# DIAL M FOR MONOSEMANTICITY: RECONSTRUCTION METRICS CANNOT CERTIFY SAE BASIS QUALITY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reconstruction metrics are the most common way to certify a sparse autoencoder (SAE), yet any functional of the reconstruction, MSE, FVU, CE-recovered and patched task accuracy among them, is provably constant under an invertible remix of the code: it cannot tell a dictionary from one strictly more mixed in the targeted latent basis. We use this identity constructively, mixing the basis of one pretrained SAE post hoc with the reconstruction preserved exactly, against a clone-split control arm whose residual support confound is a step rather than a ramp and so cannot manufacture a dose-response. Reading the result requires a calibrated operating point: the identical grid returns opposite signs depending only on the steering coefficient, and at the strength our pre-registration inherited the model sits 17–31 nats of KL from unsteered behaviour. At a post-hoc strength inside the regime our diagnostics certify, increased mixing lowers top-$k$ SAE-latent steering while every reconstruction metric reports the dictionary unchanged: $-0.298$ $[-0.509, -0.089]$ on irony, negative but unresolved on SNLI, and negative in 21 of 23 cells outside the broken regime. Full-set steering is provably unaffected. We recommend steering results report answer-position KL, an off-target accuracy delta, and one calibrated strength as a fraction of the activation norm.

## 1 INTRODUCTION

Sparse autoencoders are proposed as a fix for superposition: because a language model packs more concepts than it has dimensions (Elhage et al., 2022), neurons respond to unrelated concepts and reading off "what a unit represents" becomes unreliable. An SAE (Bricken et al., 2023; Cunningham et al., 2024) learns an overcomplete, sparsity-constrained dictionary whose atoms tend to be monosemantic, shown qualitatively at scale (Templeton et al., 2024). What has not been shown is that a monosemantic *dictionary* buys something a polysemantic one does not, holding everything else fixed. The strongest existing evidence that it pays reports robustness gains but moves the representation and the procedure that produced it together (Zhang et al., 2025). That does not settle it, because SAEs differing in interpretability also differ in width, sparsity, architecture and training run at once. What is needed is an intervention on basis quality alone.

The obstacle to building one is structural. Several widely reported SAE evaluations depend on the SAE only through its reconstruction. MSE, fraction of variance unexplained, CE-loss-recovered, patch-KL (Gao et al., 2025) and downstream accuracy after substituting the reconstruction are all computed from $\hat{x}$. Section 3 proves no such quantity can distinguish a clean feature basis from a scrambled one: the reconstruction is invariant under an invertible remixing of the code with its inverse on the decoder, so the metric is constant on the whole orbit. The observation sorts the metric zoo into a class that provably cannot support a basis-quality claim, holding the most-reported reconstruction metrics, and a class that can. It also makes an intervention available: the basis can be mixed on purpose with the reconstruction held exactly fixed, removing the confound by construction.

**Research questions.** What can an evaluation see about an SAE's basis, and what follows once something can?



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.

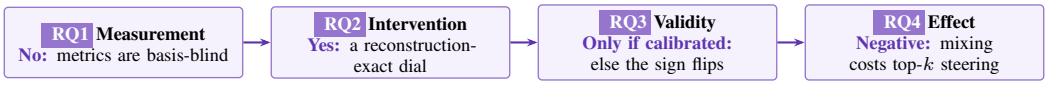
**RQ1 Measurement.** Can a reconstruction-based evaluation tell a monosemantic basis from a polysemantic one? (Sec. 3)

**RQ2 Intervention.** If not, can basis quality be intervened on directly, with the reconstruction preserved and the sparsity change controlled? (Sec. 4)

**RQ3 Validity.** At what strength can a steering result be read at all, and does the choice change it? (Sec. 5)

**RQ4 Effect.** Does basis mixing degrade what practitioners use SAEs for? (Sec. 6)

Together they form an identification chain, and Table 2 collects every number behind it.




RQ3 is needed because the same steering grid, read at two strengths, returns opposite signs, and the reversing strength is the one we pre-registered from precedent (Appendix D.5).

**Contributions.**

1. An invariance result for SAE evaluation: any functional of the reconstruction is constant on the reparameterisation orbit (Observation 1, Corollary 1), which sorts the metric zoo (Table 1).
2. A reconstruction-exact, continuously dosed intervention on basis quality with a clone-split control arm (Section 4). The residual confound is a $\lambda$-independent level (Corollary 2) and so cannot generate a dose-response; a $\lambda$-dependent code rescaling is disclosed and bounded (Appendix D.2).
3. Evidence that a steering conclusion is indexed to its operating point, with a diagnostic that separates the two readings, and four measurements any steering paper can log from its own run (Section 5).
4. A proof that full-set steering directions are invariant along the dial (Proposition 3), and a single-configuration existence result: top-$k$ SAE-latent steering is sensitive to calibrated basis mixing (Section 6).

**What is new here.** Neither the question nor the algebra is new; their combination is. Wang et al. (2026) answer our question correlationally, across SAEs differing in width, sparsity and training run at once; we intervene on the basis alone. The reparameterisation identity is classical, and the SAE literature takes it up on the training side (Locatello et al., 2019; Cain, 2026); new here is its consequence for *evaluation*, and its constructive use, since an admissible subset of the orbit that voids the certification supplies the manipulation.

**Scope, and why one configuration carries the claim.** Every experiment uses one model, one layer and one SAE family (Section 4); three training seeds, two widths and three tasks are in Table 2, and no SAE was trained for this work. RQ4 supports a possibility claim: basis mixing *can* cost top-$k$ latent steering while every reconstruction metric reports the dictionary unchanged. One configuration establishes that. A calibrated regime with a reliably positive slope would falsify it.

## 2 Related Work

**SAE evaluation and basis quality.** Dictionary learning yields features qualitatively more interpretable than neurons (Bricken et al., 2023; Templeton et al., 2024), yet SAEs underperform dense baselines on sparse probing (Kantamneni et al., 2025) and steering (Wu et al., 2025), recovering competitiveness only under careful feature selection (Arad et al., 2025; Jørgensen & Hansen, 2026). Proxy improvements do not map to practical performance (Karvonen et al., 2025), two of that benchmark's metrics fail reliability lenses (Chanin, 2026), and trained SAEs are matched by random or frozen-decoder baselines (Korznikov et al., 2026) and by randomly initialised transformers (Heap et al., 2026). One response is supervised or synthetic ground truth (Makelov et al., 2025; Chanin & Garriga-Alonso, 2026); ours keeps the real model and makes the ground truth a manipulation. Corollary 1 gives a mechanism for one large family of these failures: the reconstruction-derived subset cannot rank dictionaries by basis quality at all.

| ✗ **Reparameterization-invariant** *constant on the whole $GL_F$ orbit* | | ✓ **Potentially basis-sensitive** *can vary along the orbit* | |
|---|---|---|---|
| MSE / FVU / expl. variance | Cor. 1 | $k$-sparse probe (small $k$) | Prop. 2 |
| CE-loss recovered | Cor. 1 | Units to steer at fixed $k$ | reads the code |
| Patch-KL (Gao et al., 2025) | Cor. 1 | Absorption / splitting | property of atoms |
| Accuracy after patching $\hat{x}$ | Cor. 1 | Cross-task feature reuse | atom identity |
| Dense probe, unregularized | Prop. 2 | Any "name the top-$k$" protocol | reads latents |
| Steering dir., full set ($k$=$2P$) | Prop. 3 | Steering dir., sparse ($k$<$2P$) | Prop. 3 |

Table 1: The metric sort, *derived* from Corollary 1 and Propositions 2–3 rather than observed. The left block cannot rank dictionaries by basis quality; right-block membership is necessary, not sufficient (Chanin et al., 2025b).

**Identifiability in representation spaces.** That unsupervised latent-variable models are identifiable only up to distribution-preserving reparameterisation is classical (Locatello et al., 2019), and dictionary learning and NMF have conditions under which a sparse or nonnegative factorisation is unique up to permutation and scale (Donoho & Stodden, 2003; Gribonval & Schnass, 2010; Spielman et al., 2012; Hillar & Sommer, 2015); a reconstruction-only evaluation cannot check whether they held. The SAE literature takes this up from the *training* side: near-identifiable variants under an approximate restricted isometry condition (Nelson et al., 2026), a rate–distortion cost for monosemanticity (Mencattini et al., 2026), a limit on recovering ground-truth features unless those are extremely sparse (Cui et al., 2026), and the argument that SAE latents are not canonical units (Leask et al., 2025). Closest in spirit, Cain (2026) treats representation spaces as carrying a $GL$ gauge freedom and shows geometric similarity measures distorted by function-preserving invertible maps. Empirically, different seeds learn different features (Paulo & Belrose, 2026), unstable ones concentrate in reproducible subspaces (Gerasimov et al., 2026), and a parallel result holds for steering vectors (Venkatesh & Kurapath, 2026). Our object and use differ: the casualties are the most-reported reconstruction metrics, the route assumes no generative model or ground-truth dictionary, and an admissible subset of the orbit is exhibited constructively so the dial can be built on it. Nothing here claims two independently trained SAEs are related by such an $M$; Lemma 1 shows the training objective is not orbit-invariant.

**Validity of causal interventions.** Braun et al. (2025) measure steering as a continuous logit difference and report that for roughly a third of samples it moves opposite to the intended direction; our primary DV follows their measure. Duan (2026) predicts side effects from decoder geometry in advance, and Taimeskhanov et al. (2026) derive laws, some non-monotone, for how strength affects next-token probability, concept presence and cross-entropy. Grant et al. (2026) show causal interventions routinely push representations off the model's natural distribution and separate harmless from "pernicious" divergences; theirs is the closest precedent for what we do, applied to the model under intervention where ours is applied to the evaluation that reads it, and Asiaee (2026) supplies statistical apparatus for certifying such claims. Those laws govern one steered model's quantities as strength varies; a *comparison* between two dictionaries can invert only if the arms cross.

## 3 Reconstruction-Based Evaluations Are Invariant to Basis Remixing RQ1

Fix a transformer, a layer $\ell$, and a corpus $\mathcal{X} \subset \mathbb{R}^d$ of residual-stream activations. An SAE is a triple $(E, W, b)$ with encoder $E : \mathbb{R}^d \to \mathbb{R}^F_{\geq 0}$, decoder $W \in \mathbb{R}^{F \times d}$ whose *rows* are feature directions, and bias $b \in \mathbb{R}^d$. Codes are $a = E(x) \in \mathbb{R}^{1 \times F}$ and the reconstruction is $\hat{x} = aW + b$. Two structural properties are normally imposed: codes are **nonnegative** and decoder rows are **unit-norm**.[1]

**Definition 1** (Reparameterization)**.** For invertible $M \in \mathbb{R}^{F \times F}$, the *$M$-reparameterization* of $(E, W, b)$ is $(E_M, W_M, b)$ with $E_M(x) := E(x)M$ and $W_M := M^{-1}W$.

**Observation 1** (Reconstruction invariance)**.** *For every SAE $(E, W, b)$, every invertible $M \in GL_F(\mathbb{R})$ and every $x \in \mathbb{R}^d$, $\hat{x}_M(x) = \hat{x}(x)$ exactly.*

*Proof.* $\hat{x}_M(x) = \big(E(x)M\big)\big(M^{-1}W\big) + b = E(x)\big(MM^{-1}\big)W + b = \hat{x}(x).$ □

[1]We follow the `sae_lens` row convention; readers preferring column vectors should transpose.

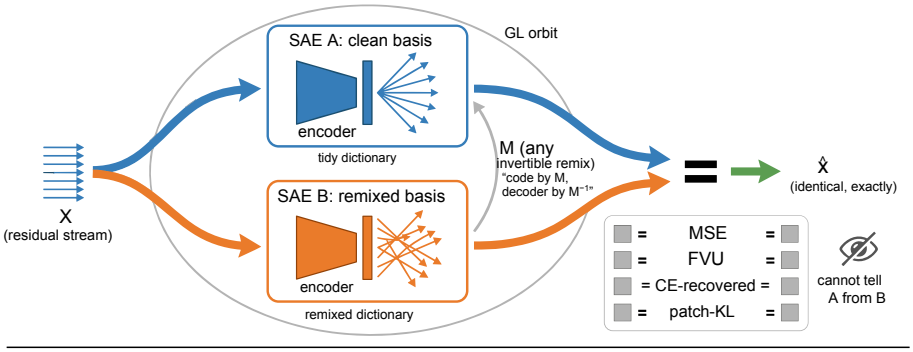



Figure 1: The mechanism behind Corollary 1: same $\hat{x}$, same metrics, different basis.

**Implication.** Reconstruction metrics stay valid measures of reconstruction; what fails is ranking bases with them, since they cannot distinguish members of a reconstruction-equivalence class.

**Corollary 1** (Blindness)**.** *Let $\Phi$ be any functional depending on the SAE only through $x \mapsto \hat{x}(x)$, expressible as $\Phi(x, \hat{x}(x))$ and possibly composed with any fixed downstream computation. Then $\Phi$ is constant on the orbit $\{(E_M, W_M) : M \in GL_F(\mathbb{R})\}$.*

The reach of Corollary 1 is the point: it covers MSE, FVU and explained variance directly, and patch-KL (Gao et al., 2025), CE-loss-recovered, patched downstream accuracy and *any* behavioural evaluation of the patched model, because in each case $\hat{x}$ enters the forward pass and the SAE is never consulted again. A natural objection is that $E_M$ need not be a legitimate SAE, since $aM$ need not be nonnegative or sparse and $M^{-1}W$ need not have unit-norm rows. For Corollary 1 that is immaterial, the metric being a function of $(x, \hat{x})$, which the transformation fixes. It bites only against the stronger claim that licenses the experiment, that one can construct an object strictly more mixed in the targeted latent basis that still has the four structural properties defined next.

**An admissible family of mixings.** Call $M$ *admissible* if it is invertible, keeps the code nonnegative, inflates mean $L_0$ by at most a constant factor, and leaves every decoder row unit-norm, the four structural properties an SAE is normally required to have (Definition 2). A continuum of them exists. Fix $P$ disjoint index pairs and let $M_\lambda$ be the identity outside those pairs and, on each pair, blend the two coordinates by $\lambda \in [0, \frac{1}{2})$, a continuous mixing parameter that blends disjoint feature pairs without altering the reconstruction (Definition 3). Composing with the positive diagonal $S$ that records the norms of the de-mixed decoder rows gives $\widetilde{M}_\lambda = M_\lambda S$, which satisfies all four (Proposition 1), since a positive diagonal preserves both nonnegativity and the support while $S^{-1}$ restores unit rows and the two cancel in $\hat{x}$. Without that rescaling a decoder-norm change would masquerade as a monosemanticity effect. All experiments below apply $\widetilde{M}_\lambda$.

**What $\lambda$ moves besides the basis.** Mixing raises mean $L_0$ whenever the two members of a pair do not always fire together (Lemma 1), so "you broke sparsity rather than the basis" is a valid rebuttal to a bare $\lambda$-sweep, and it is why the sparsity penalty makes the training objective non-invariant: $\widetilde{M}_\lambda$ is reachable by construction but not by gradient descent. Section 4 answers the rebuttal with a clone-split control arm. The rebuttal is also provably weaker than it looks. Strictly positive weights on nonnegative codes cannot cancel, so a pair fires as a block whenever either member does and the transformed support is the *same* set at every $\lambda > 0$. The inflation is a single step taken as $\lambda$ leaves 0 rather than a function of the dose (Corollary 2), and a confound that does not vary with $\lambda$ cannot generate a slope in $\lambda$. Measured, mean $L_0$ is constant to four decimals at every $\lambda > 0$ in both arms on all three tasks (Appendix D.2).

