA

Corollary A.3. For the optimal fixed-position PCHA reconstruction,

$$\big(R^2_{\text{ST}} - \rho^2_H\big) - R^2_{\text{cen}} \;=\; \iota - \frac{N d_S^2}{\text{TSS}} \;\geq\; 0.$$

Proof. The left side equals $(\text{RSS}_{\text{SF}} - \text{RSS}_{\text{ST}} - Nd^2)/\text{TSS}$. Substitute Lemma A.2 for Nd^2 and the definition of ι. Non-negativity is the last step of Proposition 2. ∎

The theorem thus accounts for the out-of-span part of location through λ and for the in-span part only up to its geometric minimum $N d_S^2/\text{TSS}$. What remains, $\iota - N d_S^2/\text{TSS}$, is the additional cost of fitting a spread-out target cloud inside a hull that cannot move, which is the source of the slack reported in Appendix A.7 and Section 5.

A.5 Remarks

1. *Order-free.* All four terms are defined from the four residuals directly, so no ordering of "remove location first, then shape" is involved. The two families are $s + c = 1 - R^2_{\text{ST}}$ and $\lambda + \iota = R^2_{\text{ST}} - R^2_{\text{cen}}$, both immediate from the definitions.
2. *Scope.* Lemmas A.1–A.2 and Proposition 2 use only that H is the convex hull of finitely many points, with affine hull $\bar{a} + S$; they do not use how the points were fitted. They therefore apply to any convex-hull dictionary, including the source-summary controls of Section 6 (for $K = 1$, $S = \{0\}$ and $c = 0$). The paper analyzes the split empirically only for PCHA.
3. *Rank.* Nothing requires $\dim S = K - 1$. In our fits $\dim S = 31$ for $K = 32$; the in-span enrichment of Appendix B.2 uses the observed rank r.
4. *Numerical check.* The closure identity holds in the reported aggregates up to rounding (Table 1; Section 6). Given Lemma A.2, which is exact, the bound $\iota \geq N d_S^2/\text{TSS}$ is equivalent to Theorem 1 for the optimal fixed-position reconstruction, so the zero-violation counts reported in Appendix E.3 also count violations of this bound.

5. *Main-text form.* The statement above gives Proposition 2 with s and λ in the closed forms of Lemma A.1 (λ written with h^*, since $P_\perp(\mu - h^*) = P_\perp(\mu - \bar{a})$ as in the proof of Lemma A.2), with $\rho_S^2 = N d_S^2/\mathrm{TSS}$, and with c and ι as the remainders of the two families. These definitions agree with those of Appendix A.0, and Corollary A.3 is the statement that the slack of Theorem 1 equals $\iota - \rho_S^2$.

A.6 Proof of Lemma 1

Lemma 1 (recentered fit under convex projection). If each row of $\hat{X}$ is the nearest point of a closed convex set H to the corresponding row of X, then $R^2_{\mathrm{shift}} \geq \|C\hat{X}\|_F^2/\mathrm{TSS}$, which is positive unless all reconstructions coincide.

Proof (firm nonexpansiveness). Here $C = I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ as in Section 2. For the projection P onto H, $\langle Px - Py,\, x - y\rangle \geq \|Px - Py\|^2$. With $r = x - Px$ this gives $\|r_x - r_y\|^2 \leq \|x - y\|^2 - \|Px - Py\|^2$. Summing over all pairs of rows and using $\sum_{i,j}\|a_i - a_j\|^2 = 2N\|CA\|_F^2$ gives $\|C(X - \hat{X})\|_F^2 \leq \mathrm{TSS} - \|C\hat{X}\|_F^2$. Dividing by TSS gives the bound, since $R^2_{\mathrm{shift}} = 1 - \|C(X - \hat{X})\|_F^2/\mathrm{TSS}$. The bound is zero only if $C\hat{X} = 0$, that is, only if every reconstruction equals their mean. ∎

The proof uses only firm nonexpansiveness of the reconstruction map, $\langle Px - Py,\, x - y\rangle \geq \|Px - Py\|^2$, so the bound holds for any firmly nonexpansive map; projection onto a closed convex set is the case used here. It covers the PCHA reconstructions, held-out self and every comparison hull of Appendix E.6, but not the sparse autoencoders of Section 7, whose reconstruction maps are not projections and need not be firmly nonexpansive.

A.7 Proofs of Proposition 1, Lemmas 2 and 3, and Theorem 1

Sections 2 and 3 use these statements; this section proves them.

Proof of Proposition 1. Write $E = X - \hat{X}$. Because C is the orthogonal projection that removes column means, $\|E\|_F^2 = \|CE\|_F^2 + \|(I - C)E\|_F^2$, and every row of $(I - C)E = \frac{1}{N}\mathbf{1}\mathbf{1}^\top E$ equals $\mu - \hat{\mu}$, so $\|(I - C)E\|_F^2 = N\|\mu - \hat{\mu}\|^2$. Dividing by $\mathrm{TSS} = N\sigma^2$ gives $1 - R^2_{\mathrm{cen}} = (1 - R^2_{\mathrm{shift}}) + \hat{\rho}^{\,2}$. ∎

Lemma 2. If every row of $\hat{X}$ lies in H, then $R^2_{\mathrm{shift}} \leq R^2_{\mathrm{ST}}$.

Proof. The pair consisting of the fixed reconstruction $\hat{X}$ and translation $t = \mu - \hat{\mu}$ is feasible for the free-translation problem. Its residual is $C(X - \hat{X})$, so the optimum cannot have larger residual. ∎

Lemma 3. If H is convex and every row of $\hat{X}$ lies in H, then $\rho_H \leq \hat{\rho}$.

Proof. The mean $\hat{\mu}$ is an average of points in H, so $\hat{\mu} \in H$ by convexity. Its distance from μ is therefore at least the distance from μ to the nearest point in H. ∎

Theorem 1 (translation bound). Let H be nonempty, closed and convex, let $\hat{X}$ be any reconstruction of X whose rows lie in H, and let $\mathrm{TSS} > 0$. Then

$$R^2_{\mathrm{shift}} - \hat{\rho}^{\,2} \;=\; R^2_{\mathrm{cen}} \;\leq\; R^2_{\mathrm{ST}} - \rho_H^2.$$

Proof. The equality is Proposition 1. For the inequality, make two relaxations. First, allow the row reconstructions to re-optimize after translation, replacing R^2_{shift} by the no-smaller R^2_{ST}. Second, replace the realized reconstruction mean by the closest point in H, replacing $\hat{\rho}^{\,2}$ by the no-larger ρ_H^2. Each change can only increase the expression. ∎

The slack is also readable:

$$(R^2_{\mathrm{ST}} - \rho_H^2) - R^2_{\mathrm{cen}} = (R^2_{\mathrm{ST}} - R^2_{\mathrm{shift}}) + (\hat{\rho}^{\,2} - \rho_H^2).$$

The two non-negative terms are exactly the re-optimization gain and the excess mean-offset penalty. If H is a singleton, both relaxations are tight: $R^2_{\mathrm{ST}} = R^2_{\mathrm{shift}} = 0$ and $R^2_{\mathrm{cen}} = -\rho_H^2$.

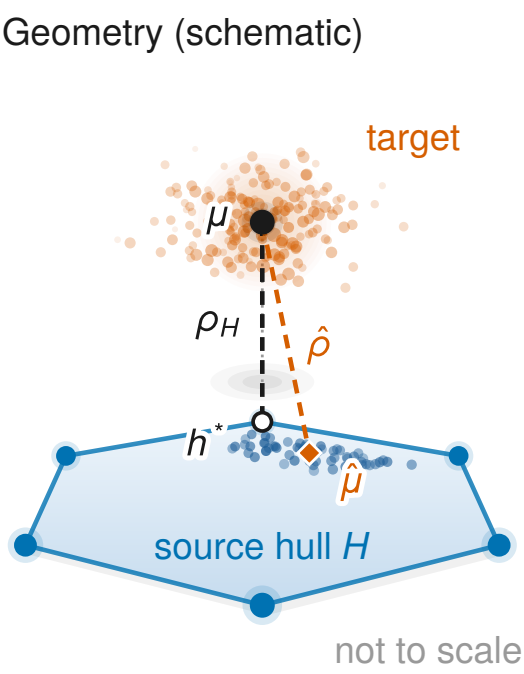

Figure A.1: **Geometry of Theorem 1 (schematic, not to scale).** Every reconstruction lies in the source hull H, so their mean $\hat{\mu}$ does too, and its distance $\hat{\rho}$ from the target mean μ is at least the distance ρ_H from μ to the nearest point h^* of H (Lemma 3). For CMV→Craigslist, $\rho_H = 0.87\sigma$ and $\hat{\rho} = 0.89\sigma$ (Section 3.1).

A.8 What the reference identity settles, and what it cannot

Rankings on one target cannot change. For a fixed target, m is constant and the identity is a strictly increasing affine transformation whenever the target has nonzero centered variation. Any two reconstructions of the same target are therefore ordered identically under the two references. The five source-summary methods in Section 6, for example, necessarily have the same within-cell ordering under R^2_{unc} and R^2_{cen}.

Comparisons across targets can change. Each target j has its own m_j. For two reconstructions evaluated on that target,

$$\Delta R^2_{\text{unc},j} = (1 - m_j)\Delta R^2_{\text{cen},j}.$$

Aggregating over heterogeneous corpora, layers, or distributions therefore applies different target-specific scalings under the two references. Aggregate orderings are not protected by the fixed-target result. Whether they change is empirical; in the five-distribution SAE suite of Appendix D.5 they happen not to, although the numerical spread between variants changes.

The residual is untouched. The identity is denominator-side accounting. Once RSS is known, it converts one score into the other; it says nothing about why RSS has that value. In particular, it cannot tell us why a reconstruction has RSS $>$ TSS, which is exactly the condition for $R^2_{\text{cen}} < 0$, or whether useful target-centered structure is present inside that residual.

A.9 Further remarks on the accounting

A reconstruction whose mean lies one σ from the target mean loses exactly 1.0 from its mean-referenced score relative to its recentered score. Fitted in-sample, a simplex-constrained reconstruction is translation-equivariant: shifting all the data shifts the archetypes, and hence every reconstruc-

tion, by the same vector and leaves the residual unchanged, so a common translation of the fitting and evaluation data creates no mean offset. A mean offset can arise whenever a fixed fitted model is evaluated on data whose location differs from its fitting data, including held-out data from the same corpus.

Convex hulls of archetypes are the case studied here, but linear and affine subspaces, convex combinations of fixed prototypes, linear decoders with convex constraints on the codes (a simplex or an ℓ_1 ball), and convex cones such as unrestricted non-negative codes are covered in the same way.

We call a setting sign-fragile when limited translated fit leaves little margin before an offset forces the score below zero. The term refers only to the score's sign. The certificate boundary has a direct scale: $\rho_H = \sqrt{R^2_{\mathrm{ST}}}$, equivalently $d(\mu, H) = \sigma\sqrt{R^2_{\mathrm{ST}}}$. A model with $R^2_{\mathrm{ST}} = 0.20$ reaches the boundary at $d(\mu, H) \approx 0.45\sigma$, whereas one with $R^2_{\mathrm{ST}} = 0.80$ reaches it only near 0.89σ. For the matched $K = 32$, 785-move PCHA study, the boundary sits near $\rho_H \approx 0.47$ ($\rho^2_H \approx 0.22$). This value is specific to this capacity and fitting budget; it is not a universal constant (other K: Appendix G.2).

B Further analyses for Section 3

B.1 Controlled translations

Exact values for the sweep of Section 3.2. Take the 60 fitted CMV dictionaries of the running example (six embedding configurations, one per model with text-embedding-005 at 768 dimensions; two splits; five seeds) and translate each hull rigidly by αv, where $v = \mu - h^*$ runs from h^*, the point of the hull nearest the Craigslist mean, to that mean, so that $\alpha = 1$ places the target centroid exactly on the hull and $\rho_H = 0$. R^2_{ST} stays constant at 0.153, while mean R^2_{cen} rises from -0.683 at $\alpha = 0$ to $+0.110$ at $\alpha = 1$ as mean ρ^2_H falls from 0.767 to zero, and falls again beyond it (-0.042 at $\alpha = 1.5$). The same sweep was run on all 360 directed source-to-target evaluations in these six configurations (six directed pairs $\times$ 60; each source dictionary serves two targets): placing each target centroid on its hull raises mean R^2_{cen} from -0.271 to $+0.149$, against a translated fit of 0.181, and makes the mean score positive for every one of the six directed pairs. With the centroid-to-hull distance removed ($\rho_H = 0$), the score rises to within about 0.03–0.04 of its translated-fit ceiling on average and no further; the remaining gap is the theorem's slack.

Repeating the sweep on all eight configurations and ten splits (400 CMV dictionaries; 2,400 evaluations over the six directed pairs; 19,200 translated-hull states), translated fit stays at 0.1615 for CMV→Craigslist and 0.1890 over all six pairs. Mean R^2_{cen} rises from -0.6840 to $+0.1129$ for CMV→Craigslist, where all 400 dictionaries change sign, and from -0.2711 to $+0.1541$ over all six pairs, where all 2,400 evaluations are positive at $\alpha = 1$. No state violates Theorem 1.

Two further sweeps. The first breaks a good score and uses no cross-corpus reconstruction at all. It is illustrative: one embedding configuration (Gemini embedding-2), one held-out split, and one dictionary seed. We start from a held-out CMV self cell whose $K = 32$ dictionary has $R^2_{\mathrm{cen}} \approx +0.19$ and $R^2_{\mathrm{ST}} \approx 0.20$. We then translate every archetype by $\beta\sigma u$, where u points along the observed CMV–Craigslist centroid direction, while leaving the target fixed. The observed corpus-centroid separation is about 1.02σ in this cell's units, so $\beta \approx 1$ corresponds to one real inter-corpus displacement.

As explained in Section 3.2, R^2_{ST} is constant along the sweep, while the mean-referenced score moves from about $+0.19$ at $\beta = 0$ to -0.04 at $\beta = 0.5$ and -0.73 at $\beta = 1.0$. The centroid-to-hull distance moves from $\rho_H \approx 0.06$ to 0.46 and 0.92 at the same points. Thus a displacement on the scale of a real corpus offset can turn a positive held-out-self score into a strongly negative one without changing translated fit.

The second repairs a bad score along a different direction from the sweep of Section 3.2, the mean-difference direction, on natural CMV→Craigslist transfer. Each fitted CMV dictionary is translated rigidly by $\beta\sigma$ along the unit vector from the CMV fitting-partition mean to the Craigslist evaluation mean, so that $\beta = 1$ moves the hull by one target σ toward the target; the natural separation is 0.95–1.21σ in the configurations used. Averaged over 19 dictionaries (configuration, split and seed combinations), $R^2_{\mathrm{ST}} = 0.1521$ at every β, while mean R^2_{cen} rises from -0.6098 at $\beta = 0$ to $+0.1402$ at $\beta = 1$ as mean ρ^2_H falls from 0.6882 to 0.0077; past the target the score falls again (-0.5876 at

$\beta = 2$). With location nearly repaired, the score approaches its translated-fit budget: at $\beta = 1$ the bound $R^2_{\mathrm{ST}} - \rho^2_H$ is 0.1444, only 0.0042 above the achieved score. The 19 units are those, among 60 candidates (six configurations $\times$ two splits $\times$ five seeds), in which this mean-difference direction is not nearly collinear (cosine below 0.9) with the direction v from the nearest hull point to the target centroid. The rule depends on geometry, not on scores; the qualifying units come from four configurations. Because this direction differs from v, the two repair sweeps step on different scales (β in target σ; α as a fraction of v). Figures H.1 and H.4 plot these sweeps.

Table B.1 gives the full α grid for both sweeps, and Table B.2 the per-pair values at α = 1, where each target centroid lies on its hull.

Table B.1: **Rigid translation along $v = \mu - h^*$: mean R^2_{cen} by step α.** Left: the 60 CMV→Craigslist dictionaries (six configurations $\times$ two splits $\times$ five seeds); $R^2_{\mathrm{ST}} = 0.1532$ at every α, and mean $\rho^2_H = 0.7675$ at α = 0. Right: all 360 directed evaluations of the same six configurations (six directed pairs $\times$ 60; each source dictionary serves two targets); $R^2_{\mathrm{ST}} = 0.1814$ at every α. At α = 1 the gap to R^2_{ST} is 0.0436 and 0.0325.

