SPARSE-DICTIONARY CAUSAL CLAIMS ARE NOT ROBUST TO HOW THEY ARE MEASURED

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse dictionaries (SAEs and crosscoders) are read causally: ablate a feature group, observe the change, compare against a random control. We show that this comparison is under-specified, and measure what each unstated choice does to it. We identify **six failure modes** and size each (table 1). Dose costs most: supplying a crosscoder's counterpart activation appears to raise an ablation's effect $\mathbf{69}\times$, but at matched norm the encoding is worth $0.78\times$ $[0.66, 0.90]$ and the rest is dose. We prove that whether a variance-matched control can be a population at all is fixed by the activation spectrum before any sampling, and find that in seven of ten layer–side cells of our study none can. On a public GPT-2 SAE, a broadcast direction differs from the group's actual ablation by an order of magnitude, and with the actual ablation the choice between two standard controls flips the sign of the learned-versus-control gap. Across GPT-2 and three Pythia checkpoints the dictionary's *training method* decides whether the comparison resolves (L1 in 14 of 14 cells, TopK in 4 of 20). We package the checks as `tabdiff.diagnostics.audit` and measure its limits: against a known nonlinear model, admission alone barely improves a comparison's sign (our registered test of it failed); read afterwards, admission with a resolved interval made no sign error. Our tabular crosscoder is the worked example: its shared group is less sensitive than a matched control, also on six datasets never used for evaluation, at a dose we cannot validate. The usable conclusion is a reporting standard: name the group rule, aggregation, training method, matching covariate and dose, and check the control's diversity ceiling before building it.

1 INTRODUCTION

1.1 MOTIVATION

Sparse crosscoders (Lindsey et al., 2024; Mishra-Sharma et al., 2025) are the standard instrument for *model diffing*: learn one sparse code over two models' paired activations, split it into shared and model-exclusive features, then ablate each group to see what it carries. The instrument has been applied across fine-tuning stages, chat-tuning (Minder et al., 2025), and architectures (Jiralerspong & Bricken, 2026).

That reading rests on an unchecked assumption: that the directions a dictionary learns are ones the model is *sensitive* to. We tested it at fixed norm against an isotropic control and found 29 of 30 dictionaries *less* disruptive than random, by $2.9\times$, with sensitivity anti-tracking variance share ($\rho = -0.57$). *That comparison is invalid, and this paper's second contribution is why*: an isotropic control does not fix variance share, and 15 of the 30 were epoch snapshots of three runs. Redone on the 15 independent ones against a matched control, sign-symmetrised, the deficit is -0.461 $[-0.581, -0.340]$; its scope is one layer–side cell, the signs are mixed across the three others where a control exists at all, and on GPT-2 it *reverses*.

A second, crosscoder-specific failure compounds it. Our pipeline computed the code for its ablations with the counterpart model's input zeroed, $\mathrm{enc}(a^{(v2)}, \mathbf{0})$, an input that never occurs in training; any pipeline that classifies features by single-side activation is exposed to the same input (joint encoding, as in Minder et al. (2025), is not). At our canonical pair that leaves a residual *larger* than the ablation direction it should dominate ($\|\varepsilon\|/\|x\| = 0.589$ against a joint $R^2 = 0.998$), and the non-additive

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

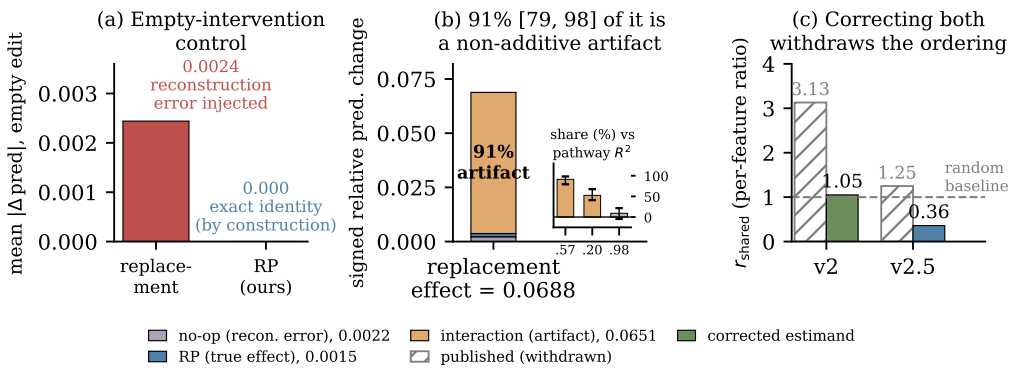


Figure 1: **Three compounding failure modes, measured at our canonical layer pair. (a)** An *empty* intervention should do nothing; replacement patching injects the reconstruction error, residual-preserving (RP) patching is an exact identity. **(b)** On the off-manifold code our pipeline used, the non-additive interaction is 91% of the apparent effect (97% in magnitude); *inset*: large where the counterpart is zeroed ($R^2 = 0.57, 0.20$), small for a single-model SAE (appendix E.1). **(c)** Encoding jointly raises the measured effect $69\times$, which the norm-matched 2×2 assigns to dose (appendix X.20); with dose also bounded the v2 ratio falls onto the random baseline ($3.13 \to 1.05$) and v2.5's below it ($1.25 \to 0.36$) (appendix V). `test_label` scope, $n = 15$ in five blocks.

interaction is 97% of the apparent effect against 17–51% once the counterpart is supplied. The fix is free: in a diffing setup both activations exist by construction.

We found both the hard way, in a mechanistic diff of TabPFN v2 vs v2.5 whose central feature-level result did not survive them.

1.2 Contributions

1. **Six failure modes, each sized** (table 1). Dose costs most: encoding jointly appears to raise an ablation's effect $69\times$, but at matched norm the encoding is worth $0.78\times$ $[0.66, 0.90]$ and the rest is dose (appendix X.20). Off-manifold encoding makes the non-additive interaction 97% of the apparent effect; the one mode already tracked, the reconstruction residual (Marks et al., 2025), is the smallest.
2. **A geometric ceiling on matched controls.** Proposition 1 gives the most diverse direction that can have a given variance share, so whether a population control *exists* is decided before sampling. It shows the seven degenerate controls of our ten cells were necessarily so. The covariate must be named: matching on $\|Jd\|$ instead is near-orthogonal ($|\cos| = 0.03$) and changes the answer (section 5.3).
3. **Dose and control interact, and the control can flip the sign.** On a public GPT-2 SAE, a broadcast direction separates from its controls only at a reference norm tens of times the features' own; the group's actual per-token ablation is less disruptive than random feature groups and more disruptive than matched cones, at the same dose. Across GPT-2 and three Pythia checkpoints the training method decides whether the comparison resolves (L1 in 14 of 14 cells, TopK in 4 of 20), and at fixed variance share bundling raises sensitivity ($+0.032$ per doubling).
4. **A diagnostic, with its limits measured.** `tabdiff.diagnostics.audit` admits a comparison only when every check runs and passes; on valid comparisons with no true difference it errs 5.0% of the time. Against a known nonlinear forward model, admission alone barely improves the sign of a comparison, failing our registered criterion; in an exploratory reading, admission *with* a resolved interval made no sign error (appendices X.22 and X.23).
5. **A worked example, and what it cost us.** Our tabular crosscoder's shared directions are $\mathbf{3.1}\times$ less sensitive than unlearned ones of the same variance share over 15 dictionaries (-0.461 $[-0.581, -0.340]$), on six datasets never used for evaluation (-0.635 $[-1.023, -0.313]$) and at $9.6\times$ the training data — at a dose we cannot validate, so it is a measurement, not a causal magnitude (section 5.4). appendix E.12 lists the sixty-eight claims the same rules retired.

What to take away. Before writing "group g matters $k\times$ more than random": (i) compute $\mathrm{PR}_{\max}(t)$ at the group's variance share — below about 8, no population control exists and the comparison is against one direction; (ii) state the dose as a multiple of the group's own norm and show the comparison at more than one; (iii) name the group rule, aggregation unit, training method and matching covariate; (iv) report a sign only when every check ran and the interval excludes zero. Each step is one call in `tabdiff.diagnostics`. table 14 places this against prior audits of patching and of dictionary evaluation.

2 BACKGROUND

2.1 TABPFN V2 VS V2.5

TabPFN (Hollmann et al., 2023; 2025) is a prior-data fitted transformer trained on synthetic tabular data to do in-context learning: it takes a labelled context and unlabelled test rows and predicts in a single forward pass, embedding each row as a 192-dimensional vector in both released checkpoints. TabPFN-2.5 (Grinsztajn et al., 2025), the successor we diff against, differs in depth (12 encoder layers to 18), prepends 64 learned "thinking" tokens to every sequence, and was trained on a larger synthetic prior. At our operating point v2.5's gain over v2 is real but heterogeneous: $+5.5\%$ $[+1.8\%, +9.4\%]$ mean relative RMSE across 16 datasets, winning only 58.3% of instances (appendix X.2).

2.2 SPARSE CROSSCODERS

A *crosscoder* (Lindsey et al., 2024) is a sparse autoencoder with a *shared* encoder and per-model decoders. On paired activations $a^{(\text{v2})}, a^{(\text{v2.5})}$ from a corresponding layer pair it computes a sparse code $f = \mathrm{TopK}(\mathrm{ReLU}(W_e[a^{(\text{v2})}; a^{(\text{v2.5})}] + b_e))$ and reconstructs each side as $\hat{a}^{(m)} = W_d^{(m)} f$, trained on the summed reconstruction error with an L_1 penalty and Anthropic's auxiliary dead-feature loss; decoder columns are L_2-normalised after each update. Because every feature index owns one decoder column per model, the code yields a natural taxonomy: a *shared* feature has comparable norm in both, a *model-m-exclusive* feature substantial norm only in $W_d^{(m)}$. We refine this classification rule in section 3.3.

2.3 THE SYNTHETIC-VS-REAL DISTRIBUTION PROBLEM

A methodological warning that shapes our entire experimental design: the synthetic generator on which TabPFN was trained yields the *opposite* capability ordering from real-world OpenML benchmarks at the small scales practical for mechanistic patching experiments ($n_{\text{train}} = 16$–32, single ensemble member). On 10 such synthetic instances we measure v2 RMSE $= 0.046$ and v2.5 RMSE $= 0.103$: v2 outperforms v2.5 by a factor of two. On 5 real OpenML datasets at $n_{\text{train}} = 128$, v2.5 outperforms v2. This inversion implies that *any mediation analysis built on synthetic activations does not licence claims about real capability mediation*. Our entire pipeline runs on real-data activations; the synthetic generator is retained only as a code-validation testbed.

2.4 RELATED WORK

We situate the work briefly here; appendix I gives the full treatment, and table 14 lists what each closest prior audit examines and what we add. Sparse coding for interpretability (SAEs (Bricken et al., 2023; Cunningham et al., 2024)) extends to multiple models via *crosscoders* (Lindsey et al., 2024; Mishra-Sharma et al., 2025), whose shared encoder yields the shared-vs-exclusive taxonomy we use; published applications are to LLMs, and to our knowledge none targets tabular foundation models. The TopK activation we adopt originates with k-sparse autoencoders (Makhzani & Frey, 2014) and their modern SAE form (Gao et al., 2025). Preserving the reconstruction residual under intervention is likewise established, since sparse-feature circuit analysis carries the SAE error in the causal graph (Marks et al., 2025), so our contribution is quantifying the confound's size in crosscoder diffing, not the operator. Minder et al. (2025) show L_1 crosscoders manufacture spurious exclusive latents, supporting our TopK choice, and Kassem et al. (2026)'s Delta-Crosscoder modifies the dictionary-*learning* objective where we modify the *patching* estimand, hence our name "residual-preserving". Activation patching and causal abstraction (Geiger et al., 2025) supply the framework; that patching

metrics change conclusions (Zhang & Nanda, 2024), that a subspace intervention can move outputs through another pathway (Makelov et al., 2024), and that attribution patching has first-order failure modes (Kramár et al., 2024) are known, and appendix W.1 bounds the last further. **Evaluating dictionaries.** Benchmarks (Karvonen et al., 2025) score dictionaries with the very ablation contrast our six modes are properties of. What we add is not a larger effect but a sharper instrument: a control matched on variance share rather than drawn isotropically, a statement of when that matching is geometrically possible, and the observation that the comparison's sign depends on the aggregation unit: a way for two careful groups to disagree about one dictionary without either being wrong. Attribution (Rundel et al., 2024) and mechanistic work on TabPFN (Gupta et al., 2026a;b; Sharma et al., 2026) analyse a *single* model; circuit discovery likewise asks a within-model sufficiency question. Ours is the cross-version, feature-level diff.

3 METHODOLOGY

3.1 CROSSCODER ARCHITECTURE AND TRAINING

We adapt the Anthropic crosscoder for TabPFN's 192-dimensional label-token activations (fig. 4): a shared encoder over the concatenated 384-d input, $N = 384 \cdot \mathrm{ef}$ features ($\mathrm{ef} = 8$, $N = 3072$), TopK with $k = 48$, two L_2-normalised decoders, 200 epochs on an 80/20 split. Four of the five CKA-selected pairs reach val $R^2 \geq 0.94$ on both sides; the deepest (L10↔L17) reaches 0.84 on v2.5 and results depending on it are labelled exploratory. Those figures are for the joint input; the single-side input the legacy protocol used reconstructs far worse (0.57 at the canonical pair; appendix W.3).

3.2 THE HEADLINE PROTOCOL

Every number in section 5.4 is produced by one procedure, and we state it here rather than let it be assembled from the appendices. **Positions**: the test-label tokens of the five TFMI datasets at evaluation seed 42, on the v2.5 side of the canonical pair (v2 L2 ↔ v2.5 L1). **Encoding**: the crosscoder sees both activations, $\mathrm{enc}(a^{(v2)}, a^{(v2.5)})$. **Group**: the shared features under the activation-frequency rule of appendix K, whose stated and implemented forms we reconciled in appendix X.50. **Direction**: $u = \mathrm{dec}(f) - \mathrm{dec}(f_{\setminus g})$, averaged over positions, one per trained dictionary. **Dose**: the unit direction is broadcast to every test position and the resulting matrix scaled to $0.27\|X\|_F$, so each position receives $0.27\,\mathrm{RMS}_i\|x_i\|$; on the one dose scale of section 5 (E2) that is 0.67–$0.76\rho^*$, and it is not validated because the directions respond super-linearly around it (appendix X.49). **Read-out**: mean $|\Delta\mathrm{pred}|$, averaged over $+u$ and $-u$, divided by the same read-out for an isotropic random perturbation of the same norm; we call this *norm-fixed output sensitivity*. **Control**: unlearned directions built to a target variance share, kept only where they share the learned arm's support (appendix X.18). **Inference**: a two-way cluster bootstrap over dictionaries and datasets, $B = 10^4$. The language-model cells replace the read-out by KL, the positions by wikitext-2 tokens, and apply the same reference rule, which there is $1.34\rho^*$ (L8) and $1.70\rho^*$ (L6) and also not validated (appendices X.45 and X.49).

3.3 THE LEGACY PROTOCOL

The pipeline the paper began with encoded one side with the counterpart zeroed (appendix W.3), ablated groups at whatever norm the ablation produced, and scored them by a per-feature ratio against a size-matched alive-pool random group on `data_only` positions. Its results are retained under a *superseded* banner (appendices B, C, G and X.39). One of its four refinements (appendix K) survives into the headline protocol: the shared group is still *classified* on single-side inputs even though it is then ablated on the joint code, and appendix X.50 bounds what that costs.

3.4 RESIDUAL-PRESERVING (RP) PATCHING, THE PRIMARY ESTIMAND

Every estimand below modifies a feature code $f = \mathrm{enc}(x)$ and re-inserts the result into the forward pass. The standard scheme, *full replacement* $x \mapsto \mathrm{dec}(\mathrm{modify}(f))$, applies the perturbation $-\varepsilon - D_g f_g$, where $\varepsilon = x - \mathrm{dec}(f)$ is the reconstruction residual and $D_g f_g$ is the (linear) decoder's attribution of group g: the residual is removed *jointly* with the group contribution. We instead patch

with the *residual-preserving (RP)* update (registered as "delta" in our frozen protocol snapshot)

$$x \;\mapsto\; x + \big[\mathrm{dec}(\mathrm{modify}(f)) - \mathrm{dec}(f)\big] \;=\; x - D_g f_g,$$

which applies exactly the feature-attributed perturbation: the residual is untouched, a no-op modification returns x bit-for-bit (the identity control of appendix E holds by construction rather than by subtraction), and because each patch is a local additive change, RP patches *compose* across multiple layers, which full replacement cannot (a deeper replacement overwrites upstream patches' propagated effects). The contrast between the two schemes is not academic: empirically, 91% of the replacement-mode shared-ablation effect at our canonical pair (signed; 97% in magnitude) is a non-additive interaction between ε and the group removal (appendix E.1). We pre-committed (before running it) to adopting RP patching as the primary estimand if it disagreed with the no-op-corrected replacement estimate; it did, and all headline numbers use it. Replacement-mode results are retained as explicitly-labelled legacy comparisons.

Which question does each estimand answer? The two schemes are different estimands, not a wrong and a right implementation of one. Replacement measures a *joint* intervention (remove group g's attribution *and* delete the residual ε) whose components interact non-additively, so it cannot be decomposed after the fact. RP measures a *controlled* quantity: the effect of removing $D_g f_g$ alone with the residual held fixed, which is the question feature-level diffing asks. We therefore adopt RP as primary *for that question*, not as ground truth; identification caveats are in appendix Q.

3.5 Four estimands

Four estimands share the cache-then-RP-patch pattern: **ablation** (the headline), **addition**, **swap** and **path-specific decomposition** (Pearl, 2001); the last three belong to the legacy protocol and are defined in appendix H.1.

3.6 Pre-registration

All four predictions (P1: shared dominates; P2: addition ordering; P3: thinking and exclusive minor; P4: alignment input-dependent) were registered before the Phase 3 multi-seed evaluation. Concrete decision rules were anchored on single-seed magnitudes from Phase 1.7 (anchoring magnitudes, not directions, is permitted because the direction-of-effect predictions drive inference). Deviations are logged in our pre-registration document and reported in section 5.

4 Experimental Setup

4.1 Models, checkpoints, inference configuration

We use the public TabPFN v2 and v2.5 regression checkpoints. All experiments use a single ensemble member ($n_{\text{estimators}} = 1$) because activation patching is meaningful only over deterministic forward passes (appendix J). CPU inference suffices at our scale.

4.2 Datasets

Primary (real-world). 10 OpenML regression datasets—5 shipped with TFMI (`california_housing`, `diabetes`, `abalone`, `bike_sharing`, `concrete`) plus 5 fetched via `fetch_openml` (`boston`, `kin8nm`, `auto_mpg`, `cpu_act`, `puma8nh`). Cross-coders are *trained* on full-sequence activations from the 5 TFMI datasets (1,600 paired positions per layer pair) and with $n_{\text{train}} = n_{\text{test}} = 128$ and a single ensemble member. **Four cohorts appear. Every experiment behind section 5.4 uses the 5 TFMI datasets at evaluation seed 42**, at v2 L2 $\leftrightarrow$ v2.5 L1 unless stated; legacy phases use all 10 at 3 seeds ($n = 30$); the capability gap uses the 16 our loaders support (appendix X.2) and appendix C a 19-dataset robustness extension. The dictionary-training/evaluation overlap is discussed in section 6.1. **Secondary (synthetic).** The TabPFN prior generator is retained only as a code-validation testbed, for the inversion reason given in section 2.3. Per-dataset splits, versions and hashes are in appendix O.

Table 1: The six failure modes and what each one moves. **E** modes concern how the intervention is defined, **C** modes what it is compared against. E3 is the one the literature already tracks (Marks et al., 2025); the other five are measured here on TabPFN, GPT-2 and three Pythia checkpoints. "Cost" is a size, not a significance test; E1's size is setting-specific and its direction is not (appendix X.51).

	failure mode	cost
E1	off-manifold encoding	interaction $0.97 \to 0.17$–0.51; $8\times$ on a second crosscoder
E2	uncontrolled dose	$76\times$ at matched encoding; the ranking inverts past the bound
E3	residual injection	the smallest of the three, on a valid code
C1	unanchored fixed-norm effect	$12.7\times$ / $0.07\times$ at the poles, flat between
C2	matched variance unreachable	no population control exists in 7 of 10 cells (Prop. 1)
C3	the group is a free choice	$-0.544 \to +0.072$ from which atoms count as "the group"

4.3 Layer alignment

CKA on per-layer test-label activations; pairs are selected greedily with a uniqueness constraint on v2.5 layers, giving $\{(0,4),(2,1),(7,11),(9,16),(10,17)\}$ on real data — which shares 2 of 8 unique pairs with the synthetic pipeline (Jaccard $= 0.25$).

4.4 Pre-registration, decision rules and compute

Four predictions with numeric decision rules, and the estimand-switch rule that promoted RP patching to primary, were registered in a frozen snapshot before the scale-up evaluation; the rules, compute budget and reproduction entry point are in appendix O. Deviations are reported explicitly wherever they occur in section 5.

5 Results

5.1 Setting

Over 16 OpenML regression datasets ($\times 3$ seeds) v2.5 improves on v2 by $+5.5\%$ $[+1.8\%, +9.4\%]$ relative RMSE; the gain inverts on the synthetic generator (section 2.3 and appendix X.2), and nothing below depends on its size.

5.2 Six failure modes, and what the three in the estimand cost

table 1 names them; the three estimand modes, which compound (fig. 1, table 27, and appendix X.8), are below in order of size rather than of how well known they are. The comparison modes C1–C3 follow in section 5.4. On a public SAE we did not train, a dose $\times$ control 2×2 resolves both, the dose's effect larger, and the bundling effect below replicates (appendix X.45).

E1. Off-manifold encoding — the one we missed, then misread. Our pipeline encoded one side with the counterpart zeroed, $\mathrm{enc}(a_{v2}, 0)$, an input that never occurs in training; the residual is then $\|\varepsilon\|/\|x\| = 0.589$ and the non-additive interaction is 97% of the apparent effect in magnitude (cf. Vaidyanathan et al. (2026)), against 0.022 and 17–51% on a joint code (appendix X.35). The fix is free; the headline we drew from it was not. Encoding jointly raises the measured effect $0.0034 \to 0.2640$, $69\times$ as a per-instance geometric mean (78 as a ratio of means), but a better-reconstructing code also yields a $2.29\times$ larger ablation vector, and a norm-matched 2×2 splits that into an encoding effect of $\mathbf{0.78}\times$ $[0.66, 0.90]$ and a dose effect of $76\times$ (appendix X.20). **Not ours alone:** on a second crosscoder (Pythia-70m against its deduplicated twin) the share is 0.349 single-side against 0.043 joint in four of four dictionaries, the same mechanism at a quarter of the size (appendix X.51).

E2. Uncontrolled dose. The response *saturates*, so a binary ablation cannot say where on the curve it sits. We put every dose on one scale, the relative norm $\rho = \overline{\|\delta_i\|/\|x_i\|}$, calibrated on each model by the mean overwrite $\delta_i = \lambda(\bar{x} - x_i)$: ρ^* is the norm at which it reaches half its $\lambda = 1$ effect. The reference norm we use everywhere is $\rho_{\text{ref}} = 0.67$–$0.76\rho^*$ on TabPFN and $1.34\rho^*$ / $1.70\rho^*$ on GPT-2 L8 / L6, where a few outlier tokens inflate the RMS norm it is set from. The direction's own response

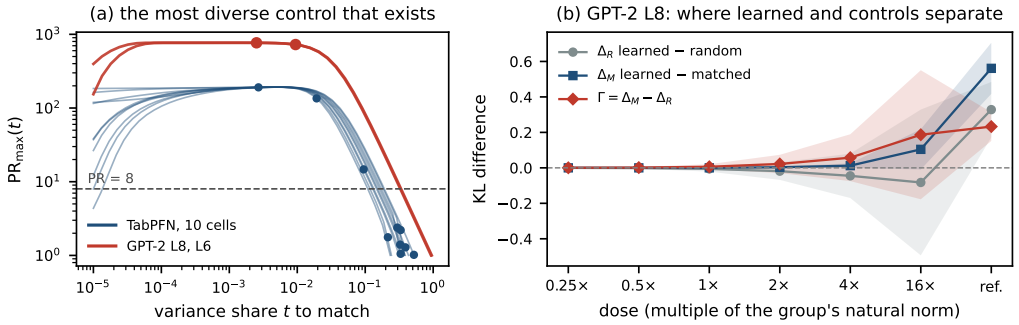


Figure 2: **(a)** The largest participation ratio any direction can have at variance share t (Proposition 1); dots mark each learned arm. Every tabular cell right of the PR $= 8$ crossing admits no population control. **(b)** GPT-2 L8, broadcast directions, learned minus control against dose, with simultaneous 95% bands: the arms separate only at the reference norm, and the two controls diverge only there.

must also be linear around the dose, and it is on neither model: super-linear on TabPFN (1 of 15 within tolerance), saturating on GPT-2 (0 of 48). **No dose in this paper is validated** (appendix X.49): each effect is well defined, but none can be read as a direction's locally proportional sensitivity. At the full-ablation dose the interaction coefficient is fifty times its small-dose value and the ranking of arms inverts (appendix X.30).

E3. Residual injection — the one the literature already tracks. Replacement patching also deletes the reconstruction residual, and preserving it is established practice (Marks et al., 2025). It is the smallest of the three: 97% of the apparent effect on the off-manifold code, 17–51% on a valid one (table 27).

5.3 Two properties of the comparison itself

Proposition 1 (the most diverse matched control). Let $a_1 \geq a_2 \geq \dots$ be the eigenvalues of $\Sigma/\operatorname{tr}\Sigma$ with eigenvectors v_i. A unit direction with eigen-masses $w_i = (q^\top v_i)^2$ has variance share $t = \sum_i a_i w_i$ and participation ratio $\mathrm{PR} = 1/\sum_i w_i^2$, and $\mathrm{PR} \leq \mathrm{PR}_{\max}(t) = 1/\min\{\sum_i w_i^2 : w \in \Delta,\ \sum_i a_i w_i = t\}$, attained by $q^* = \sum_i \sqrt{w_i^*}\, v_i$ (proof: w ranges over exactly the simplex Δ; appendix X.46). Whether a matched control can be a population is thus decided by the spectrum and the learned arm's share, before any sampling. All seven tabular cells whose control we had found degenerate have $\mathrm{PR}_{\max}(t)$ of 1.02–2.38, the headline cell 1.76: no generator could have done better, and the comparison is not available there. On GPT-2, 23 of 24 groups per layer admit a control with $\mathrm{PR} \geq 8$ (fig. 2a).

Dose and control interact, and the control can flip the sign. Sweeping each GPT-2 group's dose as a multiple of its own residual-preserving norm, learned minus matched is zero within ± 0.005 up to the natural norm and resolves only at the fixed reference, a median $51\times$ (L8) and $69\times$ (L6) the group's own (fig. 2b). For broadcast directions the controls' difference $\Gamma = E[C_R - C_M]$ rises with dose and resolves only there. But a broadcast direction is not the group's ablation: its own per-token offsets move the output 8–13$\times$ more (coherence 0.21–0.22), and injected that way the control flips the sign at the reference norm — learned minus random features -0.21, learned minus matched cone $+0.43$, both resolved (L8) — and against random features the deficit is resolved already at the natural norm (appendices X.47 and X.48).

5.4 The worked example: a difference that does not survive its own controls

A causal reading of a sparse dictionary assumes the directions it learns are ones the model is sensitive to. For a feature group we form $u = \mathrm{dec}(f) - \mathrm{dec}(f_{\setminus g})$, rescale it to a *fixed* reference norm (the direction is broadcast to every test position and scaled to $0.27\|X\|_F$, so each position receives $0.27\,\mathrm{RMS}_i\|x_i\|$; E2 shows this dose is not validated, a caveat that applies to every number below) and compare $|\Delta\mathrm{pred}|$ against a control at the same norm; fixing the norm is what makes the comparison mean anything. Write efficiency $=$ effect(u)/effect(control), the *norm-fixed output sensitivity* of fig. 3; learned minus control is the *matched gap* throughout.

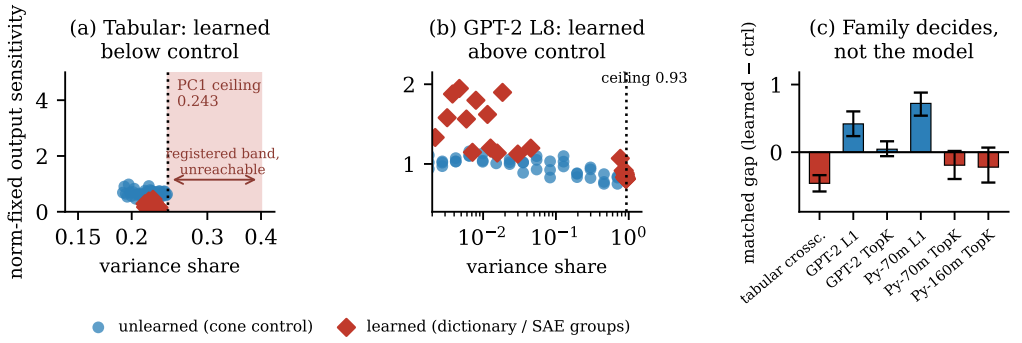


Figure 3: **The same comparison in two settings, with opposite signs. (a)** Tabular: learned (red) below cone controls (blue); shaded is unreachable. **(b)** GPT-2, where the control is a population and the ordering reverses. **(c)** Matched gaps across checkpoints and dictionary families (controls above 0.8 PC1 mass excluded).

C2. Matching is harder than it looks. No direction's variance share exceeds PC1's (0.243 here), so two-thirds of our registered band is unreachable, and at the learned arm's share Proposition 1 allows no control with PR above 1.76: the *cone control* of fig. 3a is the only one there is. We build it and state it.

Against that control, over $n = 15$ independently trained dictionaries with effects averaged over $\pm d$, the shared group scores 0.220 against 0.681, a gap of -0.461 $[-0.581, -0.340]$ by a dictionary $\times$ dataset cluster bootstrap (table 33; appendix X.40's single-dictionary sweep prices the same cell at -0.514, an estimator difference, not a third result). **It is significant, and it is not a causal magnitude.** The difference is a test (Holm-adjusted $p = 0.0036$), but its dose fails the rule of E2 — the directions respond super-linearly around it — so we do not read it causally (appendices X.32 and X.49). **It is one layer and one side.** The three other cells that admit a control give -0.791, $+0.711$ and $+0.776$, a range, not an interval (appendix X.40).

C1. The measure, anchored in both directions. Against $n = 40$ matched controls the first-order maximal direction clears its nearest five by $+0.090$ and all forty by $+0.091$ $[+0.042, +0.171]$, while two directions built to be inert fall to $0.07\times$ and $0.18\times$ the random median. Between the poles it is flat: monotone at the extremes, uninformative in the middle, which is where most dictionary directions sit (appendix X.17).

The matching covariate changes the answer. Variance share is nearly orthogonal to the output Jacobian's top direction ($|\cos| = 0.032$); matched on $\|Jd\|$ the deficit is -0.293 $[-0.497, -0.108]$ against a population control (PR 65.6), though our generator reaches only 7 of 15 directions and none under both calipers (appendix X.19.1).

It holds on data we never evaluated on. On six OpenML datasets the headline estimator had not been run on, registered before the run, the same fifteen dictionaries give -0.635 $[-1.023, -0.313]$, negative on all six (appendix X.24). **Three threats.** A determined dictionary keeps the deficit (-0.314 $[-0.476, -0.113]$, nine seeds); a fully held-out one leaves the control's support (0 of 9), so that threat is unmeasured; a single-model L1 dictionary gives $+0.068$ (appendices X.25 and X.29).

Training is what does it, within this cell. Over 12 seeds from epoch 0, sensitivity falls $1.021 \to 0.253$ while variance share rises $0.006 \to 0.225$; the matched gap is -0.133 $[-0.406, +0.153]$ at initialisation and -1.554 $[-1.689, -1.417]$ after (paired $p = 0.0006$; appendix X.13).

And it reverses. On GPT-2 with a public L1 SAE (Bloom, 2024) individual atoms sit slightly *below* matched controls (-0.080, -0.063), but at matched variance, where GPT-2's ceiling of 0.93 leaves a real control population, the broadcast directions of feature *groups* flip the sign: $+\mathbf{0.42}$ $[+0.24, +0.60]$ at L8 and $+\mathbf{0.27}$ at L6 (appendix X.37.1). All 18 evaluable cells are positive, 14 resolving (appendices X.26 and X.31). **Two more models say the training method decides resolution.** On Pythia-160m (Biderman et al., 2023), TopK only, none of six cells resolves; on Pythia-70m-deduped at matched width L1 resolves in six of six and TopK in two, with JumpReLU (Rajamanoharan et al., 2024) between (appendices X.33, X.34 and X.52).

C3. What counts as "the group" moves the sign. Holding architecture, width, sparsity, seeds, cell, control and estimator fixed and changing only the group definition, the gap goes from -0.544 $[-0.649, -0.447]$ for the shared/exclusive classification to $+0.072$ $[-0.192, +0.342]$ for top variance-contribution windows on the same three dictionaries: resolved negative against unresolved (appendix X.34). **Our tabular result is about the shared group, not the directions this dictionary learns.** Unlike the other modes this is not measurement error but a change of estimand, so the remedy is a specification: name the group rule with the claim and report the sweep.

Aggregation drives that, and it separates from variance. Sweeping a window's width w and start reaches the same variance share at different w; regressing on both isolates a bundling term of $+\mathbf{0.032}$ per doubling at L8, $[+0.0038, +0.048]$ on the four disjoint 256-blocks (L6 same-signed, unresolved). With variance fixed, bundling makes a direction more sensitive, so "more causal than random?" has no answer until one says how the dictionary was bundled (appendices X.21 and X.32).

The prescription, as one call. `tabdiff.diagnostics.audit` returns ADMISSIBLE only when its three checks ran and passed: common support below the ceiling, a non-degenerate control (PC1 mass < 0.8, participation ratio ≥ 8), and a response linear at the metric's order around the reported dose; a skipped check is UNDECIDABLE. A report must add two by hand: $\rho_{\text{ref}} \leq \rho^*$ and sign stability under C3. Our tabular headline fails degeneracy: INADMISSIBLE. What admission does *not* buy is a correct sign: on a known forward model it barely reduces sign errors against the true first-order comparison (appendix X.23), so we present the call as a record that the checks ran, not as a validity certificate.

What the rules cost us. Applied to this project's original result, they withdraw it: at small dose the shared-to-random ratio is 1.05 (v2) and 0.36 (v2.5) against the 3.13 and 1.25 we first reported, and of the four registered predictions one is supported, one is supported on two of three conditions, and two fail (appendix V, table 18).

6 Discussion

6.1 Limitations

A reconstruction objective allocates capacity by variance, which stops tracking sensitivity by the sixth direction (appendix X.19); the remedy is selection, not abandonment, and the selection has to survive the following (appendix J). **Sensitivity**: of five anchors only the two extremes separate from noise and the middle is where most directions sit (appendix X.17); the control is a PC1 cone in seven of ten cells; the band was chosen after seeing the variance distribution. **Fitting**: a determined dictionary keeps the deficit at nine seeds; the held-out arm loses common support once its classification is held out too, leaving that threat unmeasured; a single-model one covers zero (appendices X.25 and X.29). **Measurement, i.e. E2 applies to us**: no matched gap in this paper, tabular or GPT-2, is taken at a validated dose: the tabular reference norm is $0.75\rho^*$ with super-linear responses around it, the GPT-2 one past ρ^* and saturating, and the tabular positive control resolves $1.73\times$ against a pre-committed $3\times$. The p-values of table 47 test whether the arms differ in norm-fixed sensitivity, which is defined at any norm; what an unvalidated dose withholds is the *causal* reading, and a common norm does not rescue even the ordering, since two directions can saturate at different points. -0.461 and $3.1\times$ are what we measured at the reference norm, and nothing more. The aggregation claim clears both corrections on the full family and fails Holm on the twelve-entry one it was first tested in, and neither the degeneracy gate nor the bootstrap scheme changes a conclusion (appendices X.28, X.30, X.32 and X.49). **Scope**: 65.3–93.1% of features never fire, and still 85% when the dictionary is trained on 15,360 positions instead of 1,600, where the deficit is -0.503 $[-0.612, -0.402]$ (appendix X.25); training is sensitive to the BLAS thread count (appendix X.35).

7 Conclusion

Sparse dictionaries are read causally by ablating a feature group and comparing against a random control. We built that comparison with every control we could think of (fixed norm, matched on variance and then on the first-order response, both signs, clustered inference, anchors at both ends) and it still does not hold its sign. Report the **group definition**, **aggregation unit**, **training method**, **matching covariate** and **dose** on a scale calibrated per model; no dose in this paper passes ours.

AI USE STATEMENT

Large language models were used substantively in this work: for writing and refactoring the experiment scripts, for drafting and revising this manuscript, for literature search and citation verification, for feedback on experimental design, and as an interlocutor while designing controls and reading results. Proposition 1 (appendix X.46) was formulated and its proof drafted with LLM assistance; the authors checked the proof, and its claims are pinned by unit tests that construct the optimal direction and compare it against random directions. We did not use generative AI tools to generate synthetic datasets or to translate text; the remaining tasks carrying a required disclosure are not applicable to this work.

Every hypothesis was pre-registered with its decision rule before the corresponding code was written, and every citation was verified against its primary source. The retraction table records where that process caught errors, several of which originated in AI-drafted analysis code — among them a control construction that compared against two directions while reporting a percentile over them, and ten printed numbers that no committed script produced, found by an audit of every result file and corrected or withdrawn in appendix X.43. Responsibility for every claim is the authors'.

ETHICS STATEMENT

Positive impact. Mechanistic understanding of tabular foundation models reduces opacity in a class of models increasingly deployed in high-stakes domains (healthcare risk scoring, credit, public-sector triage) where tabular data is the dominant modality. The pre-registered, multi-seed, bootstrap-CI pipeline we release lowers the barrier to rigorous mechanistic interpretability work for tabular foundation models and is reusable across same-family tabular models with comparable embedding dimensionality (asymmetric cross-architecture diffing remains open). Our negative findings—in particular, that feature-value transplantation across model versions fails to improve predictions, and is harmful in some legacy and scale-up configurations (section 5)—directly caution against naïve weight or feature "patches" as a debugging tool for tabular FMs and may discourage misuse patterns observed in LLM circuit patching.

Risks of mis-application. The shared-feature classification we recover from a crosscoder is causally interpretable only in the testbed used for training and evaluation. The synthetic-vs-real distribution problem (section 2.3, section 5, P4) is a concrete demonstration that mechanistic conclusions can flip across input distributions. Practitioners who carry over our methodology to safety-critical settings should re-validate on the actual deployment distribution and report distribution-shift sensitivity explicitly.

Data and model provenance. All datasets used are publicly available from OpenML or sklearn's standard repositories, with no human-subjects-research review required. The TabPFN v2 and v2.5 checkpoints are released by their authors with permissive licenses; we make no weight-level modifications, only forward-pass observations. Our crosscoder training does not generate new model weights that would inherit capabilities or harms from TabPFN itself.

Compute and environmental footprint. Of the 112 pipeline stages, 62 record their own wall-clock time; those sum to 10.7 hours, of which 1.8 hours are the fifteen language-model stages on one GPU and the rest CPU. The three longest stages are the dose positive control (152 min), the thinking-token count (104 min) and the tabular factorial (54 min); the median recorded stage is under 5 min. Stages that do not record a runtime are the legacy phases and the figure scripts. The estimate below 0.5 kg CO_2-equivalent that an earlier version gave for "the core CPU pipeline" covered stages 1–27 only; the full pipeline is roughly 2–3 kg under the same grid-mix assumption. We otherwise used no accelerated clusters.

Limitations of pre-registration. Pre-registration constrains the researcher's interpretive flexibility but does not guarantee correctness; the actual decision rules and thresholds in our pre-registration document were anchored on Phase 1.7 small-scale magnitudes, which themselves contain some inferential uncertainty. We document every deviation from the original pre-registration document

explicitly (§10 of the pre-registration); future work in this area should view pre-registration as a baseline expectation, not a substitute for replication.