**Probes split along the same line.** Because $\langle aM, w \rangle = \langle a, Mw \rangle$ and $w \mapsto Mw$ is a bijection, an unregularized dense probe on the latents has the same infimum risk before and after remixing, while a $k$-sparse probe with $k < F$ need not (Proposition 2); our implementation approximates that limit at $C = 10^3$ and uses the dense probe as a falsifiable positive control, predicted flat in $\lambda$. Together

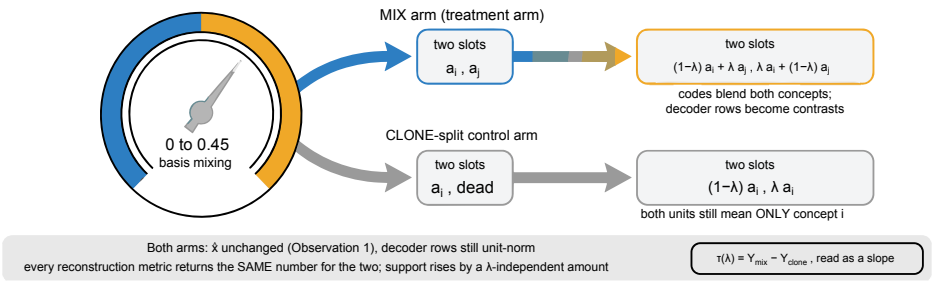



Figure 2: The dial at the centre of the instrument. The arms differ in basis structure and in nothing a reconstruction metric reads. In the mix arm the codes blend both concepts and the decoder rows become contrasts.

these results sort the metric zoo into a class that provably cannot support a basis-quality claim and a class that can (Table 1).

Figure 1 makes the mechanism concrete; Figure 4 is its empirical face.

**The sort, applied to a benchmark.** Of SAEBench's eight evaluations (Karvonen et al., 2025) exactly one is invariant by Corollary 1: loss recovered, computed by substituting $\hat{x}$ into the forward pass. The other seven read latents. The sort therefore *endorses* SAEBench's design and localises the problem on the sparsity–fidelity curve it calls the "primary evaluation metric in new architectures": that curve's fidelity axis is the one number new architectures are compared on (Rajamanoharan et al., 2024a; Bussmann et al., 2025; Mudide et al., 2025), and at fixed reconstruction it cannot register a change of basis. The remedy is to stop reading that axis as basis evidence, leaving the seven latent-reading evaluations or a calibrated behavioural axis to carry such a claim.

**Finding 1 · Answer to RQ1: no**

*Every metric in the left block of Table 1 returns an identical value for an SAE and for one strictly more mixed in the targeted latent basis with nonnegative codes, unit decoder rows and bounded $L_0$, a structural insufficiency rather than an artifact of noise, sample size or benchmark design. The mixing is detectable, by $L_0$ at matched MSE (Lemma 1), decoder coherence (Section 4) and within-pair co-firing, but by no reconstruction metric: $\widetilde{M}_\lambda$ passes the entire* `encode`/`decode` *surface, so a reconstruction-metric certification cannot detect a dictionary degraded on purpose.*

## 4 A Reconstruction-Preserving Basis-Mixing Intervention RQ2

We use **basis mixing** for the formal operation defined below and *injected polysemanticity* for its interpretation when the paired source latents encode distinguishable content; the theoretical results require only the former. Two arms make this a controlled experiment rather than a demonstration (Figure 2).

**Mix arm.**
Apply $\widetilde{M}_\lambda$ to $P$ disjoint feature pairs, sweeping $\lambda$ over an eight-point grid from 0 to 0.45. Each target's code is progressively blended with its partner's; on the decoder side the mixed rows are contrasts, $(1-\lambda)w_i - \lambda w_j$, nearly antiparallel at $\lambda = 0.45$ (Appendix D.2). Reconstruction is preserved exactly; $L_0$ rises by Lemma 1. The failure mode dosed here is documented rather than invented: narrow SAEs merge correlated features into shared latents (Chanin et al., 2025a) and hierarchical features absorb one another under a sparsity objective (Chanin et al., 2025b).

**Clone-split control arm.**
A control arm that inflates the support by the same mechanism while injecting no basis mixing. Replace $a_i$ by $\big((1-\lambda)a_i, \lambda a_i\big)$, the second slot a *dead* latent (zero activation frequency on the fitting corpus, asserted per cell) given a decoder row identical to $i$'s. Reconstruction is again exact and $L_0$ rises by the same mechanism, but both units still mean feature $i$, and the dead slots come from a pool that is 98–99% of the dictionary on every task tested (Appendix B).

| | Irony (TweetEval) | | | | SNLI | | | | BoolQ | Tally | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Seed / width | 42/16k | 123/16k | 999/16k | 42/65k | 42/16k | 123/16k | 999/16k | 42/65k | 42/16k | neg. | excl. 0 |
| **RQ1 Reconstruction metrics along the dial (Obs. 1)** | | | | | | | | | | | |
| MSE, FVU, CE, patched acc. | $0^\dagger$ | $0^\dagger$ | $0^\dagger$ | $0^\dagger$ | $0^\dagger$ | $0^\dagger$ | $0^\dagger$ | $0^\dagger$ | $0^\dagger$ | 9/9 flat | |
| **RQ2 Controls and validity gate (Sec. 4)** | | | | | | | | | | | |
| $\Delta L_0$, mix / clone | 9.2/12.8 | – | – | – | 8.5/12.5 | – | – | – | 8.4/8.5 | ctrl $\geq$ mix | |
| Untrained random dict. | $-0.189$ | $-0.055$ | $-0.238$ | $-0.241$ | $-0.316$ | $-0.384$ | $-0.305$ | $-0.124$ | $-0.181$ | – | – |
| Gate margin[§] | $+0.072$ | $-0.100$ ✗ | $+0.209$ | $+0.254$ | $+0.171$ | $+0.371$ | $+0.264$ | $+0.351$ | $+0.170$ | 8/9 pass | |
| **RQ3 RQ4 Steering slope of $\tau$ in $\lambda$ (Sec. 5–6)** | | | | | | | | | | | |
| $c = 0.02$ *(intact)* | $\mathbf{-0.298}$ | $\mathbf{-0.187}^\ddagger$ | $+0.060$ | $-0.008$ | $-0.232$ | $-0.258$ | $-0.123$ | $\mathbf{-0.948}$ | $-0.014$ | 8/9 | 3 |
| $c = 0.05$ *(intact/mild)* | $\mathbf{-2.206}$ | $\mathbf{-0.901}$ | $-1.037$ | $-0.338$ | $-0.617$ | $-0.161$ | $+0.658$ | $\mathbf{-1.979}$ | $-1.061$ | 8/9 | 3 |
| $c = 0.10$ *(mild/broken)*[¶] | $\mathbf{-2.162}$ | $\mathbf{-3.101}$ | $\mathbf{-3.502}$ | $-0.651$ | $-1.131$ | $+1.768$ | $-0.067$ | $-0.098$ | $\mathbf{-1.484}$ | 5/5 | 4 |
| $c = 1.0$ *(pre-reg., broken)* | $\mathbf{+12.479}$ | – | – | – | $\mathbf{+7.455}$ | – | – | – | – | 0/2 | 2 ✗ |
| **RQ4 Probing, Stage A, secondary DVs; headline dictionary only, value [95% CI] (App. D.1)** | | | | | | | | | | | |
| Dense probe (control) | $-0.010$ [$-0.017$, $0.001$] | | | | $-0.001$ [$-0.005$, $0.003$] | | | | $-0.012$ | 3/3 | 0 ✓ |
| $k$=8 sparse probe | $\mathbf{-0.252}$ [$-0.339$, $-0.066$] | | | | $\mathbf{+0.012}$ [$0.001$, $0.035$] | | | | $\mathbf{-0.036}$ | 2/3 | 3 |
| FMS (Härle et al., 2025) | $\mathbf{-0.064}$ [$-0.080$, $-0.032$] | | | | $\mathbf{-0.072}$ [$-0.110$, $-0.009$] | | | | $\mathbf{-0.082}$ | 3/3 | 3 |
| **Absolute steering effect at $c = 0.02$, $\lambda = 0$; headline dictionary, value [95% CI] (App. D.3)** | | | | | | | | | | | |
| SAE top-10 latents | $-0.028$ [$-0.044$, $-0.011$] | | | | $-0.010$ (CI contains 0) | | | | – | – | – |
| Class-mean direction | $+0.132$ [$0.081$, $0.183$] | | | | $+0.460$ [$0.388$, $0.533$] | | | | – | – | – |

Table 2: **Every result in one place.** Columns are the nine run × task cells ( = headline dictionary, shaded); slopes are the mix-vs-clone diff-in-diff. Green: 95% CI excludes zero in the H1 direction; orange: excludes zero against it. $^\dagger$Exact by Observation 1 (mix) and the dead-slot construction (clone), verified to $< 10^{-4}$ relative on every arm. ¶Tally excludes the four broken SNLI cells (grey). §Minimum of the run's own $\lambda$-sweep minus the random dictionary; negative fails. ‡Fails that gate, so not an independent replication. Regimes from the damage meter (Sec. 5); other CIs in Tables 11 and 7.

The estimand is therefore a difference-in-differences in $\lambda$ between the arms rather than either arm's slope. Writing $Y$ for the outcome,

$$\tau(\lambda) = \big[Y_{\text{mix}}(\lambda) - Y_{\text{mix}}(0)\big] - \big[Y_{\text{ctrl}}(\lambda) - Y_{\text{ctrl}}(0)\big] = Y_{\text{mix}}(\lambda) - Y_{\text{ctrl}}(\lambda),$$

the second equality holding because the $\lambda = 0$ cell is shared by construction. We summarise the dose response by the slope of $\tau(\lambda)$, and every estimate below is that slope. H1, fixed before these runs, predicts that slope is negative: more basis mixing, less steering utility.

**How the control handles the support increase.** The control need not match the absolute support increase exactly. The estimand requires only that the residual support difference not vary with $\lambda > 0$, and that where the arms differ the control be the more inflated. Both hold. By Corollary 2 the residual difference is a level, constant in $\lambda$, while the estimand is a slope; measured, it runs *against* the hypothesis, the control arm inflating more than the treated arm (Table 2). Support is not the only non-basis quantity $\lambda$ moves; the other is a code rescaling that the control does not match, and the same appendix bounds what it can do to a slope.

**A dynamic-range precondition on the readout.** Before a run is read, the readout must separate the trained dictionary from an *untrained random* one, selected and scored identically. Where that anchor falls inside a run's own $\lambda$-sweep, no rung's movement can be read as basis quality, since a wholly untrained basis is not separated either. The gate has teeth: it rejects one of our nine run × task cells (Section 6). A reconstruction metric offers no comparable signal about the basis: it returns the same number on a dictionary and on every reconstruction-exact remix of it (Observation 1).

**Finding 2 · Answer to RQ2: yes**

*The arms agree exactly on everything a reconstruction metric reads (Observation 1) and, up to a $\lambda$-independent level (Corollary 2), on the one thing it does not read that $\lambda$ also moves. The dial therefore estimates how sensitive a readout is to basis structure under a controlled perturbation, not the cost of naturally occurring polysemanticity (Lemma 1).*

**Dependent variables and configuration.** Stage A scores probing on cached activations, with a $k$-sparse probe on the codes as the pre-registered sensitive DV and an unregularized dense probe as the invariance positive control. Stage B scores steering, which needs live forwards over the whole grid. Its primary DV, following Braun et al. (2025), is the signed target-logit shift: the change in the raw logit (not a logit difference) of the class the model did *not* predict at baseline, sign-flipped per example so that positive means steering worked; generic perturbations lower it (Appendix D.3).

Best flip rate is a secondary DV, being what the steering literature most often reports. The steering direction is the unit-normalised $\Delta$-weighted sum of the decoder rows of the top $k \in \{10, 20\}$ class-discriminative latents, ranked within the cell's own manipulated set so untouched features cannot replace damaged ones as $\lambda$ grows (Appendix B.4).

**Model, dictionary and data.** All experiments use `gemma-3-4b-pt` (Gemma Team, Google DeepMind, 2025) at layer 17 with the Gemma Scope 2 (Google DeepMind, 2025) 16k-width JumpReLU SAE (Rajamanoharan et al., 2024b) at medium sparsity, seed 42. The tasks are three public English classification benchmarks spanning pragmatic, semantic and reading-comprehension inference: binary irony detection in tweets (TweetEval-irony (Barbieri et al., 2020), from SemEval-2018 Task 3 (Van Hee et al., 2018)), three-way natural language inference (SNLI (Bowman et al., 2015)) and yes/no question answering over a Wikipedia passage (BoolQ (Clark et al., 2019)). Each is posed few-shot (four or six demonstrations) with a class verbalizer and read at the final `Answer:` token; per task, 300 label-balanced fitting examples select and rank latents and 200 held-out examples are steered. Irony and SNLI are primary; BoolQ is a partial third task, run in Stage A and later added to Stage B. Intervals are percentile bootstraps over pairing draws with the 200 steered examples fixed, and the vector is added at every token, few-shot demonstrations included (Appendices B.4, C).

## 5 Calibrating a Valid Steering Regime RQ3

Before the dial's readout can be interpreted, it has to be taken somewhere on the steering-strength axis. Four things need separating here, because this paper's own history ran them together.

First, what was pre-registered. Stage B's analysis plan fixed $c = 1.0$, one mean activation norm, taken from precedent rather than from any calibration. An audit of six representative works shows the precedent supports no such inference: reported strengths span over two orders of magnitude in heterogeneous, mostly non-interconvertible parameterisations (Turner et al., 2023; Rimsky et al., 2024; Templeton et al., 2024; Wu et al., 2025; Taimeskhanov et al., 2026), an operating point usually cannot be recovered from the paper reporting it, and where it can be it is large (Appendix D.5).

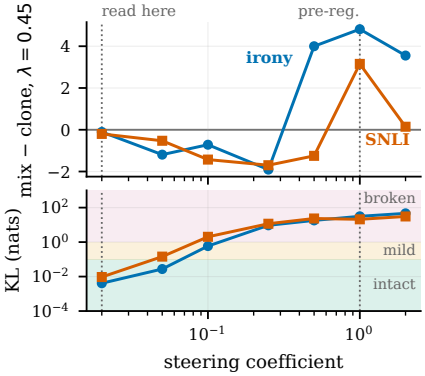



Figure 3: Where the sign flips: the $\lambda$=0.45 estimand against steering coefficient (top) and the diagnostic that assigns intact/mild/broken (bottom, Table 10). Negative supports H1; it flips positive only well inside the broken regime.

Second, what the diagnostic is. It was built to check a different concern, whether the steered model was still computing the task at all, and is not tuned to the comparison it is used on here. Sweeping eight coefficients from 0 to 2.0, it logs at the answer position the full-vocabulary KL(base $\|$ steered), next-token entropy, top-1 preservation, the softmax mass left on the verbalizer tokens, and off-target accuracy. It first confirms it is looking at the same object, reproducing Stage B's per-cell shifts to three decimals at shared coefficients. We call it the damage meter.

Third, what makes a reading valid. We bin answer-position KL into *intact* ($< 0.1$ nat), *mild* ($< 1$) and *broken*, descriptive bins for this analysis rather than universal cutoffs; Table 10's regime labels are read off the diagnostic, not assigned post hoc to rescue a result. At $c = 0.02$ the arms the dial compares sit at 0.0015–0.0094 nats of KL, keep 0.978–0.983 of the verbalizer mass against a 0.982–0.984 baseline, and preserve top-1 on 0.86–0.99 of examples. At $c = 1.0$ the same arms sit at 17.0–31.3 nats, top-1 preservation is 0.00, and verbalizer mass has collapsed by three to six orders of magnitude, to $6 \times 10^{-7}$–$1.5 \times 10^{-3}$. The direction is unit-norm and scaled by the mean activation norm, so $c = 1.0$ adds a vector as large as the layer-17 residual stream itself; we do not read the target-logit shift there as a measure of steering utility.