α	CMV→Craigslist (60)	All six pairs (360)
−0.5	−1.6457	−0.7790
0	−0.6825	−0.2712
0.25	−0.3437	−0.0921
0.5	−0.0994	+0.0373
0.75	+0.0513	+0.1175
1	+0.1096	+0.1489
1.25	+0.0778	+0.1323
1.5	−0.0418	+0.0686

Table B.2: **The same sweep at α = 1, by directed pair: mean R^2_{cen} / R^2_{ST}.** Every pair's mean score is positive once its target centroid lies on the hull.

Directed pair	R^2_{cen} at α = 1	R^2_{ST}
CaSiNo → Craigslist	+0.2257	0.2571
Craigslist → CaSiNo	+0.2121	0.2303
CaSiNo → CMV	+0.1243	0.1552
CMV → CaSiNo	+0.1288	0.1642
Craigslist → CMV	+0.0929	0.1284
CMV → Craigslist	+0.1096	0.1532

B.2 Geometry of the offset

Let h^* be the nearest point in H to the target mean, let $d = \|\mu - h^*\| = d(\mu, H)$, and let P_S denote orthogonal projection onto the source direction space S. Define

$$d_S = \|P_S(\mu - h^*)\|, \qquad \varphi = d_S^2/d^2,$$

with $\varphi = 0$ when $d = 0$. The centroid geometry then connects directly to the four-term split of Proposition 2 (Appendix A, Lemma A.2):

$$N d(\mu, H)^2 = \mathrm{RSS}_{\mathrm{AF}} - \mathrm{RSS}_{\mathrm{AT}} + N d_S^2,$$

and hence

$$\rho^2_H = \lambda + \frac{N d_S^2}{\mathrm{TSS}}.$$

Thus λ is the unavoidable out-of-span part of centroid distance, while Nd_S^2/TSS is its in-span part. We call $\varphi/(r/D)$ the in-span *enrichment*, where r/D is the expected squared share for an isotropically oriented direction in $\mathbb{R}^D$. These quantities ask how much of the model-set location mismatch lies along directions the source dictionary itself spans.

Combining this identity with the family split makes the slack of Theorem 1 exact for convex-hull dictionaries:

$$(R^2_{\mathrm{ST}} - \rho^2_H) - R^2_{\mathrm{cen}} = \iota - \frac{Nd_S^2}{\mathrm{TSS}}.$$

The theorem therefore removes only the geometric minimum of the in-span location interaction; the remaining slack is the additional simplex cost of keeping a dispersed target cloud inside a fixed hull.

In our $K = 32$ PCHA fits, H lies in an affine space of rank at most 31 inside embedding spaces of dimension 256–1,024, so μ lies outside H generically, including for held-out self cells. Membership itself is not informative; the distance is.

C Further analyses for Section 6

C.1 Fourth corpus: full results

The blueprint analysis uses 1,076 selected moves from 758 dialogues, the same eight embedding configurations as the three-corpus study, $K = 32$ PCHA, ten dialogue-level splits, five dictionary seeds, and a 538-move fitting budget matched across all four corpora.

The sign pattern is widespread: 46 of 48 PFG-involving cells have negative median R^2_{cen}, and 41 are negative in every one of their 50 split-seed evaluations. Yet the sign count alone makes PFG transfer look harsher than its typical magnitude. On the same 538-move budget and with the same eight configurations, the median PFG-involving cell is **less** negative than the original-pair cells (-0.1157 versus -0.2645), has slightly higher translated fit ($R^2_{\mathrm{ST}} = 0.1907$ versus 0.1742), nearly unchanged recovery (0.587 versus 0.602), and a smaller median location share (68.8% versus 74.0%). The larger negative fraction, 46/48 versus 39/48, therefore reflects where the two panels sit relative to the zero boundary, especially the near-zero negotiation cells in the original pairs, and not uniformly worse PFG reconstruction. The median comparison is also partly compositional: 32 of the 48 original-pair cells involve CMV, whereas only 16 of the 48 PFG-involving cells are in the CMV↔PFG directions. We therefore do not interpret the less-negative PFG median as evidence that PFG is intrinsically easier. More PFG cells are negative, but by less, so the count and the size of negative scores should both be reported.

The fixed-map accounting itself is unchanged. Across the 2,400 PFG cross-corpus evaluations, the centered residual after recentering remains below target variance: $1 - R^2_{\mathrm{shift}}$ has median 0.840 and ranges from 0.775 to 0.928. Thus the same fixed maps retain positive recentered fit, as Lemma 1 guarantees, even though almost every mean-referenced score is negative.

The convex-hull certificate behaves similarly. Of the 46 median-negative cells, 37 satisfy $\rho^2_H > R^2_{\mathrm{ST}}$, so the hull's distance from the target mean is enough to force the sign in most cases. The theorem has zero violations; its slack has median 0.0497, with 5th and 95th percentiles 0.0346 and 0.0756. Freeing one global translation still leaves substantial fit: cell-level recovery spans 0.359–0.743. The fourth corpus therefore reproduces the central accounting while showing that a high negative-cell count need not imply lower translated fit or lower recovery.

The continuous location-versus-shape comparison is unambiguous: location excess is larger than shape excess in all 48 PFG cells. Its share of the cross-versus-self excess has median 68.8%, with range 59.2%–86.5%. The directed-pair means make the gradient visible without thresholding it: 60.7% for CaSiNo→PFG, 66.3% for Craigslist→PFG, 71.5% for CMV→PFG, 68.4% for PFG→CaSiNo, 67.9% for PFG→Craigslist, and 83.5% for PFG→CMV. In particular, CMV-related directions remain the most location-heavy, while CaSiNo→PFG is much closer to an even location/shape split.

The stricter Section-4 classification is useful here because weakening that claim is itself part of the fourth-corpus result. It labels 28 of 48 PFG cells location-dominated and 20 mixed, compared with 44 of 48 location-dominated and 4 mixed for the original three-corpus pairs recomputed at the

same 538-move budget. Thus the fourth corpus preserves the direction of the location imbalance but makes strong, split-consistent dominance substantially less universal.

Shape mismatch remains substantial throughout: $\Delta s/\Delta T$ has median 0.393 and ranges from 0.207 to 0.551. Raw Δc is negative in all 48 cells, but the conditional convexity quantity $c/(1-s)$ is worse under transfer in 46 of 48, reproducing the compositional caution from Appendix E.3. Location is the larger excess family, not the only source of transfer loss.

The fourth corpus also reproduces the in-span geometry of Section 5. The in-span share φ has median 0.258, and its enrichment relative to r/D has median 4.97. The interaction term ι tracks the normalized in-span centroid-distance term Nd_S^2/TSS closely: slope 1.138, intercept 0.041, with correlation 0.955 across the 2,400 PFG cross-corpus evaluations. The location effect is therefore again not well described as a purely orthogonal domain offset.

Translated-fit recovery remains graded and directional. Pair-mean recovery is 0.577 for CaSiNo→PFG, 0.532 for Craigslist→PFG, 0.489 for CMV→PFG, 0.613 for PFG→CaSiNo, 0.611 for PFG→Craigslist, and 0.660 for PFG→CMV. Figure H.9a places these beside the six original directed pairs recomputed at the same 538-move budget across all eight configurations.

Table C.1 gives the six PFG-involving directed pairs.

Table C.1: **PFG-involving directed pairs at the 538-move budget.** Grand means over all split-seed evaluations of the pair across the eight configurations, at five decimals as computed; recovery is the pair-level value quoted above. $R^2_{\mathrm{cen}} = R^2_{\mathrm{shift}} - \hat{\rho}^{\,2}$ closes before rounding.

Directed pair	R^2_{unc}	R^2_{cen}	R^2_{shift}	$\hat{\rho}^{\,2}$	R^2_{ST}	Recovery
CaSiNo → PFG	0.77103	−0.03405	0.17215	0.20619	0.19905	0.577
Craigslist → PFG	0.74720	−0.14173	0.16030	0.30204	0.18358	0.532
CMV → PFG	0.71816	−0.27814	0.12965	0.40779	0.16941	0.489
PFG → CaSiNo	0.77323	−0.08368	0.18000	0.26369	0.20758	0.613
PFG → Craigslist	0.80879	−0.08268	0.17280	0.25548	0.21673	0.611
PFG → CMV	0.72037	−0.26367	0.12178	0.38545	0.16036	0.660

C.2 Fourth corpus: controlled translations and cells near zero

The controlled translation of a held-out-self dictionary (Appendix B.1) extends to PFG. For each of the eight configurations we take one PFG dictionary (first split, first seed), translate it rigidly toward the fitting-partition centroid of each of the three other corpora, and score it on held-out PFG data, giving 24 paths. R^2_{ST} is invariant along every path (largest within-path range 3.9×10^{-13}) and equals 0.295–0.375 across configurations. At zero displacement all 24 scores are positive, 0.288–0.363. At the real inter-corpus displacement, 0.49–0.88σ, 17 of the 24 are negative, three more lie between 0 and $+0.01$, and four are between $+0.02$ and $+0.12$. Displacement toward CMV is the largest (0.63–0.88σ) and gives the most negative scores: seven of eight paths are negative, from about -0.12 to -0.34. A real-scale displacement can therefore remove most or all of a positive held-out-self score while translated fit is untouched, the same causal isolation as in the three-corpus sweep.

The cells nearest the natural zero boundary are correspondingly fragile to sampling variation. The seven closest negative 50-evaluation medians run from -0.020 to -0.003; the closest positive cell is $+0.021$. Four of these seven are Craigslist→CaSiNo cells and three involve PFG. Dialogue-group bootstrap resampling with 200 replicates per cell places zero inside the 95% interval for four of the seven and outside for three (538-move budget). We therefore describe these cells by their proximity to the boundary rather than treating their signs as qualitatively distinct regimes.

C.3 Source summaries: full comparison

The comparison summaries use the same source fitting records as PCHA. Extreme Support is a deterministic greedy selection: its first point is the record farthest from their mean, and each of the other 31 is the remaining record farthest from the convex hull of the points already chosen (squared Euclidean distance in the embedding space, computed with the projection solver of Appendix G.3), with ties broken by a fixed record order.

Figure H.3 plots the fixed-map split of PCHA's advantage over random exemplars by directed pair. PCHA outperforms the random-draw mean, PAM (partitioning around medoids; Kaufman & Rousseeuw, 1990), and $K = 1$ in all 60 cells and sits at the 95.7th percentile of the random-draw distribution on average (minimum 81.4th), but it is not the best source summary: Extreme Support beats PCHA in 46 of 60 cells. Most directly, in all 51 cells in which PCHA itself has negative R^2_{cen}, it still beats the random mean, PAM, and $K = 1$. A negative absolute score therefore does not imply that the source summary is uninformative relative to these source-only alternatives.

The accounting of Sections 2–3 also says why one summary beats another, and two exact splits of a method difference are available. The *fixed-map split* (Proposition 1) compares the reconstructions actually produced: $\Delta R^2_{\text{cen}} = \Delta R^2_{\text{shift}} - \Delta\hat{\rho}^{\,2}$. The *free-translation split* compares the model sets after re-optimization: $\Delta R^2_{\text{cen}} = \Delta R^2_{\text{ST}} - \Delta(R^2_{\text{ST}} - R^2_{\text{cen}})$. Here the second term is the change in the score gain from freeing location; we reserve *PCHA location family* for the PCHA-specific quantity $\lambda + \iota$. Table C.2 carries the columns for both. Under the fixed-map split, PCHA exceeds the random mean by $0.0821 = 0.0094$ in recentered fit $+$ 0.0727 from a smaller mean-offset penalty, so the mean-offset contribution is 88.5% of the grand-mean gap and the majority in all 12 directed pairs. Under the free-translation split the same gap is $0.0821 = 0.0127 + 0.0694$, and the free-translation location contribution is 84.6% of it. On CMV→Craigslist, random has slightly better fixed-map centered fit, by about 0.0046, yet PCHA still wins R^2_{cen} by about 0.099 because its mean-offset penalty is lower by about 0.104. Extreme Support shows the opposite pattern relative to PCHA. Its edge of 0.0168 is $+0.0174$ in recentered fit and -0.0006 from a slightly larger mean-offset penalty; under the free-translation split it is $0.0124 + 0.0044$, about 74% translated fit. The direction of each conclusion is the same under both splits: the same headline fidelity difference can arise for different geometric reasons.

Sign fragility is also not specific to PCHA in this benchmark. Negative cells number 51 of 60 for PCHA, 50 for Extreme Support, 58 for random exemplars, 59 for PAM, and 60 for $K = 1$; the sufficient certificate $\rho^2_H > R^2_{\text{ST}}$ is met in 43, 42, 54, 54 and 60 of those cells respectively. The theorem has zero violations across all 64,800 source-summary evaluations, 60,000 of which are individual random-exemplar hulls. The $K = 1$ summary is the source mean, which is the one-archetype solution of Cutler & Breiman (1994, Proposition 1). The $K = 1$ control is the equality case of Theorem 1: $R^2_{\text{ST}} = 0$ exactly, so its score is $-\rho^2_H$. We offer no mechanism for why one summary is better placed than another. Among the three methods that select literal source points, mean hull radius (in source-σ units) orders Extreme Support (1.166) $>$ random (0.980) $>$ PAM (0.899), and mean ρ^2_H improves in the same order. PCHA breaks that one-dimensional relation: its radius is only 0.807 while its mean ρ^2_H is 0.317, slightly better than Extreme Support's 0.321. Scalar extent moves with centroid reach within the selection methods, but not across method families.

Table C.2: Alternative source summaries (full columns) on the four-corpus, 538-move benchmark. Entries are means over the 60 directed-pair-by-configuration cells. $R^2_{\text{cen}} = R^2_{\text{shift}} - \hat{\rho}^{\,2}$ is the fixed-map split. "Negative / certificate" gives the number of cells with negative R^2_{cen} and, of those, the number satisfying $\rho^2_H > R^2_{\text{ST}}$; the negative counts are identical under split means and split medians, and the Extreme Support certificate count is 41 rather than 42 under split medians. "Gain over $K = 1$" is R^2_{cen} minus the $K = 1$ value; under the free-translation split it equals the R^2_{ST} column plus the free-translation location part shown, which is the reduction in $R^2_{\text{ST}} - R^2_{\text{cen}}$ relative to $K = 1$. Entries are rounded independently; all identities close before rounding.

Method	R^2_{cen}	R^2_{shift}	$\hat{\rho}^{\,2}$	R^2_{ST}	ρ^2_H	Negative / certificate	Gain over $K = 1$	of which free-translation location
PCHA	−0.1798	0.1549	0.3347	0.1877	0.3171	51 / 43	0.3628	0.1752
Extreme Support	−0.1630	0.1722	0.3353	0.2001	0.3207	50 / 42	0.3796	0.1795
Random exemplars	−0.2619	0.1454	0.4073	0.1750	0.3927	58 / 54	0.2808	0.1057
PAM medoids	−0.2895	0.1410	0.4305	0.1716	0.4159	59 / 54	0.2531	0.0815
Source mean ($K = 1$)	−0.5427	0	0.5427	0	0.5427	60 / 60	0	0

Table C.3 gives the fixed-map split of PCHA's advantage over the random-draw mean for every directed pair, and Table C.4 the hull-extent comparison.

Table C.3: **PCHA minus random-draw mean, fixed-map split, by directed pair.** Means over the five configurations (split means). $\Delta R^2_{\text{cen}} = \Delta R^2_{\text{shift}} + (-\Delta\hat{\rho}^{\,2})$; the last column is the mean-offset contribution as a share of ΔR^2_{cen}, which exceeds 100% for CMV→Craigslist because random exemplars have slightly better recentered fit there. Entries are rounded independently from six-decimal values.

Directed pair	ΔR^2_{cen}	$\Delta R^2_{\text{shift}}$	Mean-offset contribution	Share
CaSiNo → CMV	+0.0779	+0.0068	+0.0711	91.28%
CaSiNo → Craigslist	+0.0722	+0.0271	+0.0451	62.49%
CaSiNo → PFG	+0.0845	+0.0145	+0.0701	82.89%
CMV → CaSiNo	+0.0690	+0.0007	+0.0683	98.97%
CMV → Craigslist	+0.0991	−0.0046	+0.1037	104.60%
CMV → PFG	+0.0756	+0.0044	+0.0713	94.25%
Craigslist → CaSiNo	+0.0616	+0.0267	+0.0349	56.61%
Craigslist → CMV	+0.1224	+0.0012	+0.1212	99.00%
Craigslist → PFG	+0.0669	+0.0114	+0.0555	82.96%
PFG → CaSiNo	+0.0658	+0.0127	+0.0530	80.62%
PFG → CMV	+0.0806	+0.0048	+0.0758	94.06%
PFG → Craigslist	+0.1092	+0.0073	+0.1019	93.31%
Grand mean	+0.0821	+0.0094	+0.0727	88.52%

Table C.4: **Hull extent and centroid reach by source summary.** Mean and root-mean-square hull radius in units of the source σ; mean ρ^2_H over the 60 cells.