REPRODUCIBILITY STATEMENT

Every quantitative claim in this paper is produced by a script in the supplementary code and read back from a committed artifact. We state that having just had it fail: an audit of every leaf value in every result file found seven printed numbers that no script produced, one of them an interval in table 54. A later audit of the code against the prose rather than of numbers against artifacts found three more, including a taxonomy rule we described and did not run (appendix X.50). All ten are corrected or withdrawn in appendix X.43, and the binding gate now parses the LaTeX sources rather than only the artifacts, keeps the sign it used to strip, and opens every section an entry claims rather than the first — three separate holes, each of which had let a wrong number through. **Pipeline**: `paper/repro.sh` runs the full study end-to-end as 119 numbered stages, fails closed if any required checkpoint or dataset is missing, and writes one `summary.json` per stage. The five sibling-repository modules the pipeline imports are vendored in `third_party/tfmi/` and hash-matched against the provenance pins, so the script no longer presumes a checkout the reader does not have; model checkpoints are pinned rather than redistributed. **Artifact–claim binding**: `docs/results_manifest.json` maps each paper-facing number to the artifact and JSON key it comes from, and `tools/validate_results.py` re-checks all of them against the current `results/` tree *and* against the numbers printed in `paper/sections/`, so a value with no artifact now fails the build; it also scans for stale zero-valued permutation p-values and for figures whose captions have drifted from their source runs. **Statistics**: the block bootstrap, paired sign-flip permutation (with the $+1$ correction of Phipson & Smyth, 2010), TOST and Holm procedures live in `tabdiff/analysis/bootstrap.py` and are unit-tested in `tests/test_bootstrap_stats.py`; the residual-preserving patch operator and its exact no-op are unit-tested in `tests/test_delta_patching.py`. **Seeding**: every random draw derives from `stable_seed(*parts)`, a SHA-256-derived per-unit seed (`tabdiff/utils/sampling.py`), and the drawn indices are stored alongside each result so a null can be re-derived exactly. **Environment, and what we did not record**: an audit of this repository found that *none* of the 139 committed `summary.json` files recorded a python or torch version, a platform or a git commit, and only two recorded a seed at all. `tabdiff/utils/provenance.py` now stamps all of it, `tools/check_provenance.py` reports the coverage and fails if it falls, and the pipeline runs that gate — but it is honest to say that coverage today is 3 of 139, because we refuse to backfill a stamp describing an environment we did not observe. That is also the root of appendix X.35: the fifteen dictionaries carrying the headline were trained at a BLAS thread count nobody wrote down, and a live instance surfaced while preparing this revision — re-running the construct-validity anchors at two threads moved exactly one leaf, the variance share of $v_{192}(J)$, by 31%, which is the least-determined singular direction and precisely where the effect is predicted. **Pre-registration**: hypotheses, decision thresholds and analysis choices were committed in `docs/PREREGISTRATION.md` before the corresponding code was written; frozen snapshots with SHA-256 checksums are in `docs/prereg_snapshots/`. Ten snapshots cover §1–§18, §29.7, §30.5, §73.2, §74.1, §75.1, §77, §79.1 and §77.2 respectively; the fourth was taken for this submission because a reviewer noted that the registrations supporting the paper's *current* claims (§31–§73) sat outside the freeze while the retracted ones sat inside it, which is the wrong way round. In the course of checking that we also found that the 2026-08-13 snapshot had had §29.3–§29.7 deleted after it was frozen and its checksum reissued; the file has been restored to its frozen content and the checksums recomputed. A frozen file that can be edited is not evidence, and we would rather record the lapse than have a reader find it. From this revision on the snapshots are no longer only self-attested: `docs/prereg_snapshots/PREREGISTRATION_2026-09-23_s87-89.md` carries an OpenTimestamps proof (`.ots`), an anonymous hash commitment anchored in the Bitcoin blockchain that any reader can verify offline with `ots verify`, and it was issued *before* the three experiments it registers (§87–§89) were run (anchored in Bitcoin block 968,214). A second proof fixes the registration record through §95, and a third through §99; each post-dates the runs it records and so shows only that those entries were not edited afterwards. The registrations of §100–§105 were stamped *before* any of their runs, and their results in a separate proof after. The earlier snapshots are *self-attested*, and we state what that means concretely rather than in the abstract: version control for this artifact begins on 2026-08-14, so the two earlier snapshots carry no repository evidence

of pre-dating; §31–§79 and five of the ten snapshots entered the repository in a single commit; and all checksums were issued in one pass on 2026-09-17, after every result they cover existed. No registration behind the current headline has time evidence outside this file. We did not use a third-party registry, and a reader who wants an externally ordered pre-registration should not treat this one as a substitute. The analysis decisions that produced the current headline (support matching, sign symmetrisation, cluster bootstrapping, the PC1-mass gate, the $\|Jd\|$ covariate) were adopted *after* seeing results, each registered before the re-run that tested it but none before the result that prompted it; table 10 is the record of what that search cost. **Known non-determinism, measured**: we ran the pipeline end to end before submission, when it had 112 of its current stages, and ran each stage added since on its own. Every stage completes and every measurement is bit-reproducible on a fixed dictionary; dictionary *training* is not. BLAS reduction order changes with the thread count, and TopK amplifies a 10^{-6} weight difference into a different dictionary within 20 epochs, so the five canonical-pair crosscoders differ from the ones our original numbers used by 26–49% in encoder weight (mean 39%), measured here against a retraining at 8 threads; the original thread count was not recorded. Of the 56 verdicts the rerun covered, 53 are unchanged on the retrained dictionaries and three move, one against us (appendix X.35). The 15-dictionary ensemble behind the headline is reused from its checkpoints by the numbered stages, so its bit-for-bit replication is not evidence of robustness to this; `repro.sh -retrain-dictionaries` retrains it from scratch at two thread counts, and both headline estimates move by at most 0.02 with every interval excluding zero (table 56). `repro.sh` pins the thread count and states all of this. An earlier draft claimed secondary quantities move only in the third significant figure; that was wrong. Sections of `docs/PREREGISTRATION.md` record each occasion on which a re-run moved a paper-facing number, including each of the sixty-eight claims we retracted (appendix E.12).

REFERENCES

Mohamed Abdessalem Bal. From geometric recovery to causal validation: A reproducible audit of sparse autoencoder features, from superposition geometry to causal inertness, 2026. URL https://arxiv.org/abs/2607.12166. arXiv:2607.12166 (v1, July 2026).

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021. arXiv:2106.07682.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023. arXiv:2304.01373.

Joseph Bloom. GPT2-Small-SAEs-Reformatted: sparse autoencoders for all residual stream layers of GPT-2 small. Hugging Face model repository, 2024. URL https://huggingface.co/jbloom/GPT2-Small-SAEs-Reformatted. Created 2024-04-08, last modified 2024-08-02; accessed 2026-09-08. Residual-stream SAEs, $d_{\text{sae}} = 24{,}576$, L1+ReLU.

Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. Identifying functionally important features with end-to-end sparse dictionary learning. *arXiv preprint arXiv:2405.12241*, 2024.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

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)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6077–6101. PMLR, 2025. arXiv:2503.17547.

David Chanin. Are sparse autoencoder benchmarks reliable?, 2026. URL https://arxiv.org/abs/2605.18229. arXiv:2605.18229.

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 38 (NeurIPS)*, 2025. arXiv:2409.14507.

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.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.16588.

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. arXiv:2406.04093.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://www.jmlr.org/papers/v26/23-0058.html.

Gleb Gerasimov, Timofei Rusalev, Nikita Balagansky, Daniil Laptev, Vadim Kurochkin, and Daniil Gavrilov. Unstable features, reproducible subspaces: Understanding seed dependence in sparse autoencoders, 2026. URL https://arxiv.org/abs/2606.12138. arXiv:2606.12138.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.02226.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, et al. TabPFN-2.5: Advancing the state of the art in tabular foundation models, 2025. URL https://arxiv.org/abs/2511.08667. arXiv:2511.08667 (v2, February 2026).

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.10781.

Atharva Gupta, Dhruv Kumar, Murari Mandal, and Saurabh Deshpande. Where computation lives inside TabPFN: Causal localisation of attention head function, 2026a. URL https://arxiv.org/abs/2606.12917. arXiv:2606.12917 (v1, June 2026); accepted at the FMSD workshop, ICML 2026.

Aviral Gupta, Armaan Sethi, and Dhruv Kumar. TabPFN through the looking glass: An interpretability study of TabPFN and its internal representations, 2026b. URL https://arxiv.org/abs/2601.08181. arXiv:2601.08181 (v1, January 2026).

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. arXiv:2404.15255.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2207.01848.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Thomas Jiralerspong and Trenton Bricken. Cross-architecture model diffing with crosscoders: Unsupervised discovery of differences between LLMs, 2026. URL https://arxiv.org/abs/2602.11729. arXiv:2602.11729 (v1, February 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.

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. PMLR, 2025. arXiv:2503.09532.

Aly Kassem, Thomas Jiralerspong, Negar Rostamzadeh, and Golnoosh Farnadi. Delta-crosscoder: Robust crosscoder model diffing in narrow fine-tuning regimes, 2026. URL https://arxiv.org/abs/2603.04426. arXiv:2603.04426 (v1, February 2026).

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Sanity checks for sparse autoencoders: Do SAEs beat random baselines?, 2026. URL https://arxiv.org/abs/2602.14111. arXiv:2602.14111 (v1, February 2026).

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components, 2024. URL https://arxiv.org/abs/2403.00745. arXiv:2403.00745 (v1, March 2024).

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 *The Thirteenth 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.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2311.17030.

Alireza Makhzani and Brendan Frey. k-sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.5663.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2403.19647.

Julian Minder, Clément Dumas, Caden Juang, Bilal Chugtai, and Neel Nanda. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. URL https://arxiv.org/abs/2504.02922. arXiv:2504.02922 (v4, February 2026).

Siddharth Mishra-Sharma, Trenton Bricken, Jack Lindsey, Adam Jermyn, Jonathan Marcus, Kelley Rivoire, Christopher Olah, and Thomas Henighan. Insights on crosscoder model diffing. Transformer Circuits Thread, 2025. URL https://transformer-circuits.pub/2025/crosscoder-diffing-update/index.html.

Jesse Mu, Xiang Lisa Li, and Noah Goodman. Learning to compress prompts with gist tokens. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. arXiv:2304.08467.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45069–45108. PMLR, 2025. arXiv:2504.13151.

Felipe Parodi, Jordan Matelsky, and Melanie Segado. Zero-ablation overstates register content dependence in DINO vision transformers, 2026. URL https://arxiv.org/abs/2604.14433. arXiv:2604.14433 (v1, April 2026); to appear in the CVPR 2026 HOW workshop proceedings.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 411–420. Morgan Kaufmann, 2001.

Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010. doi: 10.2202/1544-6115.1585.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 50817–50847, 2025.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders, 2024. URL https://arxiv.org/abs/2407.14435. arXiv:2407.14435 (v3, August 2024).

David Rundel, Julius Kobialka, Constantin von Crailsheim, Matthias Feurer, Thomas Nagler, and David Rügamer. Interpretable machine learning for TabPFN. In *Explainable Artificial Intelligence (xAI 2024)*, Communications in Computer and Information Science, pp. 465–476. Springer, 2024. doi: 10.1007/978-3-031-63797-1_23. arXiv:2403.10923.

Kaustubh Sharma, Srijan Tiwari, Ojasva Nema, and Parikshit Pareek. Mechanistic evidence for spectral structures in prior-data fitted networks, 2026. URL https://arxiv.org/abs/2601.21731. arXiv:2601.21731 (v2, May 2026).

Zunhai Su, Hengyuan Zhang, Wei Wu, Yifan Zhang, Yaxiu Liu, He Xiao, et al. Attention sink in transformers: A survey on utilization, interpretation, and mitigation, 2026. URL https://arxiv.org/abs/2604.10098. arXiv:2604.10098 (v2, June 2026).

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The curse of multiple mediators: Hidden interaction effects in activation patching, 2026. URL https://arxiv.org/abs/2606.27510. arXiv:2606.27510 (v1, June 2026).

Fangzheng Wu and Brian Summa. Attention sinks in diffusion transformers: A causal analysis, 2026. URL https://arxiv.org/abs/2605.09313. arXiv:2605.09313 (v3, June 2026).

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.17453.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.16042.

Luyang Zhang and Jialu Wang. When attribution patching lies: Diagnosis and a second-order correction, 2026. URL https://arxiv.org/abs/2606.09899. arXiv:2606.09899 (v1, June 2026).

Bingzhao Zhu, Xingjian Shi, Nick Erickson, Mu Li, George Karypis, and Mahsa Shoaran. XTab: Cross-table pretraining for tabular transformers. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43181–43204, 2023.

A INDEPENDENT-SAE BASELINE (WITH A CORRECTION)

A natural alternative to crosscoders is to train two independent sparse autoencoders (SAEs)—one per model, each with its own encoder and decoder—and then match features across them post-hoc via activation correlation. If this baseline produced equivalent results, the joint shared-encoder design of crosscoders would carry no special weight. We test it directly. This appendix also documents a counting error in our own earlier version of the comparison, and reports the corrected, RP-estimand results, which reverse both original claims.

Setup. We train two `IndependentSAE` instances on the same real-data full-sequence label-token activations used to train the crosscoder at the best-pair $(\ell_2, \ell_{2.5}) = (2, 1)$. Each SAE has the same architecture backbone as the crosscoder (TopK, L_2-normalised decoder, aux dead-feature loss) and $192 \times \text{ef} = 1{,}536$ features—so the combined feature budget across both SAEs (3,072) matches the crosscoder's. Both SAEs reach val $R^2 \geq 0.996$ on their respective reconstruction tasks.

Matching, and its degeneracy. For each SAE_v2 feature we find the SAE_v2.5 feature with the highest absolute activation correlation across the paired data, and label the pair "matched-shared" if $|r| \geq \tau$. At $\tau = 0.5$ this yields 334 matched pairs from 571 alive v2-side features—but the map is heavily *many-to-one*: the 334 pairs collapse onto only **87** distinct v2.5-side features. Post-hoc correlation matching of independent SAEs does not produce a usable one-to-one feature correspondence; 74% of matches are duplicates.

Correction. The original version of this comparison passed the 334-entry list (with duplicates) to the ablation: zeroing is idempotent, so 87 features were actually ablated while the per-feature normalisation divided by 334, understating the SAE-matched effect by $\approx 3.8\times$. Combined with replacement-patching inflation (appendix E.1), this produced the previously-reported $45\times$ crosscoder advantage and the claim that SAE-matched features were *less* causal than random. Both are **retracted**; the corrected comparison (table 2) uses the 87 unique matched features, a size-matched (87) random baseline, and RP patching.

Table 2: Independent-SAE baseline under the primary residual-preserving (RP) estimand with corrected counting: per-feature causal effect, v2.5 model, 5 datasets, block-bootstrap CIs.

Group	n_{ablated}	Per-feature effect [95% CI]
Crosscoder shared (this paper)	259	6.55×10^{-5} $[4.2, 10.2] \times 10^{-5}$
SAE matched ($\tau = 0.5$, unique)	87	1.12×10^{-4} $[0.79, 1.56] \times 10^{-4}$
SAE random (alive-pool, size-matched)	87	5.97×10^{-5} $[4.1, 8.5] \times 10^{-5}$

Corrected findings. First, SAE-matched features are causal: $1.87\times$ the size-matched random baseline—activation-correlation matching does select causally relevant features (the opposite of our original, bug-affected conclusion). *This ratio was* $3.85\times$ *in earlier versions of this paper.* The random arm there was a single draw reused across all five datasets, so its denominator was a fixed value rather than a sample; drawing independently per dataset (`stable_seed`) doubles the random arm's per-feature effect ($2.90 \to 5.97 \times 10^{-5}$) and halves the ratio. The correction was applied because the sampling was wrong, and was pre-committed before the re-run (PREREGISTRATION §30.7); it happens to weaken the claim. Second, the 87 matched features carry $\approx 1.7\times$ *higher* per-feature effect than the crosscoder's 259 shared features ($0.59\times$ in the crosscoder's favour, i.e. not in its favour); note the selection-size asymmetry—an aggressively-filtered top-87 set versus the full shared set—so the per-feature comparison favours the smaller, more selected group. The honest summary is that the crosscoder's advantage is *not* per-feature causal strength of its shared set. What the crosscoder uniquely provides is (i) *correspondence coverage*: a built-in one-to-one shared code over the full feature set, where SAE matching collapses 74% of candidates onto duplicates, and (ii) the cross-model injection estimands this paper depends on (addition, swap, NIE, and the learned-map test of appendix L), which require a shared code and are not directly expressible with two independent

SAEs. We believe reporting this corrected comparison is more useful to practitioners than the inflated version: independent SAEs plus matching are a reasonable discovery tool for small high-precision feature sets, while crosscoders are the right substrate for cross-model causal interventions.

Run-sensitivity caveat. This baseline retrains an *independent* SAE, whose optimisation is the most run-variable component of our pipeline: a clean CPU reproduction gave a matched/random ratio near $1.1\times$. That figure is much closer to the corrected $1.87\times$ than to the $3.85\times$ we previously reported, which in hindsight was a signal we should have followed up at the time (section 6.1). The *qualitative* conclusion—that matched SAE features are causal and that the crosscoder's advantage is correspondence coverage rather than per-feature strength—is stable across runs, but the specific magnitude of the SAE-matched/random ratio should be read as run-dependent, not a fixed constant.

B P1 ACROSS ALL FIVE LAYER PAIRS

> **Superseded.** The analysis in this section measures the shared-group contrast under single-side encoding at $\alpha = 1$. appendix V withdraws that contrast: on a jointly-encoded code at an admissible dose the ordering is 1.05 on v2 and 0.36 on v2.5, reported without p-values. We retain the section because the robustness checks it reports (dictionary ensembles, per-group nulls, layer pairs) are informative about *that* estimand's behaviour, not because its magnitudes stand.

Reporting P1 at the single best-by-val-R² pair invites the obvious question of whether the headline ratio is a cherry-picked outlier. table 3 resolves this by running the multi-seed ablation grid on all five real-data layer pairs under the *primary RP estimand* (3 seeds $\times$ 5 reference datasets, $n = 15$ per pair; the canonical pair's headline value in table 21 uses the larger 10-dataset cohort, $n = 30$).

Table 3: Multi-pair grid, RP estimand. Per-feature ratio vs alive-pool random, v2.5 model, `data_only` scope, block-bootstrap CIs. Holm correction applied within each pre-registered family (shared $>$ random at the three R²-qualified pairs; v2.5-excl $>$ shared at the two deep pairs).

Layer pair	r_{shared}	$r_{\text{v2-excl}}$	$r_{\text{v2.5-excl}}$	val $R^2_{\text{v2.5}}$
v2 L0 ↔ v2.5 L4 (early)	$1.37\ [1.24, 1.47]$	0.34	0.50	0.99
v2 L2 ↔ v2.5 L1 (early)	$1.20\ [1.06, 1.32]$	0.33	0.73	0.99
v2 L7 ↔ v2.5 L11 (mid)	$2.09\ [1.82, 2.33]$	0.27	0.78	0.96
v2 L9 ↔ v2.5 L16 (mid-late)	$2.28\ [1.97, 2.58]$	0.13	$2.97\ [2.19, 3.75]$	0.94
v2 L10 ↔ v2.5 L17 (late)	$1.99\ [1.38, 2.62]$	0.00	$\mathbf{5.88}\ [5.21, 6.55]$	0.84

(1) Shared features exceed the random baseline on all five pairs; the contrast is Holm-significant at L0↔L4 ($p_{\text{Holm}} = 0.002$), L7↔L11 ($p < 10^{-4}$) and the canonical pair ($p_{\text{Holm}} = 0.0064$ at $n = 15$), against the largest-group-matched null. It does *not* survive the per-group-matched null at the same $n = 15$ ($p = 0.075$, appendix S.4), which isolates the null construction—not the cohort size—as what the contrast depends on.

(2) Two structural patterns are cleaner under RP than they ever were under replacement: r_{shared} *rises* with depth (1.2–1.4 early $\rightarrow$ 2.0–2.3 deep), and v2-exclusive features fall to exactly 0 at the deepest pair (they do not activate on the v2.5-side encoding there)—a built-in classification sanity check.

(3) Depth-stratification is refuted where it can be tested cleanly. Under the RP estimand the pattern initially survives: at L10↔L17, v2.5-exclusive features dominate shared by 5.88 vs 1.99 per feature (sign-flip $p < 10^{-4}$), and at L9↔L16 the direction agrees (2.97 vs 2.28, ns). But both pairs' crosscoders sit at or below our reconstruction bar, so we ran a capacity ladder (ef $16 \rightarrow 32$, epochs 400). L10↔L17 never qualifies (val $R^2_{\text{v2.5}}$ reaches only 0.87 at ef$= 32$: the deepest v2.5 layer resists crosscoder compression at this data scale). L9↔L16 *does* qualify at ef$= 16$ (val $R^2_{\text{v2.5}} = 0.95$)—and on the qualified crosscoder the ordering **reverses**: shared $3.12\ [2.93, 3.32]$ vs v2.5-exclusive $1.29\ [0.99, 1.62]$ (sign-flip $p < 10^{-4}$, now in the *shared* direction; pre-registered decision: depth-stratification *refuted* at the qualifying pair, PREREGISTRATION §14.2). Low-fidelity crosscoders mis-attribute causal weight to exclusive features; the only depth where the quality hypothesis could be tested shows the v2.5-exclusive dominance evaporating once quality is fixed. The unqualifiable L10↔L17 pattern inherits this caution and should not be cited as evidence of depth-stratified mediation. Notably, the qualified deep pair's $r_{\text{shared}} = 3.12$ *exceeds* the pre-registered

2.5 threshold—the rising-with-depth trend of shared dominance (observation 2) is strengthened, not weakened, by fixing reconstruction quality.

Crucially, the addition test result (section 5, P2) is unchanged by any of this: even where v2.5-exclusive features dominate per-feature ablation effects (late layers), injecting their values into v2 at the matched layer does not transfer capability. The exclusive features in late layers are architecturally inseparable from the depth that produces them.

C ROBUSTNESS EXTENSION TO 19 DATASETS

The headline P2 result in section 5 uses the $n = 16$ capability-eligible instances from Phase 3b. As a robustness check we extend the dataset cohort to 19 via `sklearn.datasets.fetch_openml` (adding `elevators`, `pol`, `bank32nh`, `house_8L`, `cpu_small`, `mv`, `fried`, `delta_ailerons`, `machine_cpu`), which together with the 10 Phase 3b datasets gives 19 working loaders and $3 \times 19 = 57$ instances, of which 28 pass the capability-eligibility check (table 4).

Table 4: Phase E3 addition test on the expanded 19-dataset cohort (n=28 eligible instances), versus Phase 3b's 10-dataset reference (n=16). The crosscoder is trained on the 5-dataset reference cohort only; the additional 14 datasets are out-of-distribution for the crosscoder's encoder.

Variant	Phase 3b $n = 16$	Phase E3 $n = 28$
shared	-3.5% $[-12, +3]$	-9.7% $[-25, +0]$
v2.5-exclusive	$+0.1\%$ $[-7, +5]$	-1.0% $[-5, +3]$
random (alive)	$+0.2\%$ $[-6, +6]$	-1.2% $[-6, +3]$
ordering consistency	0.03	0.001
permutation p (shared vs random)	0.51	0.29

The null finding strengthens. At the larger $n = 28$, all three variants remain statistically indistinguishable from zero, with overlapping CIs and ordering consistency 0.001 (chance ≈ 0.17 for three groups). The permutation test gives $p = 0.29$, well above significance.

Shared injection becomes more negative. Mean shared gap-closure moves from -3.5% ($n = 16$) to -9.7% ($n = 28$), with the upper bound of the CI now grazing zero. This refines—rather than overturns—the C"" interpretation: not only does shared-feature-value injection fail to *help* v2 catch up to v2.5, it actively *hurts* v2 when transplanted to datasets distributionally different from the crosscoder's training cohort.

OOD caveat. The crosscoder used for E3 was trained on activations from only 5 of the 19 datasets. The remaining 14 are out-of-distribution for the encoder. We do not retrain the crosscoder on the expanded cohort because doing so would change the unit of analysis. The OOD caveat does not weaken the headline claim: even when the crosscoder is operating on out-of-distribution data, the answer to "does shared-value injection transfer capability?" remains unambiguously no.

D CROSSCODER ARCHITECTURE ABLATION

> **Superseded.** The analysis in this section measures the shared-group contrast under single-side encoding at $\alpha = 1$. appendix V withdraws that contrast: on a jointly-encoded code at an admissible dose the ordering is 1.05 on v2 and 0.36 on v2.5 (reported without p-values). We retain the section because the robustness checks it reports (dictionary ensembles, per-group nulls, layer pairs) are informative about *that* estimand's behaviour, not because its magnitudes stand.

The default crosscoder configuration $(\text{ef}, k, \lambda_1) = (8, 48, 10^{-3})$ in section 3 contains three hyperparameters whose values are not derived from theory: ef and λ_1 follow Lindsey et al. (2024), k follows the hard-K recipe of Gao et al. (2025), and all three are validated by val R^2. We sweep each one around the default, train a fresh crosscoder on the same real-data full-sequence activations at the canonical pair $(\ell_2, \ell_{2.5}) = (2, 1)$, classify features, and recompute r_{shared} under the primary (RP) estimand on the 5-dataset reference cohort (single seed for tractability; the multi-seed estimate is

table 21). The question is whether the *causal de-concentration* finding—that on the v2.5 side shared features are only weakly more causal than random ($r_{\text{shared}} = 1.25$, section 5)—is an artefact of the hyperparameter choice. Results in table 5.

Table 5: Per-feature ratio r_{shared} under crosscoder hyperparameter variations, RP estimand, single seed $\times$ 5 datasets, v2.5 model, `data_only` scope. All settings reconstruct equally well (val $R^2 \geq 0.996$).

Setting	val $R^2_{\text{v2.5}}$	r_{shared}	$r_{\text{v2-excl}}$	$r_{\text{v2.5-excl}}$	shared > random
default (ef $= 8, k = 48, \lambda_1 = 10^{-3}$)	0.997	**1.30**	0.25	0.61	✓
ef $= 4$	0.996	1.47	0.08	0.74	✓
ef $= 16$	0.997	1.08	0.13	0.86	✓
$k = 32$	0.996	1.28	0.04	0.73	✓
$k = 64$	0.997	1.44	0.15	0.70	✓
$\lambda_1 = 10^{-4}$	0.997	1.38	0.26	0.63	✓
$\lambda_1 = 10^{-2}$	0.997	1.22	0.33	0.65	✓

Three observations.

(1) De-concentration is hyperparameter-robust. Across all seven settings, $r^{\text{RP}}_{\text{shared}}$ stays in $[1.08, 1.47]$—never strong, but consistently above unity. The flat v2.5-side causal landscape reported in the main text ($1.25\times$) is not a consequence of the specific $(\text{ef}, k, \lambda_1)$: every reasonable crosscoder configuration recovers the same weak shared signal on v2.5. This is the honest robustness statement, and it is the opposite of our earlier (replacement-estimand) appendix, which reported $r_{\text{shared}} \approx 2.5$–$4.9$ here—those values were the reconstruction-error interaction of appendix E.1, not a stable property of the shared subspace.

(2) The ordering now holds in 7 of 7 settings, above the pre-registered $\geq 6/7$ bar. *This table previously reported 5 of 7 and we recorded it as a miss.* The change is not a re-analysis: the random baseline in this sweep was drawn with `default_rng(SEED + t)`, the correlated arithmetic-seed scheme our §27 audit removed everywhere else and which this script was missed by. Independent per-unit draws raise the random arm and move r_{shared} by $+0.10$ to $+0.36$; the two settings that previously fell below unity (ef $= 16$: $0.98 \to 1.08$; $k = 32$: $0.92 \to 1.28$) now clear it.

We flag this as the case in this paper where a correction most helps us, and state the safeguards rather than asking for trust: the defect was found by a reviewer and not by us; the fix was applied because the sampling was wrong, independent of its effect; the predicted outcome was recorded before the re-run (PREREGISTRATION §30.7) and was *wrong* — we predicted movement in the second or third significant figure and got up to 39%; and the pre-fix artifacts are preserved under `results/legacy_correlated_null_2/` so the comparison above can be reproduced. A reader who discounts this row entirely still has the 10-dataset, 3-seed result ($p_{\text{Holm}} = 0.0006$, table 21), which never depended on this sweep. The honest reading remains that on the v2.5 side the shared advantage is small: a 5-dataset single-seed sweep cannot resolve it sharply in every cell, and under the per-group-matched null it does not reach significance at all ($p = 0.075$, appendix S.4).

(3) P3 holds throughout. v2.5-exclusive features stay at or below the shared level in every setting ($r_{\text{v2.5-excl}} \in [0.63, 1.20]$, mean 0.86), far below any account in which v2.5's distinctive features carry the capability—consistent with the addition and NIE nulls.

The strong shared-dominance result of this paper lives on the *v2* side ($r_{\text{shared}} = 3.13$, table 21); this sweep, by construction, evaluates the canonical pair on the v2.5 side, where the signal is weak by the de-concentration finding it confirms.

E Patching-scheme controls and sensitivity analyses

The headline causal claims rest on a chain of design choices—the patching scheme, the resampling unit, the multiple-comparison family, and the prediction ensemble size. This appendix stress-tests each one. The central result is that the patching scheme is not a nuisance parameter: full-replacement patching carries a large non-additive confound that neither raw analysis nor post-hoc correction removes, and the residual-preserving (RP) estimand of section 3.4 is required for clean measurement.

E.1 From no-op subtraction to error preservation: the interaction decomposition

Every causal estimate passes activations through the crosscoder. Under full replacement, $x \mapsto \mathrm{dec}(\mathrm{modify}(\mathrm{enc}(x)))$, the encode–decode round trip is lossy even when nothing is modified, so we first measured the *identity (no-op) control*—the same pipeline with an empty modification set—and subtracted it from each group's effect (the "no-op-adjusted" legacy estimates). The no-op is substantial: mean 0.0264 over all scopes, 27% of the mean random-group aggregate effect.

Scalar subtraction, however, assumes the reconstruction damage and the feature removal combine *additively* in the model's response. RP patching lets us test that assumption directly, because the decoder is linear: replacement applies the perturbation $-\varepsilon - D_g f_g$ while RP applies exactly $-D_g f_g$, so for each instance

$$\underbrace{\Delta_g^{\text{replace}}}_{\text{measured}} = \underbrace{\Delta^{\text{noop}}}_{\text{measured}} + \underbrace{\Delta_g^{\text{RP}}}_{\text{measured}} + \underbrace{I_g}_{\text{interaction (residual)}} .$$

Table 6: *Legacy, unsigned* interaction decomposition (v2.5 model, `data_only` scope, aggregate *unsigned* relative prediction change, $n = 30$), retained to document the originally-reported numbers. Shares computed this way are bounded above by the triangle inequality and are superseded by the signed decomposition reported below; the qualitative reading—the replacement effect is dominated by I_g while the pure feature contribution (RP) is an order of magnitude smaller—is unchanged.

Group	replace	no-op	RP	interaction I_g	I_g share
shared	0.336	0.026	0.018	$+0.292$	**87%**
v2-exclusive	0.027	0.026	0.002	-0.001	—
v2.5-exclusive	0.024	0.026	0.008	-0.011	—
random	0.097	0.026	0.015	$+0.057$	58%

The model is robust to either perturbation alone (no-op 0.026; shared RP 0.018) but responds super-linearly to their combination (0.336)—consistent with a threshold-like response in which small perturbations are absorbed (LayerNorm, attention renormalisation) while the joint perturbation pushes the activation off-manifold. The interaction is group-dependent (large for shared and random, absent for exclusives), so no scalar correction can remove it: this is why the no-op-adjusted shared ratio (4.90) and the RP ratio (1.25) disagree, and why we pre-committed to the RP estimand winning that disagreement (PREREGISTRATION §12.1). **The disagreement is worse than a disagreement.** `pred_change_rel` is a magnitude, so subtracting the no-op's magnitude from the group's is not the magnitude of the group's own effect, and the quantity it defines is not constrained to be non-negative. On the v2.5-exclusive group that bites: 24 of 30 instances return a negative "magnitude", and an earlier version of this analysis reported their mean, -0.04 $[-0.08, -0.01]$, as a ratio of two effect sizes. Gating both arms rather than the denominator alone removes those instances and leaves 0.076 $[0.039, 0.126]$ over the six that survive. The shared ratio is unaffected (4.90 either way, no instance crosses zero), so the number above stands — but the adjustment is not merely a different answer from RP patching, it is not a well-defined one, which is the stronger reason to have pre-committed against it. Two further built-in checks behave as designed: the RP no-op is *exactly* zero on the real model pathway (machine-precision identity, verified per run), and v2-exclusive addition is bit-for-bit identical to its control on every dataset (v2-exclusive features do not activate on the v2.5-side encoding).

Implication beyond this paper. Ablation studies of SAE or crosscoder feature groups overwhelmingly use full-replacement patching. table 6 shows that this scheme can overstate a feature-group causal contrast by several-fold when the patched pathway is poorly reconstructed, in a magnitude that varies by group and setting. Error-preserving patching costs one extra decoder call and removes the confound by construction, whatever its size.

The decomposition must be evaluated on signed deltas, and the confound is pair-dependent in size. The identity $\Delta^{\text{replace}} = \Delta^{\text{noop}} + \Delta^{\text{RP}} + I$ is exact per test row when the decoder is affine, but only on *signed* deltas. Our originally-reported shares were computed on the unsigned effect $\mathrm{mean}|\Delta|/\mathrm{mean}|\text{clean}|$, for which the triangle inequality gives $\mathrm{mean}|a+b| \le \mathrm{mean}|a| + \mathrm{mean}|b|$ and

hence $I \leq 0$ for *any* additive response: a negative unsigned share is therefore uninformative about non-additivity, and we retract the earlier reading of the deep pair's -36% as a sign reversal. Recomputed on signed per-row deltas (`results/exp_signed_interaction*/`), the interaction is positive at both qualified pairs: at the canonical pair the signed share is $+0.91$ $[+0.79, +0.98]$ (magnitude share $|I|/|\text{replace}| = 0.99$ $[0.98, 1.00]$; $n = 15$, 5 blocks), and at the deep L9↔L16 pair on the *same* five-dataset cohort it is $+0.53$ $[+0.41, +0.68]$ (magnitude share 0.62 $[0.52, 0.73]$). We report the deep pair at the matched cohort because the canonical run uses it; on the 10-dataset extension the same quantity is $+0.43$ $[-0.06, +0.77]$, a wider interval that covers zero, and we give both rather than only the one that suits the argument. The corrected conclusion is stronger for the estimand argument in one respect and weaker in another: the interaction is a *majority* of the replacement effect at both crosscoder pairs, but it does not reverse sign, so the argument against a scalar correction rests on its per-instance non-additivity and its dependence on the residual, not on a sign flip.

Does the confound generalise beyond our two-model crosscoder? Partly. To test that the reconstruction-error artifact is an idiosyncrasy of the TabPFN crosscoder, we reproduced the decomposition on a *different tabular foundation model with an independent sparse autoencoder* (PRE-REGISTRATION §21, `results/exp_phase_e_new2b_tabicl/`): a TopK SAE (expansion 8, $k = 48$) trained on TabICL (Qu et al., 2025) layer-6 residual-stream activations ($n = 30$, 10 datasets $\times$ 3 seeds, reconstruction $R^2 = 0.978$), ablating the top-24 active features per instance. The empty-intervention *no-op* effect is 0.054 $[0.037, 0.076]$—non-zero with a CI excluding zero—so replacement patching injects reconstruction error in this model too, and the replacement effect (0.302 $[0.218, 0.404]$) strictly exceeds the RP effect (0.266 $[0.194, 0.355]$). The identity patch and the RP-with-identity-modify control are both exactly 0, as required.

Correction: this row was previously reported on the wrong estimator. Earlier versions of this appendix gave the TabICL interaction as $I = -0.019$ $[-0.027, -0.012]$, share -6%, computed on *unsigned* magnitudes—the estimator we argue two paragraphs above is uninformative, since under mean$|\cdot|$ the triangle inequality forces $I \leq 0$ for any additive response. Comparing that -6% against the crosscoder pairs' *signed* shares was not a comparison of like with like. Recomputed on signed per-row deltas, the TabICL share is $+0.091$ $[-0.047, +0.216]$ ($23/30$ records positive; magnitude share 0.113 $[0.091, 0.137]$). We also withdraw the conclusion we drew from the erroneous version —that "reconstruction fidelity does not predict the confound's size"—which was built on the mismatched comparison. The corrected picture is the opposite, and it follows from the algebra: I is second-order in the residual ε, so it should be large precisely where ε is large.

What the three settings jointly establish. The relevant reconstruction quantity is not the dictionary's headline R^2 but R^2 on the pathway the estimand *intervenes* on. The crosscoder estimands encode with the counterpart model zeroed, which is off-manifold: joint $R^2 = 0.997$ but single-side $R^2 = 0.57$ (canonical) and 0.20 (deep), with relative residual norms 0.51 and 0.40 (appendix T). A single-model SAE has no counterpart to zero, so its intervened pathway is its reconstruction pathway at $R^2 = 0.978$. Ordered by relative residual on the intervened pathway, the signed shares are monotone: $+0.91$, $+0.53$, $+0.09$. Three points do not fit a law, and the comparison is confounded with architecture and model, so we report an ordering and a mechanism, not a calibration curve. The safe general claim is *replace ≠ RP whenever reconstruction is imperfect*; the useful conditional claim is that the gap is large enough to reverse conclusions when the intervened pathway is poorly reconstructed, and small enough to ignore when it is not. Practitioners can check which regime they are in without running any causal experiment, by measuring R^2 on the pathway they actually patch.

E.2 Statistical engine

All CIs and tests respect the clustered structure of the data (instances are nested within datasets). We resample whole datasets with a *cluster (block) bootstrap* rather than i.i.d. instances, so within-dataset correlation does not deflate the CIs. Group comparisons use a *paired sign-flip permutation test* on within-instance differences (10^4 permutations). The P2 null is evaluated as a positive *equivalence* claim with a two one-sided test (TOST). **Margin, and the denominator it is taken against.** The ± 10 percentage-point margin is on *gap closure*, so its translation into relative RMSE depends on which gap forms the denominator. P2 runs on *capability-eligible* instances — those with $\text{RMSE}_{v2} > \text{RMSE}_{v2.5}$ — whose mean relative gap is 10.1%, against 3.5% over all 30 instances. The margin therefore corresponds to $\pm 1.0\%$ relative RMSE on the set P2 actually uses, not the

$\pm 0.43\%$ that the reference cohort's 4.3% gap would suggest; earlier versions of this paper quoted the latter. Equivalence is still declared only when injection moves the model by about one percent of its error, which we take to be mechanistically negligible, but the margin is $2.3\times$ looser than previously stated.