Fourth, and this is the load-bearing fact, what the whole sweep does. The estimand is negative throughout the low-divergence region on both tasks and reverses only after severe divergence (Figure 3, Table 10), so the two readings this paper contrasts are one point inside the valid region and

one point far outside it. Stage B was run twice on identical grids, same code, same 200 examples, same pairings, differing only in the coefficient at which the readout is taken. At $c = 1.0$ the verdict reverses and mixing the basis appears to *improve* steering, slopes $+12.479$ (irony) and $+7.455$ (SNLI), both with intervals excluding zero; at $c = 0.02$ the same grid supports H1. The choice does not scale the answer — it signs it.

**The two diagnostics can disagree.** On SNLI the class-mean direction at $c = 0.02$ is *intact* by answer-position KL (0.068 nats) while already costing 15 accuracy points off-target, and 25 at $c = 0.05$. They capture different failure modes, so we report both rather than collapsing them into one score and adopt the more conservative threshold where they disagree.

**The same grid, a second dependent variable.** Maximum flip rate reverses the comparison ($+1.646$ on irony) because it saturates, an untrained random dictionary scoring its maximum (Appendix D.8).

**Finding 3 · Answer to RQ3: only inside a validated regime**

*Outside a validated regime, as in our sweep, a steering result can invert. We recommend any steering result carry* (i) *answer-position KL from unsteered behaviour,* (ii) *an off-target accuracy delta,* (iii) *one strength inside the intact regime, and* (iv) *that strength relative to the activation norm, which is what makes operating points comparable across papers. The bins are ours, but each of the four is read against the run's own baseline, so the procedure transports where the numbers would not.*

## 6 Basis Mixing and Top-$k$ Steering in the Calibrated Regime RQ4

Figure 5 carries the headline dose-response; Figure 8 below extends it to every stored seed, width and coefficient. Two readouts answer RQ4, and they are not equal. Steering is primary, being the use practitioners actually make of SAE latents and what Section 5 calibrated; probing is secondary, reported because it was pre-registered, and it does not agree with the primary readout on every task. We rank them rather than reporting whichever agrees.

**Stage A (probing), in brief.** The pre-registered $k$=8 sparse probe falls with $\lambda$ on irony and BoolQ but rises on SNLI, evidence *against* H1 there; the dense-probe control stays flat on all three tasks, as Proposition 2 predicts; and an external metric we did not design, the Feature Monosemanticity Score (Härle et al., 2025), falls on all three, every interval excluding zero (Table 2; all eight DVs in Appendix D.1). The two basis-sensitive readouts disagree on SNLI, and that disagreement is the finding: they are not interchangeable, and we do not use probing to adjudicate RQ4.

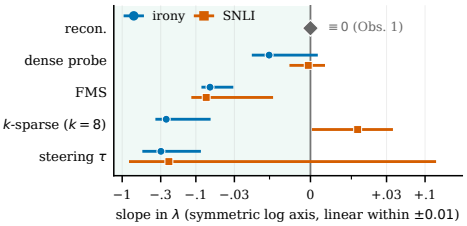



Figure 4: Slope in $\lambda$ of each readout, 95% CI. Reconstruction metrics sit at exactly 0; the code-reading readouts move (Corollary 1).

**Stage B (causal steering).** Read at $c = 0.02$, a post-hoc point chosen once the damage meter invalidated the pre-registered $c = 1.0$ (Table 10), the estimand is the slope of $\tau$ over the eight-point $\lambda$ grid on the logit-shift DV: $\mathbf{-0.298}$ $[-0.509, -0.089]$ on irony, excluding zero, and $-0.232$ $[-0.773, 0.135]$ on SNLI, negative but not excluding zero (Figure 5, Table 6). Because every reconstruction metric scores the two arms identically, a nonzero $\tau$ measures sensitivity to the mixing under the controls of Section 4. It is specific to sparse selection: the direction from the whole manipulated set ($k$=$2P$=20) is the same at every $\lambda$ in both arms, so $\tau \equiv 0$ there (Proposition 3); the experiment's own direction code confirms this on synthetic SAEs to float32 round-off (Appendix B.4).

**The slope is not a property of the chosen coefficient.** Every Stage B cell stores the DV at all three grid coefficients, so the estimand recomputes at $c \in \{0.02, 0.05, 0.10\}$ for all four seed/width configurations plus BoolQ. Of those 27 cells 24 are negative. Setting aside SNLI at $c = 0.10$, which the damage meter places in the *broken* regime, 21 of 23 are negative, 10 exclude zero, and none

excludes zero in the positive direction (Table 2; Appendix D.7). At $c = 0.02$, 8 of 9 cells are negative (one-sided sign test $p \approx 0.02$, optimistic since a run's two tasks share a dictionary).

**Which task is stable, and which is not.** At $c = 0.02$ SNLI is negative in every configuration but resolved only at 65k; irony is configuration-sensitive, resolved in one run and in a replication that fails the dynamic-range precondition of Section 4. We therefore claim a consistent sign on SNLI (4 of 4, too few to resolve alone: sign test $p = 0.06$) and a single-run interval excluding zero on irony, and we do not claim that irony reproduces. The stages also swap which task is the clean one: irony's dose-response is carried by its high-$\lambda$ rungs while SNLI, where the probe reversed, falls steadily to $-0.108$ (Figure 5). In absolute terms the SAE's top-10 latents lose to a class-mean direction but beat generic perturbations (Table 2); the estimand contrasts two SAEs.

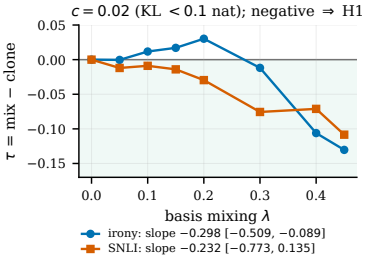



Figure 5: $\tau$ against $\lambda$ at $c = 0.02$.

**Finding 4 · Answer to RQ4**

*Within the calibrated regime, increasing basis mixing lowers top-$k$ SAE-latent steering in the headline configuration, excluding zero on irony; SNLI is negative in all four configurations but resolved in one, and BoolQ agrees in sign. Across seeds, widths, tasks and non-broken coefficients the direction is predominantly negative (21 of 23 cells) while magnitude varies; full-set steering is provably unaffected. We read this as evidence that top-$k$ steering is causally sensitive to SAE basis structure, not as a configuration-independent effect size for monosemanticity.*

## 7 Conclusion

Any metric computed from an SAE's reconstruction is constant on a reparameterisation orbit whose admissible members mix the basis deliberately (RQ1), so a reconstruction-exact, support-controlled dial is what makes basis quality answerable at all (RQ2); in our sweep, where the answer is read decides its sign (RQ3), and read inside a calibrated regime, basis mixing reduces top-$k$ latent steering in the tested configuration while full-set steering is provably unchanged (RQ4). The reusable principle is that a certification and a readout each have a domain of validity and neither reports when it has left one. If a safety case rests on basis quality certified by reconstruction, it is basis-blind; a causal check at an uncalibrated strength can return whichever sign the strength selects.

## Limitations

Five threats to the empirical estimate. *Scope:* the estimate is an existence result in the configuration Section 1 states, with what would generalise or falsify it. *Interval construction:* the intervals are mix-only, because the clone arm is identical across draws at every $\lambda$ on both tasks, and slopes are fit over the full grid including the rungs where irony is positive; the $\lambda$=0.45 endpoint contrast, which assumes no linearity, is $-0.130$ $[-0.167, -0.096]$ (irony) and $-0.108$ $[-0.194, -0.032]$ (SNLI), computed post hoc (Appendix D). *Effective sample size:* irony's twelve draws cluster into five distinct outcomes at $\lambda \geq 0.3$ (Appendix A.1), so its interval is not an honest $n$=12; SNLI's draws stay distinct, and BoolQ's replication is partial. *Calibration:* the dynamic-range precondition rejects one of nine run $\times$ task cells, seed 123 on irony, and we report the rejection rather than dropping the cell (Appendix A.2). *Effect size:* at $\lambda$=0.45 the effects are 69% (irony) and 34% of the untrained random-dictionary distance (Appendix D.3), small in logits though not on the instrument's own scale.

Two limits on the paper's *meaning* sit outside that list. Observation 1 and Corollary 1 hold for any SAE, so only RQ4 carries the scope above; generation, refusal, base-model changes and norm-preserving steering are untested. And $\widetilde{M}_\lambda$ is constructed, not gradient-descended (Lemma 1): the dial bounds sensitivity to basis structure, not the cost of natural polysemanticity.

### Ethics Statement

This work uses publicly released base models and public datasets, trains no models, and involves no human subjects, no annotation of personal data and no deployed system. Its principal ethical relevance is indirect and methodological: where reconstruction fidelity is offered as evidence for basis quality in a safety argument, Observation 1 shows that evidence is insufficient on its own. The steering interventions studied here are conducted on base models for measurement purposes; we note that the validity diagnostics we recommend are also safety-relevant instruments, since a steering intervention that has silently destroyed a model's task competence is not a controlled intervention.

### Reproducibility Statement

Every number, figure and table in this paper traces to a script and a stored JSON artifact under a stable per-experiment identifier (Appendix G). One attempted run is reported, in Appendix A.2, with *no* numbers at all, precisely because it left no artifact we can stand behind. The observation, corollaries, propositions and lemma of Section 3 are proved in full in Appendix E. Model, dictionary, task, prompting and compute details are in Appendix B, and the statistical protocol (intervals, pre-registration, and the floating-point verification of the reconstruction identity) is in Appendix C. The claims that survived our own audit, and the ones that did not, are in Appendix F. End-to-end determinism was verified against an unplanned replication: a Stage B run interrupted mid-grid was repeated on a fresh machine with a fresh install, and returned irony's slope and every per-$\lambda$ cell identical to the interrupted run in all stored digits. That configuration used 4 pairing draws and was later extended to the 12 draws reported here; the determinism check is unaffected. Anonymised code, run artifacts and logs are in the supplementary material, and will be released publicly upon deanonymization.

### AI Use Statement

We used generative AI tools (Claude, Anthropic) for three tasks requiring disclosure under the ICLR 2027 AI Policy for Authors: drafting and typesetting the written proofs in Appendix E, whose formal statements (Section 3) and proof arguments are the authors' own; feedback on the design of the Stage A and Stage B protocols and their validity checks; and implementing the analysis, experiment, data-loading and figure code from author specifications, including, during revision, checking reported numbers against the stored artifacts and computing one post-hoc statistic (the $\lambda{=}0.45$ endpoint contrast) with the existing analysis code. We did not use them to develop the theoretical framework, formulate the claims, supply critical ingredients of their proofs, propose or refine hypotheses, or interpret results; synthetic data, translation and qualitative analysis do not apply. Under recommended disclosure, we also used them to edit and structure the text, create figures and code, summarise related literature and format references. An author traced every number, table and figure to its stored run artifact (Appendix G), checked each proof line by line, re-ran the code to reproduce the reported outputs, and verified every bibliography entry against its primary source. We take responsibility for the final content, including anything produced with the aid of generative AI.

## References

Dana Arad, Aaron Mueller, and Yonatan Belinkov. SAEs are good for steering – if you select the right features. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 10241–10259, 2025.

Amir Asiaee. Certified interventional fidelity: Anytime-valid, adaptive evaluation of causal claims in mechanistic interpretability. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2026. arXiv:2607.08349.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un)reliability of steering vectors in language models. *arXiv preprint arXiv:2505.22637*, 2025. Presented at the ICLR 2025 Workshop on Foundation Models in the Wild.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Jericho Cain. Gauge freedom and metric dependence in neural representation spaces. *arXiv preprint arXiv:2603.06774*, 2026.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. Improving steering vectors by targeting sparse autoencoder features. *arXiv preprint arXiv:2411.02193*, 2024.

David Chanin. Are sparse autoencoder benchmarks reliable? *arXiv preprint arXiv:2605.18229*, 2026.

David Chanin and Adrià Garriga-Alonso. SynthSAEBench: Evaluating sparse autoencoders on scalable realistic synthetic data. *arXiv preprint arXiv:2602.14687*, 2026.

David Chanin, Tomáš Dulka, and Adrià Garriga-Alonso. Feature hedging: Correlated features break narrow sparse autoencoders. *arXiv preprint arXiv:2505.11756*, 2025a.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b. Oral; arXiv:2409.14507.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

Jingyi Cui, Qi Zhang, Yifei Wang, and Yisen Wang. On the limits of sparse autoencoders: A theoretical framework and reweighted remedy. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2506.15963.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.08600.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2003.

Evan Duan. Pre-intervention prediction of sparse autoencoder steering side effects. *arXiv preprint arXiv:2606.08365*, 2026.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. arXiv:2209.10652.

Moghis Fereidouni, Muhammad Umair Haider, Peizhong Ju, and A. B. Siddique. Evaluating sparse autoencoders for monosemantic representation. In *Findings of the Association for Computational Linguistics: EACL 2026*, 2026. arXiv:2508.15094.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2025. Oral presentation. arXiv:2406.04093.

Gemma Team, Google DeepMind. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Gleb Gerasimov, Timofei Rusalev, Nikita Balagansky, Daniil Laptev, Vadim Kurochkin, and Daniil Gavrilov. Unstable features, reproducible subspaces: Understanding seed dependence in sparse autoencoders. *arXiv preprint arXiv:2606.12138*, 2026.

Google DeepMind. Gemma scope 2: Helping the AI safety community deepen understanding of complex language model behavior. Technical report, Google DeepMind, 2025. URL https://deepmind.google/blog/gemma-scope-2-helping-the-ai-safety-community-deepen-understanding-of-complex-language-model-behavior/. Technical paper accompanying the suite of sparse autoencoders and transcoders released for the Gemma 3 model family.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations (ICLR)*, 2026. Oral; arXiv:2511.04638.

Rémi Gribonval and Karin Schnass. Dictionary identification—sparse matrix-factorization via $\ell_1$-minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

Ruben Härle, Felix Friedrich, Manuel Brack, Stephan Wäldchen, Björn Deiseroth, Patrick Schramowski, and Kristian Kersting. Measuring and guiding monosemanticity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.19382.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Automated interpretability metrics do not distinguish trained and random transformers. In *International Conference on Learning Representations (ICLR)*, 2026.

Christopher J. Hillar and Friedrich T. Sommer. When can dictionary learning uniquely recover sparse data from subsamples? *IEEE Transactions on Information Theory*, 61(11):6290–6297, 2015.

Mikkel Godsk Jørgensen and Lars Kai Hansen. Steering LLMs? actually, sparse autoencoders can outperform simple baselines. *arXiv preprint arXiv:2605.31183*, 2026.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? A case study in sparse probing. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29018–29049, 2025. arXiv:2502.16681.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29223–29264, 2025. arXiv:2503.09532.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Y. Rogov, Ivan Oseledets, and Elena Tutubalina. Sanity checks for sparse autoencoders: Do SAEs beat random baselines? *arXiv preprint arXiv:2602.14111*, 2026.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2502.04878.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, 2019.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. In *International Conference on Learning Representations (ICLR)*, 2025.

Tommaso Mencattini, Francesco Montagna, and Francesco Locatello. The rate-distortion-polysemanticity tradeoff in SAEs. *arXiv preprint arXiv:2605.14694*, 2026.