Method	Mean radius	RMS radius	Mean ρ^2_H
Extreme Support	1.1662	1.1675	0.3207
Random exemplars	0.9797	0.9853	0.3927
PAM medoids	0.8987	0.9068	0.4159
PCHA	0.8070	0.8140	0.3171
Source mean ($K = 1$)	0	0	0.5427

D Further analyses for Section 7

D.1 Setup, identity check and fixed-output margin

We use five text distributions: FineWeb, OpenWebMath, Spanish text, StarCoder (Python), and UltraChat. Documents are packed into a continuous token stream and partitioned into alternating 1,024-token sequence blocks. Each block is a BOS token followed by 1,023 tokens of the stream. Position 0 and all BOS, EOS and padding tokens are excluded before any statistic is computed; no other tokens are filtered (no high-norm or attention-sink filter), so the first token of a document that begins inside a block is kept. Reference and evaluation blocks are token-disjoint but not guaranteed to be document-disjoint: a document crossing a block boundary can contribute tokens to both halves. Reference blocks define distribution means, shift directions, and directional-variance controls; all reconstruction scores are computed on evaluation blocks.

We accumulate R^2_{shift} directly from per-token centered residuals and compute $\hat{\rho}^{\,2}$ independently from the reconstruction and target means. Proposition 1 is therefore checked between separately measured quantities rather than used to define either one. The maximum absolute identity error is 8.06×10^{-15}.

Across the 15 layer-by-distribution conditions, R^2_{cen} ranges from 0.741 to 0.829 and R^2_{shift} from 0.742 to 0.830, while $\hat{\rho}^{\,2}$ never exceeds 0.001576. Origin-referenced scores span 0.830–0.880 on the same conditions.

Proposition 1 also gives a useful fixed-output distance to zero. Because

$$R^2_{\text{cen}} = \left(\sqrt{R^2_{\text{shift}}} - \hat{\rho}\right)\left(\sqrt{R^2_{\text{shift}}} + \hat{\rho}\right),$$

the score reaches zero when $\hat{\rho} = \sqrt{R^2_{\text{shift}}}$ (for $R^2_{\text{shift}} \geq 0$, as in every natural condition). If the centered residual component is held fixed, then $\sqrt{R^2_{\text{shift}}} - \hat{\rho}$ is the additional reconstruction-mean offset, in units of σ, required to reach the sign boundary in the worst-case parallel direction. The smallest such margin across the 15 natural conditions is 0.837σ.

Table D.1: **Gemma Scope SAEs on natural data, by layer (full columns).** Ranges are over the five distributions. "Fixed-output margin" is $\sqrt{R^2_{\text{shift}}} - \hat{\rho}$. "Input-shift crossing" is the first estimated value of the imposed FineWeb shift at which $R^2_{\text{cen}} = 0$ along one of the four forward between-distribution directions (Section 7). "Largest real separation" is the largest measured FineWeb-to-other-distribution centroid separation at that layer. The final column is the drop in R^2_{cen} after removing b_{dec} with codes fixed (Section 7).

Layer	R^2_{unc}	R^2_{cen}	R^2_{shift}	max $\hat{\rho}^2$	min fixed-output margin (σ)	Input-shift crossing (σ)	Largest real separation (σ)	Drop without b_{dec}
5	0.830–0.880	0.764–0.829	0.765–0.830	0.0009	0.852	0.676–0.996	0.343	0.327–0.347
12	0.830–0.867	0.741–0.772	0.742–0.774	0.0016	0.837	0.627–1.513	0.378	0.505–0.588
20	0.855–0.874	0.767–0.790	0.768–0.791	0.0015	0.850	0.540–1.244	0.496	0.561–0.712

Table D.2 lists the 15 natural conditions.

Table D.2: **Gemma Scope SAEs on natural data, all 15 conditions.** R^2_{shift} is computed directly from per-token centered residuals, and $R^2_{\text{cen}} = R^2_{\text{shift}} - \hat{\rho}^2$ closes before rounding.

Layer	Distribution	R^2_{unc}	R^2_{cen}	$\hat{\rho}^2$	R^2_{shift}
5	FineWeb	0.8660	0.8153	0.000763	0.8161
5	OpenWebMath	0.8597	0.8055	0.000355	0.8059
5	Spanish	0.8300	0.7644	0.000523	0.7649
5	StarCoder (Python)	0.8776	0.8250	0.000261	0.8253
5	UltraChat	0.8804	0.8292	0.000871	0.8301
12	FineWeb	0.8499	0.7664	0.000893	0.7673
12	OpenWebMath	0.8374	0.7566	0.000374	0.7569
12	Spanish	0.8304	0.7414	0.000581	0.7420
12	StarCoder (Python)	0.8472	0.7673	0.000314	0.7676
12	UltraChat	0.8668	0.7720	0.001576	0.7736
20	FineWeb	0.8659	0.7863	0.000742	0.7870
20	OpenWebMath	0.8551	0.7690	0.000396	0.7694
20	Spanish	0.8584	0.7698	0.000556	0.7703
20	StarCoder (Python)	0.8591	0.7674	0.000332	0.7677
20	UltraChat	0.8742	0.7898	0.001529	0.7913

For an activation x, the released SAEs compute

$$\hat{x} = f(x)W_{\text{dec}} + b_{\text{dec}}, \qquad f(x) = \text{JumpReLU}(xW_{\text{enc}} + b_{\text{enc}}).$$

D.2 Input-shift sweeps: crossings, their decomposition, and matched real-distribution translations

In the input-shift sweeps of Section 7 the encoder sees the shifted input, so the codes and R^2_{shift} can change. Proposition 1 nevertheless continues to account exactly for each resulting fixed reconstruction, to within 1.1×10^{-15} at every grid point.

Across the 12 forward layer-by-direction sweeps, the first zero crossings of R^2_{cen}, located by linear interpolation between the two grid points that bracket the sign change, range from 0.54σ to 1.51σ. The room is not uniform: at layer 20 the earliest crossing is 0.540σ, while the largest measured FineWeb separation is 0.496σ; in the UltraChat direction the crossing is much farther away, at 1.513σ for layer 12 and 1.244σ for layer 20.

Because the codes change, both terms of the identity move, and the identity says how much each contributes. At $\beta = 0$, R^2_{cen} is 0.77–0.82 and the mean-offset penalty is below 0.001; at the crossing the two terms are equal. The split of that fall between lost recentered fit and gained mean-offset

penalty depends on the direction and repeats across layers. Toward StarCoder, lost fit accounts for 74%–80% of the fall. Toward Spanish, gained mean-offset penalty accounts for 64%–70%. Toward OpenWebMath the two are comparable (offset 37%–51%), and toward UltraChat the offset share is 76%–83% at layers 12 and 20 but 39% at layer 5. Values at the crossing are linearly interpolated between the two bracketing grid points. Just past some crossings recentered fit is itself negative (toward StarCoder at layers 12 and 20, toward OpenWebMath at layer 20), so the SAE's reconstructions then lose to the target mean even after recentering. An imposed shift therefore degrades these SAEs by moving the reconstruction mean and by losing centered structure, in direction-dependent proportions.

The comparison to real distributions is not only a comparison of scalar distances. At 18 points translated by the reference-mean displacement of a real target distribution, the translated evaluation centroid lies within 0.009–0.017σ of the target evaluation centroid. The resulting R^2_{cen} differs from the score on the corresponding real distribution by a median of -0.016, with range -0.162 to $+0.003$. These translations therefore reproduce the target location closely while still allowing non-location distributional differences to remain.

Table D.3 gives each sweep's crossing and the split of the fall in R^2_{cen} at the crossing.

Table D.3: **Input-shift sweeps: first zero crossing and the split of the fall at the crossing.** Crossings are in units of σ, located by linear interpolation between the two bracketing grid points (11 values of β per forward sweep). At $\beta = 0$, R^2_{cen} / R^2_{shift} / $\hat{\rho}^2$ are 0.8153 / 0.8161 / 0.00076 at layer 5, 0.7664 / 0.7673 / 0.00089 at layer 12 and 0.7863 / 0.7870 / 0.00074 at layer 20. "Offset share" is the gained mean-offset penalty as a share of the fall from the $\beta = 0$ score to zero; "lost-fit share" is the lost recentered fit, and the two sum to one. Shares are given to four decimals. The six reverse sweeps (StarCoder or UltraChat activations shifted toward FineWeb) cross at 0.874 and 1.054 (layer 5), 1.006 and 1.084 (layer 12), and 0.906 and 1.012 (layer 20), respectively.

Layer	Forward sweep	Crossing (σ)	Offset share	Lost-fit share
5	FineWeb → OpenWebMath	0.763	0.4581	0.5419
5	FineWeb → Spanish	0.676	0.7015	0.2985
5	FineWeb → StarCoder	0.689	0.2000	0.8000
5	FineWeb → UltraChat	0.996	0.3852	0.6148
12	FineWeb → OpenWebMath	0.776	0.5131	0.4869
12	FineWeb → Spanish	0.627	0.6774	0.3226
12	FineWeb → StarCoder	0.740	0.2646	0.7354
12	FineWeb → UltraChat	1.513	0.7553	0.2447
20	FineWeb → OpenWebMath	0.617	0.3739	0.6261
20	FineWeb → Spanish	0.903	0.6387	0.3613
20	FineWeb → StarCoder	0.540	0.2245	0.7755
20	FineWeb → UltraChat	1.244	0.8348	0.1652

The fixed-output margin holds recentered fit fixed, whereas an input shift changes the activations presented to the SAE, so the codes and R^2_{shift} can respond; the two distances need not agree.

We therefore also translate FineWeb evaluation activations toward the reference mean of each other distribution, and StarCoder and UltraChat activations back toward FineWeb, and run the shifted activations through the SAE.

The first zero crossings lie at 0.54σ–1.51σ forward (Figure H.7) and 0.87σ–1.08σ in reverse, while the largest measured separations from FineWeb are 0.34σ–0.50σ; no sweep is negative at or below its layer's largest measured separation, although at layer 20 the earliest crossing, 0.54σ, is close to that layer's 0.50σ.

D.3 Directional tracking of measured shift directions

Shift magnitude alone does not determine the SAE response. The measured between-distribution directions carry much more activation variance than generic random directions, so a raw natural-versus-random comparison would confound direction with local spread. We therefore use control directions matched to each measured direction's variance in the reference data and constrained outside the span of the measured directions.

We measure the response by projected mean tracking,

$$a_{\parallel}(\beta) = \frac{u^{\top}(\hat{\mu}_{\beta} - \hat{\mu}_0)}{\beta\sigma},$$

the component along u of the change in the mean reconstruction, as a fraction of the imposed shift, where $\hat{\mu}_{\beta}$ is the mean reconstruction of the shifted evaluation activations and u is estimated on the reference half. A value of one means the reconstruction mean follows the shift fully; the value is not clipped. At imposed shifts of 0.25σ and 0.5σ, the median tracking over the four measured directions is 0.38–0.86, against 0.13–0.23 for their variance-matched controls; within each of the six layer-by-shift cells the ratio of the two medians is 2.77–5.17. Thus the response is direction-specific beyond what local directional variance alone explains. This is an empirical property of these reconstruction maps along these directions; it is not evidence of semantic understanding or adaptation to the named datasets.

This response is input-dependent: the shifted activation is re-encoded, so the latent codes are allowed to change. Section 7 isolates a different source of location correction while holding those codes fixed. Figure H.2a plots the six layer-by-shift cells.

Table D.4 gives the six layer-by-shift cells.

Table D.4: **Projected mean tracking $a_{\parallel}$: measured directions versus variance-matched controls.** Medians over the four measured between-distribution directions and over their four controls.

Layer	Imposed shift	Median, measured directions	Median, controls	Ratio
5	0.25σ	0.3751	0.1263	2.970
5	0.5σ	0.6002	0.1738	3.452
12	0.25σ	0.4727	0.1673	2.826
12	0.5σ	0.6386	0.2304	2.772
20	0.25σ	0.5359	0.1388	3.860
20	0.5σ	0.8643	0.1673	5.166

D.4 Decoder-bias removal: what the intervention does and does not show

Figure H.2b plots the intervention for all 15 conditions. Direct recomputation confirms that R^2_{shift} is invariant to numerical precision, while mean-referenced fidelity changes sharply.

Because the codes are fixed, removing b_{dec} moves the reconstruction mean by $-b_{\text{dec}}$, so the drop in R^2_{cen} equals $\|\mu - \hat{\mu} + b_{\text{dec}}\|^2/\sigma^2 - \hat{\rho}^2$, approximately $\|b_{\text{dec}}\|^2/\sigma^2$ when $\hat{\rho}$ is small: it measures the size of the released bias relative to the target's spread.

In the released parameterization of Section 7, b_{dec} appears only at the output. That is a property of the parameterization and not of training: Gemma Scope trains with b_{dec} subtracted before the encoder, as in Bricken et al. (2023), and folds that term into the encoder bias for release (Lieberum et al., 2024, Section 3.2 and Appendix B). The training-time centering by b_{dec} therefore still shapes the codes: the folding transformation carries it into the released encoder bias.

With the learned codes held fixed, the released decoder bias therefore supplies substantial location correction: deleting it moves the reconstruction mean far from the target mean and lowers mean-referenced fidelity accordingly. The experiment does not show what the codes would be if the bias-related centering were removed and the SAE rerun, and it does not isolate the role the parameter played in training. Because the codes are held fixed, it is a pure change of output location, and the same fixed-code intervention is available for any SAE with an additive output bias. It also isolates a contribution distinct from Appendix D.3: direction-specific tracking is a response of the full encoder–decoder map to shifted inputs, whereas the decoder bias provides an additive location correction even when the codes are held fixed.

Table D.5 gives the intervention for every condition.

Table D.5: **Removing b_{dec} with the learned codes held fixed.** R^2_{shift} is unchanged in every condition (direct recomputation). "$\hat{\rho}$ without" is the reconstruction-mean offset in target-σ units after removal.

Layer	Distribution	R^2_{cen}	R^2_{cen} without b_{dec}	Drop	$\hat{\rho}$ without b_{dec}
5	FineWeb	0.8153	0.4782	0.3371	0.5812
5	OpenWebMath	0.8055	0.4766	0.3289	0.5738
5	Spanish	0.7644	0.4371	0.3273	0.5726
5	StarCoder (Python)	0.8250	0.4924	0.3326	0.5770
5	UltraChat	0.8292	0.4823	0.3470	0.5898
12	FineWeb	0.7664	0.2310	0.5354	0.7323
12	OpenWebMath	0.7566	0.2515	0.5051	0.7110
12	Spanish	0.7414	0.1532	0.5882	0.7673
12	StarCoder (Python)	0.7673	0.2470	0.5203	0.7215
12	UltraChat	0.7720	0.2566	0.5154	0.7190
20	FineWeb	0.7863	0.2098	0.5764	0.7597
20	OpenWebMath	0.7690	0.1515	0.6175	0.7861
20	Spanish	0.7698	0.0577	0.7121	0.8442
20	StarCoder (Python)	0.7674	0.0849	0.6826	0.8264
20	UltraChat	0.7898	0.2285	0.5612	0.7502

For Spanish at layer 20, for example, R^2_{cen} falls from 0.770 to 0.058 while recentered fit is unchanged.

D.5 Reference choice across SAE sparsity variants

For a single fixed target, Section 2 shows that origin- and mean-referenced scores cannot reorder reconstructions: both are monotone functions of the same residual. Averaging across targets is not constrained in the same way because the reference transformation differs by target.

We therefore evaluate five SAE sparsity variants at each of the three layers, averaging each variant over all five distributions. The origin- and mean-referenced suite orderings are identical at every layer, with no near-ties. What changes is the numerical spread. At layer 12, for example, the five variants span 0.787–0.925 under the origin reference and 0.668–0.883 under the target-mean reference.

In this tested suite, reference choice changes how far apart the SAE variants appear, not which variant ranks above which.