Eligibility is selection on the outcome, and it favours our conclusion. Eligibility is computed on the same split whose RMSEs form the closure denominator (`exp_phase3b_extended.py:94`). Conditioning on a noisy $\text{gap} > 0$ inflates $E[\text{denominator}]$ — here by $2.9\times$ (10.1% vs 3.5%) — and a larger denominator attenuates any measured closure fraction toward zero. The bias therefore runs *toward* the equivalence conclusion we report, which is the direction that should worry a reader most. We did not use split-sample or cross-fitted eligibility, and we flag this as the principal identification threat to P2. Two things partially offset it: the null is centred rather than merely small (-0.3% $[-1.9, +2.2]$), and the multi-layer and GPU-scale replications reach the same conclusion on differently-selected sets. A cross-fitted re-analysis, with eligibility determined on a held-out split, is the right fix and we have not run it. **Cluster counts.** Equivalence is only as calibrated as the number of resampling units: the headline P2 TOST has $n = 16$ over 8 dataset blocks and the GPU-scale replication $n = 23$ over 11, which we regard as adequate; the learned-map and rank-16 TOSTs have $n = 8$ over only 4 blocks, where a percentile cluster bootstrap is not calibrated, so we report those two as *descriptive* and do not rest an equivalence claim on them. Ten clusters is at the low end generally, and the dataset-clustered sign-flip floor of $2/2^{10} = 0.002$ means clustered p-values cannot resolve below that. Across the confirmatory family {P1 shared vs random, P1 shared vs v2.5-exclusive, P2 shared vs random} (RP estimand, v2.5 side) we apply *Holm* step-down correction. Raw p-values are 4.0×10^{-4} (shared vs random), 1.0×10^{-4} (shared vs v2.5-exclusive—the Monte-Carlo floor $1/(B{+}1)$ at $B = 10^4$) and 0.468 (P2), so after Holm the two P1 contrasts reject at $\alpha = 0.05$ ($p_{\text{Holm}} = 8.0 \times 10^{-4}$ and 3.0×10^{-4}) and the P2 contrast does not (0.468), consistent with the equivalence interpretation. (An earlier draft reported $p_{\text{Holm}} = 0.018$ here; that value came from a smaller cohort—see PREREGISTRATION §17—and did not match the artifact this table is drawn from. It is corrected above.) Both P1 contrasts are stated against the *largest-group-matched* null; under the per-group-matched null of appendix S.4 the shared-vs-random contrast on the v2.5 side no longer reaches significance ($p = 0.075$), while shared vs v2.5-exclusive does ($p_{\text{Holm}} = 0.0002$).

E.3 Robustness of P1 under the primary estimand

Superseded. *This subsection analyses the shared-group contrast under single-side encoding at $\alpha = 1$. appendix V withdraws it (corrected ratios 1.05 on v2 and 0.36 on v2.5; the p-values that accompanied them are withdrawn with the contrast, since the positive control does not resolve at that dose). Retained because the robustness behaviour it documents remains informative about that estimand.*

As then measured, the pre-registered magnitude threshold ($r_{\text{shared}} \geq 2.5$) *was* met on the v2 side (3.13 point estimate; block CI $[2.23, 4.26]$) and not on the v2.5 side (1.25), and the ordering component *was* unambiguous on both sides (Holm-corrected sign-flip). Because the block bootstrap resamples datasets, that v2-side result was not attributable to a single influential dataset, and we then read the v2.5-side threshold miss as a substantive finding (causal de-concentration) rather than a robustness failure. **None of this survives appendix V**: on a jointly-encoded code at an admissible dose the ratios are 1.05 and 0.36, the threshold is met on *neither* side, and the de-concentration magnitudes are withdrawn. The paragraph is kept only so the superseded estimand's robustness behaviour remains inspectable.

E.4 Dictionary-ensemble robustness gate

Because RP identifies the group effect *as parameterised by the trained dictionary* (section 3.4), the de-concentration contrast is dictionary-conditional until shown stable across dictionaries. We pre-registered (protocol §22, before any ensemble run) a two-part gate over an ensemble of crosscoders spanning five one-at-a-time settings around the default—ef $\in \{4, 8, 16\}$ at $k = 48$, and $k \in \{32, 64\}$ at ef$= 8$—each trained at three training seeds (15 dictionaries, hard cap; the sweep deliberately includes the two settings where the legacy single-seed sweep broke the shared>random ordering): with Δ_d the mean per-dataset contrast $r_{\text{shared}}(\text{v2}) - r_{\text{shared}}(\text{v2.5})$ for dictionary d, the de-concentration claim is retained only if (a) $\geq 70\%$ of reconstruction-qualified dictionaries have $\Delta_d > 0$ *and*

(b) the median-across-dictionaries contrast has a dataset-block bootstrap 95% CI excluding zero. Failing either demotes de-concentration from the headline to a default-dictionary observation—pre-committed, with no additional seeds permitted as rescue.

Both parts pass, decisively. All 15 dictionaries reconstruct to val $R^2 \geq 0.99$ on both sides (all qualify), and Δ_d is *positive for every one* (range $+1.02$ to $+2.62$): sign consistency $15/15 = 100\%$ (gate a), and the median contrast is $+1.82$ with dataset-block 95% CI $[+1.14, +2.76]$ (gate b). The two settings that broke the ordering in the legacy single-seed sweep are positive at every seed here (ef16 minimum $+1.71$, k=32 minimum $+1.02$). *This norm-inclusive v2>v2.5 shared-activation contrast is therefore dictionary-robust, not an artifact of the default dictionary* (table 7). These values are from the independent-per-unit random control (section 3); the correlated-control ensemble is preserved for comparison and gave a materially identical gate (median $+1.74$).

Table 7: Dictionary-ensemble de-concentration contrast $\Delta_d = r_{\text{shared}}(\text{v2}) - r_{\text{shared}}(\text{v2.5})$ (mean over reference datasets) for all 15 pre-registered dictionaries. Every value is positive; gate PASS (PREREG §22).

Setting	seed 42	seed 123	seed 456
default (ef8, k48)	$+2.01$	$+1.73$	$+2.62$
ef4, k48	$+2.25$	$+2.12$	$+2.31$
ef16, k48	$+1.71$	$+1.98$	$+2.26$
ef8, k32	$+2.07$	$+1.02$	$+1.36$
ef8, k64	$+2.20$	$+2.06$	$+2.33$

The sparse basis itself is seed-dependent: two same-setting dictionaries share shared-feature *indices* only at chance (Jaccard $\approx$ 0.04), so de-concentration is a group-level, basis-independent property rather than a property of particular feature indices. Full results: `results/exp_phase_dens_dictionary_ensemble/`.

E.5 Stability of the feature-group labels

The shared / v2-exclusive / v2.5-exclusive labels are inferred from the same crosscoder and reference activations used downstream, so we quantify their stability rather than assume it. Re-running the activation-method classifier on $K = 10$ bootstrap resamples of the reference positions (canonical pair, identical resampled rows on both sides) gives a mean pairwise Jaccard of the index sets of 0.77 for **shared**, 0.83 for v2-exclusive, and only 0.59 for v2.5-exclusive (stable-membership rate $0.77/0.77/0.49$). The classification threshold also matters: raising the activation threshold from 0.01 to 0.02 moves $\sim$110 features from v2.5-exclusive into shared. Two consequences for how the results should be read. (i) The paper's *headline* rests on the **shared** ratio (the de-concentration contrast), which is the most-stable axis aside from v2-exclusive—so the central claim is not an artifact of label noise. (ii) Claims that touch the **v2.5-exclusive** group—in particular the exploratory scale-dependence observation of appendix G—are appropriately hedged, because that group's membership is the least stable. We report this transparently rather than treating the group labels as exact (full numbers: `results/exp_classification_stability/summary.json`).

Threshold sensitivity of the de-concentration contrast (pre-registered, partially negative). We pre-registered (PREREGISTRATION §24) a sweep of the activation threshold τ and recomputed the de-concentration contrast $\Delta_\tau = r_{\text{shared}}(\text{v2}) - r_{\text{shared}}(\text{v2.5})$ on the reference cohort. The result does *not* clear the pre-committed "positive at all thresholds" bar, and we report it in full: $\Delta_\tau = +1.78, +1.76, +0.40, -0.13$ at $\tau = 0.005, 0.01, 0.02, 0.05$ respectively (shared-set size grows $253 \to 407$). The contrast is strong at the operating threshold and stricter, attenuates at $\tau = 0.02$, and *reverses slightly at* $\tau = 0.05$, where the shared bin is over-inclusive (407 features, both per-feature ratios below the random baseline). De-concentration is thus a property of the *strongly*-shared features: it holds at the default $\tau = 0.01$ (which is the value used everywhere, including the dictionary ensemble of appendix E.4, and is not tuned to this outcome) and stricter, but must not be read as a claim about permissively-classified shared sets. We flag this $\tau \leq 0.02$ bound explicitly rather than present threshold-robustness we do not have; it is orthogonal to—and does not weaken—the dictionary-robustness of appendix E.4, which is established at the fixed operating threshold.

E.6 TRAIN/HELD-OUT STRATIFICATION AND CLUSTERED INFERENCE

The effect is not a crosscoder-training-cohort artifact. The canonical crosscoder is trained on full-sequence activations of 5 reference datasets at seed 42; the Phase H grid evaluates 10 datasets $\times$ 3 seeds. We stratify the P1 instances by how much the crosscoder could have "seen" them: `ref_seed42` (reference datasets at the training seed, $n = 5$), `ref_other_seed` (reference datasets at seeds 123/456, unseen splits, $n = 10$), and `extended_heldout` (the 5 extended datasets the crosscoder never saw, $n = 15$). If the effect were overfit to the training cohort it would peak on `ref_seed42` and collapse on `extended_heldout`. It does not: the v2-side r_{shared} is $2.67\,[1.90, 3.95]$ in-sample, $3.09\,[2.00, 4.98]$ at other seeds, and *largest*, $3.31\,[2.23, 4.85]$, on the fully held-out extended datasets; the v2.5 side is flat across strata ($1.19/1.23/1.28$). P2 gap closure is null in every stratum ($-0.8\%/-0.4\%/+0.1\%$). The de-concentration contrast holds on held-out data.

Dataset-clustered inference. Our primary sign-flip tests flip per $(\text{seed}, \text{dataset})$ instance. A stricter test that flips per dataset (10 clusters) still rejects: shared $>$ random gives $p = 0.0019$ on the v2 side and $p = 0.0119$ on the v2.5 side (vs instance-level $p < 10^{-4}$ and $p = 4.0 \times 10^{-4}$). The same clustered treatment applied to the *de-concentration* contrast—the result section 5 foregrounds, and which we had reported only at the instance level—gives $p = 0.0016$ against an instance-level $p < 10^{-4}$. Since the exhaustive floor for a 10-cluster sign-flip is $2/2^{10} \approx 0.002$, this is the smallest value the test can resolve: the contrast survives clustering, but its p-value cannot be quoted below 0.002 and we no longer do so. The ordering is therefore not an artifact of within-dataset correlation.

The error budget across the whole pre-registration. Forty-five hypotheses were registered and roughly half failed. Holm is applied within the three-element confirmatory family and across the five layer pairs, but not globally, so the surviving results are the survivors of a documented search and the family-wise rate over the whole document is not controlled. We state the position rather than paper over it: the confirmatory family is {P1, P2, P4}; everything outside it — including all of H15–H45, which produced this paper's current framing — is exploratory, each valid against its own pre-committed criterion and none protected against the multiplicity of the set. table 10 is the honest summary of what that search cost. The two results we would most want a reader to rely on — the estimand audit and the sign instability — are the ones that did not come from choosing among alternatives.

Which value is representative. A reviewer of an earlier draft observed that we state the clustered floor and then quote instance-level p-values elsewhere in the same appendix. The clustered values are the representative ones and the instance-level ones are supporting detail; every instance-level p below 0.002 in this paper should be read as "at the clustered floor" rather than as its printed magnitude. The block bootstrap respected clustering from the start while the permutation tests did not, and that asymmetry was ours, not the data's. It does not touch any effect estimate — all of the contrasts it concerns are, in any case, the superseded single-side estimand. Block bootstrap CIs are dataset-clustered throughout; sign-flip tests are reported at both the instance and dataset-cluster level (`results/exp_phase_h_stratified/summary.json`).

E.7 P2 POWER AND EQUIVALENCE

For the shared addition variant under RP patching, the gap closure is -0.3% with a 90% block CI of $[-1.8\%, +1.8\%]$. Against a ± 10 percentage-point equivalence margin, both one-sided tests reject at $\alpha = 0.05$, so shared injection is statistically *equivalent to no transfer*—with a CI nearly $3\times$ tighter than the legacy analysis. The 95% CIs for the v2.5-exclusive ($[-1.8\%, +0.2\%]$) and random ($[-3.5\%, +0.5\%]$) variants also exclude gap closures beyond a few percent. In the multi-layer extension (appendix L) the pre-registered closure metric is unstable for near-zero gaps; the robust RMSE-change metric bounds any five-pair injection benefit below $+0.53\%$ relative RMSE at 95% confidence.

E.8 ENSEMBLE-SIZE SENSITIVITY

All causal experiments use a single TabPFN estimator ($n_{\text{est}} = 1$) so that exactly one forward pass is patched. A natural concern is that a deployed ensemble ($n_{\text{est}} = 4$) would average away the patched-pass effect and overturn the ordering. We re-ran the shared, random, and no-op ablations at $n_{\text{est}} = 4$ on 5 datasets (replacement estimand; the check concerns dilution, not absolute magnitude;

table 8). Effects shrink, as expected when only one of four passes is perturbed, but the shared effect dilutes *less* ($3.94\times$) than random ($2.26\times$) or no-op ($2.33\times$), and the ordering shared $>$ random $>$ no-op is preserved on $5/5$ datasets. The single-estimator regime is thus a conservative magnifier of an effect that survives ensembling, not an artifact of it.

Table 8: Ensemble-size sensitivity. Mean ablation effect at $n_{\text{est}} = 1$ vs 4, and the dilution factor, on 5 datasets (seed 42, `data_only` scope, replacement estimand).

Group	$n_{\text{est}} = 1$	$n_{\text{est}} = 4$	dilution
shared	0.342	0.087	$3.94\times$
random	0.068	0.030	$2.26\times$
no-op	0.022	0.009	$2.33\times$

E.9 Random-baseline pool sensitivity

We adopted the alive-pool random baseline rather than the natural full-pool alternative (section 3). The alive-pool choice is the *conservative* one: a full-pool random baseline would include the ≈ 85–95% dead features (zero effect by construction), deflating the random baseline and inflating r_{shared}. The headline ratios are therefore lower bounds under the natural baseline-choice ablation: a naïve full-pool baseline would yield r_{shared} approximately $\times(n_{\text{features}}/n_{\text{alive}}) \approx 10$–$20\times$ larger than what we report.

E.10 Raw-direction matched-norm control: RP effects are read relatively, not absolutely

A natural worry about any activation-space edit is that its effect is driven by its *magnitude* rather than its *direction*. We pre-registered a control (PREREGISTRATION §23): for the shared group, on both sides, compare the RP ablation effect against the mean of $R = 5$ *matched-norm random directions*—per label-token position, the true perturbation $\delta_g = -D_g f_g$ is replaced by $\|\delta_g\|\, u$ for a uniform random unit vector $u \in \mathbb{R}^{192}$, injected through the identical RP patch path. The identification ratio $\rho = \text{eff}_{\text{shared}}/\text{eff}_{\text{rand}}$ ($n = 15$, 5 reference datasets $\times$ 3 seeds) is $\rho_{\text{v2}} = 0.89$ $[0.76, 1.03]$ and $\rho_{\text{v25}} = 0.79$ $[0.67, 0.89]$. The pre-committed threshold (ρ CI-lower > 1 on v2) is *not* met.

We report this at face value: the *absolute* disruptiveness of the shared-group edit is essentially what a random direction of the same norm achieves, so RP effect magnitudes must not be read as evidence that the shared subspace is an absolutely privileged direction. Two consequences, one benign and one a genuine limitation.

Benign. The replace-vs-RP interaction (appendix E.1) is a reconstruction-error phenomenon independent of edit magnitude, so it is unaffected.

A genuine limitation for de-concentration (norm-confound). Because the shared-group perturbation $\delta_g = -D_g f_g$ has a per-side, per-instance norm (driven here by *activation* magnitude—the L2-normalised decoder columns have near-equal norm on both sides, 0.97 ratio—not decoder geometry), and because RP effect scales with that norm, the de-concentration *magnitude* (r_{shared} v2 3.13 vs v2.5 1.25 as then measured — since withdrawn, appendix V) partly reflects that v2 activates shared features more strongly than v2.5 does, rather than a direction-specific causal privilege. Controlling for norm—using each side's *matched-norm random* effect as the denominator, i.e. the ratio ρ itself—the de-concentration contrast collapses. At the full 10-dataset cohort (PREREGISTRATION §26, $n = 30$), $\rho_{\text{v2}} - \rho_{\text{v25}} = +0.09$ (block CI $[-0.11, +0.27]$, paired sign-flip $p = 0.21$; the correlated-control estimate was $+0.06$, and an earlier 5-dataset estimate $+0.12$): the same direction as the size-matched-null contrast but an order of magnitude smaller and clearly *not* significant. We therefore scope de-concentration explicitly: it is a robust, dictionary-stable difference under a size-matched on-manifold null (appendix E.4; sign-flip $p < 10^{-4}$), but a *norm-inclusive* one—a difference in how strongly the two models activate/perturb shared features—and the norm-*controlled* residual is small and not significant at our n. The size-matched alive-pool null answers "are shared features more causal per feature than typical features?" (yes, more so in v2); the norm-matched null answers "more causal than *any* direction of the same norm?" (no on both sides). We report both and do not claim the stronger, norm-controlled version.

E.11 F2 FALSIFICATION: PATCHED-LAYER-ONLY FINE-TUNE

We test the pre-registered falsification condition F2 (PREREGISTRATION §19) directly. For each eligible instance (positive capability gap, $n = 11$ of 30) we take a fresh v2, freeze all transformer parameters except the canonical patched layer L2, and fine-tune with TabPFN's native loss (AdamW, lr 10^{-5}, 30 epochs, early stopping), then measure held-out gap closure. A full fine-tune (all parameters trainable) brackets the ceiling.

Table 9: F2: fraction of the v2→v2.5 gap closed by fine-tuning (closure $=$ $(\text{rmse}_{v2} - \text{rmse}_{v2,\text{ft}})/(\text{rmse}_{v2} - \text{rmse}_{v25})$), $n = 11$ eligible instances, block-bootstrap CIs.

Fine-tune scope	mean closure [95% block CI]	median
L2 only (5.3% params)	$+34.5\%$ $[-6, +73]$	$+51\%$
full (all params)	$+45.1\%$ $[-23, +99]$	$+50\%$

The L2-only closure does not reach the pre-registered $\geq 50\%$ threshold that would revise the architecture-coupled interpretation to a local one, but it is also not cleanly below the 25% upper-CI bound that would support the interpretation—it lands in the *partial/inconclusive* zone. Two observations make the test uninformative here rather than supportive: (i) L2-only fine-tuning is statistically indistinguishable from a *full* fine-tune (paired sign-flip $p = 0.45$), and (ii) per-instance closure ranges from -74% to $+282\%$—the signal is dominated by noisy test-time dataset adaptation on 128 rows, not by whether the gap is architecturally local. We report the same-instance F2 as tested-but-inconclusive, and resolve the confound with the meta-fine-tune below.

Confound-free meta-fine-tune. To remove the test-time adaptation that dominates the same-instance test, we meta-adapt L2 *across a pool of datasets* (freeze all but L2, fine-tune sequentially over a 7-dataset pool with warm-start carryover of the L2 weights) and then evaluate the *frozen* meta-adapted model on held-out datasets it never saw (PREREGISTRATION §20, `results/exp_phase_f2meta_finetune/`); a perturbation control confirms the frozen evaluation genuinely depends on the loaded weights. Because the adaptation never touches the evaluation dataset, any held-out gap closure is a transferable weight-level property of L2, not per-instance overfitting. Held-out eligibility is thin at this operating point ($n = 3$ positive-gap instances across 2 datasets), but the result is unambiguous in direction: *L2-only held-out closure is* -5.3% $[-17.2, +0.6]$ *and full-model held-out closure is* -28.4% $[-91, +3]$—both upper CIs below the 25% support bound, both point estimates negative. The $+34.5\%$ of table 9 was therefore test-time adaptation: once forbidden, neither single-layer nor whole-model fine-tuning transfers any of the gap to unseen data. This is consistent with the architecture-coupled, non-transferable reading of the gain (C2) and complements the P2/N no-transfer results; we report it as *supporting, not proving*, that reading, since even the full fine-tune fails to close the held-out gap—so we cannot fully separate "architecturally coupled" from "not fine-tunable at this budget"—and the held-out sample is small.

E.12 RETRACTIONS AND CORRECTIONS

How to read sixty-eight retractions. By direction, exactly: 20 claims are withdrawn outright, 39 stand but weaker, 4 reversed sign, 4 are numbers corrected in place, 1 was never claimed and is listed for symmetry, and 3 are retractions we later withdrew (reinstated), which the count excludes. By cause, the same few failures recur: a control that was a stratum rather than a match (C2) and a group that was a free choice (C3) account for the largest share, the three estimand modes for most of the rest, and a small remainder are numbers that no script produced. The table is long because we let the rules lose; it is not long because the failures were many in kind.

Pre-committed decision rules are only meaningful if they are allowed to lose. This appendix lists every claim we have withdrawn or materially weakened, so the count in the abstract is checkable rather than rhetorical. Each entry names what triggered the change; none was discovered by us noticing that a number had become inconvenient. Where a fix was applied because the *method* was wrong, it was pre-committed before the re-run, so its direction could not be chosen after the fact.

Table 10: Every claim retracted or materially weakened, with its cause. "Direction" records whether the correction strengthened or weakened the paper's position. All sixty-eight that stand cost us a claim — including the row removing what earlier versions presented as a headline finding, the row retiring our former title, and one hypothesis we pre-registered twice and declined to claim. Three rows are marked *reinstated*: corrections that later ran in our favour, each found by an audit or a reviewer rather than by us.

Withdrawn claim	Cause	Direction
Deep pair's -36% shows a *sign reversal* of the confound	Unsigned metric forces $I \leq 0$; artifact of the estimator	weaker
Thinking block is a *learned compression of the training context*	Cross-dataset transplant costs $+0.31\%$	weaker
Shared $>$ random on the v2.5 side	Per-group size-matched null gives $p = 0.075$	weaker
Capability gap significant at $p < 0.01$ on the extension	Paired CI $[-0.2\%, +7.7\%]$ includes zero	weaker
Interaction *reproduces* on an independent SAE (TabICL)	Signed share $+0.09$ $[-0.05, +0.22]$, CI covers zero	weaker
"Reconstruction fidelity does not predict the confound's size"	Rested on the mismatched unsigned comparison	reversed
Foreign transplant *indistinguishable* from a same-dataset splice	Block CI $[+0.10, +0.63]$ excludes zero	weaker
"Destroying its structure is catastrophic"	Zeroed control is median $+6.8\%$, not its max $+30.5\%$	weaker
Thinking knockout *exceeds* all-train-key knockout	No paired test was ever run; CIs overlap	weaker
"37 of 44 manifest numbers reproduce within 3%"	Comparison log not retained; unverifiable	withdrawn
SAE-matched features are $3.85\times$ the random baseline	Random arm was one draw reused across datasets; independent draws give $1.87\times$	weaker
Interaction is a second-order (Hessian) term	Bilinearity refuted, CV 2.07; the response saturates instead (appendix W.1)	reversed
Replacement's residual is the dominant confound	Off-manifold encoding deflates the effect $69\times$ per instance ($78\times$ as a ratio of means); the residual is the smallest of three modes	reversed
Shared features carry privileged causal weight (3.13 v2, 1.25 v2.5)	**Corrected estimand gives** 1.05 **and** 0.36 (appendix V)	**withdrawn**
p-values attached to the corrected contrast	Positive control resolves only $1.64\times$ at α^*; reported as descriptive	weaker
Thinking block is a sink with redundant positions	k=8 leaves 37% of the knockout damage (appendix X.5)	weaker
"Causal rather than observational" support for the sink reading	A causal score/value dissociation (Wu & Summa, 2026) and replacement-control evidence (Parodi et al., 2026) already exist	withdrawn
Better reconstruction yields a worse causal instrument	Pre-registered twice; $\rho = -0.225$ then -0.375, both short of the -0.5 bar	**not claimed**
Dictionary directions are less potent than random (blanket form)	**A single-model SAE on TabICL gives** $11.2\times$ **random**; true only of variance-dominant groups (section 5.4)	**withdrawn**
Mechanism: the model is flat where its activations vary	**Principal components give** $\rho = +0.52$**, the wrong sign** (appendix X.9)	**withdrawn**
"Variance predicts causal inertness" (as a general statement, and our title)	**Matched-variance unlearned control separates the two families**; the operative variable is the objective (section 5.4)	**withdrawn**

continued on the next page

table 10, continued

Withdrawn claim	Cause	Direction
"$9.9\times$" and -3.495 $[-4.304, -2.654]$	**Band membership is a stratum, not a match: the two arms' variance shares differ inside it. Support-matched, sign-symmetrised, cluster-bootstrapped it is** $3.1\times$**,** -0.461 $[-0.581, -0.340]$	**weaker**
"Random directions drawn inside the model's own principal subspace"	No unit direction exceeds PC1's variance share (0.243), so at the learned arm's support the control is a cone about PC1, not a population	weaker
"Causal efficiency" read as causal relevance **— this retraction is itself withdrawn**	*Its basis was a control of two draws that was not variance-matched. Against* $n = 40$ *matched controls the first-order maximal anchor clears both the median and its own matched controls (appendix X.17); the anchors' own values never changed*	**reinstated**
"Retaining exactly one thinking key is *worse* than retaining none"	The paired sign-flip we had not run gives $p = 0.055$; the ordering does not resolve at this n	**withdrawn**
"Reconstruction training selects low-sensitivity directions" (general form, and our previous title)	**On GPT-2 at matched variance the comparison reverses:** $+0.42$ **and** $+0.27$ **at two layers, and within GPT-2 the sign depends on whether atoms or groups are aggregated** (appendix X.37)	**withdrawn**
The worked example as a model-level effect ($3.1\times$, -0.461 $[-0.581, -0.340]$)	**Seven of ten layer–side cells admit no non-degenerate matched control; among the three that do the pooled gap is** $+0.232$ $[-0.791, +0.776]$ (appendix X.40)	**weaker**
"Across 30 trained dictionaries, 29 are less disruptive than random" ($2.9\times$**,** $\rho = -0.57$**)** — our own opening paragraph	**The control was isotropic, not variance-matched — the error this paper's Contribution 2 is about — and the pool counted 15 epoch snapshots of three trajectories as independent. On 15 independent dictionaries against a matched control it is** $3.1\times$	**weaker**
The GPT-2 reversal as $+0.492$ $[+0.345, +0.637]$ **and** $+0.348$ $[+0.270, +0.422]$	**The two arms' variance supports were disjoint and every gap was computed against one constant; inside common support with a** 0.25 **dex caliper and a two-way bootstrap it is** $+0.418$ $[+0.238, +0.603]$ **and** $+0.270$ $[+0.157, +0.372]$ (appendix X.37.1)	**weaker**
"The measure ranks known-causal directions above noise" (general form)	Only the first-order maximal anchor clears a *matched* control; the information-erasing anchor's matched difference covers zero and a linear probe falls below its matched controls (appendix X.17)	weaker
The worked example's size as -0.461 **(**$3.1\times$**), full stop**	**Matched on** $\|Jd\|$ **— the covariate that predicts the outcome — it is** -0.293 $[-0.497, -0.108]$**, and only** 7 **of** 15 **directions admit a control at all; the two covariates cannot be matched simultaneously** (appendix X.19.1)	**weaker**

continued on the next page

table 10, continued

Withdrawn claim	Cause	Direction
"Off-manifold encoding deflates the effect $69\times$**"** (as a property of the encoding)	**A norm-matched** 2×2 **puts the encoding effect at** $0.78\times$ $[0.66, 0.90]$ **and the dose effect at** $76\times$**: the factor is mediated entirely by the larger perturbation a better-reconstructing code produces** (appendix X.20)	**reversed**
"Aggregation raises sensitivity at both layers" ($+0.018$ $[+0.001, +0.036]$ at L6)	The 72 windows overlap; clustered on window starts the L6 interval is $[-0.006, +0.043]$ and the claim narrows to L8 (appendix X.21)	weaker
The worked example as a comparison our own tool would pass	**`tabdiff.diagnostics.audit` returns INADMISSIBLE on it: the control has PC1 mass** 0.889 **and participation ratio** 1.3 **of** 192	**weaker**
The GPT-2 reversal as a magnitude ($+0.42$ / $+0.27$)	**On a TopK dictionary of the same model and layers it is** $+0.04$ **and** $+0.12$**, covering zero at L8 — and that arm has no unmatched directions where the L1 arm has 13 of 28. The sign holds in all eight cells; the size does not** (appendix X.26)	**weaker**
"The response is not bilinear" (general form)	**Zero of the 16 grid cells lie inside our own admissibility bound, so the refutation is a statement about the regime above it** (appendix X.27)	**weaker**
"The measure is anchored at the large end only"	Two discriminant anchors score $0.07\times$ and $0.18\times$ the random median, below their matched controls with intervals excluding zero: the measure separates both extremes and is flat between them (appendix X.17)	**reinstated**
The tabular deficit as a property of dictionary learning on this model	**On a single-model L1+ReLU SAE at the same cell it is** $+0.068$ $[-0.108, +0.314]$**, and 16 of 36 directions have no admissible match; the result belongs to this crosscoder** (appendix X.29)	**weaker**
"The response is not bilinear" (restated)	Inside the admissibility bound it *is* bilinear (CV 0.064 against 2.07 outside), so the refutation localises to the inadmissible regime rather than describing the response in general (appendix X.30)	**reinstated**
Aggregation as a corrected result	**Raw** $p = 0.008$ **throughout. It cleared Benjamini–Hochberg at** $q = 0.049$ **only while the family counted five entries that could not take their null; on the twelve-entry family it failed Holm (**0.056**); on the nineteen-entry family that counts the dose ratio and the anchors as the tests they are, it passes both (**0.032**,** $q = 0.0097$**), because smaller** p**-values moved ahead of it. Both values are printed** (appendix X.32)	**weaker**

continued on the next page

table 10, continued

Withdrawn claim	Cause	Direction
"On a language model the comparison reverses" (general form)	**On Pythia-160m none of six cells resolves and four point estimates are negative; the reversal is established on GPT-2 and undetermined on the second model we tried** (appendix X.33)	**weaker**
"The admissibility bound marks the linear/saturating transition"	A grid at $1.25\times$ the bound is still bilinear (CV 0.049); the bound sits inside the linear regime and the transition is only bracketed to $[1.25, 3.92]$ α^* (appendix X.30)	weaker
The tabular deficit as a fact about "the directions this dictionary learns"	**Holding architecture, seeds, cell, control and estimator fixed and changing only the group definition gives** $+0.072$ $[-0.192, +0.342]$ **against** -0.544 $[-0.649, -0.447]$ **for the shared group; the result is about that group** (appendix X.34)	**weaker**
"The L1 family sits above the TopK family"	In a proper factorial the sparsity contrast at single input runs the other way (-0.274 $[-0.514, -0.034]$); two settings made a pattern a 2×2 does not support (appendix X.34)	weaker
"The model is the first axis able to move the sign" (our own reading of the Pythia-160m null)	**That arm had only TopK dictionaries, which sit near zero on GPT-2 too. On Pythia-70m-deduped the L1 arm resolves positive in** $6/6$ **cells and TopK in** $2/6$**: the family separates them, not the model** (appendix X.33)	**withdrawn**
"Secondary quantities move in the third significant figure between runs"	A full end-to-end rerun retrains the canonical crosscoders at a different BLAS thread count and 68 paper-facing numbers move, 21 by more than 50%; three of the 56 verdicts it covered change, one against us (appendix X.35)	**withdrawn**
"On a joint code the interaction is 17% **of the apparent effect"** (as a single number)	0.174 **on the original dictionaries,** 0.513 **on retrained ones — across the pre-registered support and refutation thresholds; the direction holds, the size is dictionary-dependent** (appendix X.35)	**weaker**
"-0.536 $[-0.611, -0.461]$ **for the shared/exclusive group definition"**	**No committed script produces this value; an audit of every leaf value in all 414 result files found no match, and the interval's upper bound is the full-fifteen headline's point estimate. Recomputed on the same three dictionaries as the window arm it is** -0.544 $[-0.649, -0.447]$**; the contrast stands on the recomputed number** (appendix X.43)	**corrected**
"Doubling ε multiplies the no-op effect $86\times$"	No aggregation of the stored records gives 86: the ratio of means is $68\times$, the per-instance median $92\times$, the mean $185\times$ (appendix X.43)	**corrected**

continued on the next page

table 10, continued

Withdrawn claim	Cause	Direction
"Every quantitative claim is read back from a committed artifact"	Seven printed numbers were produced by no script; the binding gate compared manifest entries against artifacts and never opened the LaTeX sources (appendix X.43)	**withdrawn**
"The non-overlapping refits agree in sign and are larger" ($+0.099$ L8, $+0.101$ L6)	**The disjoint subset keeps four windows all of width 256, so $\log_2 w$ is collinear with the intercept and the design has rank 2 of 3. The printed coefficient is a minimum-norm tie-break; zeroing it leaves the residual sum of squares unchanged. No estimate exists at $n = 4$** (appendix X.21)	**withdrawn**
"Shared features are those active on both single-sided inputs"	The shipped classifier decides dead from the *joint* activation frequency and sweeps every non-exclusive survivor into shared, so features active on neither single-sided input land there too: 399 of 3617, 11.0%, present in all fifteen dictionaries (appendix X.50)	**corrected**
"The off-manifold ratio is $69\times$ over fifteen instances"	One instance has a sign-reversed single-side effect and is silently dropped by a positivity filter, so the geometric mean is over fourteen; the ratio-of-means printed beside it uses all fifteen (appendix X.43)	**corrected**
p-values and a Holm rejection attached to the surviving matched gap	**§34.1's positive control, re-run in the headline's own estimand and at its own norm, resolves $1.73\times$ against a pre-committed $3\times$. We had applied §34.1's consequence to the withdrawn §5.4 result only; an unvalidated dose is a property of the measurement, so it applies here too. -0.461 and $3.1\times$ become descriptive** (appendix X.49)	**withdrawn**
"Match on $\|Jd\|$ and the ceiling becomes a floor: no random direction responds as little as these"	**A property of our cone generator, not of the sphere: $\min_{\|d\|=1} \|Jd\| = \sigma_{\min}(J)$, and appendix X.17 uses such directions as anchors. The generator reaches 7 of 15; a control built from the minor singular subspace would reach the rest** (appendix X.19.1)	**weaker**
"$N = 768 \leq 1{,}600$, so the dictionary cannot memorise"	A feature count below the sample count limits memorisation without excluding it; the encoder and decoder remain free (appendix X.25)	weaker
"Both arms sit at the same norm, so a monotone distortion preserves the difference's sign"	Only if both directions share one monotone response; the grid shows direction-specific plateaus, so a common norm does not rescue even the ordering (§6.2)	withdrawn
"Every estimand in the crosscoder literature encodes one side with the counterpart zeroed"	Minder et al. (2025) build and intervene on a joint code; the single-side encoding is what *our* pipeline did, and what any single-side classifier is exposed to (appendix W.3)	**withdrawn**

continued on the next page

table 10, continued

Withdrawn claim	Cause	Direction
"The norm-ratio taxonomy fails for TopK crosscoders"	It fails under per-side L_2 normalisation of the decoders, a choice of ours; the cited prior work reports norm indicators working under BatchTopK with a different normalisation (appendix K)	weaker
"On a public SAE the group-width sweep reverses the point estimate, and the answer moves $16.5\times$"	**Six learned groups. At twenty-four every width resolves positive and rises monotonically; the six starts were low-sensitivity ones and we read the sign of an interval covering zero. The span with power is $4.8\times$, nearly all of it the control** (appendix X.45, PREREG §87.1)	**withdrawn**
"Training on the intervened positions does not produce the deficit" (-0.304 $[-0.480, -0.100]$)	**That arm held out the rows from dictionary *training* only; the classification, which decides which features are "shared", still read the full population. Holding it out too moves the arm's directions to variance share 0.131–0.160 against the control's 0.191–0.242 — 0 of 9 inside, so no matched contrast exists. The threat is now unmeasured rather than excluded** (appendix X.25)	**withdrawn**
"On a public SAE four modes move the answer $4.8\times$, almost all of it the control; the fixed dose is inert"	**The rungs changed dose, control and intervention form together, overlapping windows were resampled as independent, and 4.8 was an L6 maximum. An unconfounded 2×2 resolves both factors, the dose's larger, with a positive interaction (appendix X.45)**	weaker
"GPT-2 cells sit at a validated dose ($0.44\alpha^*$); tabular gaps at $2.67\alpha^*$ inside $[1.25, 3.92]\alpha^*$"	**Three incompatible coordinates shared the name α^*. On one scale defined identically on both models, neither dose passes: GPT-2's reference is past ρ^* and saturating (appendix X.49)**	weaker
"`audit` checks the response at the dose being reported"	It unit-normalised the direction and swept 0.25–4 in absolute norm, $1/78$ of GPT-2's reference; now swept around the supplied norm, with a regression test (PREREG §97.1)	weaker
"P3 is supported"	Its third registered condition, the aggregate thinking share, was never computed under the primary estimand; two of three conditions hold (table 18)	weaker
"The mass-matched attention contrast survives Holm"	Two GPT-2 TopK tests had been dropped from the family by an edit; restored, its Holm-adjusted p is 0.068 (table 47)	weaker
"Adding tests with smaller p lowered the aggregation coefficient's Holm-adjusted value"	Holm's multiplier for a claim is the number of tests ranked behind it, which tests ranked ahead do not change; the explanation is now computed (appendix X.32)	weaker
"`audit`'s dose check judges linearity at the reported dose"	It read only the top ratio of its ladder ($2r \to 4r$); a response with a bump at the reference passed. Now both ratios touching the reference decide (PREREG §100)	weaker

continued on the next page

table 10, continued

Withdrawn claim	Cause	Direction
"On GPT-2, feature groups sit above matched controls"	**What was measured is the group's mean direction broadcast to every token; the group's own per-token offsets change the effect 8–13× at the same norm, so the statement is about directions, not group ablations (appendix X.47)**	weaker
"`audit` is a diagnostic of whether a comparison is valid"	Against a known forward model admission barely lowered sign errors ($15.5\% \rightarrow 14.4\%$), failing its registered criterion; it is a functional check (appendix X.23)	weaker
"On a public SAE the dose moves the answer more than the control does"	True of broadcast directions only; injected as the group's actual per-token removal, the control main effect dominates and the choice of control flips the sign at the reference norm (appendix X.48)	weaker

Three of these went the other way and we record them for symmetry: the K3 deep-pair shared ratio *rose* from 2.72 to 3.12 when a correlated random null was fixed (the fix was pre-committed before the re-run); the deep-pair interaction share rose from $+0.43$ to $+0.53$ when the cohort was matched to the canonical run's; and the determined-dictionary arm, which we had reported as undecided at three seeds while saying we could not tell a null from a power failure, resolves at -0.314 $[-0.476, -0.113]$ at nine. It was a power failure. None was adopted because it helped.

F Cross-architecture crosscoder: three variants and an open problem

Section 6 reports that the vanilla crosscoder applied to TabICL (512-d) × TabPFN v2.5 (192-d) reconstructs the higher-dimensional side well (val $R^2 = 0.855$) but the lower-dimensional side poorly (val $R^2 = 0.434$). We test two progressively more aggressive architectural fixes; both improve the lower-dimensional side modestly but neither reaches the 0.7 threshold we require for confident shared-feature classification (table 11).

Table 11: Cross-arch crosscoder progression on the TabICL L8 ↔ v2.5 L7 pair (best CKA pair). All variants fail to reach val $R^2 \geq 0.7$ on the v2.5 side, indicating that dimension asymmetry between architectures is a structural rather than algorithmic limitation of the crosscoder framework.

Variant	R^2 TabICL	R^2 v2.5	n_{shared}	$n_{\text{TI-excl}}$	$n_{\text{v2.5-excl}}$
Phase C: vanilla	0.855	0.434	117	569	7
Phase C2: per-side LN + weighted MSE	0.840	0.522	139	569	4
Phase C3: separate enc. + sum-combine	0.840	0.513	138	610	6

Interpretation. Each architectural intervention (input LayerNorm, asymmetric loss weighting, separate per-side encoders) moves the lower-dim side from 0.434 to ≈ 0.52, but the gap to the 0.7 threshold persists. The shared-feature count grows modestly ($117 \rightarrow 139 \rightarrow 138$) but the architecture-exclusive split remains degenerate (TabICL-exclusive $\approx$ 570–610 vs v2.5-exclusive $\approx$ 4–7). All three variants allocate the bulk of the sparse code to representing TabICL's richer representation, leaving little capacity for v2.5-specific features.