Anish Mudide, Joshua Engels, Eric J. Michaud, Max Tegmark, and Christian Schroeder de Witt. Efficient dictionary learning with switch sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2025.

Walter Nelson, Theofanis Karaletsos, and Francesco Locatello. Toward identifiable sparse autoencoders. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2605.31245.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2501.16615.

Gonçalo Paulo, Stepan Shabalin, and Nora Belrose. Transcoders beat sparse autoencoders for interpretability. *arXiv preprint arXiv:2501.18823*, 2025.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024a.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024b.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828.

Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Conference on Learning Theory (COLT)*, 2012.

Magamed Taimeskhanov, Samuel Vaiter, and Damien Garreau. Towards understanding steering strength. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.02712.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

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

Cynthia Van Hee, Els Lefever, and Véronique Hoste. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pp. 39–50, 2018.

Sohan Venkatesh and Ashish Mahendran Kurapath. On the non-identifiability of steering vectors in large language models. *arXiv preprint arXiv:2602.06801*, 2026.

Xu Wang, Yan Hu, Benyou Wang, and Difan Zou. Does higher interpretability imply better utility? a pairwise analysis on sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2026.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, 2025. Spotlight. arXiv:2501.17148.

Qi Zhang, Yifei Wang, Jingyi Cui, Xiang Pan, Qi Lei, Stefanie Jegelka, and Yisen Wang. Beyond interpretability: The gains of feature monosemanticity on model robustness. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.21331.

# A Replication Across Seeds, Widths, and Tasks

Section 6 reports one configuration. This appendix collects the four others we ran, together with the calibration check every run and task must pass before its number is read. It comes first because the scope of the evidence is what a reader has least reason to take on trust; the proofs deferred from Section 3 are in Appendix E.

## A.1 Seed and Width Robustness

The headline in Section 6 is a single seed on the 16k dictionary. Three post-hoc runs, otherwise identical (same grid, same pairings-per-cell, same $c = 0.02$ operating point), each change one factor: a second and third training seed (16k, seeds 123 and 999) and a second dictionary width (65k, seed 42). A fourth attempt changed the base model and produced no usable data at all; what survives of it, together with the random-dictionary calibration check we apply to every run and task below, is in Appendix A.2, which flags one cell of this table.

| Run | Irony slope [CI95] | SNLI slope [CI95] |
|---|---|---|
| Headline (seed 42, 16k, 12 draws) | $\mathbf{-0.298}$ $[-0.509, -0.089]$ | $-0.232$ $[-0.773, 0.135]$ |
| Seed check (seed 123, 16k) | $\mathbf{-0.187}$ $[-0.208, -0.144]$ | $-0.258$ $[-0.472, 0.016]$ |
| Seed check (seed 999, 16k) | $+0.060$ $[-0.091, 0.279]$ | $-0.123$ $[-0.361, 0.256]$ |
| Width check (seed 42, 65k) | $-0.008$ $[-0.073, 0.067]$ | $\mathbf{-0.948}$ $[-1.424, -0.354]$ |
| Task check (BoolQ, seed 42, 16k) | $-0.014$ $[-0.226, 0.208]$ (BoolQ only) | |

Table 3: Mix-vs-clone diff-in-differences slope on the primary DV, headline vs. two further training seeds, a second dictionary width, and a third task. **Bold** excludes zero. The headline row is the 12-draw extension of the run reported in Section 6, superseding an earlier 4-draw version whose irony CI was narrower than honest (Appendix B). The sign of the effect is reproduced on SNLI (and on BoolQ) by every configuration; its magnitude, and on irony its significance, depend on the configuration, which is what a basis-sensitive readout should be expected to do.

**The direction of the effect is the part that reproduces.** SNLI's point estimate is negative in every configuration we ran: $-0.232$ (headline) $\rightarrow$ $-0.258$ $\rightarrow$ $-0.123$ across three seeds, and $-0.948$ at 65k; BoolQ, a different task on the same pipeline, is also negative ($-0.014$). In five runs spanning three seeds' worth of retraining, a four-fold width change and a task swap, the dial never once reports polysemanticity *helping* SNLI or BoolQ. What moves is the magnitude and with it the significance: only the width check's SNLI cell excludes zero now, against two of four before the 12-draw extension, because the more-powered headline run resolved its own irony interval at the cost of no longer resolving SNLI's.

**Irony behaves as Section 6 said it would, with one caveat fixed and one left.** We flagged there that irony was the noisy task in Stage B on two independent grounds: its mix arm collapsing to two distinct draws rather than four, and its whole slope sitting on the top three $\lambda$ rungs while weakly positive below $\lambda = 0.30$. The first ground is now only partly true: extended to 12 draws, the same run resolves to five distinct outcomes rather than two, still significant and still passing the random-dictionary check (Table 4). The second ground is untouched, and a quantity with that profile should still be expected to be seed-sensitive, and it is: it reproduces with a tighter interval at seed 123 (which fails the calibration check, so does not independently confirm it) and returns a small positive, zero-crossing slope at seed 999. The third seed therefore confirms a caveat the paper already carried rather than contradicting a claim it made; what it rules out is reading irony's headline interval as a fixed property of the task. The **width swap** points the same moral in the other direction: at 65k irony collapses to noise while SNLI roughly doubles.

We read the three checks as consistent with Corollary 1 rather than as undermining it. The corollary itself says only that reconstruction functionals are constant on the orbit; it predicts nothing about how a basis-sensitive readout moves across seeds, and a paper claiming one seed's interval was a property of SAEs in general would be overreading its own evidence. What the dial buys is not a configuration-free constant but the ability to *measure* the quantity at a stated configuration at all, which no reconstruction metric can do at any configuration, since it is constant on the entire orbit by Observation 1.

### A.2 A Configuration That Produced No Result, and the Check the Others Must Pass

**What was attempted.** A fourth post-hoc run swapped the base model (`gemma-2-2b`, layer 12, 16k) with the grid and $c = 0.02$ fixed, and produced no usable data: its launch log ends inside the model load, and the one checkpoint carrying its tag records the default configuration in its signature, a fragment of an earlier attempt whose override never reached the worker. We report no number from it anywhere, and take from it only that *a run tag names a file, while the stored signature certifies a configuration*, which is why every artifact here is keyed by signature. Whether the dial transports across base models, with $c$ recalibrated on the target model (Section 5), remains the most important open extension.

**The check the runs we *do* report must pass.** The dial's readout is only interpretable where it has dynamic range, and the largest basis perturbation available is the untrained random dictionary of Appendix D.3: a dictionary that was never trained, selected and scored through exactly the same pipeline. If a run's random-dictionary anchor falls *inside* that run's own $\lambda$-sweep, then a wholly untrained basis is not separated from the trained one on that run's *ruler*, our term for the range of readout values spanned by the run's own mix and clone arms across the $\lambda$ grid, against which any other dictionary is placed; movement of one $\lambda$ rung within the band cannot be read as a change in basis quality. Table 4 applies this to all nine run $\times$ task cells behind Table 3.

| Run | Task | own $\lambda$-sweep | | random dict. | margin | verdict |
|---|---|---|---|---|---|---|
| | | min | max | (untrained) | (min − rand.) | |
| headline (16k, s42, 12 draws) | irony | $-0.117$ | $+0.018$ | $-0.189$ | $+0.072$ | separated |
| headline (16k, s42, 12 draws) | SNLI | $-0.145$ | $-0.011$ | $-0.316$ | $+0.171$ | separated |
| seed 123 (16k) | irony | $-0.155$ | $-0.033$ | $-0.055$ | $-0.100$ | **fails** |
| seed 123 (16k) | SNLI | $-0.013$ | $+0.348$ | $-0.384$ | $+0.371$ | separated |
| seed 999 (16k) | irony | $-0.029$ | $+0.012$ | $-0.238$ | $+0.209$ | separated |
| seed 999 (16k) | SNLI | $-0.041$ | $+0.085$ | $-0.305$ | $+0.264$ | separated |
| 65k (s42) | irony | $+0.013$ | $+0.133$ | $-0.241$ | $+0.254$ | separated |
| 65k (s42) | SNLI | $+0.226$ | $+0.767$ | $-0.124$ | $+0.351$ | separated |
| BoolQ (16k, s42) | BoolQ | $-0.010$ | $+0.194$ | $-0.181$ | $+0.170$ | separated |

Table 4: The random-dictionary calibration check, applied to every run and task in Table 3. Columns are the signed target-logit shift at $c = 0.02$: the range spanned by the run's own mix and clone arms across all eight $\lambda$ rungs, and the untrained random dictionary scored identically at $k$=10. A positive margin means the untrained basis lies outside the run's sweep and the ruler resolves it; a negative margin means it does not. Eight of nine cells separate. Seed 123's irony cells do not, and we report that rather than remove them.

**Eight of nine cells pass; seed 123 on irony does not, and we keep it.** On that cell the untrained dictionary scores $-0.055$, above the midpoint of the run's own sweep $[-0.155, -0.033]$. Its slope ($-0.187$ $[-0.208, -0.144]$, Table 3) is therefore not evidence for the hypothesis, and it is not evidence against it either — it is a reading taken on a ruler that did not resolve. This is costly to us and we state the cost exactly: seed 123 is the *only* one of the three post-hoc runs that reproduced irony's significance, so once its irony cells are discounted, **irony's significant negative slope rests on the headline run alone**, while SNLI's negative sign holds across all four runs with every SNLI cell passing the check. We flag the cell rather than delete it, because a cell removed after its slope was seen is a cherry-pick whatever the stated reason, and a reader who disagrees with our judgement needs the number in order to disagree.

**Why a reconstruction metric cannot do this.** Every check in this appendix, the signature mismatch that unmasked the stray fragment and the anchor margin that flags seed 123's irony cells, is available only because the dial's readout is basis-sensitive. A reconstruction metric does score the untrained random dictionary differently, but only because its reconstruction is worse; the random

dictionary is not on the trained one's orbit. What it cannot do is separate bases at fixed reconstruction: by Observation 1 an MSE, FVU, CE-recovered or patch-KL number is *identical* on the trained dictionary and on every reconstruction-exact remix of it, the deliberately degraded $\widetilde{M}_{0.45}$ included, so it carries no information about whether a ruler resolves basis differences. That is Observation 1 restated as an operational hazard, and it is the practical reason the sort of Section 3 matters.

# B Experimental Setup

Figure 6 gives the run protocol end to end.

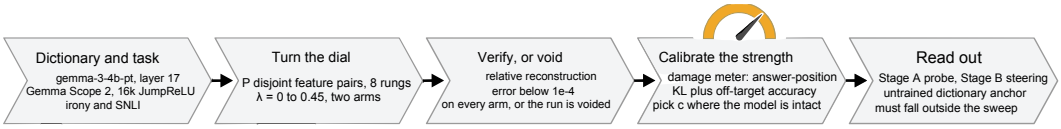



Figure 6: The run protocol end to end: dictionary and task, turn the dial, verify or void, calibrate the strength, read out. The dial itself is Figure 2; each stage is given in full below.

## B.1 Model and dictionary

The model is `google/gemma-3-4b-pt` (Gemma Team, Google DeepMind, 2025) in bf16, with the SAE from Gemma Scope 2 (Google DeepMind, 2025), release `gemma-scope-2-4b-pt-res`, 16k width, JumpReLU (Rajamanoharan et al., 2024b), medium sparsity, layer 17 (the SAE's native layer), seed 42. **No SAE was trained for this work.** A dead-latent census (300 fitting samples per task) counts what the clone arm's dead-slot pool actually is: 98.9% dead on irony (16202/16384, 182 live), 99.1% on SNLI (16243/16384, 141 live), 97.99% on BoolQ (16054/16384, 330 live). Stage B's clone arm consumes 10 of those slots per pairing ($\approx 0.062\%$ of the dead pool on every task); Stage A consumes 20 ($\approx 0.12\%$). The rarest live latent fires on 0.33% of fitting samples on every task, so the dead/live boundary is not a single fragile feature.

**Dead on the fitting corpus, and what a held-out reactivation would cost.** Both the census and the per-cell `assert_clone_slots_dead` guard that gates every clone arm are computed on the fitting corpus. A slot could therefore fire on the held-out evaluation split without either catching it, and we did not re-run the assertion there. Two things bound what that could cost. A slot with zero firings in the 300 fitting samples carries a 95% upper bound of 1% on its firing rate, and it is drawn from the 98–99% of the dictionary that never fired at all. More to the point, a reactivation enters as a level and not as a slope. The clone arm overwrites the slot's code with $\lambda a_i$ at every rung including $\lambda = 0$, so whatever that slot would have contributed to $\hat{x}$ is removed by the same amount at every $\lambda$, and the difference-in-differences subtracts terms that do not vary with the dose. To bias the estimand a reactivation would have to interact with $\lambda$.

**Grid sizes.** Stage A runs over 20 feature pairs $\times$ 10 pairing draws; Stage B over 10 pairs $\times$ 12 independently seeded draws, with $k \in \{10, 20\}$ and the 8-point $\lambda$ grid, for 384 cells per run. Every reported steering number is at $k$=10. The $k$=20 direction entered only the best-flip-rate maximum over $k$ and coefficient and was not stored separately, so no $k$=20 value, real or null, can be reported; Proposition 3 fixes it analytically.

## B.2 Tasks and data

## B.3 Prompting and extraction

Tasks are posed as few-shot multiple-choice classification with a fixed `Question: ...\nAnswer:` template and a class verbalizer, and the activation is read at the final `Answer:` token. Predictions are the argmax over the verbalizer token logits. Irony and BoolQ prompts carry four demonstrations and SNLI prompts six, two per class, drawn label-balanced from each task's training split under a fixed seed. The steered evaluation set is the 200 label-balanced held-out examples; a steering example qualifies when its baseline prediction lies in the steered class pair, which

| Name | Source | Cls. | Role |
|---|---|---|---|
| irony | `tweet_eval` | 2 | primary pragmatic |
| snli | `snli` | 3 | primary semantic |
| boolq | `boolq` | 2 | Stage A; Stage B added post hoc |

Table 5: Task provenance: `tweet_eval` config `irony` (Barbieri et al., 2020); `snli` (Bowman et al., 2015); `google/boolq` (Clark et al., 2019). All splits are label-balanced by construction. BoolQ's Stage B run (Appendix A.1) was added after an earlier Stage B design, which scored only irony and SNLI.

holds for all 200 on irony and 187 on SNLI. Unsteered accuracy, measured by the validity diagnostic of Section 5 on a label-balanced subset of the same seeded draw (100 irony and 99 SNLI examples), is 0.77 on irony (chance 0.50) and 0.68 on SNLI (chance 0.33); it was not stored for BoolQ.

### B.4 Steering-vector construction, and where the coefficient acts

Every Stage B number in this paper comes from the following intervention, stated here in full because a steering result is not interpretable without it.

**The direction.** Within a cell, latents are ranked by the same class-discriminativeness score used for target selection, and the top $k \in \{10, 20\}$ are kept.[2] The direction is the $\Delta$-weighted sum of their decoder rows, $d_k = \sum_{i \in \text{top-}k} \left(\overline{a}_i^{\text{high}} - \overline{a}_i^{\text{low}}\right) w_i$, normalised to $\|d_k\|_2 = 1$. The ranking is restricted to the cell's own manipulated index set, the targets together with their partners or dead slots, and that restriction is load-bearing rather than cosmetic. Ranking over the full dictionary lets untouched features rise into the global top-$k$ and replace the damaged targets as $\lambda$ grows, which happens at different rates in the two arms: by $\lambda = 0.4$ on SNLI only 1 of 10 damaged targets survived in the global top-10 in the mix arm against 4 of 10 in the clone arm. That asymmetry manufactures a difference-in-differences signal out of feature substitution rather than basis quality. Restricting the ranking keeps the direction tied to the features the experiment actually damaged.