Table D.6 gives the full suite.

E Further analyses for Section 5

E.1 Levels, references and the running example

The transfer-versus-self gap is much larger under the target-mean reference than under the origin reference: at the pair-mean level, the largest gap is about 0.18 against the origin but 1.04 against the target mean.

The centered residual is below the target variance in every one of the 2,400 primary cross-corpus split-seed evaluations, as Lemma 1 guarantees. Its ratio to target variance ranges from 0.72 to 0.95, with median 0.81. Equivalently, after correcting only the reconstruction mean, the same fixed weights capture about 5%–28% of target-centered variance in every cross-corpus evaluation.

The recurring CMV→Craigslist example makes the accounting concrete. Its pair-mean normalized residual is $1.684 = 0.898$ from the centered residual $+ 0.786$ from the reconstruction–target mean offset, giving $R^2_{\mathrm{cen}} = 1 - 1.684 = -0.684$. Correcting only the offset gives $R^2_{\mathrm{shift}} = 1 - 0.898 =$

Table D.6: **Gemma Scope sparsity variants: suite averages over the five distributions.** Within each layer the variants are listed in the same order under both references; there are no rank reversals and no near-ties. The rows average_l0_68, average_l0_82 and average_l0_71 are the SAEs of Section 7 and Tables D.1–D.5.

Layer	Variant	Suite R^2_{unc}	Suite R^2_{cen}	Suite mean L0
5	average_l0_309	0.9215	0.8902	317.4
5	average_l0_143	0.8907	0.8470	145.5
5	average_l0_68	0.8627	0.8079	70.0
5	average_l0_34	0.8358	0.7703	35.4
5	average_l0_18	0.8116	0.7363	19.5
12	average_l0_445	0.9250	0.8833	447.6
12	average_l0_176	0.8802	0.8135	178.7
12	average_l0_82	0.8463	0.7607	83.3
12	average_l0_41	0.8167	0.7146	43.0
12	average_l0_22	0.7870	0.6683	22.8
20	average_l0_294	0.9131	0.8586	283.3
20	average_l0_139	0.8872	0.8165	133.1
20	average_l0_71	0.8626	0.7764	72.8
20	average_l0_38	0.8371	0.7350	38.7
20	average_l0_22	0.8110	0.6926	21.9

0.102, and the fixed-map identity closes directly: $0.102 - 0.786 = -0.684$. A negative mean-referenced score therefore does not imply that the fixed reconstruction captures no target-centered structure. But recentering is not the whole explanation: the median centered residual is 0.81 under cross-corpus transfer versus 0.69 for held-out self.

Table E.1: Cross-corpus reconstruction summary (full columns). All columns are eight-configuration grand means of split-means. R^2_{shift} and $\hat{\rho}^{\,2}$ use the same fixed reconstruction as R^2_{cen}, so $R^2_{\mathrm{cen}} = R^2_{\mathrm{shift}} - \hat{\rho}^{\,2}$; the displayed components are independently rounded to four decimals, while the identity closes before rounding. Recovery is the mean of the per-configuration ratios $R^2_{\mathrm{ST,cross}}/R^2_{\mathrm{ST,self}}$ on the identical target, not the ratio of the pair-mean scores. The last three rows give the held-out-self dictionaries on each target, with the same aggregation.

Source $\rightarrow$ **target**	R^2_{unc}	R^2_{cen}	R^2_{shift}	$\hat{\rho}^{\,2}$	R^2_{ST}	**Recovery**	ρ_H
CaSiNo $\rightarrow$ Craigslist	0.8306	+0.0395	0.2389	0.1995	0.2650	0.7321	0.4227
Craigslist $\rightarrow$ CaSiNo	0.7907	+0.0014	0.2222	0.2208	0.2377	0.6868	0.4477
CaSiNo $\rightarrow$ CMV	0.7267	−0.2230	0.1278	0.3508	0.1609	0.6531	0.5761
CMV $\rightarrow$ CaSiNo	0.7317	−0.2749	0.1326	0.4076	0.1725	0.4987	0.6196
Craigslist $\rightarrow$ CMV	0.6683	−0.4854	0.0922	0.5776	0.1365	0.5544	0.7467
CMV $\rightarrow$ Craigslist	0.7026	−0.6840	0.1021	0.7861	0.1615	0.4475	0.8744
CaSiNo held-out self	0.8604	+0.3389	0.3449	0.0060	0.3459	—	0.0638
Craigslist held-out self	0.8858	+0.3561	0.3612	0.0051	0.3618	—	0.0585
CMV held-out self	0.8309	+0.2419	0.2479	0.0060	0.2485	—	0.0668

Tables E.2 and E.3 give the 48 three-corpus cells individually, and Table E.4 the 24 held-out-self cells.

Table E.2: **Three-corpus cell values of R^2_{cen}.** Each entry is the median over the ten splits of the five-seed average (Section 4), at the 785-move budget. Eleven cells are positive: CaSiNo→Craigslist in every configuration except Gemini embedding-2, and Craigslist→CaSiNo in four.

Configuration	CaSiNo → Craigslist	CaSiNo → CMV	Craigslist → CaSiNo	Craigslist → CMV	CMV → CaSiNo	CMV → Craigslist
all-MiniLM-L6-v2 (384)	+0.0534	−0.1021	+0.0532	−0.3238	−0.1294	−0.5385
all-mpnet-base-v2 (768)	+0.0924	−0.1269	+0.0947	−0.3357	−0.2957	−0.7883
e5-large-v2 (1024)	+0.0279	−0.2162	+0.0001	−0.4489	−0.2056	−0.5364
Gemini embedding-2 (512)	−0.0392	−0.3624	−0.1067	−0.6593	−0.3908	−0.9008
text-embedding-005 (256)	+0.0654	−0.2362	+0.0009	−0.5787	−0.3088	−0.7215
text-embedding-005 (512)	+0.0457	−0.2535	−0.0048	−0.5516	−0.2955	−0.7111
text-embedding-005 (768)	+0.0371	−0.2742	−0.0144	−0.5186	−0.3144	−0.6887
text-multilingual-embedding-002 (768)	+0.0481	−0.2071	−0.0120	−0.4174	−0.2664	−0.6489

Table E.3: **Three-corpus cell values of the centroid-to-hull distance ρ_H.** Same aggregation as Table E.2. Every cell with $\rho_H \geq 0.475$ is negative in Table E.2.

Configuration	CaSiNo → Craigslist	CaSiNo → CMV	Craigslist → CaSiNo	Craigslist → CMV	CMV → CaSiNo	CMV → Craigslist
all-MiniLM-L6-v2 (384)	0.3879	0.4755	0.3789	0.6410	0.5050	0.7999
all-mpnet-base-v2 (768)	0.3998	0.5098	0.3803	0.6587	0.6445	0.9456
e5-large-v2 (1024)	0.4004	0.5624	0.4202	0.7233	0.5539	0.7823
Gemini embedding-2 (512)	0.5058	0.6624	0.5573	0.8473	0.6907	0.9829
text-embedding-005 (256)	0.4231	0.6059	0.4693	0.8141	0.6584	0.9014
text-embedding-005 (512)	0.4240	0.6074	0.4629	0.7897	0.6342	0.8831
text-embedding-005 (768)	0.4317	0.6161	0.4652	0.7637	0.6464	0.8712
text-multilingual-embedding-002 (768)	0.4028	0.5738	0.4501	0.7098	0.6182	0.8590

Table E.4: **Held-out-self cells per configuration: R^2_{cen} / R^2_{ST} / centroid mass m.** Same aggregation as Table E.2 (median over the ten splits of the five-seed average). The centroid mass $m = N\|\mu\|^2/\|X\|^2_F$ of the target evaluation matrix does not depend on the seed but varies across splits (by 0.001–0.007 within a cell); the entry is its split median. It ranges from 0.61 to 0.90 over the 24 cells and from 0.71 to 0.90 for Craigslist targets (Section 2).

Configuration	CaSiNo	Craigslist	CMV
all-MiniLM-L6-v2 (384)	0.3233 / 0.3306 / 0.6681	0.3218 / 0.3275 / 0.7094	0.2534 / 0.2609 / 0.6734
all-mpnet-base-v2 (768)	0.3712 / 0.3807 / 0.7113	0.3792 / 0.3882 / 0.7810	0.3409 / 0.3476 / 0.6077
e5-large-v2 (1024)	0.2862 / 0.2920 / 0.8787	0.2998 / 0.3046 / 0.8979	0.1997 / 0.2052 / 0.8854
Gemini embedding-2 (512)	0.3401 / 0.3478 / 0.8593	0.3782 / 0.3836 / 0.8845	0.1998 / 0.2053 / 0.8527
text-embedding-005 (256)	0.3650 / 0.3726 / 0.7883	0.3912 / 0.3964 / 0.8215	0.2581 / 0.2650 / 0.7804
text-embedding-005 (512)	0.3487 / 0.3550 / 0.7791	0.3751 / 0.3798 / 0.8128	0.2308 / 0.2374 / 0.7735
text-embedding-005 (768)	0.3402 / 0.3467 / 0.7852	0.3685 / 0.3729 / 0.8131	0.2206 / 0.2266 / 0.7788
text-multilingual-embedding-002 (768)	0.3317 / 0.3386 / 0.8317	0.3380 / 0.3431 / 0.8596	0.2347 / 0.2415 / 0.8285

E.2 Recovery by direction, and what limits it

Recovery is directional: pair-median recovery is 0.67 for CaSiNo→CMV versus 0.50 for CMV→CaSiNo, and 0.58 for Craigslist→CMV versus 0.46 in the reverse direction.

The translated-fit deficit can be read through the shape family, $1 - R^2_{\text{ST}} = s + c$, written as $R^2_{\text{ST}} = (1 - s)\big(1 - c/(1 - s)\big)$: the share of target-centered variation inside the dictionary's affine span, times the share of that in-span variation the simplex constraint retains. Over the 48 cell medians, subspace mismatch s has median 0.547 for held-out self and 0.728 under transfer, so the span holds about 45% and 27% of target-centered variation, a ratio of about 0.60. The conditional convexity cost $c/(1 - s)$ has median 0.330 for held-out self and 0.392 under transfer, so the retained in-span share falls by a factor of about 0.91. Their product, about 0.55, is close to the median recovery of 0.59, and the span factor is the larger of the two. Because these are medians of separate quantities, the factorization is an approximate comparison of sizes, not a per-cell decomposition. It is also consistent with the convexity paragraph of Appendix E.3: the raw convexity term is smaller under transfer, and convexity nonetheless costs more relative to what remains in span.

For the running example, CMV→Craigslist, the pair-mean translated fit is about 0.16, against about 0.36 for a Craigslist dictionary fitted on disjoint Craigslist data.

E.3 LOCATION-VERSUS-SHAPE RATIOS, CONVEXITY, CERTIFICATE SLACK AND NUMERICAL CHECKS

Absolute values of the four terms mix transfer difficulty with model capacity. The location family exceeds the shape family in all 48 cells, by a factor of 1.53–7.48 (median 2.74). Location is the largest family of the excess, and the shape family is a substantial second part, 12%–40% of it; the subspace-mismatch term alone exceeds that total because the convexity excess is negative.

One component needs care in isolation. The raw convexity excess is negative in all 48 cells (median about -0.05), but this does not mean the simplex constraint becomes easier under transfer. Subspace mismatch rises from a median of about 0.55 for held-out self to 0.73 under transfer, leaving less in-span variation for convexity to lose. Normalizing by that available variation changes the reading: $c/(1-s)$ rises from a median of 0.33 for held-out self to 0.39 under transfer and is higher in 46 of 48 cells. Convexity therefore typically costs more relative to what remains in-span, even though the raw term is smaller. Appendix C.1 reports the same conditional quantity for the fourth corpus rather than interpreting Δc by itself.

The four negative cells where the certificate does not hold are all negotiation-pair cases near zero, with median scores between about -0.04 and -0.005; there ρ_H^2 falls just short of R_{ST}^2 and the theorem slack is about 0.03–0.05. Across the 2,400 primary cross-corpus split-seed evaluations, the slack has median 0.050 and maximum 0.132.

Numerically, the bound has no violations in any evaluated regime: all 5,850 three-corpus split-seed evaluations (3,600 at the matched budget and 2,250 at natural corpus size), all 10,439 controlled translated-hull states, and all 14,400 bootstrap resamples satisfy it, as do the 19,200 states of the full-grid sweep in Appendix B.1. In the primary size-matched evaluations, slack has median 0.037 and maximum 0.132 when cross and held-out-self cells are pooled; the larger overall maximum of 0.202 occurs only in the controlled sweeps.

Under the split-consistency rule of Appendix G.4, 42 cells are location-dominated and six are mixed at the 785-move budget; none is shape-dominated.

Equivalently, the shape family is $1/(1+7.48) = 11.8\%$ to $1/(1+1.53) = 39.5\%$ of the excess, the share Section 5 reports as 12%–40%.

E.4 GEOMETRY OF THE OFFSET

Here r/D ranges from 0.03 to 0.12 across embedding configurations, whereas the observed in-span share φ, the fraction of the squared centroid-to-hull distance lying along directions the dictionary spans (Appendix B.2), has median 0.24 and interquartile range 0.17–0.30 over the 2,400 primary cross-corpus evaluations. The row-wise enrichment $\varphi/(r/D)$ has median 4.42 and range 1.06–10.85.

That alignment matters inside the location family. Across all 48 cells, the out-of-span excess $\Delta\lambda$ is the larger location component, but the in-span interaction excess $\Delta\iota$ is still 30%–83% as large, with median 48%. Descriptively, ι tracks the normalized in-span centroid-distance term Nd_S^2/TSS closely (slope 1.18, intercept 0.035, Pearson correlation 0.989 over the 2,400 evaluations). Combined with the exact identity of Appendix B.2, this says that the in-span part of centroid displacement is not a negligible remainder: it is geometrically coupled to the extra simplex cost of keeping a dispersed target cloud in a fixed hull.

Location excess is larger than shape excess in every primary cell; within location, the out-of-span component is larger than the in-span interaction in every cell, but the latter remains substantial. We consequently use location as the stable family-level conclusion without reducing the observed offset to a purely orthogonal shift.

E.5 MATCHING THE SELF BASELINE

The matched fitting budget in Section 4 prevents a subtler baseline distortion. In the CaSiNo-sourced diagnostic comparisons, replacing the 785-move held-out-self dictionary with a same-corpus dictionary fitted on roughly 2,600 moves increases the apparent subspace-mismatch excess by 15.0%–

17.1% and the apparent translated-fit gap by 19.0%–19.3%. A stronger self model therefore makes transfer look worse even though the transferred dictionary is unchanged.

Across the 20 cross cells available in both matched- and natural-size analyses, however, the absolute source-size effect is small relative to the cross-versus-self excess: the ratio has median 2.15% and maximum 4.04%. Matching capacity is therefore important for unbiased effect sizes, but the location-versus-shape pattern is much larger than the dictionary-size perturbation.

E.6 Structure-free floors and a random-source reference

By Lemma 1, recentered fit is positive for every nearest-point reconstruction from a convex hull, and translated fit is non-negative for any hull (Section 3.1), so only their sizes can show what a transferred dictionary carries. We therefore compare each transferred PCHA dictionary with three hulls built for it and scored on the same target evaluation data. Two are structure-free floors: the dictionary's archetypes rotated at random about their mean, which keeps every distance between archetypes, and an isotropic hull of 32 vertices with independent standard normal coordinates in the embedding space, centered exactly on the archetypes' mean and rescaled so that their total variance about it equals that of the archetypes (isotropic in distribution; the variance is matched in total, not per direction). The third, a hull of $K = 32$ distinct points drawn uniformly without replacement from the source fitting partition, keeps source content but no learned archetypes, and serves as a random-source reference; duplicate 32-point subsets are rejected within a dictionary. Held-out self (Section 4) is the same-corpus reference.

The comparison covers the primary study at the 785-move budget: 1,200 dictionaries (three source corpora, eight configurations, ten splits, five seeds), each refit from its stored configuration and applied to both other corpora, which gives the 2,400 cross-corpus evaluations of Section 5. Each comparison hull is drawn five times for each source dictionary; the same draws serve both target corpora of that dictionary, and scores are averaged over the five draws within each evaluation. Comparisons are paired: within each split and seed, the comparison hulls are compared with the same refit dictionary on the same target; the per-split difference is the average over seeds, and a cell's difference is the median over the ten splits (Section 4). A cell is split-consistent when the per-split difference is positive in at least 9 of the 10 splits. Pair-level values are grand means, as in Table 1.