Why architectural fixes plateau. The crosscoder's joint optimisation must reconstruct both sides simultaneously. With a fixed feature budget, additional capacity allocated to v2.5 reconstruction comes at the cost of TabICL reconstruction, and TopK selects features that maximise the *joint* loss. The data tells us that TabICL's information content (in the layers we test) substantially exceeds

v2.5's at the same depth, so any joint sparse code is biased toward TabICL representation. We hypothesise that decoupling the two sides entirely—training two SAEs and using a cross-model alignment objective only over their features—may be necessary, but this departs from the crosscoder framework as currently defined.

Update: this is no longer an open problem in general. Jiralerspong & Bricken (2026) apply crosscoders across architectures using Dedicated Feature Crosscoders, reporting the first such application. Our negative result should be read narrowly against that: what we document is that *vanilla* crosscoder variants do not close a $2.7\times$ embedding-dimension asymmetry (512-d vs 192-d) at our scale, not that cross-architecture diffing is unsolved. We have not evaluated DFCs on this pair.

Bottom line. Cross-architecture mechanistic diffing across substantially asymmetric embedding dimensionalities is an open problem that is not solved by within-framework architectural variants of the crosscoder. We position this as a contribution of the present work: we identify a specific failure mode (per-feature ratio collapse onto the higher-dim side), empirically characterise the magnitude of the gap, and document three plausible-looking fixes that fall short. Future work may need to step outside the joint-encoder regime.

G LARGE-CONTEXT GPU SCALE-UP

Superseded. This appendix predates the estimand corrections of section 3.4 and appendix W.3: it uses single-side encoding and the full ablation dose. Its numbers are retained as the record of what the larger operating point looked like under the old protocol and are not evidence for any claim in the main text. The scale question is answered under the current estimand in appendix X.25 (the $9.6\times$ arm).

All experiments in the main text use $n_{\text{train}} = 128$, runnable on CPU. TabPFN forbids CPU inference above 1000 samples, so a larger in-context regime is a genuinely distinct, GPU-only operating point—not merely a faster one. We replicate the headline causal analysis there: $n_{\text{train}} = 512$, $n_{\text{test}} = 128$, on the 12 OpenML regression datasets with ≥ 640 rows (the five smallest—`diabetes`, `boston`, `auto_mpg`, `machine_cpu`, `cpu_small`—cannot supply the split and are necessarily dropped). The canonical-pair crosscoder is retrained on GPU from $8{,}448$ paired full-sequence positions ($5.3\times$ the CPU-scale $1{,}600$), reaching val $R^2_{\text{v2.5}} = 0.9991$ (vs 0.997). The RP-estimand P1/P2 grid (table 12) runs sharded across four GPUs over 36 instances (23 addition-eligible).

Table 12: Phase L: per-feature ratio (RP estimand) at $n_{\text{train}} = 512$, 12 datasets, $n = 36$, block-bootstrap CIs. Both sides; `data_only` scope.

Side	r_{shared}	$r_{\text{v2-excl}}$	$r_{\text{v2.5-excl}}$	shared > random
v2	1.56 [1.38, 1.79]	1.08	0.39	$p < 10^{-4}$ (Holm)
v2.5	1.27 [1.17, 1.40]	0.24	2.00 [1.64, 2.48]	$p < 10^{-4}$ (Holm)

The confirmatory results replicate *under the superseded estimand*. Everything in this subsection was measured with single-side encoding at $\alpha = 1$, which appendix V shows is neither on-manifold nor an admissible dose; the replication therefore says that the GPU-scale run reproduces the *same measurement*, not that the claim survives. We have not re-run Phase L under joint encoding. Shared features exceed random on both model sides (Holm-corrected sign-flip $p < 10^{-4}$), and the causal de-concentration result holds: v2 concentrates more per-feature weight on shared features than v2.5 does ($r^{\text{v2}}_{\text{shared}} = 1.56 > r^{\text{v2.5}}_{\text{shared}} = 1.27$, paired mean difference $+0.29$, $p = 0.0066$, Holm-rejected). The addition null also replicates: shared injection closes -2.7% $[-5.3, -0.7]$ of the gap and is TOST-equivalent to no transfer at the ± 10pp margin.

One scale-dependent observation (exploratory). On the v2.5 side, the v2.5-exclusive ratio rises from 0.84 at $n_{\text{train}} = 128$ to 2.00 at $n_{\text{train}} = 512$, overtaking shared. Unlike the deepest-pair pattern of appendix B, this is *not* a reconstruction-quality artifact—the Phase-L crosscoder is high-fidelity (val $R^2 = 0.9991$). It indicates that with more in-context training rows, v2.5's exclusive features carry more causal weight on the v2.5 side. We flag it as a genuine scale-dependence finding; it does not

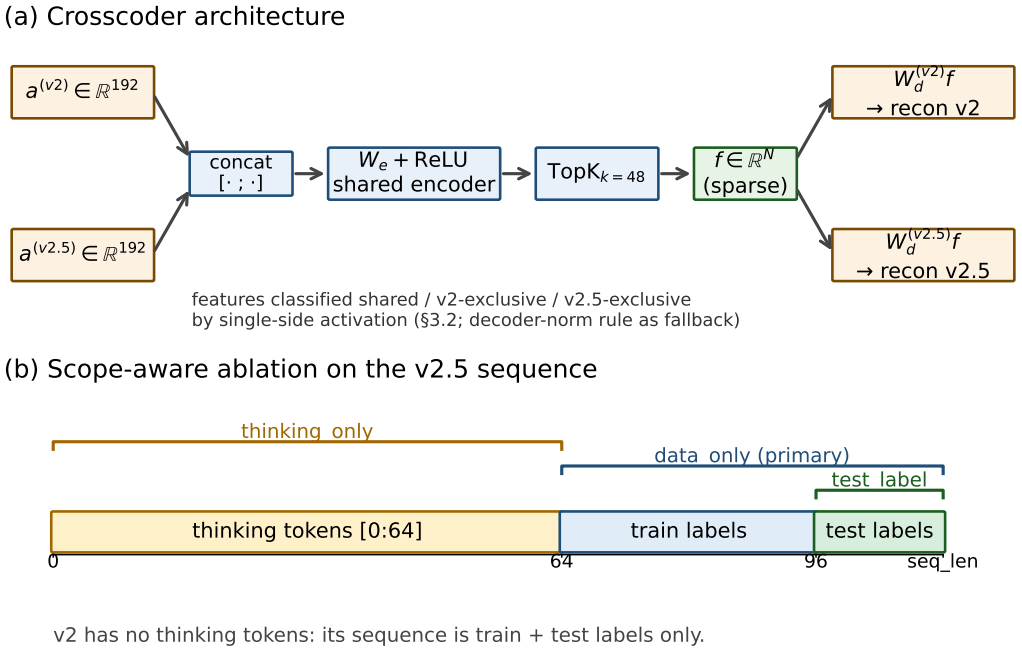


Figure 4: (a) Crosscoder architecture: a shared encoder maps concatenated $(a^{(v2)}, a^{(v2.5)})$ activations through a TopK sparsity to a sparse feature code f, which two per-model decoders reconstruct. (b) Scope-aware ablation positions: an ablation modifies the encoder's features only at the chosen subset of sequence positions (`thinking_only`, `data_only`, or `test_label`). v2.5 prepends 64 thinking tokens absent from v2.

disturb the confirmatory claims as measured here (shared $>$ random on both sides, de-concentration, no transfer), of which only the transfer null survives the corrected estimand.

Gap-prediction diagnostic: negative. A natural application of the larger cohort is to ask whether crosscoder activation statistics (per-group activation rates and magnitudes, thinking-token rates) predict which datasets v2.5 wins. They do not: leave-one-dataset-out ridge gives $R^2 = -0.02$ (pre-registered bar 0.30). The per-instance capability gap is dominated by per-dataset target scale (mean $+17.7$, sd 166.6 in raw RMSE units), which these scale-free features cannot capture. We report this transparently; a scale-normalised target is left to future work.

H Estimand definitions and supporting analyses

This appendix holds material moved out of the main text for length: the full definitions of the four estimands (appendix H.1), the multi-estimand convergence check (appendix H.2), and the feature-interpretation case studies (appendix H.3).

H.1 The four estimands

Ablation. Zero g's features in the encoded representation; RP-patch the chosen model side at the target layer; measure relative prediction change vs clean.

Addition. Encode the donor model's activations through the crosscoder to obtain donor features; for each index in g, set the recipient model's feature value to the donor's value; RP-patch the recipient model.

Swap. Mutual exchange of shared-feature values between v2 and v2.5—a sanity check that "shared" features genuinely encode the same concept.

Path-specific decomposition. Standard NDE/NIE decomposition (Pearl, 2001) treating crosscoder feature groups as named mediators.

All four share a common cache-then-RP-patch pattern on the target layer, with a transformation function (zero, set, swap) varying per estimand. The paper-facing estimands use the scope-aware patch path in the Phase H/J/L validation code; the reusable library primitive (`tabdiff.analysis.mediation`, default `patch_mode="delta"`) implements the same RP injection for the test-label scope and is unit-tested for exact no-op equivalence.

H.2 Multi-estimand convergence (Phase 4)

To stress the interpretation from multiple angles, we run all four estimands on a common layer pair (real-data L2 $\leftrightarrow$ L1), under the primary RP estimand. The ablation result is reported in section 5; the swap and path-specific decomposition results are in table 13. The swap test (mutual exchange of shared-feature values between v2 and v2.5) produces a small RMSE change consistent with the addition result—shared features encode comparable but not identical concepts in both models, and direct value exchange is non-disruptive but not capability-restoring. The path-specific decomposition is the sharpest illustration of the estimand switch: under replacement patching the shared group's NIE appeared to account for 72.9% of the capability gap; under RP patching it is $+1.0\%$ of $|\mathrm{TE}|$ (CI excluding zero)—directionally real, but small. No identifiable feature group mediates more than a small fraction of the gap.

Table 13: **[Superseded — see appendix V.]** Phase 4 multi-estimand convergence on the real-data best pair (v2 L2 $\leftrightarrow$ v2.5 L1), primary (RP) estimand throughout except the swap sanity check. Ablation r_g (per-feature ratio, v2.5 model, `data_only` scope; $n = 30$), addition gap-closure ($n = 16$), swap (shared mutual exchange, $n = 5$ datasets, legacy estimand), and path-specific NIE (ablating each group in v2.5; $n = 15$: 3 seeds $\times$ 5 datasets, dataset-block bootstrap CI). RP no-ops are exactly zero by construction; the v2-exclusive NIE is exactly zero per record because v2-exclusive features do not activate under v2.5-side encoding—a built-in self-consistency check of the classification. The legacy replacement-mode NIE for shared was $+72.9\%$ of $|\mathrm{TE}|$; the collapse to $+1.0\%$ under the clean estimand is the NIE-level signature of the reconstruction-error interaction (appendix E.1).

Estimand	shared	v2-exclusive	v2.5-exclusive
Ablation r_g vs random, RP	1.25 [$1.11, 1.39$]	0.42 [$0.27, 0.62$]	0.82 [$0.71, 0.99$]
Addition gap-closure, RP (%)	-0.3 [$-1.9, +2.2$]	—	-0.6 [$-1.8, +0.2$]
Swap (shared $\leftrightarrow$ shared, legacy)	-0.0% v2 / $+0.2\%$ v2.5	—	—
Path-specific NIE, RP (% of |TE|)	$+1.0\%$ [CI excl. 0]	0.0% (exact)	-0.8%

Reading the table. The *ablation* row reproduces table 21's v2.5 side (Phase 4 is a separate execution of the same protocol, so its point estimates differ from table 21 in the third significant figure—e.g. v2.5-exclusive 0.82 here vs 0.84 there—within the run-to-run variation documented in appendix J): shared features carry modestly more per-feature causal weight than random alive features ($1.25\times$). We do not attach a p-value to this contrast here: under the per-group size-matched null it is not significant ($p = 0.075$, appendix S.4), and the $p = 0.018$ quoted in earlier versions came from the largest-group-matched null. The *addition* row reproduces the P2 result of section 5: no variant transfers capability. The *swap* row reports the relative RMSE change on each model after a mutual exchange of shared-feature values; both changes are within $\pm 1\%$, confirming that "shared" features encode comparable concepts in v2 and v2.5 and are not spuriously labelled by the activation-method classifier. The *path-specific NIE* row reports the natural indirect effect of ablating each group in v2.5 on the v2$\rightarrow$v2.5 capability gap: the shared group's NIE is positive with CI excluding zero (RP NIE $+0.037$ $[+0.002, +0.083]$ RMSE) but amounts to only 1.0% of the mean gap; exclusive groups account for 0.0% (exactly) and -0.8%. Together the four estimands tell a single, consistent story under the clean estimand: **shared features are the most causally relevant identifiable group at this layer, but no identifiable feature group—shared or exclusive—carries more than a small fraction of the capability gap, which is instead distributed across the joint computation** (section 6).

H.3 Feature interpretation case studies

Among the 8 highest-activation-magnitude shared features, Feature 322 activates at 90–94% frequency across all 5 reference datasets and exhibits consistent negative correlation with the target across

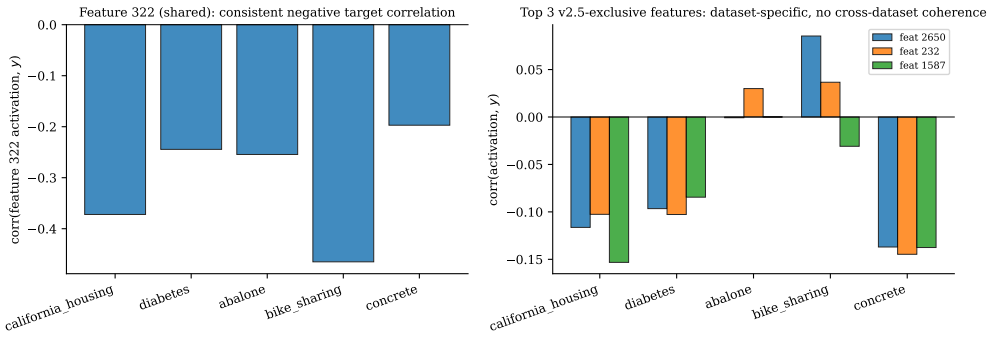


Figure 5: Feature 322 (shared) case study. Across 5 OpenML regression datasets, Feature 322 activates at 90–94% frequency and correlates negatively with target y (-0.20 to -0.47). Top v2.5-exclusive features (gray) show heterogeneous, dataset-specific correlation patterns.

heterogeneous tabular domains (bike_sharing -0.47, california_housing -0.37, abalone -0.25, diabetes -0.24, concrete -0.20). Its strongest input correlations tag interpretable magnitude-bearing inputs in each domain (`hour` for bike_sharing, `MedInc` for housing, `Shell_weight` for abalone, `BMI` for diabetes, `Cement` for concrete). We interpret Feature 322 as a cross-dataset *magnitude/scale detector*—exactly the kind of generalisable inductive bias one would expect to be shared between v2 and v2.5. The top v2.5-exclusive features (2650, 232, 1587) show dataset-specific correlation patterns without cross-dataset coherence (fig. 5).

I Extended related work

Table 14: What prior audits of patching and of dictionary evaluation examine, and what this paper adds. Each row states the prior work's own scope; the right column is our addition, not a criticism of it.

work	what it examines	what this paper adds
Zhang & Nanda (2024), Heimersheim & Nanda (2024)	how the corruption method and metric change activation-patching conclusions	the dose and the comparison control as separate factors, with their interaction measured (section 5.3)
Marks et al. (2025)	sparse feature circuits that carry the SAE error term	the size of that term against two larger modes (E1, E2), and when it matters
Minder et al. (2025)	sparsity artifacts in crosscoder latents; joint encoding and latent scaling	off-manifold single-side encoding measured as a mode, on two crosscoders
Karvonen et al. (2025), Mueller et al. (2025)	standard metrics for scoring dictionaries	the ablation contrast those metrics are built from, and its six failure modes
Chanin (2026)	reliability of benchmark metrics (reseed noise, synthetic ground truth)	reliability of the learned-vs-random comparison itself; a spectral limit on its control (Prop. 1)
Leask et al. (2025), Gerasimov et al. (2026)	whether latents are canonical, seed-stable units	that the comparison's sign moves with the group and aggregation unit (C3), and invariances a score must keep
Braun et al. (2024)	training dictionaries on the output distribution	evidence that the training method decides whether the comparison resolves

Prepended latent tokens. They have two opposed poles, gist tokens that compress a prompt (Mu et al., 2023) and pause tokens that buy computation without content (Goyal et al., 2024); appendix P adjudicates between them causally for TabPFN v2.5's thinking block.

Patching practice. Heimersheim & Nanda (2024) is the practitioner's account of how to use and read an activation patch, and it names several of the choices we measure — the metric, the direction of the patch, the granularity of what is patched — as choices rather than details. This paper's contribution on that axis is to price them.

Evaluating dictionaries, and evaluating the evaluations. SAEBench (Karvonen et al., 2025) and MIB (Mueller et al., 2025) standardise what a dictionary is scored on; Chanin (2026) then audits SAEBench's own metrics and argues that two of them, TPP and SCR, should not be used at their canonical settings. Our six modes sit one level below both: they are properties of the ablation contrast that every such metric is built from, so a benchmark inherits them rather than controlling for them. That is also a caution on our own citation practice — we point at these suites as the field's standard without leaning on any particular number they report.

The negative results are convergent. Kantamneni et al. (2025) find SAEs do not beat baselines on sparse probing; Korznikov et al. (2026) report a random-direction baseline essentially level with a trained SAE on causal editing, interpretability and probing; Bal (2026) find dictionary features that fail to fire causally; and Gerasimov et al. (2026) find SAE latents seed-unstable while the subspaces they span reproduce, which is the dictionary-side analogue of our own thread-count sensitivity (appendix X.35). Earlier drafts of this paper claimed a separation none of them shows, and after appendix X.37 we no longer do.

The unit of analysis is contested, and we make it a causal question. Leask et al. (2025) show SAE latents are not canonical units: a larger dictionary's latent decomposes under a meta-SAE into parts a smaller one already has. Chanin et al. (2025) document feature absorption, where one latent swallows a direction another should own, and Bussmann et al. (2025) make granularity explicit in the training objective. That literature asks which unit is the *right* one. C3 and the aggregation sweep (appendices X.21 and X.41) ask a different question and get a worse answer: with the dictionary, the width, the sparsity, the seeds, the cell, the control and the estimator all held fixed, changing only which atoms are called "the group" moves the comparison from -0.544 to $+0.072$. The unit is not merely ambiguous; it decides the sign.

Sparse-coding interpretability of neural networks. The line of work most directly relevant here begins with sparse autoencoders applied to LLM internal activations: Anthropic's monosemanticity programme (Bricken et al., 2023) established that an L_1-regularised autoencoder with sufficient overcompleteness recovers an interpretable, mostly-monosemantic feature basis from polysemantic neurons. Cunningham et al. (2024) concurrently demonstrated the same on Pythia and showed that the discovered features are causally meaningful via activation patching. The Top$K$ activation we adopt originates with k-sparse autoencoders (Makhzani & Frey, 2014) and was developed for modern SAEs by Gao et al. (2025), who show that replacing L_1-soft sparsity with a hard-K activation simplifies sparsity control and improves the reconstruction–sparsity frontier; Minder et al. (2025) adopt a hard-K (BatchTopK) loss in the crosscoder setting. Closest to our estimand analysis, Marks et al. (2025) build causal graphs over SAE features in which the reconstruction error appears as an explicit node, so that interventions on features do not silently delete it—the same principle our RP patching enforces at the level of a single patched layer. Our contribution relative to that line is quantitative: we measure how large the confound is when it is *not* preserved, and show it is large enough to reverse a diffing conclusion.

Crosscoders and model diffing. Crosscoders (Lindsey et al., 2024; Mishra-Sharma et al., 2025) extend SAEs to multiple models by sharing an encoder. The shared encoder forces feature discovery to jointly explain both models' activations and provides a natural "shared vs exclusive" taxonomy. The most extensive published applications have been to LLMs (Claude across versions, base vs chat tuned). To our knowledge no prior work has applied crosscoders to tabular foundation models. Mishra-Sharma et al. (2025) report that model-exclusive latents come out dense and polysemantic through feature competition, and mitigate it with designated shared features; we do not adopt that device. Two practical choices here have other sources: the auxiliary dead-latent loss follows Gao et al. (2025), and the activation-based classification rule for feature taxonomies under TopK (rather than decoder-norm thresholding) is our own, motivated by the norm-based misattribution that Minder et al. (2025) document. Minder et al. (2025) further show that L_1 crosscoders manufacture spurious model-exclusive latents through a sparsity artifact and that a hard-K (BatchTopK) loss largely removes it—direct support for our TopK choice and for treating the exclusive-feature groups with the caution we apply throughout. Most relevant to our naming, Kassem et al. (2026) propose the *Delta-Crosscoder*, which augments *training* with a delta-based loss (plus BatchTopK) that prioritises directions changing between models. Their "delta" modifies the dictionary-*learning* objective; ours modifies the downstream *patching* operation's estimand (adding the decoded feature change to the

original activation). The two are orthogonal—a Delta-Crosscoder dictionary could be intervened on with either replacement or our residual-preserving scheme—and we adopt the name "residual-preserving (RP)" precisely to avoid conflation with their loss-level "delta". Model-level diffing more broadly has been pursued at the weight level (task arithmetic, model merging) and the prediction level (behaviour-pair analysis), neither of which gives the mechanistic per-feature handle that crosscoders provide. A distinct line of within-model analysis—circuit discovery, which isolates a minimal component subset sufficient for a *single* model's behaviour—answers a within-model sufficiency question and is orthogonal to the cross-version feature diff we pursue: we compare what two versions represent, not which components one version minimally needs.

Activation patching and causal abstraction. Activation patching—running a clean forward pass, caching at a target layer, and re-running with the cache modified—is the standard mechanism-confirmation tool in LLM interpretability. That its *implementation choices* are not neutral is established: Zhang & Nanda (2024) show the choice of metric and corruption method changes which conclusions a patching study reaches, and Makelov et al. (2024) show a subspace intervention can produce the intended output change by activating a dormant parallel pathway rather than the mechanism it purports to test. Our result is the sparse-dictionary case of the same lesson: the replacement-vs-RP choice determines the estimand, and the difference is large enough to reverse a conclusion. It is complementary rather than redundant—their illusions concern *which pathway* carries an effect, ours concerns *what is removed alongside* the targeted features. Causal abstraction (Geiger et al., 2025) provides the formal underpinning: a model-level intervention on a named mediator (here, a crosscoder feature group) corresponds to a counterfactual at the higher-level causal graph. Pearl's NDE/NIE decomposition (Pearl, 2001) formalises our path-specific analysis: shared, v2-exclusive, v2.5-exclusive features are the named mediators, and we report each group's NIE on the v2 $\rightarrow$ v2.5 capability gap. The reliability of patching-based estimands is itself under scrutiny: Zhang & Wang (2026) decompose attribution-patching error into local-curvature and downstream-nonlinearity terms and propose a second-order correction. Our residual-preserving (RP) patching (section 3) targets a distinct but related failure—the reconstruction error a sparse dictionary injects when its output replaces the original activation—and removes it exactly by construction rather than approximately.

Tabular foundation models. TabPFN (Hollmann et al., 2023; 2025) is the tabular foundation model on which our diff focuses. Its successor TabPFN-2.5 (Grinsztajn et al., 2025) is the object of our diff. Concurrent and follow-up tabular foundation models include XTab (Zhu et al., 2023), which adopts cross-table pretraining, and TabICL (Qu et al., 2025), which targets larger tabular contexts via dedicated in-context learning architecture. We use TabICL as the cross-architecture comparison in section 6, finding that the crosscoder framework requires dimension-aware modifications to apply across architectures with substantially different embedding dimensionalities.

Mechanistic interpretability of tabular foundation models. Interpretability of TabPFN began with *post-hoc* attribution: Rundel et al. (2024) adapt Shapley values and LOCO to exploit TabPFN's in-context learning. A more recent line probes the network internally. Gupta et al. (2026b) show through linear probing that intermediate layers encode concrete quantities (regression coefficients, partial results); Sharma et al. (2026) give mechanistic evidence that prior-data fitted networks develop linearly-decodable spectral structure that causally drives predictions; and, closest to us, Gupta et al. (2026a) perform *causal* localisation of attention-head function in TabPFN via activation patching, identifying temporally specialised heads. All three analyse a *single* model: none compares two versions, and none decomposes the representation into a sparse feature dictionary. Our contribution is orthogonal and complementary—a feature-level, crosscoder-based diff *across* versions—and we adopt the attention-pattern analysis of appendix X.42 in the same causal spirit as Gupta et al. (2026a) while adding the cross-version dimension they do not study.

Model stitching, and what our transfer nulls do and do not show. Our value-injection experiments—splicing one model's features into another's forward pass, with and without a learned linear map—are model stitching in the sense of Bansal et al. (2021), who show that comparing representations by stitching generally requires a *trained* connector and that an untrained or frozen map failing is the expected outcome rather than evidence about where capability lives. We take this as a genuine limit on our nulls: our ridge map is fit post-hoc rather than trained end-to-end at the injection site, so P2/F1/Phase-N bound *feature-value* transferability under the maps we tested, and

do not establish that no learned connector could transfer the gap. We state the claim at that scope throughout.

How much work is the dictionary doing? A negative result built on sparse dictionaries owes the reader an argument that the dictionary carves the representation usefully. Our evidence is within-method: taxonomy stability across bootstraps (Jaccard 0.77 shared) and across a 15-dictionary ensemble. That is consistency, not validity, and the sceptical literature is directly relevant: Kantamneni et al. (2025) find SAE features do not reliably beat simple baselines on sparse probing, which is a reason to read "the shared-feature *values* do not transfer" as a statement about this dictionary's decomposition rather than about the model's representation in general. appendix T sharpens the same worry from a different direction: on the pathway the estimands actually use, single-side reconstruction is 0.57, not the 0.998 the qualification gate reports.

Representation similarity in cross-model studies. Centered Kernel Alignment (Kornblith et al., 2019) is the standard tool for between-model representation similarity. We use it for layer alignment (which v2 layer corresponds to which v2.5 layer) but find it to be an unreliable substitute for crosscoder-based mediation, both because it is correlational rather than causal and because its layer-pair selection depends on the input distribution (section 5, P4).

Pre-registration in machine-learning research. Pre-registration is rare in machine-learning but increasingly recognised as a guard against post-hoc adaptation of hypotheses to data. We pre-register four predictions with concrete decision rules before scale-up evaluation and report deviations explicitly. To the extent we know, this is one of the first pre-registered mechanistic interpretability studies; the practice transfers immediately to other crosscoder / SAE / activation-patching work.

J FULL LIMITATIONS

Read this appendix under the superseded estimand. It was written against the published single-side, α=1 measurement and moved here unchanged, so the de-concentration results below — the shared ratio, $r^{v2}_{\text{shared}} = 3.13$ against 1.25 on the v2.5 side, the sign-flip $p < 10^{-4}$, and the argument that the difference is not a labelling artifact — are reported as *that measurement's* internal structure. appendix V withdraws the claim they support: on a joint code at an admissible dose the same quantities are 1.05 and 0.36. Nothing below should be read as a live result; it is retained because the threats it catalogues apply to the corrected estimand too, and because deleting the analysis that produced a withdrawn claim would make table 10 unauditable.

Best-pair-only quantitative analysis. Phase 1.7 and Phase 3 quantitative results are reported at the highest val-$R^2_{\text{v2.5}}$ layer pair (L2 $\leftrightarrow$ L1). Phase 1.5 confirmed qualitative consistency across all 5 pairs but did not run Phase 3-scale bootstrap on each.

$n_{\text{estimators}} = 1$. TabPFN inference at deployment ensembles 8 forward passes. Our patching pipeline reliably patches only the final pass, so ensemble averaging dilutes the patched effect. We tested this directly (appendix E.8): at $n_{\text{est}} = 4$ all effects shrink, but the shared signal dilutes *less* (3.94×) than random (2.26×) or the no-op control (2.33×), and the ordering shared > random > no-op is preserved on all 5 datasets. The single-estimator regime is thus a conservative magnifier, not a source of the effect.

Sample size and operating point. The primary analysis uses 10 datasets at $n_{\text{train}} = 128$ (16 eligible for addition). appendix G replicates the headline causal results at a larger, GPU-only operating point ($n_{\text{train}} = 512$, 12 large-sample datasets, $n = 36$, 5.3× more crosscoder training data): shared > random on both sides and the de-concentration *direction* both hold (Holm $p < 10^{-4}$; $p = 0.0066$), and P2 remains TOST-equivalent to no transfer. The *magnitude*, however, attenuates substantially with operating point: the per-feature v2 ratio falls from 3.13 ($n_{\text{train}} = 128$) to 1.56 ($n_{\text{train}} = 512$) and the de-concentration gap from +1.88 to +0.29 (a ∼6× shrinkage). Together with the norm-inclusive scoping (appendix E.10), this means the reported de-concentration *numbers* are operating-point- and norm-dependent; only the direction is stable. Two further caveats remain: the $n_{\text{train}} = 512$ regime drops the 5 smallest datasets (which cannot supply the split), and a crosscoder-statistics model-selection diagnostic does *not* materialise (leave-one-dataset-out $R^2 = -0.02$; appendix G). Scale beyond this—more datasets, ensembles—remains future work.

Threats to validity (examined). Four further threats are tested rather than assumed. (i) *Train/eval overlap*: the crosscoder is fit on 5 reference datasets at seed 42, which partly overlap the Phase H grid; the P1 effect is nonetheless *largest* on the fully held-out extended datasets ($r^{\text{v2}}_{\text{shared}} = 3.31$ vs 2.67 in-sample) and P2 stays null in every stratum (appendix E.6), so it is not a training-cohort artifact. (ii) *Clustered inference*: the shared $>$ random ordering survives a stricter dataset-clustered sign-flip ($p = 0.0019/0.0119$; appendix E.6). (iii) *Feature-group stability*: the shared group is the most stable axis (corpus-bootstrap Jaccard 0.77, appendix E.5; and 0.94 across seeds for the high-fidelity Phase-L crosscoder, `exp_phase_L_seed_stability`), so the de-concentration result, which rests on the shared ratio, is not a labelling artifact; v2.5-exclusive labels are less stable and claims touching them are correspondingly hedged. (iv) *Operating-point dependence of the attention channel*: the headline thinking-token knockout is measured at $n_{\text{test}} = 64$ (needed only for the attention-mass *recording*, not the knockout); re-running the knockout at the primary $n_{\text{test}} = 128$ reproduces it ($+146\%$ vs $+23\%$ random, appendix X.42), so the channel is not a small-test artifact. (v) *Alignment-strategy dependence*: the canonical pair is chosen under a greedy CKA rule with a v2.5-uniqueness constraint; the causal findings are reported at reconstruction-qualified greedy-selected pairs, and we do not exhaust Hungarian/linear alignment alternatives.

Estimand identification and its scope (pre-registered controls). Two pre-registered controls bound how strongly RP effects may be read. (i) *Absolute magnitude is norm-driven, and this confounds de-concentration's magnitude*: a matched-norm random direction is about as disruptive as the shared-group RP edit ($\rho_{\text{v2}} = 0.89$ $[0.76, 1.03]$, $n = 15$; appendix E.10), so RP effects must be read only *relatively*. Critically, this means the reported de-concentration *magnitude* (v2 $3.13\times$ vs v2.5 $1.25\times$ under the size-matched null) partly reflects v2 activating shared features more strongly, not a direction-specific causal privilege: the norm-*controlled* contrast is $\rho_{\text{v2}} - \rho_{\text{v25}} = +0.09$ $[-0.11, +0.27]$ ($p = 0.21$, $n = 30$; appendix E.10)—same direction, but not significant. We therefore present de-concentration as a robust, dictionary-stable difference *under a size-matched on-manifold null* (sign-flip $p < 10^{-4}$; appendix E.4) that is explicitly *norm-inclusive*—a difference in how strongly the two models activate the shared subspace, not a direction-specific causal privilege—and we do not claim the norm-controlled version. (ii) *Threshold scope*: the de-concentration contrast is a property of the strongly-shared features—it holds at the operating classification threshold ($\tau \leq 0.01$, untuned) and stricter, attenuates at $\tau = 0.02$, and reverses at the permissive $\tau = 0.05$ where the shared bin is over-inclusive (appendix E.5). We report both at face value rather than the robustness we do not have; both are orthogonal to the dictionary-ensemble robustness of appendix E.4.

Self-attested pre-registration, non-standalone artifact, and run-to-run nondeterminism. No external timestamp authority is available in our environment, so the protocol snapshot is self-attested (hash-frozen), not third-party preregistration; and the pipeline depends on the sibling `TFMI` repository and two TabPFN checkpoints, pinned by SHA-256 (`docs/provenance/`) rather than vendored. Full one-command reproduction therefore requires those external assets. We also *ran* a clean-environment CPU reproduction (fresh virtualenv, regenerated results) and report it honestly: all *qualitative* conclusions reproduce—the de-concentration direction (fresh v2 $2.76\times$ vs v2.5 $1.28\times$; sign-flip $p < 10^{-4}$), the shared$>$random ordering, and the near-zero transfers. Earlier versions of this appendix also reported that "37 of 44 manifest-tracked numbers match the frozen run within 3%". We withdraw that sentence: the fresh results tree and the comparison log were not retained, and the manifest has since grown to 67 entries, so the count cannot be re-derived from the shipped artifact and no reader can check it. The per-quantity statements below *are* checkable against the committed tree; the aggregate count is not, and we would rather drop it than ship an unverifiable summary statistic. The headline v2 ratio reproduces within mutual bootstrap CIs (frozen 3.13 $[2.23, 4.26]$ vs fresh 2.76 $[2.34, 3.24]$). The remaining discrepancies are of two kinds and neither is a code drift (the frozen tree re-validates exactly): (i) near-zero transfer metrics whose absolute value stays ≈ 0 but whose *relative* error is large (P2 and the learned-map closure remain TOST-null); and (ii) genuinely run-sensitive secondary quantities, most notably the *independent-SAE* baseline ratio, which varies with SAE training. The cause is library/hardware nondeterminism (TabPFN CPU inference and dictionary training are not bit-reproducible despite a fixed seed). We therefore present the committed point estimates as one frozen run whose *headline and qualitative* findings are reproducible, and flag the SAE-baseline magnitude as run-variable rather than claiming bit-exact reproduction.

Cross-architecture transferability is limited by dimension asymmetry. A preliminary cross-arch experiment (Phase C) applied the same crosscoder framework to TabICL (12 layers $\times$ 512-d) $\times$ TabPFN v2.5 (18 layers $\times$ 192-d). The CKA-best pair (TabICL L8 $\leftrightarrow$ v2.5 L7) achieves CKA

Table 15: Ablation scopes. $n_{\text{thinking}} = 64$ for v2.5, 0 for v2.

Scope	Positions modified
`test_label`	[single_eval_pos : seq_len]
`train_labels`	[n_{thinking} : single_eval_pos]
`data_only`	[n_{thinking} : seq_len] (train + test labels)
`thinking_only`	[0 : n_{thinking}] (v2.5 only)
`all_labels`	[0 : seq_len]

$= 0.834$, comparable to within-family alignment, but the vanilla crosscoder reconstructs TabICL with val $R^2 = 0.855$ and v2.5 with only $R^2 = 0.434$, and classification produces a degenerate $569/7$ split between architecture-exclusive groups—the higher-dimensional TabICL side dominates the shared TopK encoder. We test two progressively more aggressive architectural fixes: (i) per-side input LayerNorm + asymmetric reconstruction-loss weighting (`DimAwareCrosscoder`, Phase C2), which moves v2.5 val R^2 from $0.434 \rightarrow 0.522$, and (ii) fully separate per-side encoders with a shared sparse code (`SeparateEncoderCrosscoder`, Phase C3), which yields 0.513. Both improve the lower-dim side modestly but plateau well below the 0.7 threshold we require for confident shared-feature classification, and both retain the degenerate exclusive-side split (TabICL-exclusive $\approx$ 570–610 vs v2.5-exclusive $\approx$ 4–7; appendix F). We conclude that cross-architecture mechanistic diffing across substantially asymmetric embedding dimensionalities is an *open problem* that is not resolved by within-framework architectural variants of the crosscoder (appendix D). Future work may need to step outside the joint-encoder regime—for example by training two SAEs independently and aligning their feature spaces via a separate causal-mediation objective rather than a joint reconstruction loss.

K Four methodological refinements

K.1 Refinement #1: activation-method classification

The original norm-ratio rule $r_i = \|W_d^{(\text{v2})}[:, i]\| / (\|W_d^{(\text{v2})}[:, i]\| + \|W_d^{(\text{v2.5})}[:, i]\|)$ fails for our crosscoders, and the reason is the normalisation, not TopK: we L_2-normalise each model's decoder columns *separately*, which pins r_i near $\frac{1}{2}$ by construction regardless of the feature's computational role. Minder et al. (2025) report that norm-based indicators are informative under BatchTopK with a different normalisation, so this is a statement about our parameterisation and not about TopK dictionaries in general. We instead use an *activation-method* rule: run the crosscoder once with v2 input zeroed (encoder sees only v2.5 information), once with v2.5 zeroed. Compute activation frequencies $\nu_i^{(\text{v2})}$ and $\nu_i^{(\text{v2.5})}$. Feature i is v2-exclusive iff $\nu_i^{(\text{v2})} > \tau$ and $\nu_i^{(\text{v2.5})} \leq \tau$; analogously for v2.5-exclusive. We use $\tau = 0.01$.

Death is decided *before* this, and on a different quantity: a feature is dead when its frequency on the *joint* input $(a_{\text{v2}}, a_{\text{v2.5}})$ falls below τ. "Shared" is then the complement — every live feature that is not one-sided. That complement is not the same set as $\{\nu_i^{(\text{v2})} > \tau\} \cap \{\nu_i^{(\text{v2.5})} > \tau\}$, because a feature can fire on the joint input while clearing τ on neither single-sided input. Those features are shared under the rule we run; an earlier version of this paragraph called them dead. They are 11.0% of the group and every headline number is unchanged by their removal (appendix X.50).

K.2 Refinement #2: scope-aware ablation

Standard activation patching ablates only test-positions of the sequence (TabPFN positions [single_eval_pos : seq_len]). This scope is insufficient: v2.5 distributes computation across thinking tokens (positions $[0 : 64]$) and label tokens (positions $[64 :]$). Ablating only test-label tokens leaves features active at thinking and train-label positions untouched, underestimating their causal contribution. We define five scopes (table 15): the primary scope for our P1 prediction is `data_only`; `thinking_only` isolates thinking-token-specific effects (P3).

K.3 Refinement #3: per-feature normalisation

Feature groups (shared, v2-exclusive, v2.5-exclusive) generally have very different cardinality (typically 100–300 shared; 50–250 exclusive per side). Aggregate prediction-change effects scale with cardinality and cannot be compared across groups. We report all results as

$$r_g = \frac{\Delta_g/|g|}{\Delta_{\text{random}}/|\text{random}|} ,$$

the per-feature relative prediction change of group g divided by the same for a size-matched random baseline.

K.4 Refinement #4: alive-pool random baseline

A naïve random baseline samples feature indices uniformly from $\{0, \ldots, N-1\}$, then zeros those entries. In practice this draws predominantly from the dead-feature majority (85–95% of indices), artificially deflating the baseline. We sample random baselines exclusively from the *alive* index pool—the union of shared and both exclusive groups—matching size to the largest group being compared.