The score is activation-side: it ranks latents by how their codes separate the classes, not by what they do to the output. Arad et al. (2025) find that activation-side scores tend to pick *input* features with weak output effect, and that filtering on an output score improves SAE steering two- to threefold. That bears on the weak absolute SAE effect of Appendix D.3 more than on $\tau$, whose two arms share the selection rule; whether an output-score ranking changes the slope is untested here. Chalnev et al. (2024) go further and optimise the steering vector to hit chosen SAE features while suppressing others; we deliberately use the plain decoder-row sum, since an optimised vector would adapt to each arm's basis and absorb part of the difference the dial is built to measure.

**The coefficient.** Let $s$ be the mean $\ell_2$ norm of the layer-17 activations over the fitting corpus. A forward hook replaces the layer-17 residual stream $h$ by $h + c\, s\, \sigma_e\, d_k$, where $\sigma_e \in \{\pm 1\}$ is a per-example sign that points the direction toward the class the model did *not* predict at baseline. Because $d_k$ is unit-norm and the added vector is scaled by $s$, the coefficient $c$ *is* the fraction of the mean activation norm being added, which is the property Table 9 says most reported strengths lack. There is no per-token rescaling and no interaction with layer normalisation is modelled: the addition is made to the raw residual stream, and whatever the downstream norms do to it is part of what the diagnostic measures.

**Where it is applied, and where it is read.** The vector is added at *every* token position of the sequence, broadcast over the time axis, while the dependent variable is read at the last non-padding token, which is the answer position. We did not run an answer-position-only variant, so whether a more surgical placement would move the operating-point boundary is untested here. Two things bound how much that matters for our estimand. The difference-in-differences subtracts two arms that receive the identical placement, so a placement effect has to interact with $\lambda$ to bias $\tau$. And the damage meter of Section 5 reads KL at the answer position, so the regime labels already describe the placement actually used rather than an idealised one.

[2] At $k$=20=2$P$ the ranking exhausts the manipulated set, and $d_{20}$ is then provably identical at every $\lambda$ in both arms, so $\tau \equiv 0$ (Proposition 3). The $k$=20 pass fed only the best-flip-rate DV and was not retained per cell. Running the experiment's own construction and direction code on synthetic SAEs (36 draws, both arms, the full $\lambda$ grid) gives $\max_\lambda \|d_{20}(\lambda) - d_{20}(0)\|_\infty < 2 \times 10^{-6}$, float32 round-off, against 0.12–0.26 at $k$=10.

**The clone arm's duplicated decoder rows.** The dead slot paired with target $i$ is given a copy of row $w_i$, so the two rows are exactly identical and, inheriting $\|w_i\| = 1$, exactly unit-norm. The pair is therefore perfectly collinear, which is the intended content of the arm: "two units that still mean only feature $i$" is a statement about collinear decoder directions. It creates no conditioning problem in the forward pass, since $W$ is applied and never inverted there, and $M^{-1}$ acts on the code side where it is block-diagonal with $2 \times 2$ blocks whose conditioning is governed by $1/(1 - 2\lambda)$ and is checked numerically on every arm (Appendix C). In the steering direction an exactly duplicated row contributes along $w_i$ with the summed weight of both slots rather than adding a new direction, so it changes that component's magnitude and not the span. The mix arm is built by the same routine on the same targets, which is why $\tau$ differences the construction out.

### B.5 Compute

All inference ran locally on an Apple M4 Pro (24 GB unified memory, MPS backend) except the two Stage B steering grids, which ran on a single cloud NVIDIA L4. Representative costs: the headline 128-cell steering grid took 4.23 h on the L4; the validity diagnostic that produced Table 10 took 53 min. The entire empirical corpus is reproducible on a few consumer GPU-hours plus a laptop, which we note because it bears on how cheaply the diagnostics we recommend can be adopted.

## C Statistical Protocol

**Intervals.** All confidence intervals are 95% percentile bootstrap intervals (Efron & Tibshirani, 1993), resampled over pairing draws for the dial's slopes, which is why the note on effective sample size in the Limitations is load-bearing. Where a bootstrap distribution is degenerate, as it is for the clone arm, which does not depend on the pairing draw, the code raises rather than reporting a zero-width interval as evidence.

**Pre-registration.** The dial's H1 was fixed before the corresponding runs, with numeric thresholds. One reading of it, at $c = 1.0$, is reported as having failed for a reason that turned out to invalidate the measurement rather than the hypothesis (Section 5). Post-hoc analyses are labelled as such wherever they appear. The record is a dated internal planning document, not an entry in a public registry, so a reader cannot verify its timing independently; no claim here rests on pre-registration alone.

**Verification of reconstruction exactness.** Every constructed arm asserts $\|\hat{x}' - \hat{x}\|_\infty / \|\hat{x}\|_\infty < 10^{-4}$ against the unmodified SAE and voids the run otherwise. This is the operational form of Observation 1 under floating point.

**Why that check is the empirical form of the metric sort.** It is fair to ask for the left block of Table 1 to be confirmed by measurement rather than only derived, by tabulating loss recovered or FVU at each $\lambda$ and showing a flat line. The check above already does that, and does it more stringently than a per-metric table would. Every metric in the left block is a function of the pair $(x, \hat{x})$, so a bound on $\|\hat{x}' - \hat{x}\|_\infty$ holding pointwise over the corpus certifies all of them at once, to that tolerance, including metrics nobody has proposed yet. A table of loss-recovered values is the weaker test in both directions: it constrains one scalar summary of $\hat{x}$, which can agree by cancellation at places where $\hat{x}$ itself differs, and it says nothing about the other rows. We therefore verify the input every such metric reads instead of sampling the metrics, and the reported flatness of any particular one would be a consequence of the assertion rather than independent evidence for it. The assertion is enforced on every constructed arm at every $\lambda$ rung in both arms on all three tasks, and no run in this paper was voided by it.

## D The Dial in Full

### D.1 Stage A: all dependent variables

Stage A scores eight slope-tested dependent variables per task. The body quotes two of them. All eight are in Table 7, including the two that move *against* the prediction.

**Two DVs are computed on the manipulated subset, and one of them has a bug we did not introduce but must disclose.** Six of the eight read the whole code; concept entropy and concept

| $\lambda$ | 0.00 | 0.05 | 0.10 | 0.15 | 0.20 | 0.30 | 0.40 | 0.45 |
|---|---|---|---|---|---|---|---|---|
| Irony | 0 | $-.000$ | $+.012$ | $+.017$ | $+.030$ | $-.012$ | $-.106$ | $\mathbf{-.130}$ |
| SNLI | 0 | $-.012$ | $-.009$ | $-.014$ | $-.030$ | $-.076$ | $-.071$ | $\mathbf{-.108}$ |

Table 6: Mix-vs-clone difference-in-differences on the primary DV (signed target-logit shift at $k$=10, $c = 0.02$), per $\lambda$ rung, on the 12-draw headline run. The $\lambda$=0 cell is shared by both arms, so the estimand reduces exactly to $\mathrm{mix}(\lambda) - \mathrm{clone}(\lambda)$. Negative supports H1. These are the cells underlying Figure 5 and the slopes quoted in Section 6. Because irony's response is not linear in $\lambda$, we also report, post hoc, the $\lambda$=0.45 endpoint contrast with the same bootstrap over draws: $-0.130$ $[-0.167, -0.096]$ on irony (11 of 12 draws negative, five distinct outcomes, so the interval is optimistic) and $-0.108$ $[-0.194, -0.032]$ on SNLI (9 of 12 negative). It is a robustness check on the slope's functional form, not a replacement for the pre-registered slope.

| DV | predicted | irony | SNLI | BoolQ | status |
|---|---|---|---|---|---|
| *whole code* | | | | | |
| Dense probe AUC (control) | flat | $-0.010$ $[-0.017, 0.001]$ | $-0.001$ $[-0.005, 0.003]$ | $-0.012$ $[-0.033, 0.006]$ | valid |
| FMS | ↓ | $\mathbf{-0.064}$ $[-0.080, -0.032]$ | $\mathbf{-0.072}$ $[-0.110, -0.009]$ | $\mathbf{-0.082}$ $[-0.122, -0.059]$ | valid |
| $k$-sparse AUC, $k$=1 | ↓ | $+0.050$ $[-0.008, 0.088]$ | 0.000 (degen.) | $\mathbf{+0.599}$ $[0.167, 0.775]$ | valid |
| $k$-sparse AUC, $k$=2 | ↓ | $-0.015$ $[-0.059, 0.017]$ | $+0.041$ (degen.) | $\mathbf{+0.245}$ $[0.073, 0.311]$ | valid |
| $k$-sparse AUC, $k$=4 | ↓ | $-0.026$ $[-0.065, 0.049]$ | $-0.002$ (degen.) | $-0.007$ $[-0.034, 0.013]$ | valid |
| $k$-sparse AUC, $k$=8 | ↓ | $\mathbf{-0.252}$ $[-0.339, -0.066]$ | $\mathbf{+0.012}$ $[0.001, 0.035]$ | $\mathbf{-0.036}$ $[-0.061, -0.011]$ | valid |
| *manipulated subset only; $\lambda$=0 cells not shared* | | | | | |
| Concept entropy | ↑ | $-0.515$ $[-0.515, -0.515]$ | $-0.478$ $[-0.487, -0.469]$ | $-0.466$ $[-0.474, -0.454]$ | **invalid** |
| $\lambda > 0$ only | ↑ | $\mathbf{+0.020}$ $[0.020, 0.020]$ | $\mathbf{+0.038}$ $[0.030, 0.046]$ | $\mathbf{+0.035}$ $[0.028, 0.045]$ | corrected |
| Concept separability | ↓ | $\mathbf{+0.077}$ $[0.074, 0.082]$ | $\mathbf{+0.098}$ $[0.084, 0.112]$ | $\mathbf{+0.069}$ $[0.064, 0.077]$ | suspect |

Table 7: Every Stage A difference-in-differences slope, 20 pairs × 10 pairing draws, 95% bootstrap CI. **Bold** excludes zero. "Predicted" is the pre-registered direction under H1; degenerate rows are zero-width intervals from a saturated DV and support nothing either way. The status column marks the concept-entropy row **invalid** at the point of use: it is contaminated by the dead-latent guard described below, which is why the $\lambda > 0$ row exists. No claim in this paper rests on either of the last two rows. FMS is the Feature Monosemanticity Score of Härle et al. (2025); concept separability follows Fereidouni et al. (2026).

separability average over the manipulated latents only, and that set differs between arms by construction (mix: targets ∪ partners; clone: targets ∪ dead slots). Their $\lambda = 0$ cells are therefore *not* shared, which breaks the difference-in-differences at that one rung. Concretely: our entropy implementation guards dead features with the test `tot < eps` where `tot = mass.sum() + eps`, a condition that can never fire, so a dead latent scores $H = 0$ ("perfectly monosemantic") rather than `NaN`. The clone arm's $\lambda = 0$ cell is half dead slots, reads $0.488$, and jumps to $0.976$ as soon as $\lambda > 0$ brings those slots to life. That single artifactual step is the entire $-0.515$ slope. Restricted to $\lambda > 0$, where both arms' manipulated sets are live and the guard is irrelevant, the difference-in-differences *rises* monotonically in $\lambda$ on all three tasks, which is the pre-registered H1 direction. We report both rows and rely on neither: the body's claims rest on the six whole-code DVs.

**Concept separability reverses, and we do not have a verified explanation.** Its slope is positive and excludes zero on all three tasks, on the full grid and restricted to $\lambda > 0$ alike, where H1 predicts a fall. The clone arm is exactly flat (its own dead-feature guard, a standard-deviation test, does work); the mix arm rises to a peak near $\lambda = 0.30$ and then falls back, so the DV is not even monotone in the dose. The explanation we find most likely is compositional rather than substantive: the mix arm averages over targets *and partners*, and partners are selected for minimum co-activation, hence fire rarely and separate concepts poorly on their own; mixing injects a highly separable target profile into them, which raises the subset average even as each individual unit becomes less monosemantic. We flag this as a conjecture, since the stored artifact retains per-arm means rather than per-feature values, so we cannot verify it without re-running Stage A, and we do not use this DV to support any claim. A metric that rises when polysemanticity is injected on purpose is a finding about the metric, and one an SAE evaluation suite should want to know about.

**Two further quantities are computed and do not resolve.** Label efficiency (labels needed to reach a target AUC) never reaches its target on irony at any $\lambda$ in either arm, is pinned at 8 on SNLI in both arms at every $\lambda$, and on BoolQ moves in opposite directions in the two arms for reasons the

design does not identify. Mean $L_0$ is reported in Appendix D.2. Neither is slope-tested and neither enters any claim.

### D.2 $L_0$ IN BOTH ARMS

| | irony | | SNLI | | BoolQ | |
|---|---|---|---|---|---|---|
| | mix | clone | mix | clone | mix | clone |
| $\lambda = 0$ (shared) | 50.848 | | 51.028 | | 53.517 | |
| every $\lambda \in (0, 0.45]$ | 60.056 | 63.647 | 59.527 | 63.523 | 61.912 | 62.003 |
| inflation | +9.208 | +12.798 | +8.499 | +12.495 | +8.396 | +8.487 |

Table 8: Mean $L_0$ on the fit corpus, by arm. The second row is a single number because the value is identical to four decimal places at every one of the seven nonzero $\lambda$ rungs, in both arms, on all three tasks, the empirical face of Corollary 2. Two consequences. First, the sparsity confound cannot produce a dose-response, because it does not vary with the dose. Second, where the arms mismatch it is the *control* that is more inflated, so the mismatch is conservative; and BoolQ, whose partner pool happens to match better, closes the gap to 0.09 latents (0.15%), which shows the mismatch is a property of the co-activation structure available in a given task rather than of the design.

**The near-matched task is the ablation the residual mismatch calls for.** A reader who grants Corollary 2 in principle may still want the slope shown to be insensitive to the mismatch in practice. Table 8 supplies most of an answer without extra compute, because the three tasks happen to span the mismatch by two orders of magnitude. Irony and SNLI leave gaps of 3.590 and 3.996 latents between the arms' inflations. BoolQ leaves 0.091, which is 0.15% of its mean $L_0$ and close enough to call the arms matched.

If the mismatch were producing the dose-response, the task where it nearly vanishes should behave differently from the two where it is large. It does not. BoolQ's estimand is negative at all three stored coefficients, $-0.014$, $-1.061$ and $-1.484$, and at $c = 0.10$ its interval excludes zero (Table 11), making it the third task to resolve. Direction and resolution both survive the removal of the confound. The mismatch that remains on irony and SNLI runs the wrong way to help us in any case, the control arm being the more inflated of the two, so support inflation that costs steering utility would push $\tau$ positive and against H1.

Two honest limits on this argument. It is a near-match arising across tasks rather than an auxiliary control arm matched within one, so task identity is not held fixed and BoolQ differs from the other two in more than its co-activation structure. And BoolQ's $c = 0.02$ point estimate is small enough to carry little weight on its own. What the comparison supports is the narrower claim that the sign does not depend on the mismatch being present, which is what Corollary 2 predicts and what the objection asks to see.

**What else moves besides $L_0$.** Support inflation is not the only non-basis quantity the dial disturbs, and the second one is worth naming because the control does not match it. Restoring unit decoder rows (Proposition 1) multiplies each manipulated code by the norm of its own de-mixed decoder row, $s = (1 - 2\lambda)^{-1}\|(1 - \lambda)w_i - \lambda w_j\|_2$ in the mix arm and exactly 1 in the clone arm, whose duplicated row is a copy of a unit-norm row. That factor grows with $\lambda$, so unlike $L_0$ it is not constant in the dose. It is forced rather than chosen: drop it and the mixed dictionary violates the unit-norm criterion of Definition 2, and a decoder-norm change masquerades as a basis effect.