Refitting reproduced every stored dictionary (1,200 of 1,200 archetype hashes) and every stored score (largest change in R^2_{cen}: zero over 2,400 evaluations). Across all 40,800 evaluations, covering every method and every draw, recentered fit met the lower bound of Lemma 1, with smallest slack 0.0052. The fixed-map identity of Proposition 1 held to within 1.8×10^{-15}; rotations preserved the distances between archetypes to within 3.3×10^{-16} in all 12,000 checks (6,000 rotations, each used in both evaluations of its dictionary); and isotropic hulls matched the archetypes' vertex variance to within a relative 5.1×10^{-15} in all 12,000 checks of 6,000 draws.

Tables E.5–E.8 give the results. In recentered fit, PCHA exceeds the rotated floor in all 48 cells, split-consistently, by a median paired difference of 0.096, and the isotropic floor by 0.092. Hulls of random source points also exceed the rotated floor in every cell. PCHA is ahead of them in 33 of 48 cells (27 split-consistent), by a median of 0.005. In translated fit, PCHA exceeds the rotated and isotropic floors by 0.130 and 0.127, and hulls of random source points by 0.011, ahead of them in 47 of 48 cells (44 split-consistent). On a scale from the rotated floor (0) to held-out self (1), PCHA sits at a median of 0.36 in recentered fit and 0.51 in translated fit; on a scale from hulls of random source points to held-out self, at 0.02 and 0.07.

PCHA's mean-referenced score is higher than that of hulls of random source points in all 48 cells, and in all ten splits of each; the median cell difference is 0.090, with range 0.054–0.176 (cell medians, defined as in Table E.8). Split exactly with cell means over splits, 88% of the difference comes from a smaller mean-offset penalty and 12% from recentered fit (88.4% from the penalty with cell medians, defined as in Table E.8). The share is not uniform (Table E.9; Figure H.8): it exceeds 90% in the four CMV-involving pairs and is 54%–63% between the two negotiation corpora, where learned archetypes also add recentered fit.

Per configuration, the rotated floor's recentered fit is 0.50–0.69 times r/D, where $r = 31$ is the dimension of the hull's affine span and r/D is the share of variance that a random subspace of that dimension captures in expectation (Appendix E.4). The floor falls with embedding dimension, as

Table E.5: **Pair-level recentered fit R^2_{shift} of PCHA and its comparison hulls.** Grand means over seeds, splits and the eight configurations at the 785-move budget, as in Table 1; comparison hulls are first averaged over their five draws. The last column is the lower bound of Lemma 1, $\|C\hat{X}\|_F^2/\text{TSS}$, for the PCHA reconstruction.

Source → target	PCHA	Rotated	Isotropic	Random source points	Held-out self	Lemma 1 bound (PCHA)
CaSiNo → Craigslist	0.2389	0.0380	0.0428	0.2061	0.3612	0.1493
Craigslist → CaSiNo	0.2222	0.0357	0.0405	0.1914	0.3449	0.1284
CaSiNo → CMV	0.1278	0.0347	0.0390	0.1202	0.2479	0.0739
CMV → CaSiNo	0.1326	0.0354	0.0382	0.1287	0.3449	0.0754
Craigslist → CMV	0.0922	0.0310	0.0350	0.0931	0.2479	0.0548
CMV → Craigslist	0.1021	0.0336	0.0361	0.1058	0.3612	0.0633

Table E.6: **Pair-level translated fit R^2_{ST} of PCHA and its comparison hulls.** Same aggregation as Table E.5.

Source → target	PCHA	Rotated	Isotropic	Random source points	Held-out self
CaSiNo → Craigslist	0.2650	0.0416	0.0469	0.2336	0.3618
Craigslist → CaSiNo	0.2377	0.0387	0.0441	0.2087	0.3459
CaSiNo → CMV	0.1609	0.0398	0.0450	0.1528	0.2485
CMV → CaSiNo	0.1725	0.0413	0.0447	0.1612	0.3459
Craigslist → CMV	0.1365	0.0383	0.0437	0.1321	0.2485
CMV → Craigslist	0.1615	0.0428	0.0462	0.1536	0.3618

Table E.7: **Position of PCHA between a comparison hull F (0) and held-out self (1).** For each cell, the position is $(\text{PCHA} - F)/(\text{held-out self} - F)$ on the cell values of Section 4; the table gives its median and its 10th and 90th percentiles over the 48 cells. Every denominator is positive, so no cell is excluded.

Metric	F	Median	10th percentile	90th percentile
R^2_{shift}	Rotated	0.3596	0.2204	0.6237
R^2_{shift}	Isotropic	0.3528	0.2105	0.6199
R^2_{shift}	Random source points	0.0247	−0.0191	0.2168
R^2_{ST}	Rotated	0.5098	0.3708	0.6952
R^2_{ST}	Isotropic	0.4999	0.3634	0.6873
R^2_{ST}	Random source points	0.0685	0.0276	0.2446

Table E.8: **Paired differences between PCHA and each comparison hull over the 48 cells.** A cell counts as ahead when its paired difference (median over splits) is positive, and as split-consistent when the per-split difference is positive in at least 9 of 10 splits. The median paired difference is the statistic quoted in the text; the median difference of cell values is given for comparison. The last row of each block compares hulls of random source points with the rotated floor.

Metric	Comparison	Cells ahead	Split-consistent	Median paired difference	Median difference of cell values
R^2_{shift}	PCHA − rotated	48/48	48/48	0.0959	0.0954
R^2_{shift}	PCHA − isotropic	48/48	48/48	0.0921	0.0921
R^2_{shift}	PCHA − random source points	33/48	27/48	0.0051	0.0055
R^2_{shift}	Random source points − rotated	48/48	48/48	0.0923	0.0925
R^2_{ST}	PCHA − rotated	48/48	48/48	0.1301	0.1302
R^2_{ST}	PCHA − isotropic	48/48	48/48	0.1271	0.1270
R^2_{ST}	PCHA − random source points	47/48	44/48	0.0108	0.0106
R^2_{ST}	Random source points − rotated	48/48	48/48	0.1175	0.1179

r/D does, but stays below it: the ratio is 0.50 at $D = 256$, 0.57–0.59 at $D = 384$–512 and 0.64–0.69 at $D = 768$–1024. A convex hull captures less than its full span.

Table E.9: **Where PCHA's advantage in mean-referenced score over hulls of random source points comes from, by directed pair.** Pair-level grand means as in Table E.5; "Random" is the hull of 32 random source points. The shares split the difference in R^2_{cen} exactly, using cell means over splits (Figure H.8); a share above 100% means that PCHA's recentered fit is slightly below that of the random hulls. The last row pools all 48 cells.

Source → target	PCHA R^2_{cen}	PCHA $\hat{\rho}^2$	Random R^2_{cen}	Random $\hat{\rho}^2$	Share from smaller penalty	Share from recentered fit
CaSiNo → Craigslist	+0.0395	0.1995	−0.0483	0.2545	63%	37%
Craigslist → CaSiNo	+0.0014	0.2208	−0.0651	0.2565	54%	46%
CaSiNo → CMV	−0.2230	0.3508	−0.3090	0.4291	91%	9%
CMV → CaSiNo	−0.2749	0.4076	−0.3585	0.4872	95%	5%
Craigslist → CMV	−0.4854	0.5776	−0.6281	0.7212	101%	−1%
CMV → Craigslist	−0.6840	0.7861	−0.8155	0.9213	103%	−3%
All 48 cells					88%	12%

F Further analyses for Section 6

F.1 The accounting on source-native text

On source-native text, all 60 cross-corpus cells are negative and all 3,000 split-seed cross evaluations are numerically verified negative. The fixed reconstructions nevertheless retain target-centered fit: after recentering the same fixed maps, the centered residual remains below target variance in all 3,000 evaluations, as Lemma 1 guarantees, with maximum residual ratio 0.971.

The rest of the accounting behaves as it did on blueprints. Theorem 1 has zero violations in the 3,000 raw-text cross evaluations; its slack has median 0.031, with 5th and 95th percentiles 0.020 and 0.050. The in-span interaction ι again tracks Nd_S^2/TSS closely (slope 1.09, intercept 0.026, Pearson correlation 0.990). This is the same measurement distinction as in Sections 5–6: a negative mean-referenced score can coexist with non-zero target-centered fit in the fixed reconstruction.

F.2 Pair-level comparison

Figure H.5 plots recovery by pair on both representations. Table F.1 compares the two representations at the pair level. Every entry uses the same aggregation ladder: mean over five seeds, mean over ten splits, then mean over the five full-dimensional configurations. The fixed-map identity therefore closes exactly at this aggregation level.

Table F.1: **Blueprint / source-native pair means at the 538-move budget (all quantities).** Each cell is blueprint / raw text. $R^2_{\text{cen}} = R^2_{\text{shift}} - \hat{\rho}^2$ within each representation; entries are rounded independently to four decimals, and the identity closes before rounding. Recovery is the mean of the per-configuration ratios $R^2_{\text{ST,cross}}/R^2_{\text{ST,self}}$ on the identical target, not the ratio of the pair means.

Source → target	R^2_{cen}	R^2_{shift}	$\hat{\rho}^2$	R^2_{ST}	**Recovery**	ρ_H
CaSiNo → Craigslist	+0.0414 / −0.4634	0.2290 / 0.0726	0.1876 / 0.5361	0.2517 / 0.1001	0.7384 / 0.3644	0.4112 / 0.7225
Craigslist → CaSiNo	+0.0174 / −0.8339	0.2147 / 0.0961	0.1972 / 0.9300	0.2290 / 0.1231	0.6921 / 0.2487	0.4235 / 0.9486
CaSiNo → PFG	−0.0151 / −0.5802	0.1746 / 0.0879	0.1897 / 0.6681	0.2004 / 0.1183	0.5935 / 0.3164	0.4135 / 0.8082
PFG → CaSiNo	−0.0551 / −0.7214	0.1796 / 0.0991	0.2347 / 0.8205	0.2066 / 0.1336	0.6246 / 0.2681	0.4598 / 0.8934
Craigslist → PFG	−0.1166 / −0.3780	0.1598 / 0.0992	0.2764 / 0.4773	0.1834 / 0.1170	0.5434 / 0.3137	0.5081 / 0.6817
PFG → Craigslist	−0.0784 / −0.1247	0.1688 / 0.0877	0.2472 / 0.2125	0.2121 / 0.1155	0.6227 / 0.4202	0.4740 / 0.4494
CaSiNo → CMV	−0.1883 / −0.3995	0.1296 / 0.0463	0.3179 / 0.4457	0.1610 / 0.0679	0.6558 / 0.4029	0.5478 / 0.6620
CMV → CaSiNo	−0.2529 / −0.9463	0.1312 / 0.0706	0.3841 / 1.0169	0.1690 / 0.0967	0.5113 / 0.1924	0.6008 / 0.9976
Craigslist → CMV	−0.4327 / −0.3359	0.0959 / 0.0435	0.5286 / 0.3794	0.1370 / 0.0609	0.5591 / 0.3629	0.7156 / 0.6099
CMV → Craigslist	−0.6498 / −0.4329	0.1037 / 0.0517	0.7536 / 0.4846	0.1581 / 0.0700	0.4647 / 0.2530	0.8561 / 0.6911
PFG → CMV	−0.1986 / −0.1677	0.1316 / 0.0569	0.3302 / 0.2246	0.1674 / 0.0801	0.6707 / 0.4738	0.5546 / 0.4622
CMV → PFG	−0.2289 / −0.4755	0.1398 / 0.0735	0.3688 / 0.5490	0.1765 / 0.0948	0.5195 / 0.2526	0.5886 / 0.7338

In three of the four directions whose mean-offset penalty is smaller on raw text (Craigslist→CMV, CMV→Craigslist and PFG→CMV), the smaller penalty outweighs the lost recentered fit and the

score rises (Table F.1). PFG→Craigslist shows the converse: its mean-offset penalty is also smaller on raw text, but its R^2_{shift} drops farther, making the total score more negative.

Raw-text recovery spans 0.16–0.51 with median 0.31. This is not simply a stronger held-out-self denominator under blueprints: raw held-out-self R^2_{ST} is higher in 10 of 20 self cells and lower in 10, with median raw-minus-blueprint difference -0.02. The representation effect is therefore concentrated in cross-corpus translated structure rather than a uniform change in within-corpus compressibility.

The centroid-to-hull distance ρ_H is larger on raw text in 42 of 60 paired cells, with median raw-minus-blueprint difference $+0.18$, but at the directed-pair level $\hat{\rho}^2$ and ρ_H are smaller on raw text in four of the 12 directions. The plain normalized source–target centroid separation is likewise larger in 42 of 60 cells, with median difference $+0.13$. The PCHA location-family *level*, $R^2_{\text{ST}} - R^2_{\text{cen}}$, makes the same point continuously: it spans approximately 0.24–1.04 on raw text versus 0.21–0.81 on blueprints, larger in the other eight directions and smaller in the same four. These are levels for the cross-corpus reconstructions, not cross-versus-self location shares.

The raw representation also places less of the between-corpus displacement inside the source span: the in-span share φ has median 0.17. Taken together, the paired results rule out a one-dimensional story in which blueprinting simply improves centroid alignment. The uniform change is lower cross-corpus translated fit on raw text; location is a second, direction-dependent change that can either amplify or partially offset the score difference.

The nine cells that were positive on blueprints move from cell medians between $+0.02$ and $+0.09$ on blueprints to between -0.34 and -1.20 on raw text, so these are not small movements across zero. The nine are the five CaSiNo→Craigslist cells, two Craigslist→CaSiNo cells, and the MiniLM cells for CaSiNo→PFG and PFG→CaSiNo. Craigslist→CaSiNo changes from $+0.017$ to -0.834 (Table F.1).

F.3 Cross-corpus centroid geometry

The centroid geometry rules out the simpler story that blueprints merely shrink the same corpus offsets. For each of the six unordered corpus pairs, we take the vector between the two corpus centroids, computed on the whole corpora, under each representation. The cosine between the raw-text and blueprint vectors ranges from 0.18 to 0.42. The directions retain partial positive alignment, but they are far from collinear. At the same time, the unnormalized raw-text centroid distances range from 0.48 to 0.62, compared with 0.25 to 0.46 under blueprints, and are larger in every pair.

The blueprint pipeline therefore **compresses and rotates cross-corpus geometry**. It does not simply rescale a fixed domain direction, and it does not replace the geometry with an unrelated one. This is the representation-level counterpart to the score decomposition above: the pipeline changes both where corpora sit and how much centered structure transfers between them.

F.4 Controlled translation on raw text

We repeat the held-out-self translation of Appendix B.1 on raw text, translating a PFG raw-text dictionary rigidly along the offset direction toward each of the other three corpora. One dictionary is fitted per embedding configuration (first split, one seed), and the values below are means over the five configurations. Along all three directions, R^2_{ST} stays at 0.378, unchanged to within solver tolerance. At zero displacement, $R^2_{\text{cen}} = +0.372$; at $\beta = 1$, the three scores are -0.551, -0.491, and -0.490 toward CaSiNo, Craigslist, and CMV respectively. The bound of Theorem 1 holds at every point of the sweep.

Because translated fit is held fixed, this intervention isolates the sign change to the hull's position. The same location effect demonstrated in Section 3 therefore remains available after the representation change, even though the natural raw-text geometry differs substantially from the blueprint geometry.

G FURTHER DETAILS FOR SECTION 4

Section 3 showed that location can force a negative mean-referenced score; three avoidable artifacts could mimic that effect empirically: a baseline fitted on more data, related examples on both sides of a split, and, for scores close to zero, a sign inherited from optimizer error.

G.1 CORPORA, BLUEPRINTS AND EMBEDDING CONFIGURATIONS

The primary study uses three corpora with different interaction structures and corpus-specific event definitions. CaSiNo (Chawla et al., 2021) is a multi-issue campsite negotiation corpus. Each retained record is the counterpart utterance immediately preceding a qualifying concession event: the relaxed concession magnitude must satisfy max(gross_conceded, net_change) >= 3 discrete points, and opening proposals are excluded as anchors. This yields 1,569 records from 728 dialogues. Craigslist Bargains (He et al., 2018) is a bilateral price-bargaining corpus. A retained move must be a directional price concession (seller price moves down or buyer price moves up) whose absolute change is at least 1% of the listing price; opening proposals with no prior price are excluded, and final agreement is not required. This yields 5,127 records from 2,997 dialogues. ChangeMyView (CMV; Tan et al., 2016), accessed through ConvoKit (Chang et al., 2020), is public argumentation with a thread-level human-awarded view-change outcome. CMV records lie on winning rebuttal paths in threads where the original poster explicitly awarded a delta, yielding 5,265 extraction units across 1,936 original posts. Thus the record-selection criterion is a local concession event in the two negotiation corpora but a thread-level awarded outcome in CMV.