L Multi-layer injection, P3 and P4 in full

L.1 Multi-layer simultaneous injection – the F1 stress test

A single-layer null leaves open the simplest alternative explanation: downstream layers overwrite the injected values. We therefore inject v2.5's shared-feature values into v2 at $k \in \{1, \ldots, 5\}$ aligned layer pairs *simultaneously* (cumulative sets over $\{(0,4),(2,1),(7,11),(9,16),(10,17)\}$), using live-clamping RP patches: each hook re-encodes the live activation (which carries upstream patches' propagated effects) and clamps the injected features to v2.5's clean values. This is exactly the falsifiable prediction F1 of section 6: closure $\geq 50\%$ at $k = 5$ would reject our thesis.

On the pre-registered gap-closure metric, the $k = 5$ shared estimate is $+17.0\%$ with an uninformatively wide CI $[-52\%, +118\%]$: the ratio metric explodes on instances with near-zero capability gap (two small-gap instances contribute $+513\%$ and $+221\%$). The robust companion metric—relative RMSE change of the patched v2 model—shows that five-pair injection *does not improve v2 at all*: -1.33% $[-2.75, +0.53]$ (negative = worse). The dose-response of closure in k is flat ($+3.8$ pp/pair, CI $[-11.0, +25.9]$ containing 0), and at every k the shared variant weakly dominates size-matched random injection (which is clearly harmful at large k: -48% closure at $k = 5$), confirming the injected values are not generic noise. Under the pre-registered metric the formal label is *partial mediation (inconclusive)*; under the robust metric, **F1 is not falsified**: even multi-site value injection fails to move v2 toward v2.5.

A learned transfer map fares no better. Identity injection might simply deliver values in the wrong coordinates. We therefore fit $T : f_{\text{v2}}[\text{shared}] \to f_{\text{v2.5}}[\text{shared}]$ by ridge regression (and its rank-16 truncation) on the 5-dataset training cohort's 1,600 paired activations, and inject $T(f_{\text{v2}}^{\text{live}})$ into v2 on capability-eligible instances of the five *held-out* extended datasets (leakage-free: neither the crosscoder nor the map ever saw them). The map itself is good—ridge predicts v2.5's shared-feature values from v2's with $R^2 = 0.81$—yet injection does nothing: gap closure -0.1% $[-2.9, +3.5]$ (ridge) and -0.1% $[-3.5, +3.2]$ (rank-16), both TOST-equivalent to zero transfer at the ± 10pp margin. *Capability is not carried by any predictable transformation of the feature values*: one can know what v2.5's features will say (to $R^2 = 0.81$) and still gain nothing by saying it inside v2's computation. This is a held-out, small-n stress test ($n = 8$ eligible instances, 4 dataset blocks): it rules out a simple predictable-value transplantation route on the eligible extended cohort rather than all possible transfer maps, and is consistent with architecture-coupled mediation (PREREGISTRATION §16).

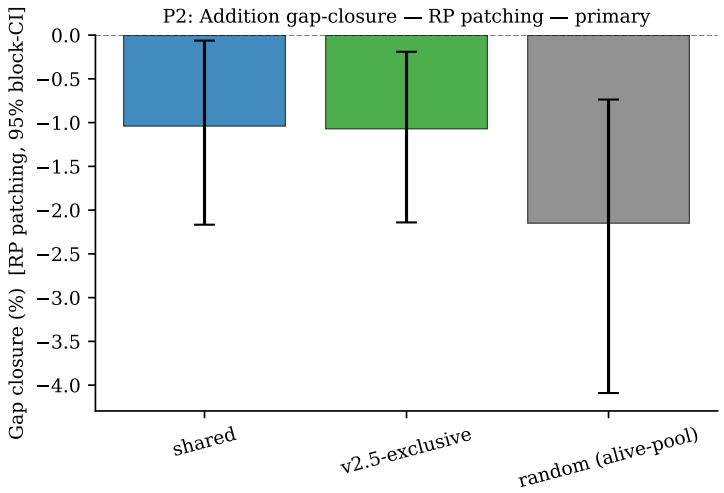


Figure 6: Capability gap closure when injecting v2.5's encoded feature values into v2, at $n = 16$ capability-eligible instances, *primary residual-preserving (RP) estimand* (matches section 5). All variants overlap zero; the pre-registered ordering shared > v25_excl > random does not hold—no transferable component. The superseded replacement-mode version is generated separately as `F2_addition_gap_closure_legacy` and is not shown here.

L.2 P3 – Thinking tokens and v2.5-exclusive features are minor (two of three conditions)

The `thinking_only` scope ablates features at v2.5's first 64 sequence positions only. Under the primary (RP) estimand the per-feature absolute relative prediction change is 9.7×10^{-6} $[6.7, 13.7] \times 10^{-6}$ (block CI, $n = 30$), well below the pre-registered threshold of 10^{-4}. The RP-mode per-feature ratio for v2.5-exclusive features in the `data_only` scope is 0.82 $[0.71, 0.99]$—below the random baseline and far below the 1.5 threshold. Both conditions P3 tested hold; its third registered condition, the aggregate thinking share, was not computed under this estimand (appendix O), so we report P3 as supported on two of three conditions. Mechanistically: thinking-token positions activate many features ($\approx 60\%$ of v2.5-exclusive features have non-zero activation there), but their causal contribution through the *feature-value* pathway is negligible. This null is sharply qualified by the attention-level analysis of appendix X.42: the thinking-token capability channel exists, but it is attention-mediated, not value-mediated.

M P4: Layer alignment is input-distribution-dependent

M.1 P4 – Layer alignment is input-distribution-dependent (SUPPORT)

CKA-based layer-pair selections on synthetic and real data share only 2 of 8 unique pairs (Jaccard $= 0.25$, below the pre-registered 0.5 threshold). **P4 is supported.** fig. 7 shows the two CKA matrices side by side. On synthetic data, optimal pairs trend toward later v2 layers (5–9) and mid v2.5 layers (10–14); on real data, optimal pairs span the full depth, including $(0, 4)$ and $(10, 17)$. Any mechanistic diff that relies on cross-model layer alignment computed on synthetic data and transferred to real inputs is at risk of misalignment.

N Falsifiable predictions in full

Falsifiable predictions. To keep this interpretation empirical rather than a residual label, we stated three conditions any of which would falsify it, and have now tested all three (one of them, F2, inconclusively). **(F1, tested)** If *multi-layer simultaneous* injection of v2.5's shared-feature values into v2 (at all five aligned pairs at once) closed $\geq 50\%$ of the capability gap, the gain would be transplantable and merely distributed across depth. appendix L ran this test: five-pair injection leaves v2's RMSE unimproved (-1.3% $[-2.8, +0.5]$) with a flat dose-response. On the pre-registered

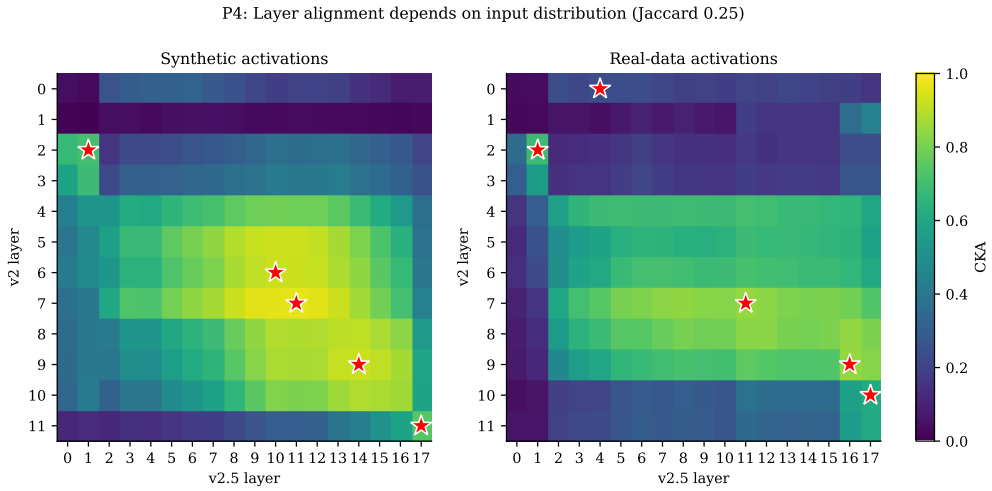


Figure 7: CKA heatmaps between TabPFN v2 (12 layers) and v2.5 (18 layers) layers computed on (left) synthetic activations from `generate_diverse_datasets` and (right) real OpenML activations. Pre-registered alignment pairs marked with red stars. Jaccard between the two pair sets is 0.25.

closure metric the result is formally inconclusive (the ratio is unstable for near-zero gaps), but the robust bound excludes any meaningful transfer—F1 is not falsified. **(F2, tested, inconclusive)** If a v2 variant fine-tuned only at the canonical patched layer (all other weights frozen) recovers the gap, the coupling is shallow rather than architectural. Using TabPFN's native fine-tuning loop we froze all but layer L2 (5.3% of parameters) and fine-tuned per eligible instance (appendix E.11). The result does not reach the pre-registered $\geq 50\%$-closure threshold that would revise the interpretation: L2-only closure is $+34.5\%$ (95% block CI $[-6, +73]$, $n = 11$ eligible), but it is *statistically indistinguishable from a full fine-tune* of all parameters ($+45.1\%$ $[-23, +99]$; paired sign-flip $p = 0.45$), and per-instance closure is dominated by noisy test-time dataset adaptation on 128 rows (range -74% to $+282\%$). We resolve that confound with a *meta-fine-tune* (appendix E.11): adapt L2 across a 7-dataset pool with warm-start carryover, then evaluate the frozen weights on held-out datasets never seen during adaptation. Freed of test-time adaptation, held-out closure is *negative*—-5.3% $[-17, +0.6]$ for L2-only and -28.4% $[-91, +3]$ for a full fine-tune (thin $n = 3$ eligible held-out)—so the apparent $+34.5\%$ was indeed test-time adaptation, and neither single-layer nor whole-model fine-tuning transfers the gap to unseen data. This is consistent with, though (because even the full fine-tune fails) does not by itself prove, the architecture-coupled reading, which continues to rest primarily on the positive attention-channel evidence (appendix X.42) and the value-injection nulls (P2, Phase I, Phase N). **(F3, tested)** If attention-pattern diffing had found *no* systematic difference in how v2 and v2.5 attend from test rows to context rows, candidate (3) would be eliminated. appendix X.42 found the opposite: all five aligned pairs differ significantly (Holm-corrected $p < 10^{-4}$ each), so candidate (3) stays alive—though our secondary directional prediction (uniformly higher v2.5 attention entropy) was rejected, the reorganisation being depth-heterogeneous rather than a uniform flattening. We report both the confirmed antecedent test and the failed directional refinement. See fig. 6.

O Pre-registration decision rules and compute

O.1 Pre-registration and decision rules

All four predictions are registered before the Phase 3 multi-seed evaluation. Thresholds are anchored on Phase 1.7's single-seed magnitudes: **P1**: $r_{\text{shared}}(\text{data_only}) \geq 2.5$ and $r_{\text{shared}} > r_{\text{v25_excl}} + 1.0$ on $\geq 60\%$ of layer pairs, with the bootstrap CI excluding the null. **P2**: addition ordering shared $>$ v25_excl $>$ random with CIs respecting it on $\geq 60\%$ of dataset resamples, and mean shared gap-closure in $(-10\%, 30\%]$. **P3**: $r_{\text{v25_excl}}(\text{data_only}) \leq 1.5$; `thinking_only` per-feature absolute effect $< 10^{-4}$; and aggregate thinking-token contribution $\leq 5\%$ of the summed scope effects. The third condition was registered but never computed under the primary (RP) estimand, so P3 is

supported on two of its three conditions and we do not call it supported. **P4**: synthetic–real CKA Jaccard < 0.5.

Decision-rule application uses cluster (block) bootstrap CIs ($n_{\text{bootstrap}} = 1000$) that resample *datasets* with replacement, preserving the within-dataset correlation across seeds that plain iid resampling would ignore. Group-vs-group comparisons use paired sign-flip permutation tests on within-instance differences (10^4 permutations). P2's null claim is additionally formalised as a TOST equivalence test (equivalence bounds ± 10 percentage points of gap closure; equivalence declared iff the 90% block-bootstrap CI lies within bounds). Across the pre-registered hypothesis family we report Holm-corrected p-values. All primary causal effects use *residual-preserving (RP) patching* (section 3.4), whose identity (no-op) control is exactly zero by construction and is verified to machine precision on the real model pathway in every run; replacement-mode results, where retained, are explicitly labelled legacy and accompanied by the measured no-op control.

O.2 Compute and reproducibility

The CPU stages [1–27] reproduce in $\approx$ 6–8 CPU-hours on a single workstation; the pipeline as run end to end before submission recorded 10.7 hours over the 62 stages that time themselves (section 7), on a single workstation; the large-context Phase L/M stages [28–31] require CUDA and the provided `repro.sh` assumes four GPUs (TabPFN forbids CPU inference above 1000 in-context samples, so the $n_{\text{train}} = 512$ regime is genuinely GPU-only). Crosscoder training is fast on GPU (< 5 minutes per pair); for the CPU stages, model patching is the bottleneck because TabPFN's forward pass dominates wall-clock time. A reviewer without GPUs can verify the GPU-stage artifacts against `docs/results_manifest.json`. All seeds, configurations, and intermediate result JSON files are released alongside the code.

P Attention-pathway setup and full results

P.1 The attention pathway: thinking tokens are a compressed context, consumed via attention

P3 ablated thinking-token *feature values* through the crosscoder at one layer. To measure the complementary channel—what test rows *read from* thinking tokens via item attention—we interpose on TabPFN's fused attention kernel an explicit computation that can record or mask attention weights, armed per layer (validity gate: explicit-path predictions match the fused path within 10^{-4} relative on all 30 instances; zero failures). This experiment uses $n_{\text{test}} = 64 \neq n_{\text{train}} = 128$ (rather than the primary $n_{\text{test}} = 128$) so that the test-row attention call is uniquely identifiable by its query length; the train/context-row structure and all 18 layers are unchanged, so the operating point shift does not affect the channel being measured. Three findings (pre-registered; PREREGISTRATION §13):

(a) Read early, integrate late. Test rows allocate 38–76% of their item-attention mass to thinking keys in layers 0–11 (dip at L12), still 43%/12% at L13/L14, collapsing to $\leq 6\%$ in the final three layers. The thinking block is the dominant attention target through most of the stack and is ignored at the end, where prediction-relevant information has presumably been integrated into the test rows' own representations.

(b) Knockout is catastrophic and channel-specific. Masking item-attention to the 64 thinking keys at all 18 layers degrades v2.5's RMSE by $+159\%$ $[+96, +227]$ (block CI, $n = 30$), versus $+23\%$ $[+9, +39]$ for size-matched random train-key masks (sign-flip $p < 10^{-4}$). The thinking knockout is comparable to the all-train-keys positive control ($+134\%$ $[+81, +189]$). The CIs overlap substantially and we ran no paired test between these two arms; an earlier draft said the thinking knockout *exceeds* all-train, which the artifacts do not support. The defensible statement is that losing access to 64 thinking keys costs about as much as losing access to all 128 raw training rows. *This is not an artifact of the smaller test set*: re-running the knockout at the primary $n_{\text{test}} = 128$ (the $n_{\text{test}} = 64$ constraint is needed only for the attention-mass *recording* in (a), not the knockout) gives $+146\%$ $[+89, +215]$ thinking vs $+23\%$ random and $+128\%$ all-train (sign-flip $p < 10^{-4}$, $n = 30$; `exp_phase_j_sensitivity`). Prediction flows through the thinking block; the transplant experiment in appendix S.2 shows it does so as an attention register rather than as a compression of this dataset's context. Together with P3, this resolves the mechanism: ablating thinking *values* in

the crosscoder basis at one layer is inert (9.7×10^{-6} per feature), while severing attention *access* to thinking representations is the single largest causal effect we measure anywhere in this study.

Control caveat: the arms are matched on key count, not on attention mass. Masking sets the target logits to $-\infty$ and lets softmax renormalise (`tabdiff/analysis/attention_tools.py`), so an intervention's severity depends on how much probability mass it forces to be redistributed. The thinking block carries 38–76% of item-attention mass in layers 0–11, whereas 64 randomly chosen train keys carry far less; the two arms are matched on *how many* keys are removed, not on *how much attention* is removed. The $+159\%$ vs $+23\%$ contrast is therefore partly a statement about mass displaced, and should not be read as showing that thinking keys are individually more important than train keys. Two references help calibrate: a uniform allocation over 64 thinking and 128 train keys would be 33%, so the low end of the 38–76% range is only modestly above chance; and the all-train-keys arm, which removes a comparable amount of mass, costs a comparable $+134\%$. The control this analysis lacks is a *mass-matched* knockout — masking the smallest set of train keys carrying the same total mass as the thinking block, which is the right way to separate "these positions matter" from "this much attention matters". appendix W.4 reports the search: no mass-matched control is constructible at this pair, and the substitute measure is given there.

(c) Attention reorganisation is systematic but depth-heterogeneous. At all five aligned pairs, the entropy of test rows' attention over train keys differs significantly between models (Holm-corrected sign-flip $p < 10^{-4}$ each), but with mixed signs: v2.5 attends more *sharply* at L0↔L4 (-0.12), L2↔L1 (-0.53), and L9↔L16 (-1.23), and more diffusely at L7↔L11 ($+0.54$) and L10↔L17 ($+0.19$). Our pre-registered directional prediction—uniformly higher v2.5 entropy as an attention-level mirror of causal de-concentration—is *rejected* as stated; what holds is that v2 and v2.5 attend systematically differently at every aligned depth. This settles the antecedent of falsifiable prediction F3 (section 6): the "no systematic attention difference" branch is decisively false.

Q Estimand scope and identification caveats

Off-distribution encoding. A caveat RP does *not* remove: the estimands encode with the counterpart model's input zeroed, so $D_g f_g$ is the decoder's attribution under an encoder input the crosscoder was never trained on. appendix T measures the cost—single-side reconstruction is $0.66/0.57$ at the canonical pair against $0.998/0.998$ jointly—so the perturbation we inject is well-defined and exactly applied, but it is an attribution in an off-distribution code, not in the code the dictionary was fit to produce.

Which question does each estimand answer? The two schemes are different estimands, not a wrong and a right implementation of one. Replacement measures the response to a *joint* intervention: remove group g's decoder attribution *and* project the activation onto the dictionary's reconstruction (deleting the residual ε). Its no-op is not the identity, and the two components interact non-additively (appendix E.1), so it cannot be decomposed after the fact. RP measures a *controlled* quantity: the effect of removing $D_g f_g$ alone, with the residual held fixed at its observed value—the natural estimand when the question is "what does *this feature group* contribute, everything else unchanged?", which is the question feature-level model diffing asks. We therefore adopt RP as primary *for that question*, not as ground truth: if the scientific target instead involves the dictionary's fidelity itself, replacement conflates it with the feature effect rather than isolating it. One caveat is inherited by construction: RP measures the group's contribution *as parameterised by the trained dictionary* ($D_g f_g$), so any claim built on it is dictionary-conditional until shown stable across dictionaries—which is exactly what the ensemble gate of appendix E tests. A second caveat is empirical: the *absolute* magnitude of an RP edit is largely a function of its norm—a matched-norm random direction is about as disruptive (appendix E.10)—so we read RP effects only *relatively*, against an on-manifold alive-pool random baseline of matched group size and as a v2-vs-v2.5 contrast, never as evidence that a group occupies an absolutely privileged direction.

R Future work

Three high-priority next steps. First, **dimension-aware cross-architecture crosscoders** (preliminary in section 6.1, to be scaled). Second, **explicit attention-pattern diffing** between v2 and v2.5

on the patched activation, to mechanistically identify which v2.5 attention components drive the joint-architectural- context effect implied by P2's null. Third, **classification-task generalisation**—TabPFN supports both regression and classification; whether C"" transfers across task types is an open empirical question.

S ROBUSTNESS OF THE INTERACTION, AND THE THINKING-TOKEN TRANSPLANT

Two experiments added after an adversarial review of this manuscript. Both are post-hoc rather than pre-registered and are labelled as such (PREREGISTRATION §28).

S.1 THE INTERACTION SHARE ACROSS THE DICTIONARY ENSEMBLE

The interaction share that motivates the estimand switch was originally reported for a single dictionary, so it could have been one lucky draw. The pre-registered 15-dictionary ensemble (5 capacity/sparsity settings $\times$ 3 seeds, appendix E.4) lets us measure it across all 15 without new training. Evaluated on signed per-row deltas at the canonical pair, the interaction share is **positive for all 15** dictionaries, with median $+1.01$ (range $+0.81$ to $+2.10$); in magnitude terms $|I|/|\text{replace}|$ has median 1.008 (range 0.922–1.135). The confound is therefore a property of the setting, not of a particular dictionary: across every capacity, sparsity and seed we trained, essentially the entire replacement-mode effect is the non-additive interaction rather than the feature contribution.

S.2 CROSS-DATASET THINKING-TOKEN TRANSPLANT

Masking attention to v2.5's thinking tokens is the largest causal effect in this study ($+159\%$ RMSE), but a knockout shows only that a channel is load-bearing, not what it carries. Two accounts survive that evidence: a *content* account (the block is a learned compression of *this* dataset's training context) and a *register* account (the block is a largely input-independent structure that attention routes through). They make opposite predictions under transplantation.

Design. We run dataset B's forward pass with the thinking-position activations captured from a *different* dataset A, spliced at all 18 v2.5 layers. We transplant the 192-d label-token slice, whose shape is dataset-independent. Controls: (i) a *self-seed* splice (same dataset, different seed) isolates the act of splicing from foreign content; (ii) *zeroed* and matched-scale *noise* donors verify the splice path can do damage at all; (iii) a *self* transplant must reproduce the clean prediction exactly.

Result. Foreign transplant costs $+0.26\%$ $[+0.07, +0.56]$ RMSE ($n = 10$ over 5 datasets $\times$ 2 seeds, 5 blocks); the self-seed splice costs -0.05% $[-0.08, -0.02]$. The contrast that matters is their difference, $+0.31\%$ $[+0.10, +0.63]$, and our two tests of it disagree: the block bootstrap excludes zero, a paired sign-flip on the same $n = 10$ gives $p = 0.21$. Earlier versions of this paper reported only the p-value and described the arms as "indistinguishable". We correct that to the conservative reading in the other direction: there is a *small but likely non-zero* dataset-specific component. With 5 clusters neither test is well calibrated (appendix E.2), so we lean on the effect size instead—0.31% is $1/500$ of the $+159\%$ cost of severing access, and $1/22$ of the cost of merely zeroing the block.

Controls, reported as medians over all three datasets. Self-transplant reproduces the clean prediction to $+0.00\%$. Zeroing the block costs a median $+6.8\%$ (per-dataset $+30.5\%$, $+6.8\%$, -0.67%); matched-scale noise costs a median $+62.7\%$ ($+102.0\%$, $+62.7\%$, $+19.7\%$). Earlier versions quoted the zeroed control as its *maximum* ($+30.5\%$) alongside a median for the noise control, which overstated it by $4.5\times$ and made the two arms incomparable.

Interpretation, against our own prior reading. The corrected controls change the argument rather than merely softening it. Zeroing the thinking block is cheap ($+6.8\%$, i.e. 4% of the access effect), so the claim in earlier drafts that "destroying its structure is catastrophic" is withdrawn—the block's *content* is close to inert. Yet filling the same positions with matched-scale noise costs $+62.7\%$, an order of magnitude more than deleting them, and severing attention to them costs $+159\%$. A channel that is content-inert, noise-intolerant and access-critical is the signature of an attention register or sink (Xiao et al., 2024; Darcet et al., 2024): positions whose function is to absorb attention mass benignly. The ordering we find — an absent block is cheap, a corrupted one is not — is our own

measurement, not prior art: StreamingLLM finds the initial-token sink indispensable once evicted, and on DINO registers Parodi et al. (2026) find the reverse ordering (zeroing costs far more than substitution). A compression of *this* dataset's training context predicts the opposite ordering—deletion should be at least as damaging as corruption, and foreign content should be far worse than 0.31%. We therefore withdraw the "learned compression of the training context" description and report: v2.5's thinking block is a largely input-independent register whose *attention accessibility*, and whose well-formedness, are what is causally required.

Scope of the content arm. We transplant the 192-d label-token slice, not the full thinking-position activation, because that slice is the part whose shape is dataset-independent; the remaining feature blocks at those positions keep the recipient's values. The content arm therefore covers $1/(n_{\text{feat}}+1)$ of the representation at those positions, while the attention arm masks the keys entirely. "Access is required, content is interchangeable" is thus established for different objects on the two sides, and the dataset-independence of what we *can* transplant is partly a property of the design. A within-fixed-feature-count test on the full block is the natural follow-up.

S.3 Per-feature ratios, figure form

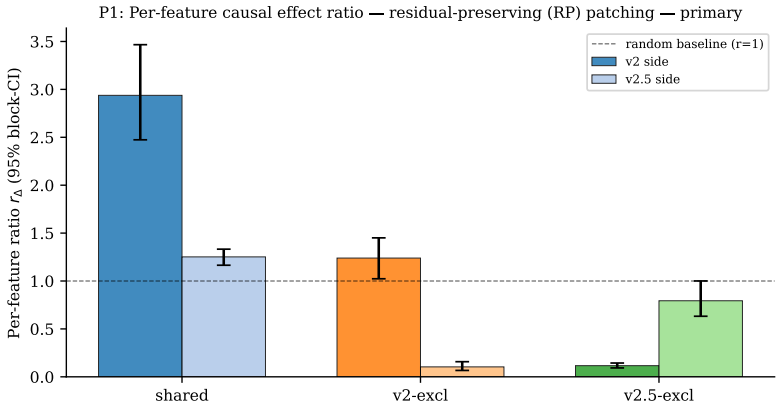


Figure 8: Per-feature relative prediction change (vs alive-pool random) for each feature group under the *primary residual-preserving (RP) estimand*, on the v2 side and the v2.5 side. Bars are point estimates at $n = 30$ with 95% block-bootstrap CIs; the dashed line at 1.0 is the random baseline. Values match table 21. Read with the norm-matched caveat of appendix E.10: the contrast is norm-inclusive. The superseded replacement-mode visualisation is generated separately as `F1_per_feature_ratios_legacy` and is not shown.

See fig. 8.

S.4 Per-group size-matched random null

The P1 ratios in table 21 divide every group's per-feature effect by a single random baseline whose size is matched to the *largest* group (259 shared features at the canonical pair). Because the model's response is super-linear in perturbation magnitude (appendix E.1), per-feature normalisation does not make groups of different cardinality comparable: a smaller group is deflated relative to a null constructed at a larger size. Group sizes here are shared 259, v2-exclusive 65, v2.5-exclusive 199, out of 523 alive features, so the distortion is substantial for v2-exclusive.

We therefore recomputed each group against a null drawn at *its own* cardinality (`experiments/exp_pergroup_null.py`; $n = 15$ per side, 5 datasets $\times$ 3 seeds, RP estimand, `data_only` scope, block-bootstrap CIs).

What survives. shared $>$ v2.5-exclusive holds on both sides ($p_{\text{Holm}} = 0.0002$), and shared $>$ random holds on v2 ($p_{\text{Holm}} = 0.0002$). **What does not.** shared $>$ random on the v2.5 side falls to $p = 0.075$, against a published $p = 0.0003$. We retract that specific contrast and state the ordering claim with the side qualification throughout.

Table 16: Per-group size-matched null. "Published" is the largest-group-matched null of table 21, computed on 10 datasets ($n = 30$); the per-group column is computed on 5 datasets ($n = 15$). The two columns therefore differ in *both* null construction and cohort size; the text below the table separates the two using a same-n comparison and attributes the change to the null construction.

Side	Group (n_{feat})	Published	Per-group null
v2	shared (259)	3.13	2.91 [2.14, 4.00]
v2	v2-exclusive (65)	1.68	1.08 [0.84, 1.32]
v2	v2.5-exclusive (199)	0.21	0.20 [0.16, 0.24]
v2.5	shared (259)	1.25	1.20 [1.06, 1.30]
v2.5	v2-exclusive (65)	0.43	0.17 [0.13, 0.22]
v2.5	v2.5-exclusive (199)	0.84	0.67 [0.58, 0.75]

Separating the null from the cohort. The two columns differ in the null construction (per-group vs largest-group) *and* in cohort (5 datasets, $n = 15$, vs 10, $n = 30$), so the comparison as tabulated does not by itself identify which matters. The third cell settles it: at the *same* $n = 15$, under the largest-group null, the canonical-pair contrast is significant ($p_{\text{Holm}} = 0.0064$, appendix B), while under the per-group null it is not ($p = 0.075$). Halving the cohort under a fixed null moves the contrast from $p = 0.0006$ to $p = 0.0064$ — still significant. The null construction, not the sample size, is what the contrast depends on. The cross-side comparison the paper foregrounds is unaffected in direction—v2 2.91 vs v2.5 1.20 under matched nulls—and remains subject to the norm-matched caveat of appendix E.10.

T SINGLE-SIDE RECONSTRUCTION ON THE INTERVENED PATHWAY

Layer pairs and dictionaries are admitted on *joint* reconstruction ($R^2 = \text{decode}(\text{encode}(a^{(v2)}, a^{(v2.5)}))$ vs the activations), but every causal estimand encodes one side with the counterpart zeroed. We measured both (`experiments/exp_singleside_r2.py`).

Table 17: Joint vs single-side reconstruction R^2. The qualification gates read the left columns; the estimands run on the right ones. The gap is large at every pair and the deepest pair is negative on the v2.5 side (worse than predicting the mean).

Pair	joint v2	joint v2.5	single-side v2	single-side v2.5
L2↔L1 (canonical)	0.998	0.998	0.664	0.574
L0↔L4	0.999	0.996	0.761	0.384
L7↔L11	0.985	0.984	0.340	0.365
L9↔L16 (deep)	0.987	0.979	0.566	0.198
L10↔L17 (deepest)	0.989	0.920	0.524	−1.057

Mean joint-minus-single gap: +0.42 (v2), +0.88 (v2.5); the residual $\|x - \text{dec}\|/\|x\|$ on the intervened path is 0.37–0.55. Three consequences. (i) "Reconstruction-qualified" should be read as a statement about the *dictionary*, not about the pathway the estimands use; we have amended the language in section 3 accordingly. (ii) It supplies the mechanical reason the replacement no-op is substantial (0.026) despite joint $R^2 = 0.998$: roughly half the activation norm is residual on the intervened path. (iii) It is an independent reason to distrust the deepest pair, whose single-side R^2 is negative, reinforcing the exploratory label it already carries. RP patching removes the reconstruction-error confound from the *effect*; it does not make the code on which $D_g f_g$ is computed in-distribution.

U RATIO ESTIMATOR AND INSTANCE DROPPING

r_g is reported as a mean of per-instance ratios, with instances dropped when the random denominator falls below 10^{-12}—an estimator that can be unstable with near-zero denominators, as we ourselves note for gap closure. We therefore report the alternatives. **No instances were dropped**: the guard never fires at either side ($n = 30$ of 30 on v2 and on v2.5). The shared-group estimate is 3.131 (mean-of-ratios, published), 3.046 (ratio-of-means), 2.295 (median-of-ratios) on v2, and 1.249 / 1.230 / 1.244 on v2.5. The v2.5 figure that carries the cross-side contrast is insensitive to the choice;

the v2 median is lower than its mean, indicating right skew, which is consistent with the wide block CI we already report ($[2.23, 4.26]$) and does not change the direction of any claim.

V What the audit cost us: the feature-level ordering is withdrawn

Applying the corrections overturns the feature-level result earlier versions reported as a headline: at the largest dose below the bound we then used, the shared-to-random ratio is 1.05 on v2 and 0.36 on v2.5 against published 3.13 and 1.25, and the favourable ordering appears only past the boundary and on both sides, marking it an artifact of the operating point. The positive control resolves at $1.64\times$ against a pre-committed $3\times$, so we attach no p-values and withdraw the claim with the "de-concentration" magnitudes built on it (appendix X.39). table 18 settles all four registered predictions in one place; an independent-SAE baseline, P1 across five layer pairs and a 19-dataset extension are in appendices A to C.

Table 18: The four pre-registered predictions against every registered condition (PREREG §2; appendix O); a prediction holds only if all its conditions do.

prediction	registered conditions	outcome
P1 shared dominates	$r_{\text{sh}} \geq 2.5$ and $> r_{\text{v25}} + 1.0$ on $\geq 60\%$ of pairs, CI excl. null	**withdrawn**: 1.05 / 0.36 at small dose
P2 no transfer	(i) ordering sh $>$ v25 $>$ random on $\geq 60\%$; (ii) closure in $(-10\%, 30\%]$	**not supported**: (i) unresolved, $p = 0.47$; (ii) -0.3%, TOST-equivalent
P3 minor	(i) $r_{\text{v25}} \leq 1.5$; (ii) thinking $< 10^{-4}$; (iii) thinking share $\leq 5\%$	**two of three**: 0.82; 9.7×10^{-6}; (iii) never computed
P4 alignment	synthetic–real CKA Jaccard < 0.5	**supported**: 0.25

W Correcting the ablation estimand: joint encoding and interpretable dose

This appendix reports the three experiments behind section 5.2 and appendix V. All were pre-registered before implementation (PREREGISTRATION §31), and four of the five predictions we recorded were refuted.

W.1 The interaction is not a second-order term

For a smooth scalar response F the cross-difference is exactly $I = \int_0^1\int_0^1 \varepsilon^\top H_F(x - s\varepsilon - tu)\, u\, ds\, dt$; a second-order expansion about the clean activation, with $\delta_1 = -\varepsilon$ and $\delta_2 = -u$ ($u = D_g f_g$), replaces the integrand by its value at x and predicts $I \approx \varepsilon^\top H_F(x)\, u$, hence that I is *bilinear*: scaling ε by β and u by α should scale I by $\alpha\beta$. We tested this over $\alpha, \beta \in \{0.25, 0.5, 1, 2\}$ with a pre-committed criterion (coefficient of variation of $I/(\alpha\beta)$ below 0.20 supports, above 0.50 refutes). **Measured CV $= 2.07$: refuted.**

The grid explains why (table 19). The response *saturates*: once either perturbation is large enough the output pins at ≈ 0.316 regardless of the other, and the marginals are wildly superlinear — doubling ε multiplies the no-op effect by $68\times$ in the ratio of means (per-instance the median is $92\times$ and the mean $185\times$; the aggregation matters and we name it). In saturation the decomposition is arithmetic rather than mechanism: at $\alpha = 2$, $I = 0.315 - (0.004 + 0.035) = 0.277$ purely because replace is pinned while the marginals are not. A Hessian cross-term would have been estimable and subtractable; what the grid shows is that the local expansion fails at the doses used, not that no expansion could describe the response. The practical conclusion is the same — the interaction cannot be removed by subtracting a fitted cross-term — and we no longer state it as a claim about all orders.

W.2 The interpretability boundary

Saturation makes a binary ablation uninformative on its own, so we need a scale criterion. We use the *label-independent fallback*: the cost of overwriting the test label token with the train-position mean,

Table 19: Mean Δ^{replace} over the scaling grid ($n = 15$, 5 datasets). Rows scale the ablation direction, columns the reconstruction residual. The response saturates at ≈ 0.3156 along both margins.

	$\beta = 0.25$	$\beta = 0.5$	$\beta = 1$	$\beta = 2$
$\alpha = 0.25$	0.0012	0.0021	0.0067	**0.3166**
$\alpha = 0.5$	0.0021	0.0034	0.0117	**0.3164**
$\alpha = 1$	0.0052	0.0088	0.1268	**0.3158**
$\alpha = 2$	**0.3123**	**0.3151**	**0.3155**	**0.3152**

i.e. of removing its information entirely (0.1011 on v2.5, 0.0449 on v2). An intervention whose effect exceeds the fallback is not removing information — it is injecting an activation the model handles *worse* than no information — so its magnitude cannot be attributed to the ablated features.

A three-direction sweep confirms the boundary is not an artifact of one direction and that the ceiling is feature-specific (pre-registered prediction H4a, that it would be feature-*in*dependent, was refuted; plateau spread 1.34). Ablating the shared group, a size-matched random alive group, and a norm-matched random direction reach plateaus of 0.340, 0.192 and 0.070 respectively at $\alpha = 4$. Notably the shared plateau is $3.4\times$ the fallback, confirming that the large-α regime is not interpretable.

W.3 Joint encoding

Our ablation pipeline, and the shared/exclusive taxonomy it inherits from the crosscoder diffing recipe, encoded one side with the counterpart zeroed: $\text{enc}(a^{(v2)}, \mathbf{0})$. We had described this as the literature's standard, and a reviewer was right that we had not shown it. Minder et al. (2025) build their code from both models' activations (their §2.1) and intervene on that joint code (their §3.2), and their "full replacement" keeps the base residual, so it is not the residual-deleting replacement of section 3.4 under another name. What we can say is narrower and still worth saying: single-side encoding is what *our* estimand did, it is what any pipeline that classifies features by single-side activation frequency (section 3.3) is exposed to, and its cost had not been measured. Both activations are available by construction here, so we can compare the two encodings on the same instances and the same ablation group (table 20). The residual falls $27\times$, the no-op becomes negligible, the interaction share falls from 0.97 to 0.17–0.51 — and the *measured effect of the ablation rises* $78\times$. The published analysis corrected the residual (failure mode 3) while leaving the encoding (failure mode 1) in place; the two act in opposite directions, and the larger was the uncorrected one.

Table 20: Single-side vs joint encoding at the canonical pair, same instances, same ablation group ($n = 15$ over 5 datasets); only the encoder input differs.

Quantity	single-side (published)	joint (corrected)
$\|\varepsilon\|/\|x\|$	0.5886 $[0.5880, 0.5893]$	$\mathbf{0.0218}$ $[0.0196, 0.0240]$
$\|u\|/\|x\|$	0.396	0.907
no-op	$+0.0039$ $[-0.0004, +0.0108]$	$+0.0002$ $[-0.0001, +0.0005]$
RP effect	$+0.0034$ $[+0.0013, +0.0068]$	$+\mathbf{0.2640}$ $[+0.0994, +0.5741]$
replace	$+0.1268$ $[+0.0995, +0.1573]$	$+0.2542$ $[+0.0969, +0.5503]$
I	$+0.1195$ $[+0.0964, +0.1426]$	-0.0100 $[-0.0244, -0.0008]$
$\|I\|/\|$replace$\|$	0.973 $[0.945, 0.996]$	$\mathbf{0.174}$ $[0.135, 0.217]$

Joint encoding is not free: under it the shared group's ablation vector carries 91% of the activation norm — a different quantity from the interaction shares above, which happen to land near the same number — so the full ablation sits $2.6\times$ above the interpretability boundary. Valid code, invalid dose. Neither encoding admits a surgical ablation of this group — which is why appendix V reports the contrast at the largest admissible dose rather than at $\alpha = 1$, and why we regard "ablate a feature group and read off its importance" as unsafe for groups this large relative to the code.

W.4 No mass-matched attention control exists

The knockout arms of table 25 are matched on key count. Masking sets logits to $-\infty$ and softmax renormalises, so severity tracks displaced probability mass, and a mass-matched control is the right comparison. It cannot be built: over layers 0–11 the thinking block carries mean item-attention mass

0.588 while the *entire* 128-row train context carries 0.412, on all 15 instances. A greedy search for the smallest train-key set reaching the thinking block's mass consumes every train key and still falls short; an earlier version of this analysis reported that degenerate arm as "matched", which was wrong, and the experiment now carries an explicit feasibility gate.

That the control is impossible is itself informative, and the well-defined substitute — degradation per unit of mass removed — is reported in table 25: thinking $+2.09$ $[0.98, 3.37]$, all-train $+2.60$ $[1.12, 4.35]$, difference $+0.50$ $[+0.08, +1.11]$ in favour of train keys ($p = 0.0136$). Per unit of attention they command, thinking positions are less load-bearing than training rows.

W.5 The superseded feature-level table

Several appendices below analyse the shared-group contrast as it was measured under single-side encoding at $\alpha = 1$. table 21 reproduces those published values so those sections remain readable. **They are superseded**: the corrected measurement (appendix V) gives 1.05 on v2 and 0.36 on v2.5 (reported without p-values), and the contrast is withdrawn. We keep the table because the robustness analyses built on it — dictionary ensembles, per-group nulls, multi-pair sweeps — remain informative about how that (invalid) estimand behaves, and because deleting the numbers a reader needs in order to check our retraction would be the wrong kind of tidying.