The geometry deserves stating plainly, because it is not what "mixing" suggests. The code side blends, $a'_i = (1 - \lambda)a_i + \lambda a_j$, but the decoder side must invert that blend, so the mixed rows are *contrasts*, $w'_i \propto (1 - \lambda)w_i - \lambda w_j$ and $w'_j \propto (1 - \lambda)w_j - \lambda w_i$.[3] With $c = \langle w_i, w_j \rangle$ the two mixed rows have cosine $\big(c(1-2\lambda+2\lambda^2) - 2\lambda(1-\lambda)\big) / \big(1-2\lambda+2\lambda^2-2\lambda(1-\lambda)c\big)$, which tends to $-1$ as $\lambda \to \frac{1}{2}^-$ whatever $c$ is. For a near-orthogonal pair ($c \approx 0$) at the top rung $\lambda = 0.45$ this is $\approx -0.98$, and the rescaling factor above is $s \approx 7.1$: the two mixed latents write nearly opposite directions,

[3] A nonnegative matrix has a nonnegative inverse only if it is monomial (a permuted positive diagonal). $M_\lambda$ is nonnegative and, for $\lambda > 0$, not monomial, so $M_\lambda^{-1}$ must carry negative entries; nonnegative codes and a decoder built only from nonnegative blends of the original rows are jointly impossible on this orbit.

each with a code about seven times its original scale. Both figures are closed-form evaluations at $c = 0$, not measurements; the realised pairs' inner products were not stored. The mix arm is therefore a stronger perturbation of the latent geometry than a convex blend would be, which is part of what $\tau$ measures sensitivity to and is a reason not to read its magnitude as the cost of natural polysemanticity.

Two things bound what it can do to the estimand. It is the same factor for both members of a pair and varies between pairs only through $\langle w_i, w_j \rangle$, since $(1 - 2\lambda)^{-1}$ and the $\lambda$-dependent weights are common to every pair; a factor common to the manipulated set cancels in the ranking that selects the top $k$ and again in the unit normalisation of $d_k$, leaving a reweighting of order the pairs' decoder inner products. And the steering coefficient is applied to that unit-norm direction and scaled by the mean activation norm, so both arms are steered at the same magnitude whatever their codes did. What we have not run is the separation itself, a sweep with the renormalisation disabled, which would isolate the two effects at the price of leaving the object outside the admissible class. We report the rescaling as a disclosed asymmetry rather than an excluded one.

### D.3 Ceiling anchors and the $\lambda$-ruler

Ceiling anchors place non-SAE bases on the calibrated $\lambda$-axis: raw class-mean-difference directions, single neurons, an untrained random SAE, and matched-norm Gaussian directions. Where an anchor's score falls inside the SAE's own $\lambda$-sweep it can be reported as a $\lambda$-*equivalent*. **The ruler largely fails to resolve, which we report rather than suppress.** All values below are computed on the headline run's 12 pairing draws. On irony every logit-shift $\lambda$-equivalent is undefined: the mix arm's sweep occupies a narrow band ($[-0.117, -0.004]$; $[-0.117, +0.018]$ with the clone arm, Table 4) and every external basis falls outside it — raw class-mean ($+0.132$) and neurons ($+0.271$) above, random SAE ($-0.189$) below. On SNLI exactly one logit-shift anchor resolves. The flip DV does resolve anchors (class-mean and neurons at $\lambda \approx 0.32$–$0.38$ on irony, the *untrained* random dictionary at $\lambda \approx 0.35$ on SNLI), but a ruler on which an untrained dictionary earns a $\lambda$-equivalent is reading saturation rather than basis quality (Figure 7), so we do not interpret it. The one logit-shift anchor is the single calibrated cross-basis statement this design produces: *on SNLI, using raw neurons at $k$=10 instead of trained SAE latents costs about as much steerability as $\lambda \approx 0.23$ of basis mixing.* It does not generalise, since on irony neurons beat the SAE at every $\lambda$, and it is stated as one task's calibration rather than a law.

**Absolute steering effects at the calibrated operating point.** At $c = 0.02$, with every direction unit-norm at equal magnitude, the class-mean difference moves the target logit by $+0.132$ $[0.081, 0.183]$ (irony) and $+0.460$ $[0.388, 0.533]$ (SNLI), and a trained probe by $+0.236$ $[0.167, 0.306]$ (SNLI), while the SAE's top-10 latents sit at $-0.028$ $[-0.044, -0.011]$ and at $-0.010$ with an interval containing zero; generic perturbations run $-0.10$ to $-0.50$. Inside the calibrated regime the SAE therefore clears the random-direction floor and loses to two simple baselines, converging with Wu et al. (2025). Call this raw level the *target-logit displacement*, the same signed target-logit shift read at $\lambda = 0$ rather than differenced across $\lambda$; the name is used only here, to keep the level apart from the estimand. On it the SAE direction is negative on irony and indistinguishable from zero on SNLI. A negative value here is not a push the wrong way. The DV is a raw logit, not a logit difference, and every generic unit-norm perturbation at this magnitude lowers it: the untrained random dictionary and matched-norm Gaussian directions sit at $-0.10$ to $-0.50$. The SAE direction lies above that floor on both tasks, so relative to generic damage it moves toward the target, only far less than the class-mean direction. This does not undercut the dial, whose estimand is a *difference* between two SAEs built by the identical selection and steering procedure and differing only in basis mixing, and is negative even where either arm's absolute effect is small. The dial makes no claim that a ranked-latent SAE direction is the efficient way to move this DV, and a class-mean direction, built from no SAE's dictionary, is not a competitor for RQ4 as posed. We report the comparison because a reader benchmarking absolute steering strength against a non-SAE baseline needs it, and because it bounds how much the identified sensitivity can matter operationally at this layer and strength.

### D.4 One grid, two dependent variables

Figure 7 shows the saturation mechanism behind the secondary-DV reversal reported in Section 5.

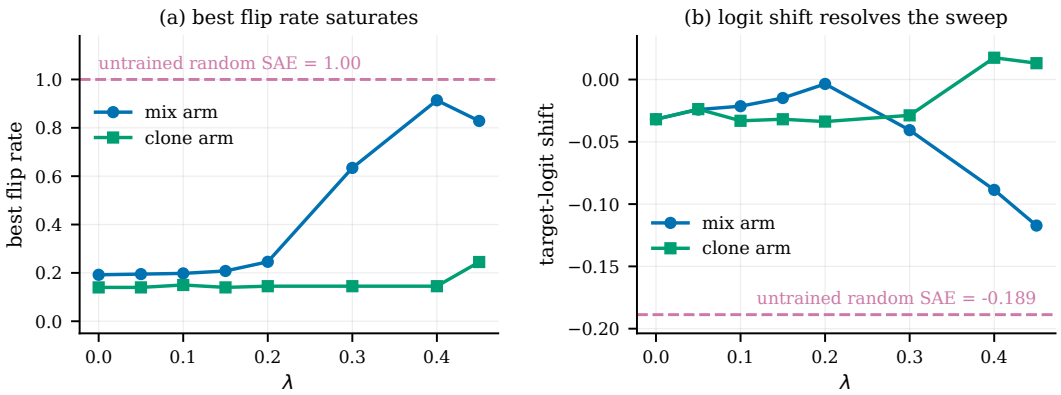



Figure 7: One grid, two dependent variables, opposite conclusions (irony). (a) Best flip rate saturates at the ceiling, and an *untrained random dictionary* scores the maximum, 1.0. (b) The signed logit shift resolves the whole sweep, and the same random dictionary scores $-0.189$, off-scale in the bad direction, below every $\lambda$ the real SAE reaches. One baseline, opposite extremes, depending only on which DV is read.

| Work | Strength reported | Units | Recoverable as a fraction of activation? | Damage diagnostic reported |
|---|---|---|---|---|
| Turner et al. (2023) (ActAdd) | $\|c\|$ typically $<15$; $c{=}+10$ effective | raw additive coefficient on a prompt-pair activation difference | **Yes**; their Appendix F puts $c{=}+10$ at $\approx 10\times$ the residual-stream norm | Perplexity on unrelated text; KL of outputs vs. a random vector |
| Rimsky et al. (2024) (CAA) | multiplier $\pm 1$ (capabilities) to $\pm 2$ (main) | multiplier on a difference-in-means vector, norms standardised across behaviours but not layers | No; layer norm deliberately not divided out | MMLU and TruthfulQA off-target; degraded open-ended text quality at larger multipliers |
| Templeton et al. (2024) | clamp to $10\times$ the feature's max activation | multiple of one feature's own maximum | No | None reported |
| Wu et al. (2025) (AxBench) | $\alpha$ swept to 100; *best* $\alpha$ reported per concept | factor $\times$ max activation over the eval set | No | Fluency and instruction-following scores; larger factors monotonically reduce them |
| Taimeskhanov et al. (2026) | $\alpha$ to 200–2000 | unnormalised difference-of-means | No | Cross-entropy law, derived theoretically |
| **This work** | $c{=}0.02$ (headline); $c{=}1.0$ (pre-registered) | $c\times$ mean activation norm, direction unit-norm | **Yes, by construction**; $c$ *is* that fraction | Answer-position KL and off-target $\Delta$accuracy, per cell |

Table 9: The coefficient audit behind Section 5. Two of six works state a strength that can be placed on a common axis; the other four are stated in units that cannot be converted without re-running the model. Four of six report some damage diagnostic, but only two report one at the answer position, which is where a task DV is read.

### D.5 Reported steering-strength parameterisations and accompanying validity diagnostics

Section 5 rests on the claim that a steering result's operating point is usually not recoverable from the paper that reports it. That claim is checkable, so we checked it. We took the steering works this paper already engages with, plus the two most-cited activation-addition methods, and recorded for each: the intervention strength as the authors state it, the units that strength is stated in, whether those units can be converted to a fraction of the activation the vector is added to, and whether any diagnostic of model damage accompanies the result (Table 9).

Three things follow, and only the third is comfortable for us. *First, there is no $c{\approx}1$ convention to indict.* An earlier draft of this paper asserted one; the audit does not support it, and we have removed the claim (Appendix F). What the audit supports is weaker in form and worse in substance: strengths span more than two orders of magnitude and are quoted in units that are mostly not interconvertible, so the field has not converged on a bad operating point so much as failed to establish a shared axis on which any operating point could be named. *Second, where the units* are *convertible, the reported strength is large.* ActAdd's own appendix estimates that at its effective coefficient "around 90% of the residual stream is determined by the steering vector", an order of magnitude past our $c{=}1.0$,

| | Irony | | SNLI | |
|---|---|---|---|---|
| $c$ | DiD | regime | DiD | regime |
| 0.02 | $\mathbf{-0.104}$ | intact | $\mathbf{-0.199}$ | intact |
| 0.05 | $\mathbf{-1.192}$ | intact | $\mathbf{-0.528}$ | mild |
| 0.10 | $\mathbf{-0.717}$ | mild | $\mathbf{-1.428}$ | broken |
| 0.25 | $\mathbf{-1.908}$ | broken | $\mathbf{-1.698}$ | broken |
| 0.50 | $+4.000$ | broken | $-1.250$ | broken |
| 1.00 | $+4.816$ | broken | $+3.154$ | broken |
| 2.00 | $+3.553$ | broken | $+0.144$ | broken |

Table 10: The same estimand ($\lambda$=0.45, pairing 0) read at seven steering coefficients. Negative supports H1. Regime is assigned by answer-position KL from unsteered behaviour. The sign is stable throughout the intact and mild regimes and flips only once the model has stopped computing the task.

which is the point at which our meter reports the model has stopped answering the question. Our pre-registered reading was not an outlier; it was ordinary. *Third, the diagnostics that do exist are mostly not at the answer position.* Perplexity on unrelated text, MMLU, and fluency scores all detect gross incapacitation, and Wu et al. (2025) in particular report that larger factors monotonically degrade instruction-following. None of them would catch the specific failure documented in Section 5, where the model retains fluent general competence while the verbalizer mass at the one position the DV is read from has collapsed by three to six orders of magnitude. That is the gap the meter of Section 5 is meant to close, and the audit is why we think it is worth the two extra numbers it costs.

### D.6 The steering-strength sweep behind Figure 3

Table 10 is the numeric form of Figure 3's top panel, with the regime labels the meter assigns. Regimes are read off answer-position KL from unsteered behaviour: intact $< 0.1$ nat, mild $< 1$ nat, broken otherwise.

**How much the bin boundaries carry.** The bins are descriptive, and it is reasonable to ask what changes if they move. Very little, for a structural reason visible in the table: the sign of the estimand does not change at either boundary. The intact/mild cut carries nothing at all, since the estimand is negative on both sides of it on both tasks. Only the mild/broken cut could matter, and the sign survives at least one full rung past it on each task. On irony the cut falls between $c = 0.10$ and $c = 0.25$ while the sign holds negative through $c = 0.25$, a rung already labelled broken, and flips only at $c = 0.50$. On SNLI the cut falls between $c = 0.05$ and $c = 0.10$ while the sign holds negative through $c = 0.50$ and flips at $c = 1.0$, three rungs into broken. Making the reversal coincide with the mild/broken boundary would take moving that boundary two- to fivefold in coefficient on irony and ten- to twentyfold on SNLI, the latter larger than the disagreement between our two diagnostics there.

We would rather practitioners not adopt our numbers as defaults. The transportable recommendation is the one in Section 5: report the answer-position KL and the off-target accuracy delta and let a reader place the operating point, since the value at which a given model stops computing a given task is a property of both and not a constant. If a single default is wanted anyway, the intact bin is the one our other diagnostics independently corroborate, verbalizer mass and top-1 preservation both being near baseline throughout it, and it is the conservative choice in the sense that every cell we read inside it agrees in sign with cells well outside it.

### D.7 Every run at every stored operating point

Section 6's slope is read at $c = 0.02$. That is one point inside the intact regime, and a fair question is whether the finding is a property of the dial or of the point. It is not extra work to answer: every Stage B cell stores the logit-shift DV at *every* coefficient in the run's grid, not only at the one selected, so the difference-in-differences can be recomputed at $c \in \{0.02, 0.05, 0.10\}$ for all four seed/width configurations of Table 3 plus BoolQ, with no additional compute.