In the blueprint analyses, each selected record is represented by a precomputed LLM-generated procedural blueprint. Blueprints were generated with the repository model identifier models/gemini-3.7-flash at temperature 0 with greedy decoding, a 25-40 word target, structured JSON output, and standardized relational placeholders such as PARTY_A, PARTY_B and target-rank tokens. The prompts deliberately expose surrounding conversational context and response evidence: CaSiNo includes prior dialogue and the responding concession; Craigslist includes prior dialogue and the counterparty response; CMV includes the root post/stance, claim graph and OP response; PersuasionForGood (PFG, introduced in Section 6) includes the preceding dialogue and subsequent reply. Automated leakage audits found no corpus-specific item tokens in CaSiNo, no prices or dollar amounts in Craigslist, no topic-entity or delta tokens in CMV, and no charity names or dollar amounts in PFG. Blueprints were generated once and held fixed across all analyses. Because the generator saw context and outcome-relevant response evidence, Section 6 treats source-native text versus blueprint as a representation-pipeline contrast, not as a causal estimate of abstraction alone.

The blueprint PCHA studies use eight embedding configurations from six models. Two are Google models from the Gecko line, text-embedding-005 (768 dimensions) and text-multilingual-embedding-002 (768) (Lee et al., 2024; Google Cloud, 2026), and text-embedding-005 is also used at its supported reduced sizes of 512 and 256. The others are Gemini embedding-2 at 512 dimensions, a truncation of its 3,072-dimensional output renormalized to unit length (Google, 2026), the sentence-transformers models all-mpnet-base-v2 (768) and all-MiniLM-L6-v2 (384) (Reimers & Gurevych, 2019; Song et al., 2020; Wang et al., 2020), and e5-large-v2 (1024) (Wang et al., 2022). All text embeddings are L2-normalized before reconstruction, and every analysis uses the same version of each embedding model. The two reduced sizes of text-embedding-005 are additional configurations of the same model, not independent model replications. The source-native representation contrast in Section 6 uses five full-dimensional configurations (text-embedding-005, MPNet, E5, multilingual-002 and MiniLM), all but the three truncated ones, on the same records and splits.

Table G.1 summarizes the three primary corpora.

G.2 BUDGET AND CAPACITY SENSITIVITY

The corpora differ substantially in size, so unrestricted 50/50 splits would give the source dictionaries very different amounts of training data. Every source fitting partition is capped at the size of the CaSiNo fitting half in the same split, whatever the target: 785 moves in six of the ten splits and 784 or 786 in the other four, the 785-move budget. Section 6 applies the same rule with the smaller PersuasionForGood fitting half, 538 moves in seven splits and 537 or 539 in the other three, the 538-move budget.

Table G.1: **Primary corpora.** Records are the selected moves; groups are the units kept whole by the splits (Section 4). The PersuasionForGood counts are given in Section 6.

Corpus	Records	Group unit	Groups	Group size (mean / median / max)
CaSiNo	1,569	dialogue	728	2.16 / 2 / 11
Craigslist Bargains	5,127	dialogue	2,997	1.71 / 1 / 9
ChangeMyView	5,265	original post	1,936	2.72 / 2 / 41

Re-evaluating the original corpus pairs at 538 rather than 785 moves changes them little: the median cell R^2_{cen} moves from -0.2600 to -0.2645, median recovery from 0.59 to 0.60, and 39 rather than 37 of the 48 cells have a negative median. Section 6 reports its own eight-configuration aggregates at the 538-move budget so that all four corpora are compared under one capacity setting.

We also retain a secondary natural-size sensitivity analysis for the three-corpus study, in which Craigslist and CMV dictionaries use their full fitting halves for the five full-dimensional configurations. It is used only to quantify the effect of sample-size mismatch; the main cross-versus-self claims use the primary size-matched analysis. Appendix E.5 reports that sensitivity explicitly.

Capacity. All other PCHA results in the paper use $K = 32$. Refitting every dictionary of the primary study at $K = 8$, 16 and 64, with the same records, splits, seeds, fitting budget and solver settings, gives 48, 45, 37 and 33 negative cross-corpus cells of the 48 at $K = 8$, 16, 32 and 64. All cells that change sign lie between the two negotiation corpora, and no cell that is negative at a larger K is non-negative at a smaller one. At $K = 64$, all 32 CMV-involving cells and one of the 16 cells between the negotiation corpora are negative; the four cells that turn non-negative between $K = 32$ and 64 are the four negative cells at $K = 32$ whose sign the certificate does not force. Median held-out-self translated fit is 0.230, 0.289, 0.341 and 0.388; recovery has median 0.504, 0.549, 0.588 and 0.639 and range 0.20–0.70, 0.25–0.73, 0.32–0.77 and 0.38–0.79; and the location family is the larger part of the excess in all 48 cells at each of the four K, with median share 0.795, 0.757, 0.729 and 0.712. No evaluation violates Theorem 1.

G.3 Solver tolerances and independent checks

We call a value or sign numerically verified when the optimizer supplies a bound tight enough to establish it. All three-corpus evaluations satisfy the numerical verification criteria. Appendices C.1, C.3, E.3 and F.1 check the implemented bound of Theorem 1 numerically in every evaluated regime.

For fixed-position reconstruction, each target vector is solved under the simplex constraint in the reduced archetype coordinates. A solution is accepted only when the gap between the residual achieved and a valid lower bound on the optimum is at most $10^{-10}\max(1, \|x\|^2)$. A mean-referenced score is called numerically verified negative only when the lower bound on the fixed-position residual itself exceeds TSS. A slightly negative point estimate is therefore not enough for a sign claim.

For the free-translation problem, we use monotone accelerated projected gradient and require a Frank-Wolfe optimality gap no larger than $10^{-7}\,\text{TSS}$. Reduced-coordinate solutions were independently checked in the original embedding space using full-dimensional objective evaluation, a separately implemented gradient and Frank-Wolfe gap, alternative initialization, and convex subproblem solves. Agreement is assessed through compatible objective bounds rather than equality of two primal iterates.

G.4 Aggregation rules and family labels

We keep optimization precision, dictionary initialization, data splitting, and finite target-sample variation separate. Within a source-target-configuration cell, the primary cell value averages the five dictionary seeds inside each split and then takes the median over the ten splits. Cross-minus-self quantities use the same paired route on the identical target matrix. Pair-level grand means used in summary tables instead average over seeds, splits, and embedding configurations so that the algebraic identities close exactly at the reported level. In the source-summary benchmark of Section

6, cell values are instead means over the ten splits (after averaging dictionary seeds or random draws within each split), so that differences between methods decompose exactly; its negative-cell counts are identical under split means and split medians. Recovery is formed for each evaluation before aggregation; shares and other ratios of excess terms are formed from the aggregated numerator and denominator, at both the cell and the pair level. We label these aggregation levels whenever a number is quoted. In the three-corpus study, the group bootstrap resamples the evaluation groups of the first split and seed of each cell, with the dictionary fixed (200 replicates), so it measures evaluation-sample variation only.

The ten group splits are a robustness distribution over one finite corpus, not ten independent population samples. We therefore report their variation but do not turn the ten splits into a confidence interval.

For cross-versus-self family labels, let τ_M denote the standard deviation across the ten split-level held-out-self values of metric M; it is a robustness scale, not a standard error. A family is called dominant only if its paired excess is positive and at least two thirds of the total excess in at least 9 of 10 splits, and its median excess exceeds twice the corresponding robustness scale. If neither family clears that rule, the cell is mixed. If the total excess is too small or a family has a negative median, the cell is left unclassified. This is the rule behind the location-dominance counts reported later.

G.5 How the later studies change the setting

The later studies keep the measurement definitions fixed while changing one part of the empirical setting. Section 6 first adds PersuasionForGood and repeats the blueprint PCHA analysis across the same eight embedding configurations at the common 538-move budget. It then compares PCHA with source-only convex summaries on the five full-dimensional configurations. Section 6 keeps the four corpora, records, group splits, reconstruction model, and 538-move budget but replaces the blueprint representation with source-native move text in those same five full-dimensional configurations. Section 7 leaves convex-hull projection entirely and asks whether the fixed-map identity also describes Gemma Scope sparse autoencoders.

Five configurations appear only where a paired design requires them: the source-summary controls and the raw-text representation contrast both use the five full-dimensional configurations common to all studies, a subset of the eight used for the blueprint panels.

The paper contains six studies that share the same measurement questions but change the data, source summary, representation, or reconstruction regime.

G.6 Three comparison objects

Three comparison objects recur and should not be conflated. A **denominator reference** (origin versus target mean) changes how the same residual is scored. A **matched held-out-self baseline** changes the fitted reconstruction model while keeping the target evaluation matrix fixed, so it calibrates transfer against same-corpus capacity. A **free-translation diagnostic** relaxes location using target information; for PCHA it also re-optimizes the row weights. The first is score normalization, the second is a matched model comparison, and the third is a diagnostic rather than zero-shot performance.

H Additional figures

I Related work and reporting practice in full

I.1 Classical decompositions of skill scores

The residual side also has a classical beginning, in forecast verification. Murphy (1988, Eq. 12) decomposes the mean-squared-error skill score of a scalar forecast against the sample mean into a squared correlation, a conditional-bias term and a squared mean bias normalized by the observed variance. Murphy & Epstein (1989, Eq. 13) apply the decomposition to spatial fields scored against climatology; a further term, the squared mean anomaly over the anomaly variance, converts a score

Table G.2: The six studies: data, models and purpose.

Study	Data / representation	Evaluation design	Scientific role
Three-corpus transfer	CaSiNo, Craigslist Bargains, ChangeMyView; LLM-generated blueprints; 8 embedding configurations from 6 models	$K = 32$ PCHA ($K = 8$, 16, 64: Appendix G.2); 785-move source budget; 10 group splits; 5 dictionary seeds	Establish the reference/ location/translated-fit accounting
Structure-free floors	Same 3 blueprint corpora; 8 embedding configurations	Each of the 1,200 transferred $K = 32$ PCHA dictionaries against its archetypes rotated at random about their mean, an isotropic hull of matched variance and a hull of 32 random source points, each drawn five times; 785-move source budget; same splits, seeds and target evaluation data	Test whether the fit that survives transfer exceeds structure-free floors and a random-source reference (Appendix E.6)
Fourth-corpus extension	Adds PersuasionForGood; LLM-generated blueprints; 8 embedding configurations from 6 models	$K = 32$ PCHA; 538-move source budget; 10 group splits; 5 dictionary seeds	Measure the same accounting on a fourth corpus
Source-summary controls	Same 4 blueprint corpora; 5 full-dimensional embedding configurations	PCHA $K = 32$; Extreme Support $K = 32$; 100 random $K = 32$ exemplar draws; PAM $K = 32$; $K = 1$ source mean; same 538-move source records and 10 group splits	Test whether sign fragility and method differences are specific to learned PCHA
Representation contrast	Same 4 corpora; source-native raw moves versus the blueprint pipeline; 5 full-dimensional configurations	$K = 32$ PCHA; 538-move source budget; identical records and group splits	Test whether the accounting survives a representation change
Sparse autoencoders	Five text distributions represented by Gemma 2 2B residual activations	Gemma Scope JumpReLU SAEs at layers 5, 12, 20; token-disjoint alternating sequence blocks	Test the fixed-map accounting in a different reconstruction regime

against the field mean into one against climatology. In our notation that term is $m/(1-m)$, so their formula combines the reference identity of Section 2 with a split of the residual into mean bias and centered error. Taylor (2001, Eq. 3) separates the mean-square difference between two fields into an overall bias and a centered pattern difference, and Jolliff et al. (2009, Eqs. 5 and 12) normalize that split and plot it as a target diagram whose unit circle is the locus where the score against the reference mean is zero. In hydrology, Gupta et al. (2009, Eqs. 3–4) write Nash–Sutcliffe efficiency, a mean-referenced R^2, in terms of correlation, relative variability and normalized bias. Related decompositions of the mean squared error into bias, variance and correlation components or into systematic and unsystematic parts are used in econometric forecasting (Theil, 1966), model validation (Willmott, 1981) and agronomy (Kobayashi & Salam, 2000).

Those fields responded in two ways. Forecast verification built diagrams that keep bias, correlation, variability and centered error visible together; hydrology went further and proposed a replacement criterion, the Kling–Gupta efficiency, because optimizing the original one rewards underestimated

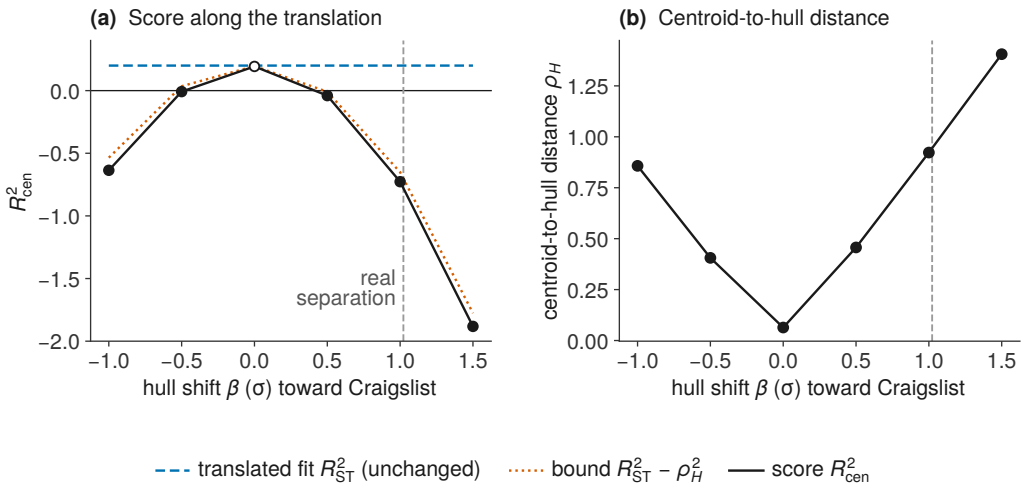

Figure H.1: **Breaking a good held-out-self score by moving only the hull.** One illustrative CMV held-out-self dictionary ($K = 32$; Gemini embedding-2, one held-out split, one seed) with every archetype translated by $\beta\sigma$ along the observed CMV–Craigslist centroid direction while the target stays fixed. (a) Mean-referenced score R^2_{cen} (filled negative, open non-negative), the unchanged translated fit R^2_{ST} and the bound $R^2_{\text{ST}} - \rho^2_H$ of Theorem 1. (b) Centroid-to-hull distance ρ_H. The dashed line marks the observed corpus-centroid separation, about 1.02σ in this cell's units. The score falls from about +0.19 at $\beta = 0$ to -0.04 at $\beta = 0.5$ and -0.73 at $\beta = 1$ as ρ_H grows from 0.06 to 0.46 and 0.92 (Appendix B.1).

variability (Gupta et al., 2009). We take the first route. Our aim is not a replacement for reconstruction fidelity but an accounting that keeps scientifically different quantities visible.

These decompositions concern forecasts that have already been produced. They do not optimize over a constrained set of feasible reconstructions, measure the distance from a target centroid to that set, or re-optimize after translation.

The difference between the two references is therefore classical, and so is the split of a squared error into mean bias and centered error. Section 3 states that split in the form we use and then turns to a question these decompositions do not address: when the geometry of a constrained reconstruction set forces the mean-referenced score below zero, and what would remain if location were freed.

Proposition 1 is the split of a squared error into mean bias and centered error from forecast verification and hydrology (Murphy, 1988, Eq. 12; Murphy & Epstein, 1989, Eq. 13; Taylor, 2001, Eq. 3; Jolliff et al., 2009, Eq. 12; Gupta et al., 2009, Eqs. 3–4).

In Figure 1b, because $R^2_{\text{cen}} = x^2 - y^2$, the diagonal is exactly $R^2_{\text{cen}} = 0$ and the grey curves are lines of equal R^2_{cen}, an analogue of the target diagram of Jolliff et al. (2009).