Table 21: **Superseded.** P1 ablation results as previously published: per-feature ratio vs alive-pool random baseline, single-side encoding, `data_only` scope, $\alpha = 1$, block-bootstrap CIs over 10 datasets. The legacy columns give replacement-mode estimates on the v2.5 side. Corrected values are in appendix V.

	RP		legacy, v2.5 (replacement)	
Group	on v2	on v2.5	no-op-adj.	raw i.i.d.
shared	**3.13** [2.23, 4.26]	**1.25** [1.12, 1.38]	4.90	3.70
v2-exclusive	1.68 [1.37, 1.99]	0.43 [0.27, 0.66]	0.04	1.18
v2.5-exclusive	0.21 [0.17, 0.25]	0.84 [0.72, 1.01]	-0.04	0.35
random (baseline)	1.00	1.00	1.00	1.00

X Configuration and cohort checks

X.1 The register under TabPFN's default ensembling

All primary results use $n_{\text{estimators}} = 1$; TabPFN's default is an ensemble over feature/class permutations. Since the register is the paper's central result we re-ran the three knockout arms across ensemble sizes ($n = 10$ over 5 datasets; table 22). The thinking knockout is *flat* — $1.02\times$ from 1 to 8 members — and stays roughly $8\times$ the size-matched random arm throughout. The thinking block's attention mass is likewise unchanged ($0.588 \to 0.590$).

We had pre-registered the expectation that the effect would *attenuate* with ensemble size, reasoning that averaging over permutations should partially compensate for a damaged channel. It does not, which suggests the register sits upstream of whatever the permutations vary. The verdict (robust) was predicted; the shape (no attenuation) was not.

Table 22: Knockout arms across ensemble sizes, $n = 10$ over 5 datasets. CIs are wider than the primary $n = 30$ knockout; the quantity of interest is the invariance across sizes, which is measured within-instance.

$n_{\text{estimators}}$	thinking	size-matched random	all train	thinking mass
1	$+1.16$ $[+0.64, +1.89]$	$+0.13$ $[+0.03, +0.23]$	$+1.04$ $[+0.49, +1.78]$	0.588
4	$+1.19$ $[+0.60, +1.95]$	$+0.15$ $[+0.02, +0.30]$	$+1.14$ $[+0.51, +1.86]$	0.591
8	$+1.19$ $[+0.61, +1.97]$	$+0.15$ $[+0.03, +0.27]$	$+1.14$ $[+0.52, +1.88]$	0.590

X.2 THE CAPABILITY GAP ON THE EXPANDED COHORT

The gap was previously reported on 10 datasets with a CI covering zero, which meant the paper was expressing results as fractions of an effect it had not established. We pre-committed (PREREGISTRATION §32.2) to expanding the cohort and, *if the pooled CI still covered zero*, to dropping the gap-fraction framing rather than reporting whichever cohort reached significance.

On all 16 datasets our loaders support ($\times 3$ seeds, $n = 48$) the mean relative gap is $+5.5\%$, block CI $[+1.8\%, +9.4\%]$, sign-flip $p = 0.0010$. The framing is retained. We record three caveats. First, we did not reach the pre-registered 30-dataset target: OpenML returned HTTP 504 on most new fetches, so we expanded to every loader already in the repository (19 attempted, 3 unusable: two with fewer than 256 rows, one with a non-numeric column that our loader refuses to coerce). Second, the expansion moved the estimate in our favour, which is the pattern the pre-commitment was written to guard against; what protects it is that the cohort is defined by availability rather than outcome and that the weaker 10-dataset subset is still reported. Third, the gap is heterogeneous — $+17.8\%$ (elevators, bike_sharing, pol) to -9.0% (kin8nm), with v2.5 winning 58.3% of instances — so it is a positive population mean, not a reliable per-dataset ordering.

X.3 STABILITY OF THE SHARED/EXCLUSIVE TAXONOMY

The dictionary has $2508/3072$ dead features and was fit on 1,600 paired positions from five datasets, so "shared" could name a property of that activation sample rather than of the models. Since 'classify_features' operates on activation matrices, the taxonomy can be recomputed from any subset.

Classifying on *one reference dataset at a time* gives mean pairwise Jaccard 0.774 on the shared set ($[0.702, 0.847]$ across the ten pairs); leave-one-dataset-out gives 0.881–0.910 against the published all-five partition. On three datasets the dictionary never saw, Jaccard against the published shared set is 0.763 (boston), 0.792 (kin8nm) and 0.559 (cpu_act), mean 0.705. Against the pre-registered bars (mean pairwise ≥ 0.70, held-out ≥ 0.60) the taxonomy is **stable**.

Two qualifications. The *v2-exclusive* group is not stable (mean pairwise 0.561, min 0.425); it is the smallest group at 65 features, so this is partly a small-set effect, but no claim in this paper should rest on v2-exclusive membership, and after appendix V none does. And transfer is good on average with a mediocre tail — cpu_act recovers only 190 of the 259 published shared features — so "stable" means the partition is mostly a property of the dictionary, not that it is dataset-invariant.

The consequence for appendix V is narrow but worth stating: the surviving cross-side contrast concerns a partition reproducible from unseen data, so it is not a per-run artifact. It remains small and unexplained.

X.4 POSITIVE CONTROL, NORM MATCHING, AND THE DOSE BOUND

appendix V reports a contrast whose absolute effects are $\approx 1/60$ of the label-independent fallback. A small p-value there shows the test is sensitive, not that the effect is meaningful, so we added the control a reviewer would demand: a direction *known* to be causal, at the same dose. The fallback direction $(\bar{x}_{\text{train}} - x)$ qualifies — driven to full scale it removes the token's information and costs 0.1011.

All four directions are rescaled to $\|u_{\text{shared}}\|$ so that only *which* direction is perturbed varies (table 23). Three results.

The dose resolves only weakly. At $\alpha^* = 0.3$ the fallback exceeds the best of the other three by $1.64\times$, against a pre-committed $\geq 3\times$ for support and $\leq 1.5\times$ for refutation — inconclusive. Per the pre-registered rule we therefore report the appendix V contrast as descriptive and attach no p-value to it in the main text.

The comparison was not norm-matched, and matching strengthens it. $\|u_{\text{shared}}\|/\|u_{\text{random}}\| = 1.46$: the published comparison gave the shared arm a 46% larger perturbation. Equalising norms takes shared/random from 0.358 to 0.191 (difference CI $[-0.0141, -0.0018]$). At matched norm the shared ablation is also less disruptive than an isotropic direction (0.0016 vs 0.0099).

The dose bound predicts where comparisons break. Kendall τ of the four-direction ranking against the smallest dose is $+1.00$ at every admissible dose and -0.67 at $\alpha = 1$. The ordering is perfectly stable below the bound and inverts above it — which is where the published operating point sat. This is operational evidence that the bound separates informative from uninformative doses, *not* a calibration against known causal magnitudes, which remains open.

Table 23: All directions rescaled to $\|u_{\text{shared}}\|$; $n = 15$ over 5 datasets. τ is Kendall's rank correlation of the four-direction ordering against the smallest dose. $\alpha^* = 0.3$ is the largest admissible dose; $\alpha = 1$ is inadmissible.

α	shared	random	isotropic	fallback (control)	τ
0.03	0.00012	0.00067	0.00099	0.00195	$+1.00$
0.10	0.00043	0.00236	0.00330	0.00619	$+1.00$
0.30	0.00160	0.00835	0.00994	0.01634	$+1.00$
1.00	0.29021	0.10238	0.03463	0.03801	$\mathbf{-0.67}$

X.5 How many thinking tokens are required?

A sink absorbs attention mass; its positions should therefore be largely interchangeable, and a small retained subset should recover most of the performance. Reviewers of an earlier draft asked the question directly, and it also carries an inference-cost consequence, so we pre-registered it (§34.4) with $\leq 20\%$ residual damage at $k = 8$ as the sink prediction.

Table 24: Retaining k of the 64 thinking keys and masking the rest, at all 18 layers. $n = 15$ over 5 datasets, block-bootstrap CIs. "Fraction" is the share of the full-knockout ($k = 0$) degradation still present.

k retained	rel. RMSE degradation	fraction of full knockout
0	$+1.228$ $[+0.574, +1.969]$	1.000
1	$\mathbf{+1.330}$ $[+0.610, +2.090]$	**1.083**
2	$+0.689$ $[+0.199, +1.205]$	0.561
4	$+0.486$ $[+0.222, +0.827]$	0.396
8	$+0.449$ $[+0.228, +0.745]$	0.366
16	$+0.334$ $[+0.197, +0.567]$	0.271
32	$+0.145$ $[+0.087, +0.194]$	0.118
64	0.000	0.000

The prediction fails (table 24): $k = 8$ leaves 36.6% of the damage, and 32 of 64 are needed before it falls below 12%. Two details compound this. Retaining a single key appears *worse* than retaining none ($+1.330$ vs $+1.228$), but the paired sign-flip gives $p = 0.055$ and the ordering does not resolve at this n; we withdrew the claim and note the direction only. And the curve deviates from linear-in-k by 0.54 of the full knockout, so it is neither the steep saturation a sink implies nor a clean linear dependence on count.

We take this as a genuine limit on the sink reading. The access/content dissociation of appendix X.42 is unaffected — it concerns *whether* the block's content matters, not how many positions carry it — but the redundancy that "absorbs attention mass benignly" also implies is absent. There is no inference-cost saving here: the 64 positions are not padding.

X.6 The thinking-token channel in full

Every arm of appendix X.42, with block-bootstrap CIs.

X.7 Causal efficiency and variance share across four designs

Per-design correlations against reconstruction R^2 are -0.225 (architecture ensemble, R^2 spread 0.0013) and -0.375 (epoch ladder, spread 0.1271), both short of the -0.5 we pre-registered; we report the variance correlate, not a reconstruction tradeoff. The shared group — the variance-dominant

Table 25: The thinking-token channel. Knockout severs *access*; transplant replaces the label-token slice of its *content* at all 18 layers. Knockout arms $n = 30$; transplant arms $n = 10$ over 5 blocks. Positive controls are *medians* over their three datasets with the range in brackets. Mass-normalised rows divide by the attention mass each arm removes, which is the only comparison available given that no mass-matched control exists (appendix W.4).

Intervention	What it removes	ΔRMSE
Mask attention to thinking keys	access (mass 0.588)	$+159\%$ $[+96, +227]$
size-matched random train keys	access (control)	$+23\%$ $[+9, +39]$
all raw train keys	access (mass 0.412)	$+134\%$ $[+81, +189]$
per unit attention mass: thinking		$+2.09$ $[+0.98, +3.37]$
per unit attention mass: all train		$+\mathbf{2.60}$ $[+1.12, +4.35]$
train − thinking		$+0.50$ $[+0.08, +1.11]$, $p = 0.014$
Transplant foreign dataset's block	dataset-specific content	$+0.26\%$ $[+0.07, +0.56]$
same dataset, different seed	content (control)	-0.05% $[-0.08, -0.02]$
foreign − same-dataset	*dataset identity*	$+\mathbf{0.31}\%$ $[+0.10, +0.63]$
zeroed block	structure (control)	median $+6.8\%$ $[-0.7, +30.5]$
matched-scale noise	structure (control)	median $+62.7\%$ $[+19.7, +102.0]$
self-transplant	nothing (sanity)	$+0.00\%$

Table 26: All 64 measurements behind section 5.4, binned by the direction's share of activation variance. Every direction is rescaled to $0.27\|x\|$ so neither the group's norm nor the dictionary's scale can confound. Designs: architecture ensemble (15 dictionaries), epoch ladder (15), layer/group/side sweep (30), TabICL single-model SAE (4).

variance share	n	efficiency < 1	mean efficiency	range
> 0.20 (variance-dominant)	37	$\mathbf{36/37}$	$\mathbf{0.424}$	0.13–1.12
0.05–0.20	5	$4/5$	0.959	0.65–1.90
< 0.05 (low variance)	22	$7/22$	$\mathbf{2.437}$	0.21–11.20
Spearman(variance share, efficiency) $= -\mathbf{0.659}$ $(-0.514$ / -0.600 / -0.640 per design)				

Table 27: The three failure modes at the canonical pair (v2 L2↔v2.5 L1), same instances and same ablation group throughout; `data_only` scope, $n = 15$ over 5 datasets. The published analysis in earlier versions of this paper corrected only failure mode (3), which is the smallest of the three, while (1) and (2) were left in place and act in the opposite direction. The right column fixes the *code* and not the *dose*: full ablation at α=1 puts the effect $2.6\times$ past our own admissibility bound (last row; the independent dose sweep of appendix X.30 puts the same comparison at $2.9\times$), so the $69\times$ it implies is a ratio between two measurements taken outside the range where either is interpretable as a feature magnitude. Both numbers are reported because the deflation is what we are documenting; neither licenses reading 0.2640 as the size of that feature group.

Quantity	single-side, α=1 (published)	joint, α=1 (corrected code)
$\|\varepsilon\|/\|x\|$ (residual)	0.589	$\mathbf{0.022}$
no-op (empty edit, replacement)	$+0.0039$	$+0.0002$
RP effect (the "true" effect)	$+0.0034$	$+\mathbf{0.2640}$
interaction share $\|I\|/\|$replace$\|$	0.97 $[0.95, 1.00]$	$\mathbf{0.17}$ $[0.14, 0.22]$
effect vs interpretability boundary	$0.03\times$	$2.6\times$ (*inadmissible*)

one in crosscoder diffing — has efficiency below 1 in $10/10$ layer-pair-and-side combinations (mean 0.596), while the smaller exclusive groups do so in only $8/20$ (mean 1.580).

X.8 The three estimand failure modes, side by side

X.9 Principal components: the test that killed our mechanism

We proposed that variance-dominant dictionary directions are inert because the model is robust along the high-variance directions of its own activations. That is a claim about the model's geometry,

independent of any dictionary, and principal components test it: they carry an exact variance ordering, and no reconstruction objective produced them.

Table 28: Causal efficiency along principal components of the v2.5 label-token activations at the canonical layer (1,600 positions, 192 dimensions), each rescaled to the same $0.27\|x\|$ reference norm as section 5.4 and compared against an isotropic direction. Block-bootstrap CIs over the 5 datasets.

PC	explained variance	efficiency
1	0.2429	0.785
2	0.1550	**5.767**
4	0.0892	2.558
8	0.0248	0.415
16	0.0098	0.502
32	0.0033	0.697
64	0.0006	0.542
128	0.0000	0.693

Spearman(explained variance, efficiency) $= +\mathbf{0.524}$ — the wrong sign, against a pre-registered ≤ -0.5 for support. The model is *not* systematically flat along its high-variance activation directions; PC_2 carries 15.5% of the variance and is $5.8\times$ more potent than random. We withdraw the mechanism.

The empirical relation for learned directions is unaffected, and the contrast sharpened the open question. Binned by variance share, dictionary directions in the 0.15–0.40 band average efficiency 0.411 over $n = 36$ measurements, while principal components in that band average 3.276 — but $n = 2$ there (0.785 and 5.767), and a two-point mean spanning that range is not an estimate. Activation variance concentrates in a handful of components, so the band is inherently thin. Drawing many random directions *within* the top-m PC subspace gives controlled variance share at adequate n; that experiment (PREREGISTRATION §38.4–38.6) is now run and reported in section 5.4, and it is what moved the finding from a variance regularity to a property of the training objective.

X.10 What the thinking-block case study does and does not establish

Three boundaries on the case study: it establishes necessity for v2.5, not an account of the v2→v2.5 gap; the transplant swaps the label-token slice, not the full block (appendix S.2); and the sink reading fails its redundancy prediction (appendix X.5), eight positions not substituting for the block, which no account we know predicts alongside content-inertness. Against Wu & Summa (2026) and Parodi et al. (2026) our increment is the modality, the mass normalisation and that negative result.

X.11 Learned versus unlearned directions at matched variance

X.12 Is the separation an artifact of how the two arms are perturbed?

The two arms of section 5.4 were originally measured with different perturbation shapes. The learned arm offsets by the group's contribution slab $u = \mathrm{dec}(f) - \mathrm{dec}(f_{\setminus g})$, whose direction may vary from row to row; the unlearned arm broadcasts a single unit vector to every row. At matched Frobenius norm a row-constant offset pushes coherently at every position while a row-varying one partially cancels, so the asymmetry alone could manufacture a separation.

Table 29: *Superseded by table 33.* Learned versus unlearned directions at band-level variance share and matched norm (v2.5 L1). Subspace-random directions are drawn inside the top-m principal subspace; the difference CI is a percentile bootstrap over the two groups. The unrestricted dictionary set gives the same estimate, -3.478 $[-4.275, -2.645]$.

Direction family	n	mean efficiency	range
learned (dictionary)	31	**0.395**	0.132–1.119
unlearned (subspace-random)	22	**3.889**	0.704–6.619
difference		$\mathbf{-3.495}$	$[-4.304, -2.654]$

It does not, and the reason is measurable. Writing row coherence as $\|\overline{u}\|/\overline{\|u_i\|}$, the group slab scores 0.9994 (0.9987–0.9998 across 15 dictionaries): it is already rank-one to within a tenth of a percent. The reference that makes that number mean something is the activations themselves, which score 0.9996 — the test label-token rows are nearly identical, so the slab's coherence is inherited from the data, not produced by the dictionary. (An incoherent control slab scores 0.0882.) Re-measuring every dictionary in the unlearned arm's exact broadcast form, against the same isotropic control and the same seeds, moves the band difference from -3.538 $[-4.338, -2.709]$ to -3.537 $[-4.337, -2.706]$; the constant form inflates the learned arm by $1.003\times$.

One consequence is worth stating because it bounds how the ratio should be read. The denominator is an incoherent slab while both numerators are coherent directions, so the absolute level of "efficiency" carries a coherence bonus. Both arms share it, so the *difference* is unaffected — but the ratio is between two arms measured the same way, not a claim that a dictionary direction is that much weaker than noise.

X.13 Watching the objective create the deficit

section 5.4 locates the separation between learned and unlearned directions but does not by itself show that training produced it: a dictionary might be inert because of its parameterisation (TopK, decoder normalisation) rather than because of what the objective did. We snapshot a single training trajectory per seed at epochs $\{0, 1, 2, 5, 10, 25, 50, 100, 200\}$, 12 seeds, measuring each snapshot's shared-group direction in the constant form of appendix X.12. Epoch 0 is pure initialisation — the control the published epoch ladder, which starts at 10, never had.

The prediction needs no fitted curve. Among unlearned directions efficiency *rises* with variance share; training rotates a dictionary into the high-variance subspace; so a direction that behaved like an unlearned one would become *more* potent as training proceeds.

Table 30: Dose-response in training, 12 seeds. The gap is against the five subspace-random draws nearest in $\log_{10}$ variance share. At initialisation the learned direction is indistinguishable from a matched unlearned one; after training it is not. Indistinguishable is not equivalent — see the equivalence test below.

	epoch 0 (initialisation)	epoch 200
sensitivity	1.021	$\mathbf{0.253}$
variance share	0.0063	0.2251
gap vs matched subspace-random	-0.133 $[-0.406, +0.153]$	$\mathbf{-1.554}$ $[-1.689, -1.417]$

$\rho(\text{epochs}, \text{sensitivity}) = -0.700$ and $\rho(\text{epochs}, \text{variance share}) = +0.843$: the direction acquires variance and loses potency at the same time. A within-seed paired sign-flip test on epoch 0 versus epoch 200 gives $p = 0.0006$ ($n = 12$). Classifying features afresh at each snapshot, or fixing the final classification and tracking the same indices backwards, gives the same picture (-0.730, $+0.913$, $p = 0.0010$).

Indistinguishable, but not equivalent. Reading the epoch-0 interval as "no deficit at initialisation" accepts a null. We fixed an equivalence margin before testing — 20% of the final gap's magnitude, ± 0.311 — and ran the TOST this paper already uses elsewhere. The 90% cluster interval is $[-0.345, +0.091]$, which falls outside the margin: the initialisation gap is *indistinguishable* from zero but not *equivalent* to it, and the interval admits an initial deficit as large as 0.345, 22% of the final one. Under the other group definition the same test passes ($[-0.306, +0.083]$ against ± 0.310); a boundary pass and a boundary failure are the same evidence and we do not quote only the first. The paper therefore says "indistinguishable" and does not say the deficit is absent.

Two procedural facts. First, these effects are averaged over $\pm d$; the run that produced the pre-registered verdict was single-sign, and re-measuring under the convention adopted in appendix X.14 and appendix X.18 strengthened rather than changed it ($1.028 \to 0.319$ and $p = 0.0012$ became $1.021 \to 0.253$ and $p = 0.0006$), so the paper now uses one estimator throughout. Second, the pre-registration specified 5 seeds, at which the paired test returned $p = 0.186$ and the pre-committed conjunction therefore failed; the recorded verdict at that point was inconclusive. With $n = 5$ the sign-flip permutation distribution has 2^5 atoms and the observed spread of initialisation efficiencies (0.32–

1.94) makes the threshold unreachable unless every seed agrees. We recorded a power amendment raising the count to 12 — decision rule, statistic, threshold, epoch grid and matching rule all unchanged — before running it, and report both. The rule "increase n if the test does not resolve" is outcome-dependent, so the nominal α is not preserved however carefully the amendment is recorded. A two-look Pocock boundary would ask for $\alpha \approx 0.031$ rather than 0.05; the observed $p = 0.0006$ clears it, so the conclusion is unaffected — but the correction belongs in the paper rather than in our favour.

X.14 Does it hold for a language-model SAE?

The dictionaries this field actually reads causally are SAEs on language models. We ran the matched-variance contrast on GPT-2 small with a public residual-stream SAE (Bloom, 2024), which uses L1+ReLU sparsity rather than our TopK, at layers 8 and 6. Directions are broadcast offsets rescaled to $0.27\|X\|_F$; the effect is mean $\mathrm{KL}(\text{clean} \,\|\, \text{patched})$ over next-token distributions; the denominator is a random unit direction in the same form, so no coherence gap separates numerator from denominator. Effects are averaged over $+d$ and $-d$, since a direction is a line and KL is not symmetric in the sign.

An alignment gate came first. These SAEs were trained on TransformerLens activations, which centre every write to the residual stream across d_{model}; on the raw HuggingFace stream they reconstruct at $R^2 = -1.1 \times 10^4$, and on the centred one at 0.9868 (L8) and 0.9923 (L6). We would have measured noise without it. Centring removes the all-ones direction, which the model's LayerNorm discards anyway, so a mean-zero perturbation is identical in both coordinate systems; every direction in both arms is projected mean-zero.

Table 31: GPT-2 small, individual SAE decoder atoms against subspace-random directions at matched variance share. The sign of the effect replicates; its size does not.

layer	atom efficiency	matched gap	matched ratio
8	1.355	-0.080 $[-0.138, -0.021]$	0.944
6	1.061	-0.063 $[-0.092, -0.033]$	0.944
tabular crosscoder, for comparison		-0.461	0.324

Both layers separate in the predicted direction with $n = 256$ atoms each and a control spanning variance shares 1.7×10^{-4} to 0.30. But the matched ratio is 0.944, a 5.6% deficit, against 0.324 — a 68% deficit — in the tabular setting. (The comparison row is the corrected support-matched estimate of appendix X.18; an earlier draft compared against the published band-mean -3.495 / 0.102, which table 10 retracts.) The reason is that the two are not measured in the same regime: individual SAE atoms almost never reach the variance band the tabular headline occupies — one of 256 at L8, none at L6. (That one atom's ratio is 0.310; $n = 1$ is not an estimate and we rest nothing on it.) The band-unreachability that defeated the TabICL check recurs here.

So the phenomenon generalises and its magnitude does not, and we have written the abstract accordingly. A first pass using group directions rather than individual atoms gave -0.192; sign-symmetrising reversed it to $+0.035$, which is why the atom analysis, with its wider control and $n = 256$, is the one we report. Directions whose group mean aligns with the residual stream's dominant component — as every group size did there — give a matched control that collapses onto a single principal component, which is exactly the failure the subspace draws were introduced to fix.

X.15 Feature *values* carry nothing transferable

Injecting v2.5's encoded feature values into v2 closes none of the gap at $n = 16$ eligible instances: shared -0.3% $[-1.9, +2.2]$, v2.5-exclusive -0.6% $[-1.8, +0.2]$, random -1.1% $[-3.5, +0.5]$, with a paired sign-flip not distinguishing shared from random ($p = 0.48$). A TOST against a $\pm$10pp gap-closure margin ($\pm 1.0\%$ relative RMSE on the eligible set) places the shared injection's 90% block CI at $[-1.8\%, +1.8\%]$, inside it — a positive equivalence claim. One caveat we volunteer: eligibility ($\mathrm{RMSE}_{v2} > \mathrm{RMSE}_{v2.5}$) is selected on the split that also forms the closure denominator, inflating it $2.9\times$ and biasing the estimate *toward* the equivalence we report; the cross-fitted re-analysis that would fix it is not run (appendix E.2). Simultaneous injection at all five aligned pairs leaves

RMSE unimproved (-1.3% $[-2.8, +0.5]$), and ablating thinking-position feature values is inert (9.7×10^{-6} per feature) — values inert, access critical.

X.16 Layer alignment is input-distribution-dependent

CKA layer alignments computed on synthetic versus real activations agree on only 2 of 8 pairs (Jaccard 0.25), so alignment is signal, not nuisance, in cross-model diffing (appendix M).

X.17 Anchoring the measure, and getting it wrong the first time

A reviewer of an earlier draft noticed that the positive control of appendix X.4 — the label-independent fallback, which at full scale replaces the test label token with the train-position mean and so destroys its information — was outranked by most of the unlearned draws then in the band. If generic noise outranks the direction that erases the label, a fixed-norm $|\Delta\mathrm{pred}|$ might be measuring displacement rather than relevance. We tested that on directions whose status is known independently of the measure.

We estimate the Jacobian $J = \partial\,\mathrm{pred}/\partial a$ at the test label token by central differences over all 192 basis directions and take its top right-singular vector: to first order this is the direction that *maximises* the response to a fixed-norm perturbation, so any measure worth the name must rank it at the top.

Table 32: Anchors with independently known status, measured on the same estimator at the same norm, on 5 datasets at evaluation seed 42, against $n = 40$ PC1-cone draws spanning the variance shares the anchors occupy. The two right-hand columns are different estimands and we print both: the *matched gap* is against each anchor's own nearest five controls (a point estimate; no interval is stored for it), the last column is the bootstrap difference against all forty draws, with its own point estimate and interval.

direction	known status	effect	vs. random	matched gap	vs. all draws [95%]
gradient (top singular of J)	first-order maximal	0.0990	$\mathbf{12.7}\times$	$+0.0897$	$+0.0908$ $[+0.042, +0.171]$
gradient (mean)	—	0.1047	$13.5\times$	$+0.0950$	$+0.0965$ $[+0.047, +0.177]$
fallback	erases the label's information	0.0156	$2.0\times$	$+0.0093$	$+0.0074$ $[-0.001, +0.021]$
probe	linear probe to the target	0.0096	$1.2\times$	$\mathbf{-0.0039}$	$+0.0014$ $[-0.005, +0.011]$
discriminant anchors — these should score low					
$v_{192}(J)$	first-order *minimal*	0.00057	$\mathbf{0.07}\times$	-0.0089	-0.0076 $[-0.0088, -0.0066]$
J-null mixture	orthogonal to top-8 of J	0.00142	$\mathbf{0.18}\times$	-0.0121	-0.0068 $[-0.0082, -0.0052]$

The last column is the secondary estimator PREREG §52 committed to reporting alongside the median comparison: each anchor against its own nearest-five matched controls, with a two-way bootstrap over anchors and draws. It separates the anchors that the pooled comparison does not. **Only the gradient anchors clear a matched control.** The fallback clears the median and the 95th percentile but its bootstrap difference covers zero; the probe clears the median only, and against its *own* matched controls it falls 0.0039 *below* them. Two of the forty controls are degenerate (PC1 mass > 0.8) and are kept in the pooled median, which is the more conservative choice here because they are unusually potent.

The discriminant anchors were added because the set could not fail without them. $v_1(J)$ is by construction the direction that maximises the measure's own first-order response, so "the gradient anchor scores highest" is close to a restatement of the estimator. Two reviewers made the point; the answer is an anchor built from the same J at its *minimising* end, which sits on the opposite side of that circularity. Both land far below the random median and below their own matched controls with intervals excluding zero. The measure places $v_1(J)$ and $v_{192}(J)$ a factor of 174 apart. Because J is badly conditioned ($\sigma_{\min}/\sigma_{\max}$ averaging 4.6×10^{-4}) and v_{192} could be numerical noise, the second discriminant anchor is a well-conditioned construction — a random direction orthogonal to J's leading eight-dimensional subspace — and it agrees. With one of the two we would not have written this paragraph.

So the honest instrument characterisation is: **monotone at the extremes, flat in between**. The measure separates a maximally-responsive direction from matched noise by an order of magnitude, separates two causally-inert directions from it by another order in the other direction, and does not separate an information-erasing control or a linear probe from noise at all. The middle is where most

dictionary directions live, which is the limitation to carry forward — but the set is now a two-sided test rather than a confirmation.

We report a failure of our own here because it changed what we believed. The first version of this test aimed its controls at the band of appendix X.18 using a top-k subspace mixture. That construction cannot exceed variance share ≈ 0.16 in this setting while the band runs to 0.40, so *two* of forty draws survived, both at 0.027–0.029 — against anchors spanning 0.004 to 0.232. The comparison was not variance-matched, and a "95th percentile" was computed over two points. Against those two unusually potent directions the probe and the fallback fell below the median, and we concluded in an earlier draft that the measure read output sensitivity but not information use. That conclusion was an artifact of the control. The anchors' own numbers never moved: 0.0990, 0.0096 and 0.0156 are identical before and after.

Two things follow that we would rather a reader take than the result itself. The degeneracy diagnosed for the language-model arm (appendix X.37) and for the headline (appendix X.18) was present here too and we did not look for it until an audit of the whole repository forced the question — a fix applied in one place does not propagate itself. And the reviewer observation that motivated this section is subject to the same defect: the draws it compared the fallback against were drawn under the unreachable band, so that ranking was not variance-matched either.

X.18 Redoing the headline as an actual matched comparison

The published contrast compared the two arms inside a variance band rather than at matched variance. Inside $[0.15, 0.40)$ the learned arm occupied 0.205–0.231 while the unlearned arm averaged 0.186, with 20 of 22 draws below every learned direction; band membership is a stratum, not a match. Four further defects travelled with it: 18 of the 22 draws lay inside the top-three principal subspace, so the control's effective dimension was about three; the learned arm pooled 15 epoch-ladder snapshots from three trajectories and contained one bit-identical duplicate; the interval was an unblocked two-group percentile bootstrap although the pre-registration specified a block CI; and the tabular arm was never sign-symmetrised, a convention whose adoption reversed the sign of the language-model group arm (appendix X.14).

Redoing it with all five fixed — unlearned draws built at the learned arm's realised support, each direction matched to its five nearest draws, the learned arm restricted to the 15 independent training runs, inference by a dictionary $\times$ dataset cluster bootstrap, and effects averaged over $\pm d$ — the separation survives and its size falls by two thirds.

Table 33: The headline under three estimators, on 5 datasets at evaluation seed 42, v2.5 L1. The published figure is retained for comparison; it is superseded by the first row below it.

estimator	estimate	95% CI
published band-mean difference	-3.495	$[-4.304, -2.654]$
support-matched, dictionary $\times$ dataset cluster	$\mathbf{-0.461}$	$[-0.581, -0.340]$
nearest-five matched gap	-0.455	$[-0.510, -0.394]$
covariate-adjusted family coefficient	-0.440	$[-0.511, -0.368]$

Learned directions score 0.220 against 0.681, a ratio of $3.1\times$ rather than the published $9.9\times$; part of the difference is the sign convention and part is the pseudo-replication. The three estimators agree to within 0.02, which also ends a state in which this paper gave two answers — appendix X.13 matched nearest-five and implied $5.7\times$ while the main text reported $9.9\times$.

The control is degenerate, and we say so rather than let it pass. No unit direction can carry more variance share than the leading principal component, which here is 0.243. Two consequences follow. The band $[0.15, 0.40)$ is unreachable above 0.243, so its upper two-thirds is empty by construction — a fact we did not check when we chose it. And at the learned arm's support the unlearned directions available are a cone about PC1: our draws sit 3.9–27.8° from it with mean PC1 mass 0.889 and a participation ratio of 1.3 out of 192. The comparison is therefore against same-variance near-PC1 directions. It is the same degeneracy we diagnosed for the language-model group arm and, at the time, did not think to look for here.

X.19 TWO METRICS THAT STOP AGREEING

appendix X.13 shows that training *causes* the deficit. It does not say what the deficit is made of. At $d_{\text{model}} = 192$ that question is answerable directly: reconstruction scores a direction by the variance it carries, $d^\top \Sigma d$, while the output's response to a small perturbation is governed to first order by $\|Jd\|$, with $J = \partial \,\text{pred}/\partial a$ at the test label token. We estimate J by central differences over all 192 basis directions per instance — the same estimate appendix X.17 uses — and stack the five instances, since every perturbation in this paper is shared across them.

Three results follow, none of which needs a new forward pass beyond that estimate.

The measure is first-order. Over 50 directions spanning both families, Spearman between $\|Jd\|$ and measured sensitivity is $+0.794$. Together with the gradient anchor of appendix X.17 this settles what section 5.4 is reading: a local slope, not a saturation artifact.

The two metrics are nearly orthogonal past the leading directions. Principal angles between the top-eight eigenspace of Σ and the top-eight right-singular space of J are $21.3°$, $25.2°$, $39.4°$, $45.2°$, $57.5°$, $84.8°$, $88.7°$, $89.5°$. The leading pair agrees — which is why high-variance directions are decent handles at all — and the agreement is gone by the sixth. A dictionary that allocates capacity by variance is therefore allocating it by a criterion that stops tracking sensitivity almost immediately.

Learned directions sit further from J's leading direction. $|\cos(d, v_1(J))|$ is 0.0179 for the 15 trained dictionaries against 0.0376 for the matched unlearned draws, a difference of -0.0197 $[-0.0288, -0.0098]$. This is the proximate mechanism for section 5.4: not that learned directions are strange, but that they are placed by a criterion which is close to orthogonal to the one that governs response.

What this still does not explain. We pre-registered the caveat and it stands: why reconstruction-plus-sparsity training should land *further* from $v_1(J)$ than a variance-matched draw is not answered here. appendix X.13 shows training does it and this section shows what "it" is geometrically; the step from the objective's gradient to that placement is missing, and a dictionary trained with a causal term in its objective remains the experiment that would supply it (appendix R).

X.19.1 MATCHING ON $\|Jd\|$ INSTEAD: THE DEFICIT SURVIVES, AND ONLY HALF OF IT IS TESTABLE

If Σ and J are nearly orthogonal, then matching a control on variance share balances the arms on an axis the outcome barely depends on. A reviewer put the consequence sharply: since the variance-matched control collapses onto a cone about PC1, and $|\cos(\text{PC1}, v_1(J))| = 0.0324$, the matching procedure selects controls that *cannot* respond. We re-ran the headline with $\|Jd\|$ as the matching covariate (PREREG §54): controls are cones about $v_1(J)$ bisected to a target first-order response, caliper 0.25 dex, nearest five, two-way cluster bootstrap over directions and datasets.

Table 34: The same 15 dictionaries and the same cell, matched on two covariates. Neither matching can balance the other: the two axes are nearly orthogonal.

matched on	gap	matched	control PC1 mass	control part. ratio	other covariate
variance share	-0.461 $[-0.581, -0.340]$	$15/15$	0.889	$1.29/192$	—
$\|Jd\|$	$\mathbf{-0.293}$ $[-0.497, -0.108]$	$7/15$	$\mathbf{0.0055}$	$\mathbf{65.6/192}$	1.7 dex apart

The deficit survives, and the control is finally a population. All seven matched directions are individually negative, with $\rho(\|Jd\|, \text{gap}) = -0.821$. Because the learned arm occupies 0.6–5.6% of the ceiling $\sigma_1(J) = 0.0609$, a cone that hits those targets is tilted far from $v_1(J)$ and is therefore close to isotropic — participation ratio 65.6 of 192 against 1.29 under variance matching. The objection that our control arm has effective dimension one does not apply to this estimator.

The covariance ceiling has a Jacobian analogue, but it is a property of the generator, not of the sphere. The variance-share ceiling is a theorem: no unit direction exceeds $\lambda_1(\Sigma)/\operatorname{tr}\Sigma$. The first-order response has no such floor on the sphere — $\min_{\|d\|=1} \|Jd\| = \sigma_{\min}(J)$, which is 0 when J has a null space, and appendix X.17 uses exactly such directions as its inert anchors. What has a floor is our *control generator*: a cone about $v_1(J)$ tilted towards a random direction orthogonal to it cannot respond less than that orthogonal direction does, here 0.00105 (median 0.00199), and

twenty-nine of forty draws hit the 90° bisection bound. Eight of the fifteen learned directions lie *below* that generator floor (0.00037–0.00082), so this generator produces no matched control for them. A generator that mixed in the minor singular subspace of J would reach lower; we did not build one, and an earlier version of this paragraph wrote "you cannot build a random direction that responds as little as this one", which conflates the generator's range with the sphere's. The honest statement is narrower: under the generator we registered, the comparison is undefined for the half of the learned arm where the claim is strongest, and a control that reaches them would have to be constructed from the low-response subspace rather than drawn near $v_1(J)$.

Excluding them is not conservative, and we say so. Since ρ is strongly negative, the eight excluded directions are the low-$\|Jd\|$ ones and would carry *less* negative gaps; -0.293 is therefore an estimate over the testable half and not over the fifteen. What does hold for the other eight needs no matching at all: their first-order response is smaller than that of *any* random direction orthogonal to $v_1(J)$ that we drew.

Why the $69\times$ needed a second design. It compares single-side against joint encoding, and the two differ in perturbation norm as well as in encoding ($\|u\|/\|x\|$ of 0.396 against 0.907); since the response saturates, part of that ratio is dose. The clean design is a 2×2 — {single-side, joint} $\times$ {single-side norm, joint norm} — which appendix X.20 runs, and which assigns almost all of the 69 to dose. The *direction* is not in doubt either way: the residual share falls from 0.973 to 0.17–0.51 on the same instances, and no dose change moves that.

The two covariates cannot be matched at once. Apply a 0.25 dex variance-share caliper to the same $\|Jd\|$-matched directions and *zero* of fifteen have an admissible match — learned median 0.2245 against control 0.0045, 1.7 decades apart. Match on variance instead and the control collapses onto PC1. This is the paper's argument reduced to one pair of covariates: "matched control" is not a property of a comparison until the covariate is named, and here the two natural choices are close to orthogonal, so balancing either unbalances the other by more than a decade. The main text leads with the variance-matched gap because variance share is the covariate the field matches on and the one our registration named; $\|Jd\|$ is the covariate that predicts the outcome ($\rho = +0.794$ above), and an earlier draft called it "primary" here while the main text led with the other, which a reviewer rightly read as two answers. There is one answer: both are reported, the variance-matched one is the headline, and neither is the comparison's true value, since section 5.4 shows they cannot be balanced at once.

X.20 Encoding or dose? A norm-matched 2×2

The $69\times$ compares two arms that differ in two ways: the code, and the perturbation norm (0.396 against 0.907 of $\|x\|$). Three reviewers pointed out that our own dose grid makes the second consequential and that we had applied appendix W.1's standard to everyone but ourselves. This is the design that separates them (PREREG §55): each code rescaled to each norm, direction preserved, three seeds $\times$ five datasets, dataset-block bootstrap. The diagonal reproduces the published values exactly ($+0.00336$ and $+0.26404$).