Two things this rules out and one it does not. It rules out the reading that the headline is an artifact of a single fortunate coefficient: at $c = 0.05$, one rung up, irony is *significantly* negative in the same

| Run | Task | $c = 0.02$ | $c = 0.05$ | $c = 0.10$ |
|---|---|---|---|---|
| Headline (42, 16k, 12 draws) | irony | $\mathbf{-0.298}$ $[-0.509, -0.089]$ | $\mathbf{-2.206}$ $[-3.762, -1.015]$ | $\mathbf{-2.162}$ $[-4.050, -0.539]$ |
| Headline (42, 16k, 12 draws) | SNLI | $-0.232$ $[-0.773, 0.135]$ | $-0.617$ $[-3.016, 0.982]$ | $-1.131$ $[-6.629, 3.011]$ |
| Seed 123, 16k | irony | $\mathbf{-0.187}$ $[-0.208, -0.144]$ | $\mathbf{-0.901}$ $[-0.978, -0.731]$ | $\mathbf{-3.101}$ $[-4.103, -2.617]$ |
| Seed 123, 16k | SNLI | $-0.258$ $[-0.472, 0.016]$ | $-0.161$ $[-0.717, 0.732]$ | $+1.768$ $[-0.486, 3.295]$ |
| Seed 999, 16k | irony | $+0.060$ $[-0.091, 0.279]$ | $-1.037$ $[-2.428, 0.114]$ | $\mathbf{-3.502}$ $[-5.661, -1.516]$ |
| Seed 999, 16k | SNLI | $-0.123$ $[-0.361, 0.256]$ | $+0.658$ $[-0.733, 2.095]$ | $-0.067$ $[-4.607, 4.112]$ |
| Width 65k, seed 42 | irony | $-0.008$ $[-0.073, 0.067]$ | $-0.338$ $[-0.620, 0.041]$ | $-0.651$ $[-2.059, 1.415]$ |
| Width 65k, seed 42 | SNLI | $\mathbf{-0.948}$ $[-1.424, -0.354]$ | $\mathbf{-1.979}$ $[-3.455, -0.036]$ | $-0.098$ $[-2.396, 2.635]$ |
| Task check (42, 16k) | BoolQ | $-0.014$ $[-0.226, 0.208]$ | $-1.061$ $[-1.799, 0.292]$ | $\mathbf{-1.484}$ $[-2.297, -0.370]$ |

Table 11: The primary estimand at all three stored operating points, for all four seed/width configurations plus BoolQ. **Bold** excludes zero. **24 of 27 cells are negative; outside the broken regime (all but SNLI at $c = 0.10$) 21 of 23 are negative and 10 exclude zero; none excludes zero in the positive direction.** Regimes (Table 10): $c = 0.02$ is intact on both irony and SNLI; $c = 0.05$ is intact on irony and mild on SNLI; $c = 0.10$ is mild on irony and *broken* on SNLI, which is where the SNLI intervals widen catastrophically ($\pm$4–7 logits), exactly the failure mode Section 5 predicts: intervals wide enough to include both signs (seed 123 has a positive point estimate there), and no cell excluding zero in the positive direction. BoolQ was not run at every regime boundary and is not separately classified.

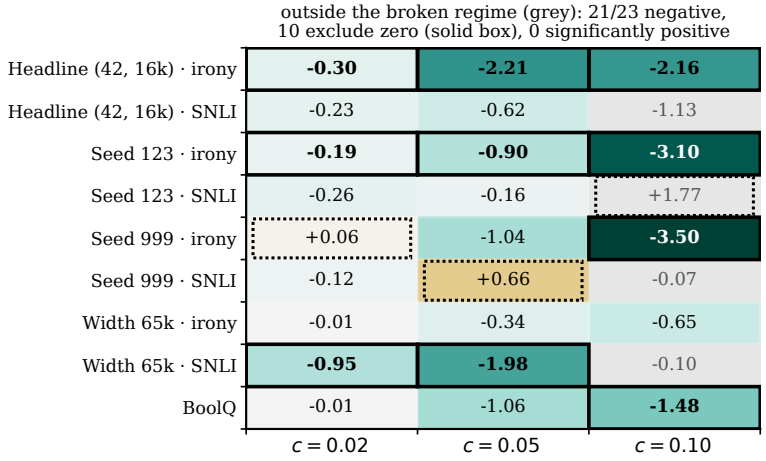



Figure 8: Table 11 as a grid. Solid box: interval excludes zero. Dotted box: positive point estimate. Colour is the sign relative to H1 (a colour-blind-safe diverging scale), not task identity. Grey cells (SNLI at $c = 0.10$) are in the broken regime and outside the tally; there intervals widen enough to include both signs, and no cell excludes zero in the positive direction.

two of four seed/width configurations as at $c = 0.02$, with a point estimate an order of magnitude larger; at $c = 0.10$ a third task, BoolQ, joins the significant column. It rules out the reading that the effect requires the model to be damaged: the largest and tightest effects are in the intact and mild regimes, and it is the *broken* column whose intervals blow up. It does not establish that any particular magnitude transports, since the magnitude varies by an order of magnitude across coefficients within a single configuration, which is the same lesson Table 3 teaches across configurations.

The same recomputation applied to the *pre-registered* run, whose grid is $c \in \{0.5, 1, 2\}$, completes the picture and is the run-level form of Table 10: on irony $+7.082$ $[5.920, 8.245]$, $+12.479$ $[11.476, 13.414]$ and $+8.670$ $[8.315, 9.043]$; on SNLI $+1.896$ $[-2.407, 7.172]$, $+7.455$ $[3.927, 10.855]$ and $+2.147$ $[-4.251, 10.269]$. Every one of those six cells is positive and four exclude zero, against *zero* significantly positive cells among the twenty-seven in Table 11. Across the design, the sign of the estimand tracks the operating point: negative in $21/23$ cells outside the broken regime ($24/27$ counting SNLI at $c = 0.10$), positive in $6/6$ at the pre-registered strengths. These are not thirty-three independent replications. The cells revisit the same few seed/width configurations and pairing draws at different coefficients, so they are strongly correlated, and we attach no pooled test to the counts; they describe one design read at two strength regimes, and the claim they support is that the sign change coincides with the regime boundary, not that it has been replicated thirty-three times.

### D.8 Secondary Stage B dependent variables

Two further quantities are computed at every Stage B cell and are reported here because Table 1 lists "units needed to steer at fixed $k$" as a basis-sensitive metric and it would be evasive to assert that row without saying what happened when we measured it.

**Specificity** (spec = best flip rate − off-target accuracy drop) inherits the saturation of its first term. On the headline run its slope is $+1.789$ $[0.736, 2.285]$ on irony against best-flip's $+1.646$ $[0.681, 2.068]$: the off-target correction moves the estimate by under $9\%$ and does not change its sign, so specificity carries no information that Section 5's flip-rate discussion does not already carry.

**Units needed** is undefined wherever no point on the units × strength grid reaches the flip-rate target, and it is undefined often: it is finite in only $35/192$ irony cells on the 12-draw headline run, $4/64$ at seeds 123 and 999, and in *seven* of the eight seed/width run × task combinations (BoolQ was not evaluated on this grid) it is either almost entirely missing ($4/64$, three cases) or present in every cell (four cases), with only the headline run's irony cells anywhere in between. A DV that is absent precisely where steering fails cannot be used to measure how badly steering fails. We therefore report the metric sort's claim about it as what it is, a statement that the quantity *reads the code* and hence *can* vary along the orbit, and not as a measured result.

## E Proofs

### E.1 Statements deferred from Section 3

The body states what each of the following establishes and defers the formalism here.

**Definition 2** (Admissibility)**.** $M$ is *admissible* with sparsity budget $\kappa \geq 1$ if **(A1)** $M$ is invertible; **(A2)** $E(x)M \geq 0$ componentwise on $\mathcal{X}$; **(A3)** $\mathbb{E}\|E(x)M\|_0 \leq \kappa\, \mathbb{E}\|E(x)\|_0$; and **(A4)** every row of $M^{-1}W$ has unit $\ell_2$ norm.

**Definition 3** (Pairwise mixing family)**.** Fix $P$ *disjoint* index pairs $\{(i_p, j_p)\}$ and $\lambda \in [0, \frac{1}{2})$. Let $M_\lambda$ be the identity outside those pairs and, on each pair, the block $\left(\begin{smallmatrix} 1-\lambda & \lambda \\ \lambda & 1-\lambda \end{smallmatrix}\right)$.

**Proposition 1** (The admissible class is a continuum)**.** *For every $\lambda \in [0, \frac{1}{2})$, $M_\lambda$ satisfies (A1)–(A3) with $\kappa = 2$, and can be made to satisfy (A4) by composition with a positive diagonal matrix. In particular* $\det M_\lambda = (1-2\lambda)^P \neq 0$*, singular exactly at $\lambda = \frac{1}{2}$; $a \geq 0$ implies $aM_\lambda \geq 0$; and with $S := \mathrm{diag}(\|(M_\lambda^{-1}W)_f\|_2)$ the composite $\widetilde{M}_\lambda := M_\lambda S$ satisfies all four.*

Proposition 1 upgrades the argument from "the metric cannot distinguish an SAE from a non-SAE" to "the metric cannot distinguish an SAE from one strictly more mixed in the targeted latent basis and satisfying (A1)–(A4)". It does not make the two indistinguishable: $L_0$ rises (Lemma 1) and paired decoder rows approach antiparallel (Appendix D.2).

**Lemma 1** (Sparsity inflation)**.** *Let $\lambda \in (0, \frac{1}{2})$ and let $q_p$ be the fraction of corpus tokens on which exactly one of $a_{i_p}, a_{j_p}$ is nonzero. Then $\mathbb{E}\|aM_\lambda\|_0 = \mathbb{E}\|a\|_0 + \sum_p q_p$, so mean $L_0$ strictly increases whenever any $q_p > 0$.*

**Corollary 2** (The sparsity confound is a step, not a ramp)**.** *For every $\lambda \in (0, \frac{1}{2})$ the support of $aM_\lambda$ is the* same *set, independent of $\lambda$. Hence $\mathbb{E}\|aM_\lambda\|_0$ is constant on $(0, \frac{1}{2})$: the inflation of Lemma 1 is a single step taken as $\lambda$ leaves 0, not a function of the dose.*

**Proposition 2** (Dense probes are blind, $k$-sparse probes are not)**.** *$\langle aM, w\rangle = \langle a, Mw\rangle$ and $w \mapsto Mw$ is a bijection, so the* unregularized *linear classes on $a$ and on $aM$ are equal and have the same infimum of any risk. For $k < F$ the constrained classes $\{\|w\|_0 \leq k\}$ are in general distinct, since $M_\lambda e_i = (1-\lambda)e_i + \lambda e_j$ has $\ell_0 = 2$.*

**Proposition 3** (At full-set rank the steering direction does not move along the dial)**.** *Let $\mathcal{I} = \bigcup_{p=1}^{P}\{i_p, j_p\}$ be the manipulated index set of Definition 3 ($|\mathcal{I}| = 2P$), and let $d_k$ be the steering direction of Appendix B.4, ranked over $\mathcal{I}$. At $k = 2P$, $d_k$ is the same vector at every $\lambda$ in both the mix arm and the clone-split control arm, so the steered forward pass, and with it the outcome $Y$, is constant in $\lambda$ in both arms and $\tau(\lambda) \equiv 0$. In our design $P = 10$, so this is the $k = 20$ direction.*

### E.2 PROOFS

*Proof of Corollary 1.* By Observation 1 the arguments of $\Phi$ are pointwise unchanged, so its value is unchanged. □

*Proof of Proposition 1.* *(a) Invertibility.* Direct computation on each $2 \times 2$ block: $\det M_\lambda^{(p)} = (1-\lambda)^2 - \lambda^2 = 1 - 2\lambda$, with blockwise inverse $\frac{1}{1-2\lambda}\left(\begin{smallmatrix} 1-\lambda & -\lambda \\ -\lambda & 1-\lambda \end{smallmatrix}\right)$. Block-diagonality (after a permutation, guaranteed by disjointness of the pairs) gives $\det M_\lambda = (1-2\lambda)^P$, nonzero for $\lambda \in [0, \frac{1}{2})$ and zero exactly at $\lambda = \frac{1}{2}$, where the two rows of each block coincide.

*(b) Nonnegativity.* The transformed coordinates within pair $p$ are $a'_i = (1-\lambda)a_i + \lambda a_j$ and $a'_j = \lambda a_i + (1-\lambda)a_j$, nonnegative combinations of nonnegative numbers since $\lambda \geq 0$ and $1 - \lambda > 0$. Coordinates outside the pairs are unchanged.

*(c) Sparsity budget.* A coordinate outside the pairs is unchanged. Within pair $p$: if both $a_{i_p}, a_{j_p}$ are zero, both stay zero; if both are nonzero, the support cannot grow; if exactly one is nonzero, the support of that pair grows from 1 to 2, an increase of one. Summing over pairs, $\|aM_\lambda\|_0 \leq \|a\|_0 + \#\{p : \text{exactly one active}\} \leq 2\|a\|_0$, the second inequality because every pair contributing an increase already had at least one active coordinate. Hence (A3) holds with $\kappa = 2$.

*(d) Unit norm.* Each $s_f := \|(M_\lambda^{-1}W)_f\|_2 > 0$, because the rows of $M_\lambda^{-1}W$ are nonzero, so $S = \mathrm{diag}(s)$ is invertible and $\widetilde{M}_\lambda = M_\lambda S$ is a product of invertible matrices. Since $S$ is positive diagonal, $aM_\lambda S \geq 0$ whenever $aM_\lambda \geq 0$, and $\mathrm{supp}(aM_\lambda S) = \mathrm{supp}(aM_\lambda)$, so (A2) and (A3) transfer verbatim. Finally $(M_\lambda S)^{-1}W = S^{-1}M_\lambda^{-1}W$, whose $f$-th row is $(M_\lambda^{-1}W)_f / s_f$, of unit norm, giving (A4). □

*Proof of Lemma 1.* Immediate from the case analysis in (c) above: the support is unchanged on every pair except those where exactly one member fires, each of which contributes exactly $+1$. Taking expectations over the corpus and writing $q_p$ for the probability of that event on pair $p$ gives the stated identity. □

*Proof of Proposition 2.* For the unregularized case, $\langle aM, w\rangle = a(Mw) = \langle a, Mw\rangle$. The map $w \mapsto Mw$ is a bijection of $\mathbb{R}^F$ since $M$ is invertible, so as $w$ ranges over $\mathbb{R}^F$ the induced $v = Mw$ ranges over all of $\mathbb{R}^F$, and conversely $w = M^{-1}v$. The two function classes are therefore equal, and equal function classes have equal infima over any risk functional; in particular the best attainable AUC is unchanged. For the $k$-sparse case, take $M = M_\lambda$ with $\lambda \in (0, \frac{1}{2})$ and $w = e_i$ for $i$ a paired coordinate. Then $M_\lambda e_i = (1-\lambda)e_i + \lambda e_j$ has $\ell_0 = 2$, so the reparameterized $k{=}1$ class contains a function the original $k{=}1$ class does not, and vice versa; the classes are distinct. □

**The regularization caveat.** Proposition 2 is exact only without regularization: an $\ell_2$ penalty contributes $\|w\|^2$, and $\|Mw\| \neq \|w\|$ in general. Our implementation approximates the unregularized limit at $C = 10^3$ and uses the dense probe as a falsifiable positive control whose difference-in-differences is *predicted* flat in $\lambda$. That prediction was tested on three tasks and the confidence interval contained zero on all three.

*Proof of Proposition 3.* At $k = 2P$ the ranking is vacuous: the top $2P$ of a $2P$-element set is the set, whatever the score. So $d_{2P} = \sum_{f \in \mathcal{I}} \Delta a_f \, w_f$ before normalisation, with $\Delta a = \overline{a}^{\,\mathrm{high}} - \overline{a}^{\,\mathrm{low}}$.

*Mix arm.* Under $\widetilde{M}_\lambda = M_\lambda S$ codes map to $aM_\lambda S$ and the decoder to $S^{-1}M_\lambda^{-1}W$ (Proposition 1(d)), so feature $f$ contributes $\big((aM_\lambda)_f s_f\big)\big((M_\lambda^{-1}W)_f / s_f\big)$ and the rescaling cancels term by term. Class means are linear in $a$, so the same holds for $\Delta a$. $M_\lambda$ is the identity off $\mathcal{I}$ and an invertible block-diagonal matrix $M_{\lambda,\mathcal{I}}$ on it, so $(M_\lambda^{-1})_{\mathcal{I}} = M_{\lambda,\mathcal{I}}^{-1}$ and, as in the proof of Observation 1,

$$d_{2P}(\lambda) = \Delta a_{\mathcal{I}} \, M_{\lambda,\mathcal{I}} \, M_{\lambda,\mathcal{I}}^{-1} \, W_{\mathcal{I}} = \Delta a_{\mathcal{I}} \, W_{\mathcal{I}} = d_{2P}(0).$$

*Clone-split arm.* On pair $p$ the dead slot carries $w_{j_p} = w_{i_p}$ and the codes $(1-\lambda)a_{i_p}$ and $\lambda a_{i_p}$, whose sum is $a_{i_p}$ at every $\lambda$; class means inherit the identity. The pair contributes $\big(\Delta a_{i_p}(\lambda) + \Delta a_{j_p}(\lambda)\big) w_{i_p} = \Delta a_{i_p} w_{i_p}$, its $\lambda = 0$ value.