I.2 How fidelity under shift is reported

This section covers a small set of papers, read in full, and reports what each does. The reference differs between papers, sometimes under the same name. Gao et al. (2025), Rajamanoharan et al. (2024b) and Lieberum et al. (2024) divide the reconstruction error by the error of always predicting the mean, whereas He et al. (2024, Eq. 6) define "explained variance" as one minus the squared error over the squared norm of the input, a ratio against uncentered energy. For downstream loss, the field has already argued over the reference: the fraction of loss recovered was first measured against zero-ablation (Bricken et al., 2023; Rajamanoharan et al., 2024a) and criticized for it by Gao et al. (2025), because zero-ablating the residual stream is so damaging that poor reconstructions still score highly.

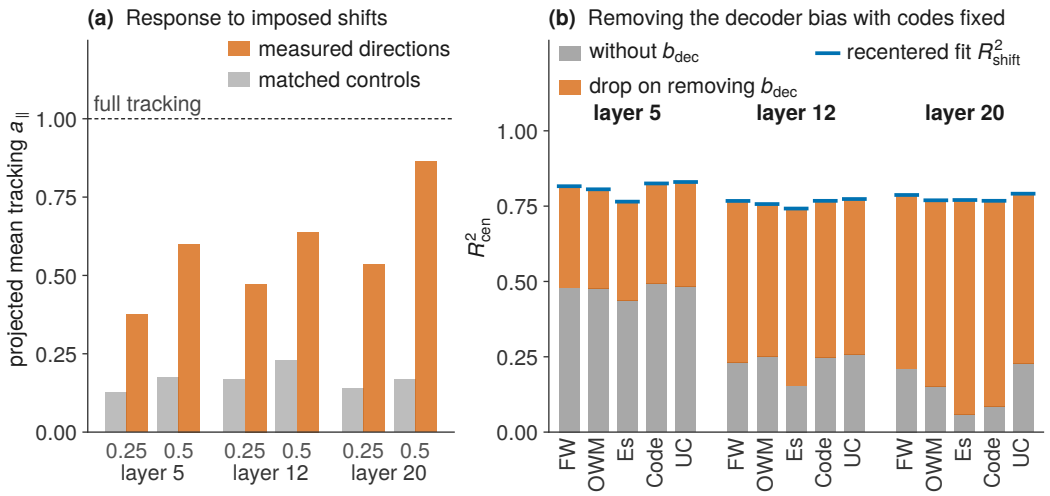

Figure H.2: **How the Gemma Scope SAEs handle location.** (a) Projected mean tracking $a_{\parallel}$, the component along the shift direction of the change in mean reconstruction as a fraction of the imposed shift, at $\beta = 0.25\sigma$ and 0.5σ: median over the four measured between-distribution directions (vermilion) and over their four variance-matched control directions (grey), by layer (Table D.4). The ratio of the two medians is 2.77–5.17. This is a response of the full encoder–decoder map to shifted inputs, with the codes free to change. (b) R^2_{cen} for each of the 15 layer-by-distribution conditions with the released decoder bias (full bar) and with b_{dec} removed from the output while the learned codes are held fixed (grey part); the vermilion part is the drop (Table D.5). Recentered fit R^2_{shift} (blue line, Table D.2) is unchanged by the removal. With the codes held fixed, the decoder bias supplies substantial location correction; the intervention is not a retrained SAE. FW = FineWeb, OWM = OpenWebMath, Es = Spanish, Code = StarCoder (Python), UC = UltraChat.

Fidelity under distribution shift is also reported, and this is where the location term matters. Lieberum et al. (2024) evaluate base-model SAEs on instruction-tuned activations and on subsets of The Pile; He et al. (2024) evaluate on longer contexts and on an instruction-tuned model, with squared error and downstream loss; and O'Neill et al. (2025) compare Gemma Scope SAEs with SAEs trained on medical text, on medical text. None of these separates the reconstruction-mean offset from centered error. In the last, fraction of variance explained is written as $1-\mathrm{Var}(x-\hat{x})/\mathrm{Var}(x)$ (O'Neill et al., 2025, Eq. 13); as written, the variance of the residual removes its mean, which makes the quantity our recentered fit R^2_{shift} and leaves the mean offset out altogether. In the dictionary-learning literature, transfer has been argued from a target reconstruction residual that falls as a source dictionary adapts, with no comparison to predicting the target mean (Ni et al., 2013, Eq. 2 and Fig. 6). None of these choices is wrong for the question each paper asks. Whatever the reference, unless the mean-offset penalty is reported, a reader of a fidelity number under shift cannot determine how much of it is location.

I.3 The reporting practices in full

Separating bias from centered error has long precedent in forecast verification (Section 2). We recommend carrying that discipline into reconstruction of learned representations under shift, adding the two distinctions that setting needs: between the reconstruction actually produced and a target-informed relaxation of the model set, and between transfer and a matched same-corpus model. We do not recommend one reference: the origin and the target mean answer different questions, and either can be the right one (Section 2). A fidelity result should instead let a reader answer all three questions of this paper.

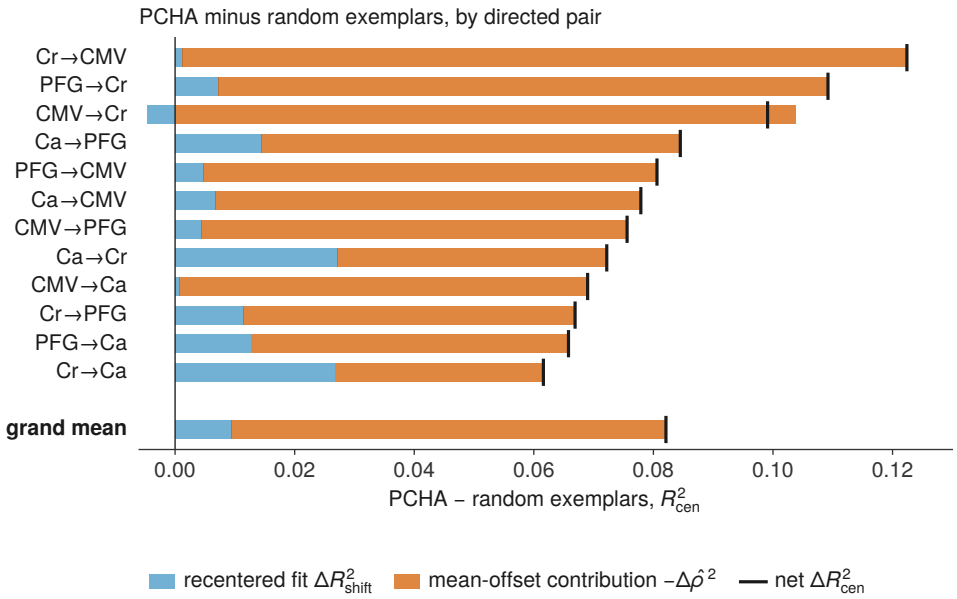

Figure H.3: **Why PCHA beats random exemplars.** Difference in R^2_{cen} between PCHA and the mean over 100 random 32-exemplar hulls on the four-corpus, 538-move benchmark, by directed pair (means over the five full-dimensional configurations of split means; Table C.3), split by Proposition 1 into the recentered-fit term $\Delta R^2_{\text{shift}}$ (blue) and the mean-offset contribution $-\Delta\hat{\rho}^2$ (vermilion); black tick: net difference. The mean-offset contribution is the larger part in all 12 pairs and 88.5% of the grand-mean difference of 0.082. For CMV→Craigslist the recentered-fit term is slightly negative.

State the reference predictor. Say in words what a score of zero means. The same CMV→Craigslist reconstructions score 0.70 against the origin and -0.68 against the target mean (Table 1), and both numbers are correct. If an origin-referenced score is scientifically meaningful, also report centroid mass m (or the corresponding R^2_{cen}), so that the amount of reference credit is recoverable rather than merely named.

Report the mean-offset penalty $\hat{\rho}^2$. It applies to any fixed reconstruction, needs only the two means and the target variance, and turns a single score into the exact split $R^2_{\text{cen}} = R^2_{\text{shift}} - \hat{\rho}^2$. It is informative both when large, 0.20–0.79 for the transferred dictionaries of Table 1, and when negligible, at most 0.0016 for the Gemma Scope SAEs of Section 7.

Report translation-corrected fit, and say which kind. R^2_{shift} describes the reconstruction that was actually produced. For model sets that can be re-optimized, R^2_{ST} may be added, labeled as a diagnostic that uses the target; it is not zero-shot performance. For convex model sets, ρ_H gives in addition the certificate $\rho^2_H > R^2_{\text{ST}}$, which says whether the hull's distance from the target mean forces a negative sign whatever the weights.

For transfer claims, compare with a matched held-out-self model on the identical target. Translated fit of 0.16 means little until it is set beside the 0.36 that a same-corpus dictionary of the same capacity achieves on the same held-out data (Section 5). The self model should match the transferred one in capacity and fitting budget and be fitted on disjoint groups: an unmatched baseline inflated the apparent loss in subspace match by 15%–17% and in translated fit by about 19% (Appendix E.5).

Do not read a negative score as an absence of target-centered fit, and do not overread signs near zero. In Section 5, the recentered fit that survived transfer exceeded structure-free floors but nearly matched hulls of random source points, and transferred dictionaries recovered 32%–77% of same-corpus translated fit. A sign within resampling variation should be reported as near the boundary, with optimization error bounded tightly enough that it cannot decide the sign. In our study, the

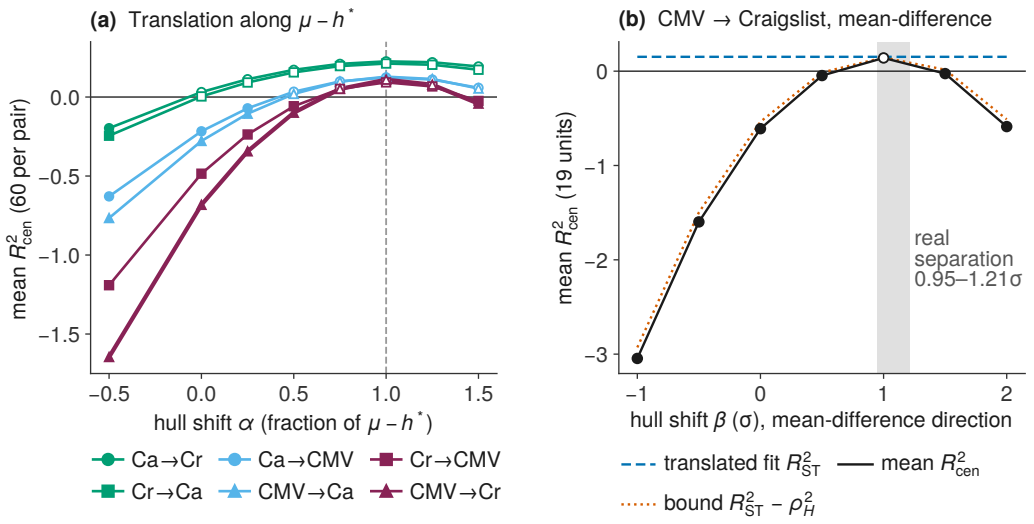

Figure H.4: **Controlled translations of the CMV and three-corpus dictionaries.** (a) Each fitted dictionary of each directed pair (60 per pair: six configurations × two splits × five seeds) translated rigidly by $\alpha(\mu - h^*)$; mean R^2_{cen} per pair (filled negative, open non-negative). At $\alpha = 1$ each target centroid lies on its hull and every pair's mean score is positive (Table B.2); the CMV→Craigslist curve and the six-pair mean reproduce Table B.1. (b) The 19 CMV→Craigslist dictionaries whose mean-difference direction is not nearly collinear with $\mu - h^*$, translated by $\beta\sigma$ along the unit vector from the CMV fitting-partition mean to the Craigslist evaluation mean. The translated fit $R^2_{\text{ST}} = 0.1521$ is unchanged; mean R^2_{cen} rises from -0.610 at $\beta = 0$ to $+0.140$ at $\beta = 1$ and falls to -0.588 at $\beta = 2$. The shaded band marks the natural separation, 0.95–1.21σ. The two sweeps step on different scales.

bootstrap standard deviation of R^2_{cen} had median 0.0067 and the widest 95% interval spanned 0.060 (Section 4). The count and the typical size of negative scores can also move in opposite directions (Section 6), so report both.

Decompose differences between methods, not only scores. The same identity applies to a difference: $\Delta R^2_{\text{cen}} = \Delta R^2_{\text{shift}} - \Delta \hat{\rho}^{\,2}$. In Section 6 (538-move budget), 88.5% of PCHA's advantage over random exemplars came from a smaller mean-offset penalty, whereas Extreme Support's advantage over PCHA came almost entirely from better recentered fit. A leaderboard of single scores would have presented both as the same kind of improvement.

Name the representation pipeline. A fidelity score under transfer describes records in a representation, not the records alone. Replacing LLM-generated blueprints with source-native text of the same records changed the sign of all nine cells that had been positive (Section 6, 538-move budget).

I.4 Related work in full

Negative scores and the choice of benchmark. The decompositions that precede our fixed-map identity are reviewed in Appendix I.1. A neighboring literature debates the benchmark itself. Schaefli & Gupta (2007) argue that the observed mean is often an uninformative benchmark and that Nash–Sutcliffe values cannot be interpreted without a suitable one, and in empirical finance out-of-sample R^2 against the historical mean is routinely negative for unrestricted predictive regressions (Campbell & Thompson, 2008; Welch & Goyal, 2008). Our question differs from theirs. We do not argue that the target mean is a poor benchmark or look for a more favorable one. We accept what a negative mean-referenced score says, that total squared error exceeds that of the mean predictor, and ask what that residual contains.

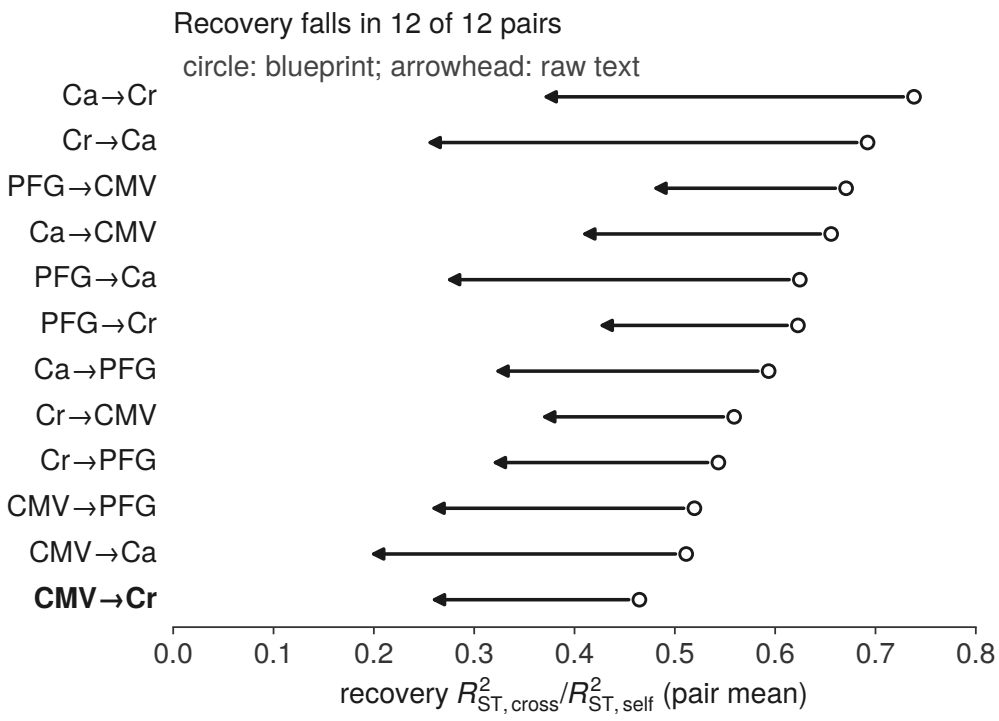

Figure H.5: **Translated-fit recovery on blueprints and on source-native text.** Recovery is the translated fit of the transferred dictionary divided by that of a same-corpus held-out-self dictionary on the identical target; values are means of the per-configuration ratios over the five full-dimensional configurations at the 538-move budget (Table F.1). Recovery is lower on raw text in all 12 directed pairs; at the cell level it is lower in all 60 paired cells, with raw-text median 0.31 and median paired difference -0.27 (Section 6).