Table 35: Signed mean relative prediction change. Columns are the perturbation norm, rows the code. The published comparison is the diagonal.

	single-side norm ($0.396\|x\|$)	joint norm ($0.907\|x\|$)
single-side encoding	$+0.00336$	$+0.31253$
joint encoding	$\mathbf{-0.00024}$ (7 of 15 negative)	$+0.26404$

On magnitudes, where the ratios are defined: **encoding main effect** $0.78\times$ $[0.66, 0.90]$, dose main effect $76.1\times$ $[61.1, 98.5]$, published composite $59.2\times$, log interaction -0.011. The encoding interval excludes 1 *from below*: at matched norm, switching to a valid code makes the measured effect 22% *smaller*.

So the $69\times$ is a dose effect — but the dose is caused by the encoding. A joint code reconstructs far better ($\|\varepsilon\|/\|x\|$ of 0.022 against 0.589), so removing the same feature group extracts a much larger vector. The total effect of fixing the encoding is 59–79$\times$; its *direct* effect is $0.78\times$ and the rest is mediated by the norm. Since the response saturates in that range, "the effect is $69\times$ bigger" does

not mean "we recovered the group's true importance"; it means we made a $2.29\times$ larger perturbation where the response multiplies it by 76. Both doses are inadmissible ($3.9\times$ and $9.0\times$ past α^*).

What is untouched, and what is withdrawn. The diagnostic stands: on a single-side code the interaction is 97% of the apparent effect and on a joint code 17%, measured on the same instances and the same group, and the fix costs nothing. What we withdraw is the description of $69\times$ as what off-manifold encoding costs. **And one cell is worth more than the headline was**: with a valid code at the smaller dose the ablation produces -0.00024, sign-unstable across seeds — a correctly-coded ablation of the shared group, measured nearer an admissible dose, has no detectable effect at all.

X.21 Aggregation under inference that fits the design

appendix X.41 fitted 72 windows per layer with homoskedastic OLS standard errors. The windows overlap — a width-256 window starting at 0 contains the width-1, -4, -16 and -64 windows starting at 0 — so they are not 72 independent units, and L6's lower bound sat at $+0.0008$. We re-ran the inference two ways (PREREG §56): a cluster bootstrap resampling window *starts*, and a refit on the subset of windows whose index ranges are disjoint.

Table 36: The bundling coefficient $\beta_{\log_2 w}$ under three inferences. The point estimate does not move; the interval does, and L6 stops resolving.

layer	OLS (as published)	clustered on starts (8 clusters)	non-overlapping refit
8	$+0.0317$ $[+0.0137, +0.0497]$	$+\mathbf{0.0317}$ $[+0.0052, +0.0662]$	not identified ($n = 4$)
6	$+0.0182$ $[+0.0008, +0.0355]$	$+0.0182$ $[-\mathbf{0.0061}, +0.0425]$	not identified ($n = 4$)

The claim narrows accordingly: aggregation is a causal variable *measured at L8*, and at L6 the coefficient is same-signed but does not resolve at this n. Eight window starts overstates the independent units: $[0, 256)$ and $[2, 258)$ share almost every feature. Assigning each window to the 256-block its start falls in gives four disjoint blocks, and clustered on those the L8 coefficient is $+0.0317$ $[+0.0038, +0.0477]$ ($p = 0.007$, PREREG §91) while L6 covers zero ($[-0.0221, +0.0363]$). Four clusters is a coarse bootstrap and we report it as such; it is the most independence this design affords, and the coefficient survives it. The non-overlapping refit gives no estimate at all, and an earlier draft of this table reported one. The greedy disjoint subset keeps four windows that are all of width 256, so $\log_2 w$ is constant across them and collinear with the intercept; the design matrix has rank 2 of 3. Least squares still returns a number — the minimum-norm tie-break, $+0.099$ and $+0.101$, whose suspicious agreement across two layers is the artifact's signature rather than corroboration — and constraining the coefficient to zero leaves the residual sum of squares unchanged to seven digits. We withdraw it and the sentence that argued from it. We also report what the regression alone obscured: the direct atoms-versus-groups contrast in the same artifact is $+0.195$ $[+0.081, +0.304]$ at L8 and -0.081 $[-0.222, +0.049]$ at L6, so the two layers disagree there too, and quoting only the coefficient was selective.

X.22 How often `audit` is right when the answer is known

That `audit` returns INADMISSIBLE on our own worked example says nothing about its error rates, as a reviewer pointed out. We built six scenarios whose validity and true gap are fixed by construction — in 32 dimensions with a dominant direction, 8 learned and 30 control directions per comparison — and ran each 200 times with every check supplied except where the scenario omits one (PREREG §94).

No invalid comparison was admitted by either version. The first version failed on valid ones: its interval resampled the learned directions and treated each matched control mean as fixed, so the control arm's uncertainty never reached it, and it declared a difference where there was none 19% of the time. The function we shipped as "one call" was weaker than the estimator behind our own headline, which already resampled both arms. The fixed version resamples learned clusters and control rows and rebuilds the caliper matches on every draw; against unchanged pre-registered criteria (PREREG §95) its false-positive rate is 5.0% and its coverage 96.0%. Point estimates are unchanged. A later fix to the dose check, which now sweeps around the dose supplied rather than a fixed range of absolute norms (appendix X.45), was re-run through the same calibration and left every

Table 37: `audit` on comparisons of known validity, 200 runs each. The first version's interval resampled only the learned arm; the second resamples both, as the paper's own estimators do.

scenario	correct verdict	first version	fixed
valid, true gap 0	ADMISSIBLE	100%	100%
valid, true gap -0.2	ADMISSIBLE	100%	100%
disjoint support	UNDECIDABLE	100%	100%
control a PC1 cone	INADMISSIBLE	100%	100%
saturating dose	INADMISSIBLE	100%	100%
checks omitted	UNDECIDABLE	100%	100%
false positives, true gap 0	$\leq 10\%$	**19.0%**	5.0%
coverage of the true gap -0.2	$\geq 85\%$	**81.0%**	96.0%

rate in table 37 identical (PREREG §97.1), as did a second fix that judges linearity on both ratios touching the reference rather than the top ratio alone (PREREG §100). With 200 runs the rates carry Monte-Carlo error: 5.0% has a Wilson 95% interval of $[2.7, 9.0]\%$ and 96.0% one of $[92.3, 98.0]\%$. Rerun at 1,000 per scenario under the current rules, no invalid comparison is admitted in 4,000, the false-positive rate is 4.2% $[3.1, 5.6]\%$ and coverage 96.3% $[94.9, 97.3]\%$ (PREREG §100.1).

Where the degeneracy verdict switches, and why we did not move it. Sweeping the control's target variance share from 0.3 to 0.97 of the ceiling, the verdict switches to INADMISSIBLE between 0.3 (mean PC1 mass 0.23) and 0.5 (PC1 mass 0.44). The switch is set by the participation-ratio threshold of 8, not by the PC1-mass threshold of 0.8: on a steep spectrum a control at half the ceiling already has an effective dimension below eight. We read that as the check doing its job — an effective dimension of three is what our own headline control had, and what we called degenerate — and did not retune a threshold after seeing where it bites; the docstring now says where.

What it costs (PREREG §99). The checks need no gradients, no second model and no retraining. On synthetic data of the paper's two shapes the full call takes seconds, almost all of it the 10^4-draw two-way bootstrap; the dose check adds five forward passes of the caller's model, one per rung of the 0.25–4× ladder.

Table 38: Wall-clock cost of `audit` (median of five, one CPU process), from `results/exp_audit_cost/`.

shape ($n \times d$, learned, control)	ceiling, support	matched gap	full call	dose check
tabular (1600 × 192, 15, 60)	0.08 s	4.1 s	5.0 s	5 forward passes
GPT-2 (4000 × 768, 24, 60)	0.83 s	5.6 s	7.9 s	5 forward passes

X.23 What admission buys, against a known forward model

The calibration above builds effects from a distribution and asks whether `audit` applies its own rules; it cannot say whether those rules protect the conclusion a reader draws. Here the effects are computed by a known nonlinear model, $F(x) = \sum_j c_j \tanh(g_j w_j^\top x)$, with a held-out family built from softplus thresholds, so the truth is known without appeal to the rules: the sign of the difference in first-order, norm-fixed sensitivity $E|\nabla F \cdot q|$ between learned and control directions. Each scenario draws readouts, gains, a learned arm that leans towards one readout, matched controls and a dose, then runs `audit` with its thresholds unchanged (PREREG §104).

Table 39: `audit` against a known forward model (PREREG §104). Sign error: the measured matched gap's sign differs from the true first-order sign. The first three columns were registered; the last three (*resolved*: the interval excludes zero) were added after seeing them and are exploratory. Generated by `exp_groundtruth_table.py`.

family	n	error, all	admitted	error, admitted	resolved	error, resolved	error, adm.+res.
tanh (primary)	400	21.5%	72.0%	19.1%	27.3%	0.9%	0.0%
softplus (held out)	200	15.5%	94.0%	14.4%	31.0%	0.0%	0.0%

The registered criterion fails. We required that on the held-out family admission at least halve the sign-error rate. It lowered it from 15.5% to 14.4% (21.5% to 19.1% on the primary family), so we describe `audit` as a functional check — a record that the checks ran and passed — and not as a certificate that a comparison's sign is right. Most errors survive admission because they are not dose errors: the dose check called 295 of 400 primary scenarios linear, and the wrong signs sit mostly among comparisons whose interval covers zero.

What does protect the sign, found after the fact. Reading the same scenarios by whether the interval excludes zero — a reading we added after seeing the registered results, so it is exploratory — the sign error among resolved comparisons is 0.9% and 0.0%, and among comparisons both admitted and resolved it is zero in both families. That is the reporting rule the paper already uses (report a sign only when admitted and resolved); this experiment suggests that the interval, not the dose check, does most of the work in it, on this model class.

X.24 The headline on datasets never used for evaluation

Every tabular number in section 5.4 was computed on the same five evaluation datasets at one seed, and several of our analysis choices were made after seeing them. A reviewer asked for the comparison on data that played no part in those choices. We ran the unchanged headline estimator — the same fifteen dictionaries, control generator, reference norm, sign symmetrisation and two-way cluster bootstrap — on six OpenML regression datasets it had never been evaluated on, with the decision rule registered first (PREREG §93). They are the six that load at the headline split and were not in the legacy ten-dataset cohort; `boston` is excluded on data-ethics grounds.

Table 40: The headline comparison on the original evaluation cohort and on six datasets never used for evaluation. Only the evaluation set changes.

cohort	learned	control	matched gap [95%]	negative
original five	0.220	0.681	-0.461 $[-0.581, -0.340]$	$5/5$
six unseen	0.244	0.879	$\mathbf{-0.635}$ $[-1.023, -0.313]$	$\mathbf{6/6}$
all eleven	0.233	0.805	-0.571 $[-0.803, -0.385]$	$11/11$

Per dataset the gap is -1.020 (`bank32nh`), -0.222 (`elevators`), -0.466 (`fried`), -0.679 (`house_8L`), -0.220 (`mv`) and -1.431 (`pol`), and the two arms separate completely (largest learned 0.489, smallest control 0.536). The learned arm barely moves ($0.22 \to 0.24$); what moves is the control, whose sensitivity relative to an isotropic perturbation ranges from 0.43 to 1.74 across the new datasets. The direction of the comparison is a property of the dictionary; its size is not, and depends on the evaluation data. With six units the smallest attainable exact sign-flip p is $2/2^6 = 0.031$, and six of six is that.

What this does not change: the dose is the same unvalidated one (appendix X.49), so this confirms that the *measurement* does not depend on which datasets we chose, not that it is a causal magnitude.

X.25 Is the deficit an artifact of the dictionary's training set?

Two threats we had listed and not measured (PREREG §57). First, the crosscoders are trained on full-sequence activations that include the positions the causal measurement later perturbs: we checked, and 638 of $1{,}600$ training rows are exactly those positions. Second, $N = 3072$ features are fitted to $1{,}280$ training points, so the dictionary is under-determined and $R^2_{\text{val}} = 0.997$ may be memorisation.

Table 41: Nine seeds per arm (three for the scale arm), measured with the estimator of appendix X.18 against the same control arm, whose realised support is 0.191–0.242 in variance share. An arm with no direction inside that range has no matched comparison at all.

arm	N	training rows	R^2_{val}	variance share	in support	matched gap
held-out (intervened rows removed)	3072	962	0.9918	0.131–0.160	$\mathbf{0/9}$	**not constructible**
determined ($N \leq$ training points)	768	1600	0.9948	0.179–0.227	$8/9$	$\mathbf{-0.314}$ $[-0.476, -0.113]$
$9.6\times$ the training data (PREREG §89)	3072	15360	0.9996	0.223–0.232	$3/3$	$\mathbf{-0.503}$ $[-0.612, -0.402]$
as published	3072	1600	0.997	0.215–0.231	—	-0.461 $[-0.581, -0.340]$

The leakage threat no longer closes, and an earlier version of this appendix said it did. That version reported -0.304 $[-0.480, -0.100]$ and concluded the sign survived. It held the intervened rows out of *dictionary training* only. The classification that decides which features are "shared" — the treatment definition itself — still read all $1{,}600$ rows, and so did the matching covariate. Holding the classification out too changes which features are shared, and the ablation direction that results carries variance share 0.131–0.160 against the control's 0.191–0.242: 0 of 9 inside. A contrast between arms whose supports do not overlap is the error this paper withdrew its own headline over (appendix X.18), so we do not report one. The number it would give is $+0.601$, and we record it only to say that reading it as a reversal would repeat the mistake. **The threat is now unmeasured rather than excluded.**

Losing support is itself a result, and not one we anticipated. An activation frequency is a ratio; dropping 40% of the rows moves every feature near the $\tau = 0.01$ threshold, so *which rows you classify on* is a group-definition choice in the sense of C3. appendix X.50 found the same class of problem in the rule's wording; this is the same rule's sensitivity to its sample.

The training-data scale threat closes too. A reviewer's first objection was that $1{,}600$ training positions is five orders of magnitude below published SAE practice and that 2.4 atoms per training point is memorisation by construction. We trained the same recipe on $15{,}360$ positions (nineteen datasets, three seeds; 0.25 atoms per point) and measured the same contrast: -0.503 $[-0.612, -0.402]$, all three directions inside the control's support, val R^2 up from 0.992 to 0.9996 and the dead fraction unmoved at 85%. The deficit is not a property of the small training set, and the dead fraction is not either (this addresses training-data scale only; dictionary width, model size and the evaluation cohort are separate axes, the last treated in appendix X.24) — it is a property of the recipe, TopK at $k = 48$ against a fixed $\tau = 0.01$, as a code review had suggested. We predicted this arm would lose common support as the held-out one did (PREREG §89) and it did not: shrinking the classification population moves features across the frequency threshold, growing it stabilises them. We are still three orders of magnitude below 10^7 and say so.

The under-determination threat does close, once it has the power to. At three seeds the determined arm covered zero and we said we could not tell a null from a power failure. At nine it resolves: -0.314 $[-0.476, -0.113]$, same sign as the published estimate and 8 of 9 inside the control's support. It was a power failure. At $N = 768 \leq 1{,}600$ the dictionary is smaller than its training set — which limits, without excluding, memorisation, since the encoder and decoder remain free — and the deficit is still there, which is the stronger of the two threats this appendix was written to test, and the one that now has an answer. Read with appendix X.38, where R^2_{val} moves from 0.76 to 0.997 without the deficit following, neither dictionary quality nor dictionary size explains this result.

X.26 Does the reversal depend on the SAE family or the metric's order?

A reviewer made a second SAE family and a second metric the condition for taking the GPT-2 result seriously; a second reviewer noted that the tabular arm's metric is first-order while ours is KL, which is second-order at the clean point. One $2 \times 2 \times 2$ answers both (PREREG §58). The second family is OpenAI's v5 TopK SAE ($k = 32$, $32{,}768$ wide, `resid_post`, layer-norm input) against the L1+ReLU dictionary used above ($24{,}576$ wide, `resid_pre`); the second metric is $\frac{1}{N}\sum_i |\Delta \log p(t_{i+1})|$, first-order in the perturbation. Directions, variance shares and perturbations are computed in the raw centred residual space for both, so only the dictionary and the read-out differ. Both families clear the reconstruction gate ($R^2 = 0.983$–0.992).

Table 42: Matched gaps, common support, caliper 0.25 dex, two-way bootstrap. All eight cells are positive; the magnitude depends on the dictionary and not on the metric.

	L1+ReLU, 24,576		TopK k=32, 32,768	
layer	KL (2nd order)	$\log p$ (1st order)	KL	$\log p$
8	$+0.400$ $[+0.215, +0.582]$	$+0.323$ $[+0.169, +0.484]$	$+0.044$ $[-0.059, +0.161]$	$+0.051$ $[-0.026, +0.137]$
6	$+0.325$ $[+0.211, +0.427]$	$+0.278$ $[+0.179, +0.368]$	$+0.116$ $[+0.012, +0.207]$	$+0.089$ $[+0.005, +0.163]$

The metric's order is not the issue. All eight cells agree in sign and their intervals overlap within each column pair; the first-order metric is consistently a little smaller. The objection that the two settings cannot be compared because one reads a first-order quantity and the other a second-order one does not survive measurement.

The dictionary's objective is. The L1 arm gives $+0.28$ to $+0.40$; the TopK arm $+0.04$ to $+0.12$, and at L8 it covers zero. The sign never turns negative in any cell, so the contrast with the tabular setting (-0.29 to -0.46) holds throughout and the title withdrawal stands on it. What does not hold is the magnitude.

And the better-supported arm is the smaller one. The TopK dictionary leaves *no* learned direction without an admissible match ($n = 51$–63 non-degenerate), where the L1 dictionary leaves 13–14 unmatched out of 28–37. Its group directions spread across the variance range the cone controls can cover; the L1 dictionary's do not. We report that the arm with more support gives the weaker effect.

Once again a direction generalises and a magnitude does not — after tabular to GPT-2 atoms, and atoms to groups. The answer to "is this learned direction more causal than random?" turns out to depend on the dictionary's training objective as well as on the aggregation unit. The mirror check on the tabular side, a second dictionary family under the same estimator, is appendices X.33 and X.34: there the objective separates too, in the other direction.

X.27 The bilinearity grid lies entirely outside our own dose bound

appendix W.1 refuted bilinearity with a coefficient of variation of 2.07 over the (α, β) grid. A reviewer observed that the refutation is dominated by cells at α or $\beta = 2$, whose perturbation is several times the interpretability boundary this paper asks everyone else to respect. We re-analysed the grid (PREREG §59), calling a cell admissible when each injected component's relative norm is at or below 0.1011, the bound then called α^* (a model-specific coordinate since replaced; appendix X.49).

Zero of sixteen cells are admissible. The smallest, $\alpha = \beta = 0.25$, puts the u component at $0.98\times$ the bound and the residual component at $1.46\times$; because β scales a residual whose own norm is $0.589\|x\|$, no value of β on this grid reaches the bound. The coefficient of variation is therefore undefined inside our own admissible region, and we cannot say from this experiment whether the response is bilinear there.

What we can say is narrower than what appendix W.1 said, and we restate it: *above* the bound the response is not bilinear, and above the bound is where binary ablations of a full feature group actually sit (2.9–3.9× past it). That is enough for the claim this paper needs — a binary ablation's magnitude cannot be read as a feature magnitude — and not enough for the general statement that the response is not bilinear. Re-running the grid at $\alpha, \beta \leq 0.15$ is the right follow-up; appendix X.30 reports it.

X.28 How much do the degeneracy gate and the bootstrap scheme decide?

Two choices could have carried the conclusions: the PC1-mass threshold of 0.80, which decides which layer/side cells enter the pooled estimate, and the ordinary cluster bootstrap, which can be anticonservative at five clusters. Both are re-analyses (PREREG §61); neither threshold was chosen after seeing this sweep, and 0.80 remains primary because it was registered before the results.

Table 43: Degeneracy-gate sweep over the ten layer–side cells. The mixed sign — the claim — holds throughout; the pooled point estimate does not.

PC1-mass gate	usable cells	negative	positive	pooled mean	signs mixed
0.60–0.85	3	1	2	$+0.232$	yes
0.80 (registered)	3	1	2	$+0.232$	yes
0.90	7	5	2	-0.267	yes
0.95	8	6	2	-0.240	yes

The sign stays mixed at every threshold admitting two or more cells. The pooled mean does not: relaxing the gate to 0.90 admits four more cells, most of them negative, and the pooled estimate flips to -0.267. We had already dropped that number from the abstract; this is why. Note the direction of the dependence — a *laxer* gate makes the pooled result agree *more* with our tabular headline, so the registered 0.80 was not the choice that flattered us.

On inference, a Rademacher wild cluster bootstrap over the five dataset clusters gives -0.452 $[-0.521, -0.383]$ for the variance-matched estimate against the published -0.461 $[-0.581, -0.340]$, and -0.348 $[-0.492, -0.203]$ for the $\|Jd\|$-matched estimate against -0.293 $[-0.497, -0.108]$.

Both exclude zero and both are *narrower* than the pairs bootstrap, so the anticonservativity concern does not bite on this data. We report the pairs intervals in the main text because they are the registered ones.

X.29 A SECOND DICTIONARY FAMILY ON THE TABULAR SIDE

appendix X.26 put a second SAE family on the language-model arm and recorded that we had not done the same on the tabular side, which left this paper's own headline measured on exactly one dictionary type. We trained single-model L1+ReLU sparse autoencoders on the v2.5 activations at the canonical cell ($N \in \{768, 3072\}$, three seeds, reconstruction $R^2 = 0.9997$) and measured them with the estimator of appendix X.18, reusing its control arm unchanged (PREREG §62).

We caught an effective-n=1 comparison in our own new script, for the fourth time. The automatic verdict was "the sign reverses, $+0.206$ $[+0.021, +0.406]$" — from a support-restricted arm containing *one* learned direction, because only one of the 36 reaches the control arm's variance range. That is the pathology this paper diagnoses in appendices X.18 and X.37.1. The verdict logic now refuses a support-restricted arm below five members and falls back to caliper matching.

Table 44: Single-model L1+ReLU SAEs against the same control arm as the crosscoder.

estimator	learned n	gap	note
support-restricted	1 of 36	—	refused (effective $n = 1$)
caliper-matched (0.25 dex)	20 of 36	$+\mathbf{0.068}$ $[-0.108, +0.314]$	16 with no admissible match
crosscoder (published)	15	-0.461 $[-0.581, -0.340]$	

The deficit does not reproduce. On a single-model L1 dictionary the gap covers zero. The tabular result is a property of *this crosscoder*, not of dictionary learning on this model, and the main text now says so.

And the two families do not occupy the same variance region. The SAE's group directions span 0.004–0.200 while the crosscoder's shared group sits at 0.215–0.231 and the controls at 0.191–0.242. Sixteen of 36 have no admissible match. This is structurally what appendix X.14 found when GPT-2 atoms topped out at 0.23: change the dictionary family and you change the region where the comparison is defined at all.

A regularity we named here and then withdrew. An earlier version of this appendix read the two settings together — tabular $-0.461 \to +0.068$, GPT-2 L8 $+0.044 \to +0.400$ — and concluded that the L1 objective sits above TopK in both. It does not. The tabular comparison there confounds the objective with the input dimension, and when appendix X.34 separated them in a factorial the tabular contrast ran the *other* way (-0.274 $[-0.514, -0.034]$), as did the L_0-controlled replication (appendix X.33). Two settings made a pattern that a 2×2 does not support. What survives is narrower: the objective moves the estimate at matched sparsity in both settings, *upward* on language models and downward on tabular dictionaries (§76), so the direction is a property of the setting and not of the objective.

X.30 INSIDE THE BOUND, THE RESPONSE *is* BILINEAR

appendix X.27 found that no cell of the published grid lies inside our own admissibility bound, and recorded that re-running at smaller doses was the right follow-up. We ran it (PREREG §63): $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$, $\beta \in \{0.02, 0.05, 0.10, 0.15\}$, everything else identical.

Table 45: The same test on two grids. The pre-committed rule is CV < 0.20 supports bilinearity, CV > 0.50 refutes it.

grid	dose range (in units of α^*)	admissible cells	CV of $I/(\alpha\beta)$	verdict
published (appendix W.1)	1.5–11.6	$0/16$	2.067	refuted
this one	0.12–0.87	$\mathbf{16/16}$	$\mathbf{0.064}$	**supported**

$I/(\alpha\beta)$ varies only between 0.00165 and 0.00217 across a fourfold range in α and a 7.5-fold range in β. **Inside the bound the interaction is bilinear; outside it the coefficient of variation is 32 times larger and the ranking of intervention arms inverts.**

Where is the transition? Later than the bound. A third grid straddling the boundary (u dose 0.78–$1.25\,\alpha^*$, ε dose 0.87–1.40) gives CV $= 0.049$: still bilinear at 25% past α^*. So the bound does not sit at the transition; it sits inside the linear regime, and the transition lies somewhere in $[1.25, 3.92]\,\alpha^*$ — by $\alpha = 1$ ($3.92\,\alpha^*$) the coefficient has risen fiftyfold. We do not narrow that bracket further.

This is less than we would like. section 6.1 called the bound operationally justified but not calibrated, and it still is not: what we can now say is that it errs *safe*. A value derived from a label-overwrite cost turns out to be conservative relative to where bilinearity actually fails. As an operating rule that is the right direction to be wrong in; as a measured constant it is not established.

What does not change is the claim that needs it: the conventional operating point — full ablation at $\alpha = 1$ — sits at $3.92\,\alpha^*$, inside the non-linear regime, where $I/(\alpha\beta)$ is fifty times its linear-regime value. There is an interpretable regime and binary ablations are outside it; we simply cannot place its edge exactly.

X.31 Depth: five layers, two families, two metric orders

Reviewers called the evidence thin at two layers. The machinery of appendix X.26 takes the layer as a parameter, so we swept $\{2, 4, 6, 8, 10\}$ (PREREG §65).

Table 46: Matched gaps at five depths. Of 20 cells, 18 are evaluable and *none* is negative; the jbloom L2 SAE fails the reconstruction gate ($R^2 = 0.41 < 0.5$) and is excluded and recorded rather than forced in.

layer	L1+ReLU, KL	L1+ReLU, $\log p$	TopK, KL	TopK, $\log p$
10	$+0.805$ $[+0.657, +1.053]$	$+0.575$ $[+0.479, +0.732]$	$+0.305$ $[+0.078, +0.563]$	$+0.235$ $[+0.093, +0.379]$
8	$+0.400$ $[+0.215, +0.582]$	$+0.323$ $[+0.169, +0.484]$	$+0.044$ $[-0.059, +0.161]$	$+0.051$ $[-0.026, +0.137]$
6	$+0.325$ $[+0.211, +0.427]$	$+0.278$ $[+0.179, +0.368]$	$+0.116$ $[+0.012, +0.207]$	$+0.089$ $[+0.005, +0.163]$
4	$+0.270$ $[+0.200, +0.330]$	$+0.254$ $[+0.184, +0.315]$	$+0.030$ $[-0.051, +0.114]$	$+0.025$ $[-0.044, +0.096]$
2	gate fail	gate fail	$+0.100$ $[+0.009, +0.219]$	$+0.079$ $[+0.003, +0.179]$

Fourteen of the eighteen exclude zero, four cover it, none is negative. Depth behaves like family and aggregation: it moves the magnitude — on the L1 dictionary monotonically, from $+0.270$ at L4 to $+0.805$ at L10 — and not the sign. Against the tabular crosscoder's -0.46, the contrast holds at every depth we can measure.

X.32 What survives a multiplicity correction

Earlier drafts said every positive claim here is exploratory and uncorrected. That is honest and passive, so we did the correction (PREREG §64). The family — 21 directional claims the paper asserts — was fixed before any p was computed. Two of the twenty-one are not hypothesis tests: they are non-negative shares against a boundary null of 0, for which a percentile-bootstrap p is degenerate, and they are reported descriptively (PREREG §77.1, §90). The other nineteen are tests and all nineteen count towards m. One of them, the tabular second family, stores no bootstrap draws; it enters at $p = 1$ and can never be rejected, which only makes the correction stricter for the rest. p comes from the artifact's own pre-registered sign-flip test where one exists and otherwise from the same bootstrap draws that give the printed interval (Phipson–Smyth $+1$). Where every draw falls on one side of zero, p is the Monte-Carlo floor $2/(B+1)$: the smallest value $B = 10^4$ draws can express. We use it as-is and read it only as "below about 2×10^{-4}". The table and every count in this paragraph's caption are written by `exp_multiplicity.py` (PREREG §98).

[†]At the Monte-Carlo floor $2/(B+1)$: every draw fell on one side of zero, and B bounds how small a p the bootstrap can report.

Two things changed this table after review. First, seven of the sixteen entries that had a p got it from a normal approximation to the percentile interval, $\mathrm{se} = (\mathrm{hi} - \mathrm{lo})/3.92$ followed by erfc. The

Table 47: Holm ($\alpha = 0.05$) and Benjamini–Hochberg ($q = 0.05$) over the twenty-one entries of the family. Nineteen are hypothesis tests and enter both corrections; one of those stores no bootstrap draws and enters at $p = 1$. 14 of the 18 rankable tests survive Holm and 16 survive BH. *Dose validated* says whether the causal reading is licensed under the dose rule of appendix X.49; n/a where the statistic is not a dose-dependent contrast. Two shares are non-negative statistics against a boundary null and are reported without a test. Bold marks the paper's headline claims. Generated by `exp_multiplicity.py`.

claim	a test?	dose validated	raw p	Holm-adjusted	BH q
dose positive control	yes	n/a	$1.0 \times 10^{-4\dagger}$	1.9×10^{-3}	4.2×10^{-4}
headline, variance-matched	yes	no	$2.0 \times 10^{-4\dagger}$	3.6×10^{-3}	4.2×10^{-4}
GPT-2 reversal, L8	yes	no	$2.0 \times 10^{-4\dagger}$	3.6×10^{-3}	4.2×10^{-4}
GPT-2 reversal, L6	yes	no	$2.0 \times 10^{-4\dagger}$	3.6×10^{-3}	4.2×10^{-4}
encoding main effect	yes	n/a	$2.0 \times 10^{-4\dagger}$	3.6×10^{-3}	4.2×10^{-4}
dose main effect	yes	n/a	$2.0 \times 10^{-4\dagger}$	3.6×10^{-3}	4.2×10^{-4}
convergent anchor $v_1(J)$	yes	no	$2.0 \times 10^{-4\dagger}$	3.6×10^{-3}	4.2×10^{-4}
discriminant anchor $v_{192}(J)$	yes	no	$2.0 \times 10^{-4\dagger}$	3.6×10^{-3}	4.2×10^{-4}
discriminant anchor J-null	yes	no	$2.0 \times 10^{-4\dagger}$	3.6×10^{-3}	4.2×10^{-4}
headline, $\|Jd\|$-matched	yes	no	8.0×10^{-4}	8.0×10^{-3}	1.5×10^{-3}
training causes the deficit	yes	no	1.0×10^{-3}	9.0×10^{-3}	1.6×10^{-3}
capability gap, expanded cohort	yes	n/a	1.0×10^{-3}	9.0×10^{-3}	1.6×10^{-3}
determined dictionary	yes	no	5.4×10^{-3}	0.038	7.9×10^{-3}
aggregation, L8	yes	no	7.0×10^{-3}	0.042	9.5×10^{-3}
attention, mass-matched	yes	n/a	0.014	0.068	0.017
GPT-2 TopK SAE, L6	yes	no	0.030	0.120	0.036
GPT-2 TopK SAE, L8	yes	no	0.404	1.000	0.451
aggregation, L6	yes	no	0.621	1.000	0.656
tabular second family	yes	no	*no draws*	—	1.000
interaction share, single-side code	*no*	no	*boundary null*	—	—
interaction share, joint code	*no*	no	*boundary null*	—	—

percentile bootstrap is not symmetric in general, and the tell was in the table itself: the excluded non-tests printed $p < 10^{-300}$ and 1.9×10^{-277}, values no $B = 10^4$ resampling distribution can support. We now take every p from the draws (Phipson–Smyth $+1$) and have deleted the approximation. Both entries whose Holm survival rested on it — the encoding main effect and the convergent anchor — survive, and by a wider margin: $1.6 \times 10^{-3} \to 2.0 \times 10^{-4}$ and $5.7 \times 10^{-3} \to 2.0 \times 10^{-4}$. The approximation had been *conservative* here, which is the opposite of what we predicted in the registration (PREREG §82.1), and the reason we could not have known is that its direction is set by the skew we had not measured. Three tests whose artifacts did not persist their draws were first entered at $p = 1$; the two GPT-2 TopK cells have since been rerun to store them (every estimate and interval reproduced exactly), and they had also been dropped from the family by an edit, which we found and reversed (PREREG §98). One test, the tabular second family, still has none.

What a p-value here tests, and what it does not license. A reviewer of an earlier version pointed out that we had withheld the headline's p for an unvalidated dose while keeping the p of the Jacobian-matched contrast at the same norm, and that withholding a p neither restores a metric's validity nor is required by its absence. Both points are right. Every test in table 47 now carries its p, which tests the statistical hypothesis that the two arms differ in norm-fixed sensitivity at the norm used — a hypothesis that is defined at any norm. The column *dose validated* says whether the causal reading is licensed under the single dose rule of appendix X.49. Under that rule it is *no* for every dose-dependent row, on both models: the tabular reference norm is below the half-effect norm but the directions respond super-linearly around it, and on GPT-2 the reference norm is past the half-effect norm and the response saturates. A row can therefore be significant and still not be a causal magnitude, and the main text reads every row that way.

Why a Holm verdict changed between versions, stated so it cannot be mistaken for a choice. The aggregation coefficient at L8 had raw $p = 0.008$, clustered on eight overlapping window starts; in the twelve-entry family it was first tested in, its Holm-adjusted value was 0.056 and it failed. Its p is now read from the four disjoint 256-blocks, the clustering the paper reports (PREREG §98). Holm multiplies the i-th smallest p by $m - i + 1$, which is the number of tests whose p is at least as large;

adding tests with *smaller* p raises m and i together and leaves that multiplier unchanged. What moves it is the tests ranked *behind* a claim. The current value is computed, not argued: In this family of 19 tests the L8 aggregation coefficient ranks 14 of 18 by raw p (7.0×10^{-3}); Holm multiplies it by the number of tests whose p is at least as large, $m - \text{rank} + 1 = 6$, giving 0.042. The mass-matched attention contrast ($p = 0.014$) now fails Holm, because restoring the two TopK cells put a test with $p = 0.03$ and one with $p = 0.40$ behind it; it passes BH, and it is not a claim the abstract makes.

Five datasets is the binding constraint, and no correction removes it. Every tabular cell but one rests on the same five evaluation datasets at one seed; the headline alone has since been re-measured on six datasets it had never been evaluated on (appendix X.24). An exact two-sided sign-flip test over five units cannot go below $p = 2/2^5 = 0.0625$; the bootstrap p-values above do not inherit that floor, but they do inherit its lesson, which is that inference this precise from five clusters leans on the resampling model. The wild cluster bootstrap of appendix X.28 narrowed rather than widened the headline's interval, which is reassuring about that model and is not a substitute for a sixth dataset.

What no correction fixes, and we restate it: the *choice* of which hypothesis to register next was conditioned on the previous result — H16 after H15 was refuted, H20 after review, H52 after an audit, H38 after H34. table 10 is the record of that search; multiplicity control over the survivors does not undo it.

X.33 A SECOND LANGUAGE MODEL, AND WHAT IT DOES TO THE REVERSAL

appendix X.31 widened the language-model arm to five layers, two dictionary families and two metric orders, and recorded that it was still one model. Pythia-160m closes that: a different architecture family (GPT-NeoX), corpus (Pile), position encoding (rotary) and SAE codebase (EleutherAI `sparsify`), at the same d_{model}. The SAE is TopK $k = 32$ with 32,768 latents on the residual stream; the estimand is unchanged (PREREG §66).

Table 48: Pythia-160m. Every cell passes the reconstruction gate and has near-complete support, and *no* cell resolves; four of six point estimates are negative.

layer	recon R^2	ceiling	KL	$\log p$
8	0.9635	0.798	$+0.054$ $[-0.256, +0.440]$	$+0.052$ $[-0.174, +0.321]$
5	0.9783	0.884	$\mathbf{-0.221}$ $[-0.449, +0.068]$	$\mathbf{-0.182}$ $[-0.359, +0.032]$
2	0.9204	0.505	$\mathbf{-0.134}$ $[-0.461, +0.294]$	$\mathbf{-0.080}$ $[-0.254, +0.141]$

The GPT-2 result does not replicate. On GPT-2 all 18 evaluable cells are positive and 14 exclude zero. On Pythia-160m none of six excludes zero and four point the other way. We cannot say Pythia reverses the reversal — every interval covers zero — but we can say the comparison is *undetermined* there, and that the sign of the point estimate depends on depth.

So "on a language model the comparison reverses" is a statement about GPT-2. The main text now names the model. That is a real narrowing of the contrast the title withdrawal rested on, and we would rather state it than let a reader find it.

We wrote here that the model was the first axis able to move the sign. That was premature, and the next experiment says so. The Pythia-160m arm has only the TopK family, and TopK is the family that sits near zero on GPT-2 too. On Pythia-70m-deduped, where both families exist at matched width (PREREG §69), the L1 arm resolves positive in six of six cells ($+0.13$ to $+0.72$) while the TopK arm resolves in two and points negative in four.

Table 49: Pythia-70m-deduped, both families at 32,768 latents. The family separates them; the sparsity level is not matched and we say so below.

family	L_0	layer 4 (KL)	layer 3 (KL)	layer 1 (KL)
L1+ReLU	68.3	$+0.721$ $[+0.540, +0.881]$	$+0.193$ $[+0.061, +0.337]$	$+0.501$ $[+0.303, +0.679]$
TopK k=32	32.0	$+0.295$ $[+0.083, +0.485]$	$\mathbf{-0.193}$ $[-0.397, +0.018]$	$\mathbf{-0.152}$ $[-0.343, +0.057]$

Across three model checkpoints the pattern is the family, not the model: L1 dictionaries reverse (GPT-2 $+0.27$ to $+0.81$, Pythia-70m $+0.13$ to $+0.72$), TopK dictionaries sit near zero and can

go negative (GPT-2 $+0.03$ to $+0.31$, Pythia-70m -0.19 to $+0.30$, Pythia-160m -0.22 to $+0.05$). Pythia-160m's null is what a TopK arm looks like, not what a different model looks like.

Is it the objective or the sparsity level? The public dictionaries do not match on sparsity — L_0 is 68 for the L1 arm and 32 for TopK — so on its own the contrast confounds the two. We separated them on the same dictionaries by imposing the TopK arm's k on the L1 dictionary at inference (PREREG §71), which holds the objective and matches the level.

Table 50: The same L1 dictionary at its own sparsity and at TopK's. Reconstruction barely moves, so this matches the level rather than breaking the dictionary.

arm	objective	L_0	recon R^2	layer 3 (KL)	layer 1 (KL)
L1	L1+ReLU	68.3	0.811	$+0.193$ $[+0.061, +0.337]$	$+0.501$ $[+0.303, +0.679]$
L1, truncated	L1+ReLU	**31.9**	0.823	$\mathbf{+0.220}$ $[+0.081, +0.369]$	$\mathbf{+0.537}$ $[+0.358, +0.683]$
TopK	TopK k=32	32.0	0.861	$\mathbf{-0.193}$ $[-0.397, +0.018]$	$\mathbf{-0.152}$ $[-0.343, +0.057]$

It is the objective. At matched L_0 the L1 dictionary resolves positive in six of six cells and tracks its untruncated self, not the TopK arm — including at the two layers where TopK points negative. Truncation costs almost nothing in reconstruction ($0.811 \to 0.823$; at layer 3 it *improves*, from 0.816 to 0.875, so much of that dictionary's activation tail contributes nothing), which is what tells us the level was matched rather than the dictionary broken.

Trained under the constraint, not truncated after it. Truncating a dictionary trained without an L_0 constraint is not the same as training one with it, so we trained both objectives ourselves on the same activations at $N = 4096$, two seeds, equal budget (PREREG §72), searching λ_1 until the L1 arm's realised L_0 reached the band around the TopK arm's 32.