In both arms the unnormalised vector is unchanged, hence so is its unit normalisation. The baseline forward pass never touches the SAE, since the hook adds to the raw residual stream, so the per-example signs $\sigma_e$ do not depend on the arm or on $\lambda$, and neither do $c$ and $s$. The hook therefore adds the identical vector to the identical residual stream at every $\lambda$, $Y_{\text{mix}}$ and $Y_{\text{ctrl}}$ are constant in $\lambda$, and since the $\lambda = 0$ cell is shared, $\tau(\lambda) \equiv 0$. □

The proposition shows the estimand is only informative below full rank, as at the primary $k = 10$: a direction that exhausts the manipulated set cannot, by construction, register basis mixing, so the dial's sensitivity lives entirely in *which* latents a sub-full ranking selects and how their weights are split. It is the steering analogue of Proposition 2, where the dense probe is blind and the sparse one is not.

*Proof of Corollary 2.* Fix $\lambda \in (0, \frac{1}{2})$, so that both $1 - \lambda > 0$ and $\lambda > 0$. Coordinates outside the pairs are untouched. Within pair $p$, the transformed coordinates $a'_i = (1-\lambda)a_i + \lambda a_j$ and $a'_j = \lambda a_i + (1-\lambda)a_j$ are sums of products of *strictly positive* scalars with *nonnegative* codes, so no cancellation is possible: $a'_i > 0$ iff $a_i > 0$ or $a_j > 0$, and the same for $a'_j$. The transformed support on the pair is therefore $\{i, j\}$ if either member fires and $\emptyset$ otherwise, a set that does not depend on $\lambda$. Composing with the positive diagonal $S$ preserves supports. Hence $\|aM_\lambda\|_0$, and so its expectation, is the same for every $\lambda \in (0, \frac{1}{2})$. □

**Why disjoint pairs, and what carries over to larger blocks.** The experiments sweep one geometry, disjoint $2 \times 2$ blocks, and it is worth separating which of the results above depend on that choice. Observation 1 and Corollary 1 need only invertibility, so they already cover $k$-way blends, rotations within a clustered subset, and dense $M$. So does Corollary 2: its proof uses nothing about the block beyond strict positivity of the entries and nonnegativity of the codes, which together forbid cancellation, so on a $k$-way block the transformed support is the whole block whenever any member fires and is again independent of $\lambda$. Proposition 1 generalises the same way, since the diagonal rescaling that restores unit decoder norms is applied row by row.

What the pair structure buys is not theoretical reach but a readable dose. Two coordinates admit a single scalar $\lambda$ with an endpoint that can be stated in words, $\lambda \to \frac{1}{2}^-$ being an even blend of the pair's codes, and they make the confound a pairwise co-activation rate that partner selection can be asked to minimise. A 3-way blend has no comparably canonical one-parameter path, and a rotation within a cluster mixes a variable number of meanings per unit at a fixed parameter value. We did not run those variants, so we state the scope plainly: the theory is geometry-independent and the estimate is not. Whether a $k$-way blend costs steering utility per unit of $\lambda$ at the same rate as a pairwise one is open, and is the cheapest extension of this work we can name, since it reuses the entire pipeline.

**Partner selection, and what it does and does not buy.** Lemma 1 makes the mix arm's $L_0$ inflation $\sum_p q_p$, which depends on the co-activation structure of the chosen pairs; the clone arm's is $\sum_p \Pr[a_{i_p} \neq 0]$. Writing $q_p = \Pr[a_{i_p} \neq 0] + \Pr[a_{j_p} \neq 0] - 2\Pr[a_{i_p} \neq 0, a_{j_p} \neq 0]$, the two inflations are equal iff

$$\Pr[a_{j_p} \neq 0] \;=\; 2\,\Pr[a_{i_p} \neq 0,\; a_{j_p} \neq 0].$$

Minimising co-activation does not achieve this and cannot: at $\Pr[\text{co-fire}] = 0$ the mix arm inflates by $\Pr[a_{i_p} \neq 0] + \Pr[a_{j_p} \neq 0]$, strictly *more* than the clone arm. What partner selection actually buys is a partner that is frequency-matched (killing the "frequent features carry more causal leverage" confound) and, among those, the one whose co-firing rate leaves the two inflations closest together given the marginals available. Because targets are the most task-discriminative latents, and therefore fire often (mean rate $0.64$ on irony), the achievable co-firing floor is bounded below by $\Pr[a_i] + \Pr[a_j] - 1$, and the arms do not in general match. Appendix D.2 reports what they actually do, which is the empirical half of the argument; Corollary 2 is the theoretical half, and is what makes the residual mismatch irrelevant to a slope.

**The selection rule in full.** Targets are the $P$ features with the largest $|\overline{a}_i^{\text{high}} - \overline{a}_i^{\text{low}}| \cdot \|w_i\|$, the magnitude of a feature's decoded class-difference contribution in residual space; the decoder-norm

factor is there because ranking on the code difference alone is not scale-invariant and silently reorders under any decoder renormalisation. For each target $t$ the partner pool is every not-yet-taken feature whose activation frequency lies in $[0.5f_t, 2f_t]$, which is the frequency match. From that pool we draw 32 candidates uniformly at random under the draw's own seed, compute for each the co-activation rate $\Pr[a_t \neq 0 \wedge a_c \neq 0]$ over the full fitting corpus at the answer position, and take the minimiser. Pairs are disjoint by construction, since both members are marked taken. Clone-arm dead slots are the least-active features with activation frequency exactly zero on that same corpus, excluding anything already taken.

The rule has one source of variation we exploit and one we do not. The 32 candidates are drawn under the draw's own seed, so each of Stage B's twelve pairing draws sees a different admissible pool and selects different partners, and the bootstrap resamples over those draws, which puts pairing-to-pairing variation inside every interval this paper reports. What we did not vary is the rule itself. Picking uniformly at random among the frequency-matched candidates, or maximising an estimate of semantic distance instead, would change the co-firing rate $q_p$ and with it the $L_0$ mismatch of Lemma 1, though not the $\lambda$-independence of Corollary 2 that makes the mismatch a level. Whether the slope's magnitude is sensitive to the choice of rule is untested.

The dead/live boundary is therefore "fired at least once at $|a| > 10^{-8}$ on the fitting corpus" rather than a tuned threshold, which is what makes the census in Appendix B interpretable: the rarest live latent fires on 1 of 300 fitting samples, so a borderline feature is one that fired exactly once, and moving the cut to admit it would change the pool by a fraction of a percent. Nothing in the clone arm depends on which dead slot is drawn, because every dead slot receives a decoder row copied from its target.

### E.3 Scope of the invariance beyond sparse autoencoders

Observation 1 is stated for an SAE, but its proof uses only that a sparse code is mapped to an output by a linear decoder. Three neighbouring families inherit it, and the third is worth separating because it is the case where the freedom is deliberately removed.

A *transcoder* reconstructs the output of a network component from that component's input through a sparse code (Paulo et al., 2025), so its prediction is again $aW + b$ and Definition 1 applies verbatim with $\hat{x}$ replaced by that prediction. Any evaluation reading only the predicted output, which includes the fidelity half of how transcoders are usually scored, is constant on the same orbit. A *crosscoder* reads and writes several sites at once (Lindsey et al., 2024), with a decoder per site; applying $M$ to the shared code and $M^{-1}$ to every site's decoder leaves every site's output fixed simultaneously, so the argument goes through jointly rather than site by site. In both cases the constructive half carries over too: a dial could be built on either, and Corollary 2 would still make its support inflation a level.

Whitening followed by independent component analysis is the instructive contrast. Whitening fixes the reparameterisation freedom down to the orthogonal group, and ICA's non-Gaussianity criterion narrows it further, to permutation and sign under the standard conditions (Comon, 1994). That pipeline is not a counterexample to the observation. It is an illustration of what the observation says is necessary, because the freedom is removed by a criterion computed *on the code*, which is exactly the right block of Table 1, and never by improving the reconstruction. An SAE post-processed this way would have a more identified basis, and the thing that identified it would still be invisible to every metric in the left block.

The same reading applies to the SAE's own sparsity penalty, which is why Lemma 1 matters beyond the control arm. Training is not orbit-invariant, so an SAE's basis is pinned down during fitting by a term that reads the code. Our claim is about what a released dictionary can be *certified* with afterwards, and there the pinning leaves no trace a reconstruction metric can recover.

## F Analysis Audit Trail and Corrected Claims

We record every error found in this paper's own pipeline or argument that changed something already written down. Each was caught the same way, by checking a written claim against something outside the text that asserted it. None was caught by rereading the prose, which is why the checks below are recommended in the body rather than merely performed here.

| Claim as written | Caught by | Status now |
|---|---|---|
| $c$=1.0 refutes H1 | the diagnostic | withdrawn; artifact of the operating point |
| $c$=1.0 is the field's convention | the literature | withdrawn; no convention exists |
| The arms are matched on $L_0$ | the algebra | withdrawn; replaced by a $\lambda$-independent level |
| Concept entropy supports H1 | the implementation | withdrawn; DV marked invalid |
| Stage A slopes from a pilot draw | the artifacts | superseded by the 10-draw grid |
| A cross-model arm produced numbers | the artifacts | withdrawn; that run produced none |
| D.6 mild/broken cuts as placed | Table 10 | corrected one rung later; sign argument holds |
| Determinism check covers the 12-draw grid | the run logs | narrowed to the 4-draw run |
| Two secondary-DV figures (4-draw values) | the artifacts | updated to the 12-draw run |

Table 12: Status of every claim this paper has withdrawn or corrected. Each row is expanded below, and each is reachable from the Limitations section.

**The operating-point inversion itself** (Section 5) was originally reported internally as a clean refutation of H1: at the pre-registered coefficient $c = 1.0$, the sign came back positive and was, for a time, taken at face value. A subsequent validity diagnostic, built to check a different concern, showed the reading to be an artifact of the coefficient at which the dependent variable was read, not evidence against the hypothesis. We regard the fact that the pre-registration inherited that coefficient from precedent, without examining the precedent, as the transferable part of the story, and it is why Section 5 is reported as a result in its own right rather than a footnote.

**We asserted a convention that does not exist.** An earlier draft described $c$=1.0 as "the coefficient the steering literature conventionally uses," in the abstract, the introduction, the roadmap table and four places in Section 5, and cited nothing for it, because the belief came from this project's own pre-registration rather than from any survey. Auditing it (Appendix D.5) shows reported strengths spanning more than two orders of magnitude in units that are mostly not interconvertible: there is no single convention to inherit or to indict. The claim has been removed everywhere and replaced by what the audit supports. We flag this as the correction most likely to have survived review unchallenged, since an uncited premise about "what the field does" is read as background rather than as a claim, and it was the load-bearing premise of the result we consider most transferable.

**The sparsity-matching claim was wrong, and the prior statement overstated the degree of matching in our own favour.** An earlier version of Section 4 and of Appendix E asserted that the mix and clone arms' sparsity inflations "coincide exactly when the partner never co-fires with the target," and presented minimum-co-activation partner selection as achieving that. The algebra gives the opposite: at zero co-activation $q_p = \Pr[a_{i_p} \neq 0] + \Pr[a_{j_p} \neq 0]$, which is strictly *larger* than the clone arm's $\Pr[a_{i_p} \neq 0]$. The corrected statement, its measured consequences, and the reason the residual mismatch does not threaten a slope (Corollary 2) are in Section 4 and Appendix D.2. The error mattered because the assertion that the arms were matched on $L_0$ was load-bearing rhetoric; the replacement argument is stronger, because it does not require the arms to match at all.

**A dead-feature guard in our concept-entropy implementation cannot fire.** The condition `tot < eps` is tested against `tot = mass.sum() + eps`, so dead latents are scored as maximally monosemantic instead of `NaN`. This inflates one Stage A dependent variable's slope by roughly $25\times$ and accounts for its entire reported value. It is disclosed with the corrected $\lambda > 0$ analysis in Appendix D.1; no claim in the paper rests on that DV. We report it because a metric bug that produces a *significant result in the predicted direction* is exactly the kind an author is least likely to look for.

**Stage A numbers in an earlier draft came from a superseded run.** That draft quoted $-0.197$ (irony), $-0.043$ (BoolQ) and $+0.003$ (SNLI) at $k$=8, and $+0.73$/$+0.27$ for BoolQ's small-$k$ reversal, from a single-pairing-draw pilot. Section 6 quotes the 10-draw full-grid run ($-0.252$, $-0.036$, $+0.012$; $+0.599$/$+0.245$). The signs and the qualitative story are unchanged; the magnitudes and intervals are not, and the pilot's SNLI value was small enough to be described as a ceiling where the full grid's is a significant reversal.

**We reported numbers for a run that had produced none.** An earlier version of Appendix A.2 quoted a cross-model (`gemma-2-2b`) slope on each task and a random-dictionary anchor for that configuration, attributing them to the run's retained stdout log. Audited against the artifacts directly, that log ends inside the model load, before any evaluation; no results file was ever written; and the checkpoint carrying the run's tag records the *default* configuration and holds four cells bit-identical to the headline run's. Those figures had no source we could stand behind, and the appendix now reports the attempt with no numbers at all. The comparison sentence built on them was also wrong where it *was* checkable: it claimed every other run separated the untrained random dictionary from the trained one by $0.58$–$1.00$, which is the SNLI range only; on irony the margins are far smaller and one is negative (Table 4). The replacement check is computed from stored artifacts for all nine run $\times$ task cells, and it costs us the only post-hoc reproduction of irony's significance.

**Three smaller corrections.** An earlier version of Appendix D.6 placed each task's mild/broken cut one rung too early relative to Table 10; the cuts now match the table, and the conclusion that the sign survives past the cut is unchanged, the margin now being two- to fivefold on irony and ten- to twentyfold on SNLI. An earlier version of Appendix G implied the 16-digit determinism check covered the 12-draw grid; it covered the original 4-draw configuration, and the text now says so. Two secondary-DV figures, the best-flip-rate slope ($+1.955$) and the count of finite units-needed cells ($12/64$), had been carried over from the 4-draw headline run after the 12-draw extension replaced it; they now read $+1.646$ and $35/192$, and the flip-rate slope still reverses the comparison.

## G Reproducibility

Every number in this paper is produced by a script in the accompanying repository and stored as a JSON artifact under a stable per-experiment identifier; every figure is regenerated from those artifacts by a single script, and every table's entries trace to a named field in a named file. Run logs, including device, dtype and per-cell progress, are retained for every run that produced an artifact; the one attempt that did not is Appendix A.2, whose log is retained but truncated and from which we report nothing. Seeded determinism was verified end-to-end: the Stage B steering grid in its original 4-pairing-draw configuration, re-run on a fresh cloud VM after a mid-run failure, reproduced its irony slope to all 16 significant digits; that configuration was later extended to the 12 draws reported here. Pre-registered decision rules, including the one that failed, are recorded with their thresholds and the dates they were fixed, in an internal planning document. Anonymised code, run artifacts, logs and figure scripts are provided as supplementary material and will be released publicly upon deanonymization.