Reconstruction fidelity in learned representations. Archetypal analysis reports a mean-referenced ratio, residual sum of squares relative to the one-archetype fit at the sample mean (Cutler & Breiman, 1994), and its scalable PCHA form reports variance explained against the total sum of squared entries (Mørup & Hansen, 2012). Archetypal analysis has also been run on unit-normalized data, as here (Mei et al., 2018). The large centroid mass of text embeddings is a known anisotropy (Ethayarajh, 2019; Gao et al., 2019; Li et al., 2020), which post-processing such as all-but-the-top or whitening reduces (Mu & Viswanath, 2018; Su et al., 2021). Sparse-autoencoder papers use both kinds of reference and have already contested the reference for downstream loss (Appendix I.2). Our concern is the step after that one: a mean-referenced score still combines mean offset with centered fit. Studies closest to our Section 7 examine SAE error and SAE behavior away from the training distribution: Engels et al. (2025) show that much of the SAE error vector is linearly predictable from the input activation, and Heindrich et al. (2025) find that SAE features transfer inconsistently across domains. Neither separates the reconstruction-mean offset from centered error. SAEs also learn a location vector, subtracted before encoding and added back after decoding (Bricken et al., 2023; Gao et al., 2025); Section 7 measures how much location correction the released Gemma Scope bias supplies. A free translation is the simplest of the alignments used in domain adaptation, which also match covariances or subspaces (Fernando et al., 2013; Sun et al., 2016); here it is a diagnostic, not an adaptation method. Transfer of SAEs between base and chat models has also been studied (Kissane et al., 2024).

What this paper adds. Proposition 1 applies this classical error accounting to learned reconstructions. The works reviewed in Appendix I.1 do not optimize over a constrained reconstruction set, measure centroid-to-set distance, or re-optimize after translation. Our results concern exactly these steps, which make model-set location and translation-corrected fit separately measurable. The cer-

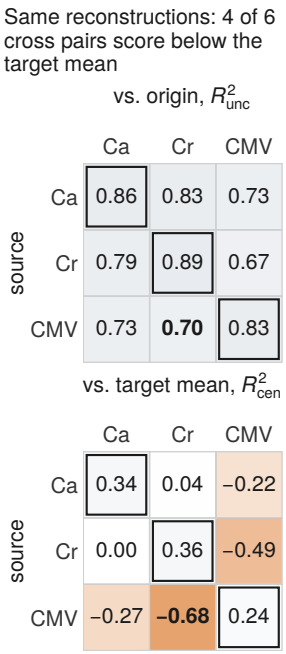

Figure H.6: **The same reconstructions under two reference predictors.** Fidelity of the nine source→target reconstructions of the three-corpus study (rows: source dictionary; columns: target corpus) against the origin (R^2_{unc}, top) and against the target mean (R^2_{cen}, bottom). Both matrices share one color scale: vermilion below zero and a faint slate above, with intensity proportional to magnitude, so cells near zero stay almost white. The boxed diagonal is the held-out-self dictionary, fitted on disjoint data from the target corpus and scored on the same target data. Every reconstruction scores 0.67–0.89 against the origin; against the target mean the four CMV-involving cross pairs are negative and the self cells stay positive. The bold entry is the running example, CMV→Craigslist. Values are eight-configuration grand means of split means at the 785-move budget (Table 1; held-out-self rows in Table E.1).

tificate of Theorem 1 depends only on the fitted model set and the target, not on the reconstruction actually produced, yet it forces the sign in 33 of the 37 negative cells of the three-corpus study (Section 3.1). Proposition 2 refines it for convex hulls into an exact split of each residual into location and shape families, with the location family at least ρ^2_H (Section 3.3 and Appendix A). Applied separately to transferred and held-out-self dictionaries on the same target, the difference between those splits apportions the transfer excess between location and shape (Section 5). And because a rigid translation leaves translated fit unchanged, the same quantities support a controlled intervention in which the hull's position alone reverses the sign (Section 3.2). Empirically, separating reference credit, mean offset and translation-corrected fit changes how results are read for transferred dictionaries, under a change of representation, between source summaries, and for sparse autoencoders. The two model families are not disjoint literatures: Fel et al. (2025) constrain SAE dictionary atoms to convex combinations of data points, in-distribution and with ordinary sparse codes. Our bridge between the families is one of measurement, not of architecture.

I.5 Limitations in full

The paper's exact results and its empirical results have different scope, and we discuss them separately.

Exact results. Proposition 1 holds for any fixed reconstruction, and Theorem 1 for any fixed reconstruction whose rows lie in a closed convex model set; neither depends on our data. The certificate is sufficient, not necessary, and the theorem can force a sign but does not apportion an observed excess; only the four-term split does that. The split holds for any convex-hull dictionary (Appendix A), but we analyze it empirically only for PCHA. R^2_{ST} uses the target to choose a translation and

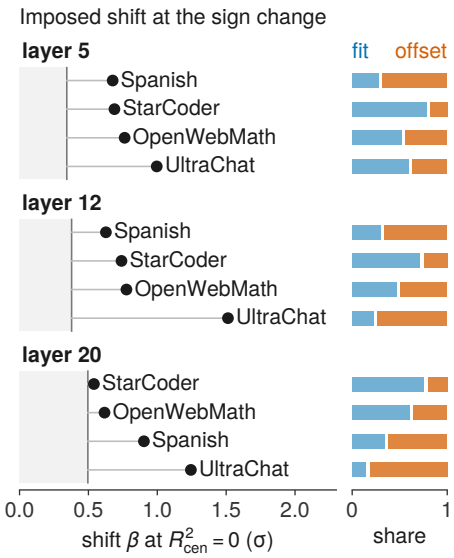

Figure H.7: **Imposed input shifts at the sign change, Gemma Scope SAEs.** Imposed FineWeb shift β at which R^2_{cen} first reaches zero along each of the four forward between-distribution directions, by layer (linear interpolation between grid points; Table D.3). Grey bands end at the largest measured FineWeb separation at that layer (Table D.1); the thin line joins it to the crossing. Right: share of the fall in R^2_{cen} at the crossing from lost recentered fit (blue) and from gained mean-offset penalty (vermilion).

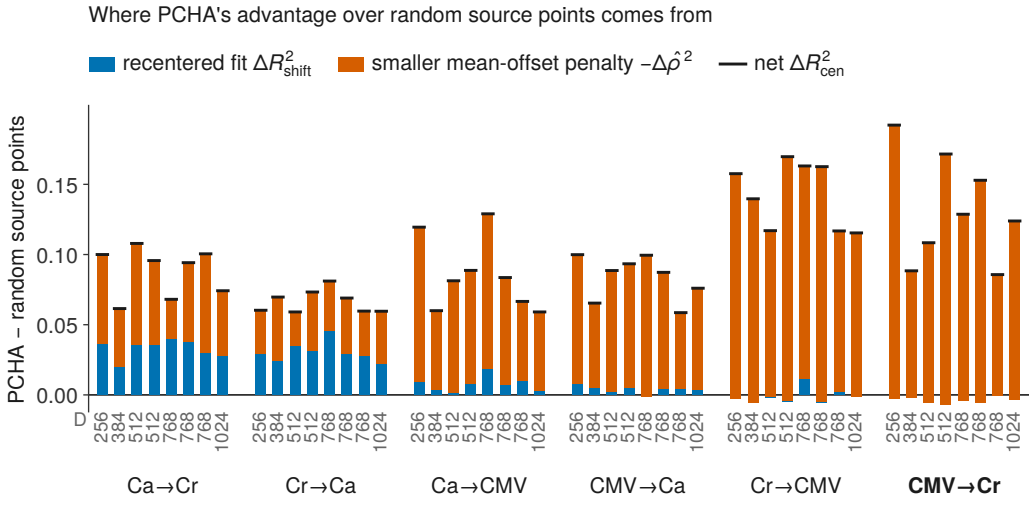

Figure H.8: **Where PCHA's advantage over random source points comes from.** For each of the 48 cells (six directed pairs × eight embedding configurations, ordered by dimension D), the difference in R^2_{cen} between the transferred PCHA dictionary and hulls of 32 random source points (black tick), split exactly into the difference in recentered fit (blue) and the smaller mean-offset penalty (vermilion); values are means over ten splits, which make the split exact. PCHA scores higher in all 48 cells. Location accounts for over 90% of the advantage in the CMV-involving pairs and for 54–63% between the two negotiation corpora, where learned archetypes also add recentered fit. Within each pair, bars are ordered by D; from left to right: text-embedding-005 (256), all-MiniLM-L6-v2 (384), Gemini embedding-2 (512), text-embedding-005 (512), all-mpnet-base-v2 (768), text-embedding-005 (768), text-multilingual-embedding-002 (768), e5-large-v2 (1024).

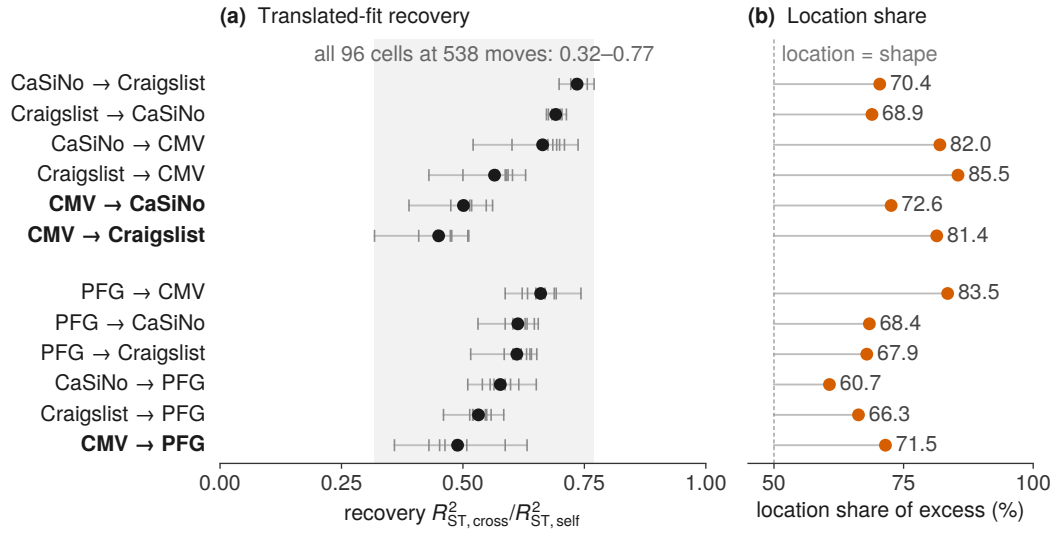

Figure H.9: **Translated-fit recovery and the location share of the excess, all 12 directed pairs.** 538-move budget; eight embedding configurations; dot = mean of per-evaluation recovery over the pair's split-seed evaluations; ticks = cell values (Section 4); shares formed from the aggregated excess terms. (a) Recovery is the translated fit R^2_{ST} of the transferred dictionary divided by that of a same-capacity held-out-self dictionary on the identical target, formed per evaluation before aggregation; cell values are the median over ten splits of the five-seed average. The shaded band spans all 96 cells, 0.32–0.77. Pairs of the three-corpus study are above and PFG-involving pairs below; bold rows have CMV as the source. Translated fit uses the target to choose the translation and is a diagnostic, not zero-shot performance. (b) Share of the cross-minus-self excess error $\Delta(1 - R^2_{\mathrm{cen}})$ carried by the location family $\Delta(\lambda + \iota)$ rather than the shape family $\Delta(s + c)$, per pair, in percent; the dashed line marks equal shares. The 48-cell statistics of Section 5 use the primary 785-move three-corpus study; this figure uses the common 538-move four-corpus benchmark of Section 6.

new weights, so it is a diagnostic of what a model set could capture, never a measure of zero-shot performance.

Corpora. The dictionary experiments cover four dialogue and argumentation corpora whose records are selected by corpus-specific criteria, not by one common definition of a successful move. The unit of generalization is the directed corpus pair, of which there are twelve. Eight embedding configurations show how results spread across representations of the same text; two are truncations of a third, and none is an independent replication. Regularities such as the ordering of recovery across pairs, or the in-span share of the centroid offset, are observations about these corpora.

Representation. All blueprint results pass through one LLM pipeline. The raw-text comparison contrasts two pipelines on the same records; because the blueprint generator also saw surrounding context and response evidence, the contrast is not an estimate of the effect of abstraction.

Source summaries. PCHA is the object we measured, not a method we recommend: Extreme Support, a deterministic boundary-seeking summary, had the higher mean-referenced score in 46 of 60 cells. The benchmark covers five embedding configurations, and we offer no mechanism for why one summary is better placed than another; scalar hull extent does not explain it across method families.

Sparse autoencoders. The study covers one base model and one SAE family at three layers. Its evaluation halves share no tokens but can share documents. The five distributions are not a calibrated ladder of shifts, so "measured separations fall short of the zero crossings" is a statement about these distributions. We do not establish that any of the five distributions lies outside the data on which these SAEs were trained, so the natural conditions are shifts in the input mix rather than tests on

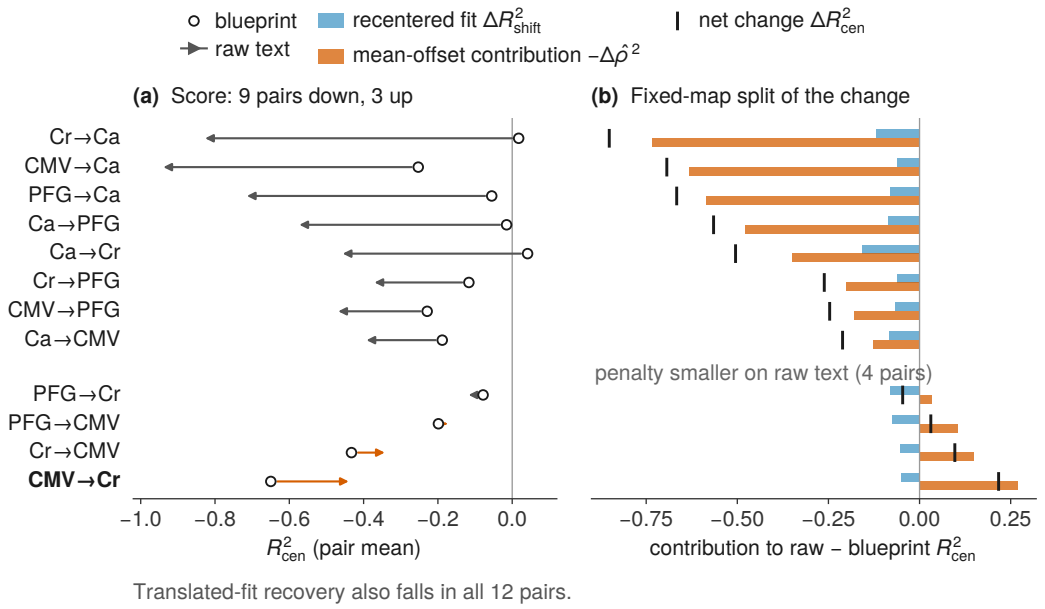

Figure H.10: **Same records, two representations.** Pair means over five seeds, ten splits and the five full-dimensional embedding configurations at the 538-move budget, for LLM-generated blueprints and for the source-native text of the same records (Table F.1). (a) Mean-referenced score R^2_{cen} of each directed pair on blueprints (circle) and on raw text (arrowhead); vermilion arrows rise. The eight pairs above the gap have a larger mean-offset penalty on raw text, the four below a smaller one. (b) The change split exactly by Proposition 1 into the recentered-fit term $\Delta R^2_{\text{shift}}$ (blue) and the mean-offset contribution $-\Delta\hat{\rho}^2$ (vermilion); black tick: net change ΔR^2_{cen}. Recentered fit is lower on raw text in all 12 pairs. The mean-offset penalty is larger on raw text in eight pairs and smaller in four; in three of those four (Craigslist→CMV, CMV→Craigslist, PFG→CMV) the smaller penalty outweighs the lost recentered fit and the score rises. Translated-fit recovery is also lower on raw text in all 12 pairs (Figure H.5).

unseen distributions. Removing the decoder bias is a fixed-code intervention, not a retrained model (Section 7). We measure direction-specific tracking (Appendix D.3), but do not explain why the reconstruction map has it.

Labels. "Location-dominated" rests on a convention about split consistency and share; we therefore report the continuous location and shape excesses wherever the label appears.

The paper establishes a measurement account and demonstrates it in two reconstruction regimes; its empirical regularities are scoped to the corpora, representations, and models tested.