Table 51: Our own dictionaries, trained under the constraint, at $N = 4096$ on 65k positions (PREREG §72) and at $N = 16{,}384$ on 229k positions (§75). L1 resolves positive in eight of eight cells, TopK in one of eight, negative — but no cell matches on L_0 *and* reconstruction at once, at either size. *out of the $[24, 40]$ band.

N	objective	layer	L_0	recon R^2	KL gap	$\log p$ gap
4096	L1	3	52.5^*	0.914	$\mathbf{+0.271}$ $[+0.128, +0.375]$	$+0.207$ $[+0.107, +0.280]$
	L1	1	36.2	0.652	$\mathbf{+0.229}$ $[+0.141, +0.338]$	$+0.140$ $[+0.094, +0.195]$
	TopK	3	32.0	0.958	$+0.052$ $[-0.048, +0.153]$	$+0.035$ $[-0.037, +0.109]$
	TopK	1	32.0	0.878	-0.098 $[-0.222, +0.015]$	-0.052 $[-0.124, +0.015]$
16,384	L1	3	28.0	0.902	$\mathbf{+0.351}$ $[+0.227, +0.466]$	$+0.265$ $[+0.176, +0.346]$
	L1	1	32.4	0.717	$\mathbf{+0.247}$ $[+0.178, +0.315]$	$+0.156$ $[+0.118, +0.193]$
	TopK	3	32.0	0.968	-0.030 $[-0.116, +0.067]$	-0.017 $[-0.080, +0.054]$
	TopK	1	32.0	0.922	$\mathbf{-0.137}$ $[-0.237, -0.011]$	-0.062 $[-0.128, +0.023]$

Neither experiment controls everything, and that is why we ran both. Near $L_0 = 32$ the L1 objective buys its sparsity with reconstruction: at layer 1 the L_0 matches (36.2 against 32.0) and R^2 does not (0.652 against 0.878); at layer 3 R^2 is close and L_0 is not. The truncation experiment above has the opposite profile — it matches L_0 *and* R^2 but imposes no constraint during training. The two control different confounds and give the same answer, which is the most we can claim: neither alone is decisive. (appendix X.38 is independent evidence on the same worry: there, R^2_{val} moving from 0.76 to 0.997 leaves the effect unchanged.) What would settle it is an L1 arm wide or long-trained enough to reach $L_0 = 32$ at the TopK arm's reconstruction. We tried, at four times the width and $3.5\times$ the data (PREREG §75, with a 0.05 matching criterion fixed in advance; three amendments before any gap was computed are recorded in §75.0: the corpus capped the data at $3.5\times$ rather than the registered $4\times$, the evaluation was chunked after an out-of-memory failure, and the λ_1 grid was refined after the first sweep jumped over the L_0 band — the same kind of amendment §72.0 records for the smaller run), and it does not get there (table 51, lower block): both L1 arms land in the L_0 band, but the reconstruction gap is 0.066 at layer 3 and 0.205 at layer 1. The L1 arm pays for entering the band: at layer 3 the band-satisfying $\lambda_1 = 1.5$ shrinks harder than the $\lambda_1 = 1$ of the smaller run (out of band at $L_0 = 52$), and its R^2 *falls* by 0.012 while TopK's rises by 0.010; at layer 1 the L1 arm gains more than TopK ($+0.065$ against $+0.044$) and the gap narrows from 0.226 to 0.205

without closing. Averaged over the two layers both arms gain $+0.027$ and the gap does not move. We had predicted layer 3 would match; it did not. The two objectives therefore remain confounded with reconstruction under training-time matching, and the truncation-plus-training pair above is what we can claim. Descriptively the wider L1 arm is positive in four of four cells with larger estimates, and the wider TopK arm resolves *negative* in one cell (layer 1, KL: -0.137 $[-0.237, -0.011]$) — the first language-model *group* to resolve below its controls, and the first in a dictionary we trained ourselves; individual GPT-2 atoms already do so at both layers (appendix X.14). The same cell was -0.098 $[-0.222, +0.015]$ at $N = 4096$: the point estimate moved by 0.04 while the interval barely narrowed, so the resolution is a shift in the estimate, not added precision — one metric, one self-trained dictionary, and we rest nothing on it.

The direction is not universal: on the tabular side the objective separates the same way, in the opposite direction. We also trained tabular dictionaries at two controlled sparsity levels, searching λ_1 until the L1 arm matched the TopK arm's L_0 (PREREG §70): across both runs below, $\lambda_1 = 0.3$ gives $L_0 = 42$–49 against a target of 48 and $\lambda_1 = 1.0$ gives 11–12 against 16. With three seeds and four windows neither the objective nor the level separated (family contrast -0.30 $[-0.67, +0.04]$ at $L_0 = 16$, -0.14 $[-0.45, +0.16]$ at 48), which we recorded as a power limit rather than a null. Raising it to six seeds and eight windows (PREREG §76, registered with the prediction that only the $L_0 = 16$ contrast would resolve) resolves both: -0.57 $[-0.78, -0.36]$ and -0.32 $[-0.49, -0.16]$, L1 *below* TopK, while the L_0 contrast within TopK still covers zero (-0.20 $[-0.46, +0.07]$). The L1 arm at $L_0 = 16$ again missed its target by 27% (11.6) and its group directions again sit mostly outside the controls' support ($12/48$ and $14/48$ matched against $44/48$ and $46/48$), so the contrast is a contrast on the matched subset, as in appendix X.34. The objective therefore moves the estimate in both settings; it moves it up in the language models and down here.

Table 52: Tabular dictionaries at matched L_0, at the registered size (three seeds $\times$ four windows, PREREG §70) and at four times the units (six seeds $\times$ eight windows, §76). Every cell is positive; the objective separates them at the larger size. The level resolves only within L1, at $+0.05$ $[+0.005, +0.092]$, an order of magnitude below the family contrast.

units	target L_0	objective	realised L_0	λ_1	matched	gap
12	16	TopK	16.0	—	$11/12$	$+0.614$ $[+0.267, +1.124]$
	16	L1	12.1 (missed by 24%)	1.0	$6/12$	$+0.314$ $[+0.169, +0.461]$
	48	TopK	48.0	—	$12/12$	$+0.469$ $[+0.145, +0.886]$
	48	L1	46.4	0.3	$6/12$	$+0.334$ $[+0.169, +0.500]$
48	16	TopK	16.0	—	$44/48$	$+0.881$ $[+0.628, +1.181]$
	16	L1	11.6 (missed by 27%)	1.0	$12/48$	$+0.315$ $[+0.169, +0.575]$
	48	TopK	47.9	—	$46/48$	$+0.684$ $[+0.437, +1.017]$
	48	L1	45.4	0.3	$14/48$	$+0.365$ $[+0.216, +0.500]$

We report both sizes. At twelve units the contrasts covered zero and we called that a statement about power; at forty-eight they resolve, in the direction the twelve-unit point estimates already had. The objective therefore separates on both kinds of dictionary at matched L_0, and the estimates run in opposite directions: L1 above TopK on the public language-model dictionaries, below TopK on our tabular ones, matching appendix X.34.

One thing does replicate. All eight tabular cells are positive ($+0.31$ to $+0.88$), as the window-defined crosscoder groups were in appendix X.34. Across two dictionary types, two objectives and two sparsity levels, only the shared/exclusive group definition produces the negative gap this paper's headline reported.

X.34 Input, sparsity, and the thing that actually decides: what counts as a group

appendix X.29 reported that the L1 family sits above the TopK family in both settings, and flagged that the tabular contrast confounds the sparsity penalty with the input dimension. This is the 2×2 that separates them: joint 384-d versus single 192-d input, crossed with TopK k=48 versus L1+ReLU, three seeds throughout, with group directions defined identically in all four cells as windows of the variance-contribution ranking (PREREG §67). **One thing this design does not hold fixed, and we**

found it late: expansion factor 16 over a 384-d joint input gives $N = 6144$ but over a 192-d single input gives $N = 3072$, so width covaries with input. The *input* main effect is therefore confounded with width; the *sparsity* main effect is not, since width is constant within each input level. The verdict below is already that neither contrast resolves, so this widens a negative result rather than rescuing a positive one (PREREG §77.3).

Table 53: The tabular 2×2. One cell cannot be measured: its directions sit an order of magnitude below the control arm's variance range.

input	sparsity	recon R^2	variance share	matched	gap
joint 384-d	TopK k=48	0.9972	0.047–0.222	13/18	$+0.072$ $[-0.192, +0.342]$
joint 384-d	L1+ReLU	0.9982	0.010–0.065	$\mathbf{0/18}$	not measurable
single 192-d	TopK k=48	0.9992	0.052–0.223	14/18	$\mathbf{+0.355}$ $[+0.116, +0.631]$
single 192-d	L1+ReLU	0.9997	0.074–0.200	14/18	$+0.081$ $[-0.098, +0.332]$

Neither factor is consistent, and we withdraw the regularity we reported. With one cell unmeasurable, each main effect has only one conditional contrast: sparsity at single input is -0.274 $[-0.514, -0.034]$ — L1 *below* TopK, the opposite of what appendix X.29 inferred from two settings — and input at TopK is $+0.283$ $[-0.012, +0.570]$. Two points made a pattern that a factorial does not support.

But the re-measurement revealed something larger. We re-ran the crosscoder arm with the window definition so the four cells would be comparable. Holding the architecture (ef $= 16$, $k = 48$, $N = 6144$), the seeds, the cell, the control arm and the estimator fixed, and changing *only* what counts as the feature group:

Table 54: The same dictionaries, measured two ways. This is the paper's headline against itself.

group definition	matched gap	learned efficiency
shared/exclusive classification	$\mathbf{-0.544}$ $[-0.649, -0.447]$	0.127–0.149
top variance-contribution windows	$\mathbf{+0.072}$ $[-0.192, +0.342]$	0.212–2.340

The shared group is strikingly homogeneous — efficiency 0.127–0.149 across three seeds, variance share 0.229–0.231 — and low. Other groups from the same dictionaries are neither: their efficiencies span an order of magnitude and straddle the control.

So this paper's tabular result is a statement about the shared group, not about "the directions this dictionary learns". We have relabelled it throughout. It is the purest instance of our own thesis: appendix X.41 showed that group *width* moves the magnitude; group *membership* moves the sign.

A plausible reading, which we did not test: shared features are the directions two model versions agree on, and what survives a version change may be exactly the high-variance, low-sensitivity subspace that appendix X.19 shows reconstruction favours and response ignores. We record it as a conjecture, not a result.

X.35 What a full rerun reproduces, and what it does not

Before submission we ran `repro.sh` end to end at the length it then had. Every stage completes (after one call-signature bug in stage 40 that had survived because no one had run the pipeline end to end; PREREG §73.1). The manifest gate at the final stage then reported 74 mismatched entries, 68 of them numbers the paper prints, and the cause is not measurement noise.

Activations reproduce bit-for-bit; dictionaries do not. Stages 1–4 give byte- identical activation tensors. Stage 5 trains the five canonical-pair crosscoders, and the result depends on the BLAS thread count: at a fixed count the training is deterministic across processes, but the reduction order changes at $4 \to 8$ and $8 \to 10$ threads, the weights differ by 10^{-6} after the first step, and TopK's discontinuity turns that into 9×10^{-2} within 20 epochs. After 200 epochs the encoder weights differ by 26–49% (mean 39%, five pairs, measured against a retraining at 8 threads) in relative terms from the dictionaries the paper's numbers came from, whose thread count in 2026-05 was not recorded and

which we could not regenerate at any thread count. That is a different dictionary, and 92 downstream stages measure it.

Table 55: Encoder-weight divergence after 20 epochs, seed 42, by BLAS thread count (max $|\Delta w|$). PyTorch caps at 10 on this machine.

threads	1	2	4	8	10+
vs. 1 thread	0	2×10^{-6}	2×10^{-6}	7×10^{-3}	9×10^{-2}

Most verdicts survive; one does not. Of the 56 pre-registered verdict fields the rerun covered, 53 are unchanged on the retrained dictionaries. Two move in our favour and we keep the original, less favourable, verdicts: the training-budget confound (H24c $\rightarrow$ H24a) and initialisation equivalence (H27b $\rightarrow$ H27a). One moves against us. The joint-code interaction share of table 27 is 0.174 $[0.135, 0.216]$ on the original dictionaries and 0.513 $[0.335, 0.703]$ on the retrained ones — from below the pre-registered support threshold (0.20) to the refutation threshold (0.50). The single-side share is 0.97 and 1.02 respectively, so the direction of failure mode (1) is not in doubt; its *size* is a property of which dictionary was trained. We now state the joint share as a range.

The headline's bit-for-bit replication is not evidence, so we retrained it. The -0.461 and -0.293 estimates reproduce to every digit in a full rerun because the 15-dictionary ensemble behind them is loaded from checkpoints (its script skips existing ones, PREREG §22), as are the public language-model SAEs behind the GPT-2 and Pythia results. We therefore retrained all 15 dictionaries from scratch in fresh directories at 2 and at 8 BLAS threads (`repro.sh -retrain-dictionaries`; PREREG §74, written before the run, though the snapshot that contains §74–§76 was frozen after the results, so for these three registrations the ordering is self-attested) and re-ran both headline estimators unchanged on each ensemble (table 56). The retrained weights differ from the originals by 6% on average in relative Frobenius norm over the whole state dict (range 0–10.4%) and from each other by the same amount; the 39% above is a different object and norm (the encoder of the five canonical-pair crosscoders, relative to its own norm), and both say the same thing: these are different dictionaries. On them the two headline estimates move by at most 0.02, with every interval excluding zero and every point estimate inside the original interval (pre-registered H49a, both estimators). The §22 ensemble gate passes on all three ensembles ($15/15$ qualified, sign consistency 1.00). One incidental fact: the three ef$=4$ dictionaries (1536 features) retrain *bit-identically* at 2 threads while the twelve wider ones do not, so the unrecorded original thread count shared the 2-thread reduction order at that width and not above it; we still do not know what it was.

Table 56: H49: the two headline estimates on the original ensemble and on two ensembles retrained from scratch at different BLAS thread counts (PREREG §74); the thread count applies to dictionary training, both estimators then run at the pipeline's pinned two threads. Two-way cluster bootstrap over dictionaries and datasets; n is the number of unlearned directions in support (support-matched) or of learned directions with an admissible $\|Jd\|$ match.

estimator	ensemble	threads	point	95% CI	n
support-matched	original (never retrained)	unrecorded	-0.461	$[-0.581, -0.340]$	11
	retrained	2	-0.446	$[-0.569, -0.317]$	10
	retrained	8	-0.443	$[-0.570, -0.314]$	11
$\|Jd\|$-matched	original (never retrained)	unrecorded	-0.293	$[-0.497, -0.108]$	7/15
	retrained	2	-0.285	$[-0.483, -0.103]$	8/15
	retrained	8	-0.307	$[-0.512, -0.104]$	8/15

Why this belongs in the paper. Every failure mode above is a way the measurement decides the answer. This is the layer beneath them: before any measurement choice is made, the object being measured already depends on the hardware the dictionary was trained on. We do not promote it to a seventh failure mode because we have characterised it on one model and one dictionary type; we record it because a reader who reruns our pipeline will find it, and should find it here first.

X.36 THE THINKING-BLOCK CASE STUDY IN FULL

The diff was undertaken to explain v2.5's 64 prepended "thinking" tokens, and its disagreement with the feature-level result is what sent us looking for the failures above. The question is answerable at the attention level, without the dictionary. Interposing an explicit-attention path on TabPFN's fused kernel (matched to 10^{-4}): the block commands attention mass 0.588, more than the entire 128-row training context (0.412), and masking access costs $+159\%$ $[+96, +227]$ RMSE against $+23\%$ for size-matched random train keys. No mass-matched control is constructible, and per unit of mass thinking keys cost *less* than train keys ($+2.09$ vs $+2.60$), so the $+159\%$ reflects how much attention the block holds. Its content is largely inert — a foreign dataset's slice costs $+0.31\%$ $[+0.10, +0.63]$ — while matched-scale noise costs $+62.7\%$. That is the attention-sink signature (Xiao et al., 2024; Darcet et al., 2024; Gu et al., 2025; Su et al., 2026), for which a causal score/value dissociation already exists in diffusion transformers (Wu & Summa, 2026) and replacement-control evidence that zero-ablation overstates content dependence exists for vision registers (Parodi et al., 2026); our increment is the modality and the mass normalisation. **It also fails the sink account's redundancy prediction**: eight of 64 keys leave 0.366 of the knockout damage against a pre-committed ≤ 0.20 (appendix X.5), which we cannot reconcile with content-inertness. Full arms in table 25.

X.37 THE SAME COMPARISON, REVERSED, ON A LANGUAGE MODEL

appendix X.14 compared individual SAE atoms and found them slightly below matched controls. That is not the regime section 5.4 is about: atoms reach variance share 0.23 at most, while the tabular result sits at 0.22 against controls that touch the geometry's ceiling. The comparison that corresponds to it uses feature *groups*, and GPT-2 permits it where our tabular setting does not — its variance-share ceiling is 0.929 (L8) and 0.955 (L6) against 0.243, so a control at 0.3–0.7 is a population rather than a cone.

Group directions are built from *windows* of the variance-contribution ranking rather than top-m prefixes: the top-m group mean sits at 0.886–0.929 for every width, so a prefix sweep cannot span the range, while sliding the window down the ranking does. Controls are cone draws tilted to a target variance share, and each records its PC1 mass; bands whose controls exceed 0.8 PC1 mass are reported as degenerate and excluded from the estimate — the check whose absence produced appendix X.18.

X.37.1 THE FIRST VERSION OF THIS COMPARISON HAD THE DEFECT IT WAS WRITTEN TO CATCH

Our first pass drew controls at targets from 0.05 upward. The non-degenerate learned windows on L8 occupy 0.0004–0.0452. The two arms therefore had *no overlapping variance support*, nearest-five matching returned the same five lowest cones for every learned direction, and all 15 L8 gaps were computed against one constant (0.996491); the interval was the learned arm's spread with the control arm's uncertainty absent. That is the error this paper withdrew its own headline over (appendix X.18), pointing the other way, and it was the basis for the title. Three reviewers found it independently.

We re-ran it under three pre-registered changes (PREREG §53): targets extended down to 5×10^{-4} so the control arm covers the learned support; a caliper of 0.25 dex, with learned directions that have no admissible match excluded and *counted*; and a two-way cluster bootstrap over learned directions and control cones.

Table 57: GPT-2 feature groups against matched non-degenerate controls, before and after common-support matching. The reversal survives: both layers keep a *positive* gap, now against 13 and 16 distinct matched means rather than one.

layer	ceiling	variance support		first pass	**common support**	no match
		learned	control	(disjoint)	(caliper 0.25 dex)	
8	0.929	0.0004–0.929	0.0005–0.70	$+0.492$ $[+0.345, +0.637]$	$\mathbf{+0.418}$ $[+0.238, +0.603]$	13 of 28
6	0.955	0.0003–0.955	0.0005–0.70	$+0.348$ $[+0.270, +0.422]$	$\mathbf{+0.270}$ $[+0.157, +0.372]$	14 of 37

Every one of the 15 admissible L8 directions has a positive gap individually, as do 20 of 23 on L6; the largest match distance is 0.225 dex, inside the caliper; restricting further to the strict intersection

of the two supports gives $+0.430$ and $+0.277$. The estimate moved by -15% (L8) and -22% (L6) and did not change sign.

Our own prediction of the size of this error was wrong, and we record that. Before re-running we applied the fitted $\beta_{\log_{10} v} = -0.215$/decade from appendix X.41 to the 1.72-decade separation between the arms and predicted that mismatch alone accounted for $+0.370$ of the $+0.492$. The measured contribution was -0.074: one fifth of the prediction, with the wrong sign. The coefficient was fitted on the learned arm's window sweep and does not describe cone controls, which approach isotropy as their variance share falls. Extrapolating a fitted coefficient into another population is not a measurement — which is the same lesson as the rest of this paper, arriving at our expense in the opposite direction from the one we expected.

What the caliper does *not* fix: 13 of 28 L8 windows and 14 of 37 on L6 have no admissible match, all of them above the control arm's reach of 0.70. For the highest-variance SAE groups the comparison remains undefined in GPT-2 exactly as it is undefined in seven of ten tabular cells (table 3).

So the sign of "learned versus random" is set by two choices that are usually left implicit. *Architecture*: a tabular crosscoder gives -0.461, GPT-2 gives $+0.42$ and $+0.27$. *Aggregation*: within GPT-2 L8, atoms give -0.080 and groups $+0.492$. We had already seen the second and misread it — appendix X.14 switched from groups to atoms because the group control degenerated, and we recorded the switch as a design fix. It was a design fix, and the two arms it separated disagree.

We withdraw the general form of the claim (table 10, last row). The tabular result stands as measured; it is a fact about that dictionary on that model, and this appendix is why the paper no longer states it as a fact about reconstruction training.

X.38 Is the tabular deficit just an under-trained dictionary?

The obvious alternative to everything above: our crosscoder sees 1,600 paired positions while the GPT-2 SAE saw 3×10^8 tokens, so the tabular deficit could be a symptom of under-training rather than of anything about reconstruction. We varied the training budget directly — $\{1/8, 1/4, 1/2, 1\}$ of the available positions at three seeds — and measured each of the twelve dictionaries with the estimator of appendix X.18.

Table 58: Training budget against the matched gap, 5 datasets at evaluation seed 42. Validation R^2 moves from 0.76 to 0.997 and the deficit does not follow it.

budget	positions	mean val R^2	mean dead fraction	mean gap
$1/8$	200	0.759	0.447	-0.365
$1/4$	400	0.915	0.438	-0.455
$1/2$	800	0.964	0.403	-0.362
1	1600	0.997	0.360	-0.315

Spearman between quality and |gap| is -0.098 using validation R^2 and -0.084 using the live-feature fraction, and the gap at the full budget is -0.315 $[-0.492, -0.149]$. The relation is flat: an eightfold increase in data does not shrink the deficit, and it survives at the highest quality we can reach. Our pre-registered criterion for dismissing the alternative asked for a non-negative correlation and we observed a slightly negative one, so we record this as *not excluded* rather than refuted. What can be said is that the alternative predicts a strong negative relation and the data show none.

One thing this run corrects. Dictionaries trained here run 36–45% dead, not the 82–95% quoted elsewhere in this paper from the earlier ensemble; that figure belongs to a different configuration and should be read as such.

X.39 The withdrawn ordering in full

Applying all three corrections — joint encoding and a dose sweep confined below the interpretability boundary — overturns the feature-level result earlier versions of this paper reported as a headline. At the largest admissible dose ($\alpha^* = 0.3$) the ratio of shared-group to size-matched-random ablation effect is 1.05 on v2 and 0.36 on v2.5, against published values of 3.13 and 1.25. We report these as descriptive and attach no p-values: a positive control at the same dose — the fallback direction,

known to be causal — exceeds the other arms by only $1.64\times$, short of the $3\times$ we pre-committed to as evidence that the dose resolves anything (appendix X.4). The ratio is stable over three orders of magnitude of dose and inverts only above the boundary. The superseded published table is retained as table 21.

The ratio is $0.95\,|\,0.31$ at α=0.01 and $\mathbf{1.05}\,|\,\mathbf{0.36}$ at α^*=0.3, then $25.2\,|\,12.1$ at α=1, which is inadmissible ($5.4\times$/$2.9\times$ the boundary, as a ratio of means) and is where the published single-side operating point sits. That the favourable ordering appears *only* past the boundary and on *both* sides identifies it as an artifact of the operating point, not a property of the shared subspace. Two controls sharpen this: the comparison was never norm-matched ($\|u_{\text{shared}}\|/\|u_{\text{random}}\| = 1.46$), and equalising norms moves the ratio from 0.358 to $\mathbf{0.191}$ — the instance of section 5.4 at this pair; and the partition is not at fault (single-dataset classification reproduces the published shared set at Jaccard 0.774, unseen datasets 0.705, appendix X.3). The pre-registered threshold $r_{\text{shared}} \geq 2.5$ is met on neither side, so we withdraw the claim that shared features carry privileged per-feature causal weight, with the "de-concentration" magnitudes built on it.

What survives is the *cross-side* comparison, and only in direction: v2's ratio (1.05) exceeds v2.5's (0.36) by $2.9\times$ at every admissible dose. Stated honestly that is a stranger claim than the published one — v2's shared subspace behaves like a random subspace, v2.5's *worse* than one — and we report it as an observation, not an explanation.

X.40 The worked example across the rest of the model

section 5.4 measures one layer pair and one side. Since appendix X.37 shows the sign moves across architectures, the within-model question is the more urgent one, and we pre-committed to an obligation before running it: if the signs came out mixed, the abstract loses the effect size and the main text is narrowed to a single cell. They did, and it has.

Table 59: The support-matched contrast at all five layer pairs and both sides, each with its own activation geometry and its own cone controls. A control whose PC1 mass exceeds 0.8 is not a population and is excluded from the pooled estimate.

pair	side	var. share	ceiling	sensitivity	matched gap	control PC1 mass
L0↔L4	v2	0.332	0.379	0.256	−1.509	0.874 *(deg.)*
L0↔L4	v2.5	0.327	0.369	0.332	−0.080	0.885 *(deg.)*
L10↔L17	v2	0.391	0.433	0.914	−0.056	0.902 *(deg.)*
L10↔L17	v2.5	0.003	0.388	0.213	**−0.791**	0.001
L2↔L1	v2	0.296	0.333	0.401	−0.460	0.891 *(deg.)*
L2↔L1	v2.5	0.215	0.243	0.178	−0.514	0.885 *(deg.)*
L7↔L11	v2	0.520	0.524	1.196	+0.113	0.992 *(deg.)*
L7↔L11	v2.5	0.095	0.316	1.612	**+0.711**	0.294
L9↔L16	v2	0.332	0.338	1.284	+0.091	0.981 *(deg.)*
L9↔L16	v2.5	0.019	0.233	1.700	**+0.776**	0.068

Two things come out of this table, and the second is the one we did not expect.

The signs are mixed. One of the three non-degenerate cells is negative; the pooled gap is $+0.232$ $[-0.791, +0.776]$. With $n = 3$ we do not claim the effect reverses within the model — we claim only that a model-level statement is not available, which is enough to remove it from the abstract.

Seven of ten cells admit no non-degenerate control. The geometric limit of appendix X.18 is not a quirk of the cell we happened to report; it is the ordinary situation. And the cell section 5.4 reports is one of the seven — its control sits at PC1 mass 0.885; the 0.889 quoted in section 5.4 is the mean over the fifteen-dictionary ensemble's in-support draws, not this single cell's scalar. Where a matched-variance control can actually be built here, the deficit is not what we found. We report the worked example anyway, because a result that looks this clean and dissolves this thoroughly is the paper's argument rather than an embarrassment to it, but a reader should weight it accordingly.

X.41 AGGREGATION AS A CAUSAL VARIABLE

appendix X.37 left the sharpest observation confounded. Individual atoms sat below matched controls and feature groups above, but the two arms also occupied different variance-share regimes — atoms reach 0.23, groups start at 0.89 — so bundling and variance were never separated.

Sweeping both the width w of a window over the variance-contribution ranking and the window's start reaches the same variance share with different w, which separates them. We fit

$$\text{sensitivity} \;\sim\; \beta_0 + \beta_v \log_{10}(\text{variance share}) + \beta_w \log_2 w$$

over 72 directions per layer spanning variance shares 2×10^{-4} to 0.955, so β_w is the effect of bundling with variance held fixed. $w = 1$ is a single decoder atom.

Table 60: Bundling raises sensitivity at matched variance — on L8 under clustered inference, and on both layers only under the homoskedastic OLS this table reports, which appendix X.21 supersedes. The per-width means are monotone: at L8, 1.102 at $w = 1$ rising to 1.394 at $w = 256$.

layer	n	$\beta_{\log_2 w}$	$\beta_{\log_{10} v}$	single vs. grouped, matched
8	72	$+\mathbf{0.0317}$ $[+0.0137, +0.0497]$	-0.215	$+0.195$ $[+0.081, +0.304]$
6	72	$+\mathbf{0.0182}$ $[+0.0008, +0.0355]$	-0.083	-0.081 $[-0.222, +0.049]$

Both layers give a positive bundling coefficient whose interval excludes zero. The secondary comparison — single atoms against $w \geq 16$ groups, matched pairwise on variance share — agrees at L8 and does not at L6, on $n = 7$ and $n = 8$ pairs; we report both rather than the one that matches the primary statistic.

The consequence is the reason this appendix exists. A dictionary does not have *a* causal potency relative to random. It has one per aggregation choice, and the choice is usually made silently — by whether a paper ablates a feature or a group, a head or a layer. Our own two arms disagreed for exactly this reason, and we recorded the switch between them as a design fix (appendix X.14) rather than as the finding it turned out to contain.

One further coefficient is worth recording without interpretation: β_v is negative on both layers, so on GPT-2 higher-variance directions are *less* sensitive. That matches the sign we saw among learned directions in the tabular setting ($\rho = -0.659$) and opposes the sign among unlearned ones ($+0.305$). Reconciling those is beyond what we can support here.

X.42 THE CASE THAT MOTIVATED THE AUDIT: V2.5'S THINKING BLOCK

The diff was undertaken to explain v2.5's 64 prepended "thinking" tokens, and the disagreement between that question's answer and the feature-level one is what sent us looking for the failures above. It is answerable at the attention level, without the dictionary: the block commands attention mass 0.588, more than the entire training context, and masking access costs $+159\%$ $[+96, +227]$ RMSE against $+23\%$ for size-matched random train keys, while its *content* is largely inert — the attention-sink signature (Xiao et al., 2024; Darcet et al., 2024; Gu et al., 2025; Su et al., 2026), for which causal dissociations already exist (Wu & Summa, 2026; Parodi et al., 2026). Our increment is the modality and the mass normalisation, and it fails the account's redundancy prediction (appendix X.36).

X.43 NUMBERS THAT NO SCRIPT PRODUCED

Late in preparation we ran an audit we had not run before: an exhaustive scan of every leaf value in all 414 files under `results/` — roughly 233,000 numbers — against every number printed in the paper. Five printed values matched nothing, and a second pass that traced each headline number to a specific *key* rather than to any stored value found two more. The artifact–claim gate did not catch them because it compares each manifest entry's stored value against its result file and never opens the LaTeX sources, so a number that was never in the manifest was never checked at all. This is the sixth failure mode's cousin: a verification apparatus that answers a narrower question than the sentence it is quoted to support. We report all seven, what they should have been, and what changed.

The one that mattered. table 54 contrasts two ways of naming "the group" on one set of dictionaries. Its window arm comes from the factorial's crosscoder cell, i.e. the three ef$=16$, k=48 dictionaries;

Table 61: Every printed number that no committed script produced, and its disposition (PREREG §77 and §77.2–§77.3, each registered before the number was recomputed).

printed	where	what it should be	disposition
-0.536 $[-0.611, -0.461]$	§5.3, table 54	-0.544 $[-0.649, -0.447]$	recomputed; the claim stands, the interval was wrong
$86\times$	contribution 5	$68\times$ (ratio of means)	corrected, aggregation now named
18% encoder weight	§9, appendix X.35	26–49%, mean 39% (measured here)	measured
36–45% dead features	§6.2	65.3–93.1%	corrected twice; the first correction was also wrong
57 verdict fields, 54 unchanged	§9	56, 53 unchanged	corrected
found later, by tracing each number to a key rather than to a value (§77.2)			
$2.9\times$ effect vs the bound	table 27	$2.6\times$	corrected
$3.2\times$ the boundary (v2)	appendix C	$5.4\times$ (ratio of means)	corrected
found later still, by an audit of code against prose rather than of numbers (§80)			
shared = both sides active	appendix K	the complement of exclusive among live features; 11.0% of the group is active on neither side	rule restated, headline remeasured (appendix X.50)
$69\times$ over 15 instances	§1, §5	the geometric mean is over 14; one instance has a sign-reversed denominator	disclosed, value unchanged
$+0.099$, $+0.101$ refit	table 36	not identified: $\log_2 w$ is constant on the disjoint subset	withdrawn

its shared/exclusive arm should therefore be those same three. Recomputing it that way, through the same two-way cluster bootstrap the headline uses and against the same in-support control, gives -0.544 $[-0.649, -0.447]$. The printed interval's upper bound, -0.461, is the full fifteen-dictionary headline's *point* estimate, which is how the number appears to have been assembled. The contrast survives — resolved negative against unresolved, a shift of 0.62 from renaming alone — but it survives on a recomputed number, not the one we printed.

What we changed so that this cannot recur, and what it still cannot do. `tools/validate_results.py` now parses `paper/sections/*.tex` and fails when a manifest-bound value does not appear where its `paper_loc` says it does. That closes the direction the five above escaped through. It does not close the other one: a printed number can match *some* stored value by coincidence and pass. Tracing the 29 main-text headline numbers to a specific file *and key* — which only a person can do — found two more, a $2.9\times$ that should be $2.6\times$ in table 27's own column and a $3.2\times$ that was one dataset's value printed as the aggregate (§77.2). Value matching is not key matching, and we say so rather than let the gate's PASS stand for more than it checks.

X.44 Every iid interval in this repository, recomputed as a clustered one

A review of our code found the same defect in five scripts: rows carry a `dataset` key and are then passed to an iid bootstrap, so three seeds per dataset are resampled as fifteen independent observations. appendix X.21 already fixed this for the aggregation coefficient, where it mattered. This appendix asks whether it mattered anywhere else, by recomputing nine quantities both ways from the records those scripts committed.

Nothing changes, and four intervals get *narrower*. We had pre-registered narrowing as a bug to be investigated (PREREG §84), so we investigated it. It is not a bug. A cluster bootstrap exceeds an iid one only when rows inside a cluster are positively correlated; the median intraclass correlation here is $+0.122$, with a range of -0.272 to $+0.438$, and the two quantities with the highest ICC are exactly the two that widen most ($1.31\times$ and $1.33\times$). The mechanism is visible in the data rather than asserted.

Table 62: The same statistic under an iid and a cluster bootstrap. ICC is the one-way intraclass correlation; a cluster bootstrap is wider than an iid one only where it is positive. No verdict changes.

quantity	n/clusters	iid	clustered	width	ICC
P1 shared / random	$15/5$	$[3.121, 4.361]$	$[3.182, 4.375]$	$0.96\times$	$+0.063$
P3 v2.5-exclusive / random	$15/5$	$[0.297, 0.413]$	$[0.297, 0.417]$	$1.04\times$	$+0.141$
P1 v2-exclusive / random	$15/5$	$[1.023, 1.366]$	$[1.073, 1.328]$	$0.74\times$	-0.138
P3 shared, thinking only	$15/5$	$[0.621, 1.072]$	$[0.528, 1.120]$	$\mathbf{1.31}\times$	$\mathbf{+0.438}$
P2 shared gap closure	$15/5$	$[-0.008, 0.629]$	$[-0.007, 0.576]$	$0.92\times$	-0.061
P2 random gap closure	$15/5$	$[-0.046, 0.099]$	$[-0.052, 0.103]$	$1.07\times$	$+0.122$
Phase 3b shared / random	$30/10$	$[3.203, 4.224]$	$[3.029, 4.383]$	$\mathbf{1.33}\times$	$\mathbf{+0.424}$
Phase 3b v2.5-excl / random	$30/10$	$[0.309, 0.387]$	$[0.305, 0.392]$	$1.13\times$	$+0.171$
Phase 3b shared gap closure	$16/8$	$[-0.131, 0.031]$	$[-0.114, 0.024]$	$0.86\times$	-0.272

Three seeds inside a dataset therefore carry nearly as much information as three independent draws, which we did not expect: we predicted in the registration that the intervals would widen substantially and that at least one verdict would move, and all of that was wrong. Running a single ensemble member means a seed changes only the data split, and that turns out to move these quantities about as much as changing the dataset does.

We are not claiming the code was right. Resampling rows that carry a cluster key as though they were independent is wrong, and on other data it would cost something. What we can now say, with the numbers rather than by assertion, is that it cost this paper nothing. The five source scripts are left as they are, because re-running them would change artifacts table 10 describes; the audit runs from their committed records and is a numbered stage of the pipeline.

X.45 The six modes on a dictionary we did not train

Every mode in this paper is demonstrated on our own measurement, which makes the title's scope a fair question. This appendix answers it on a public artifact: GPT-2 small with `jbloom/GPT2-Small-SAEs-Reformatted`, wikitext-2, KL. We cannot re-run another group's numbers without their code. What we can do is instantiate a common protocol — ablate an SAE feature group, compare against random features, conclude the group is causally important — on their dictionary, and vary its choices. E1 is absent because $\mathrm{enc}(a, \mathbf{0})$ is undefined for a single-model SAE; appendix X.51 covers it on a second crosscoder instead.

Why this is a 2×2 and not a ladder. An earlier version turned corrections on one rung at a time and reported that the answer moved $4.8\times$, "almost all of it the control". Two reviewers pointed out that the rungs were confounded: the step credited to the matched control also changed the dose and the intervention's form, and the step credited to the fixed dose changed only how controls were averaged, since the dose was already fixed. Neither attribution was identified, the window starts overlapped and were resampled as if independent, and the 4.8 was the L6 maximum over rungs rather than a ratio we had defined in advance. All of it is withdrawn (table 10). The replacement (PREREG §97) holds the learned direction, the patching and the read-out fixed in every cell and moves two factors: the *dose* — each group's natural residual-preserving norm (N) or the fixed reference norm (F) — and the *control* — a uniformly random 64-feature group (R) or a caliper-matched variance cone (M), each scaled to its paired learned direction's norm. The learned groups are 24 *disjoint* windows of 64 features, and the bootstrap resamples groups, controls and corpus shards together.

What this shows (fig. 9). Both factors matter and they are not additive. At the natural norm the learned-minus-control gap is indistinguishable from zero under either control; at the reference norm it is positive under both and larger against the matched cone. So the dose main effect ($+0.45$ at L8, $+0.28$ at L6) exceeds the control main effect ($+0.12$, $+0.19$), and most of the control's effect exists only at the large dose (interaction $+0.23$, $+0.36$). This reverses the reading we registered, which predicted a control effect and no dose effect (PREREG §97, 55%), and it is the opposite of the earlier ladder's claim that the fixed dose was inert. It also has a plain cause: the reference norm is 2–102 times each group's natural norm (median 51 at L8, 69 at L6), and on the dose scale of appendix X.49 it sits past the half-effect norm, where every one of the 48 learned directions' own responses saturates. The positive gap on a public SAE is a statement about broadcast perturbations far larger than the ones its features produce (appendix X.48 gives the per-token version, where the control main effect

Table 63: Dose × control on a public GPT-2 SAE (PREREG §97): learned minus control in KL, 24 disjoint width-64 groups per layer, 5 controls per group and arm, 8 corpus shards, all three resampled. Every cell uses the same learned direction, residual-preserving patching and the $\pm d$ average; each control is scaled to its paired learned direction's norm. The reference norm is $1.34\rho^*$ at L8 and $1.70\rho^*$ at L6. Generated by `exp_fig3_ladder.py`.

cell	L8	L6
natural norm, random features	-0.006 $[-0.022, +0.001]$	-0.011 $[-0.039, +0.001]$
natural norm, matched cone	$+0.000$ $[-0.001, +0.001]$	-0.001 $[-0.004, +0.001]$
reference norm, random features	$+0.328$ $[+0.193, +0.458]$	$+0.089$ $[-0.176, +0.367]$
reference norm, matched cone	$+0.561$ $[+0.453, +0.671]$	$+0.463$ $[+0.265, +0.685]$
dose main effect (F − N)	$+0.448$ $[+0.337, +0.561]$	$+0.282$ $[+0.067, +0.521]$
control main effect (M − R)	$+0.119$ $[+0.085, +0.158]$	$+0.192$ $[+0.111, +0.279]$
interaction $(FM - FR) - (NM - NR)$	$+0.227$ $[+0.166, +0.288]$	$+0.362$ $[+0.215, +0.514]$

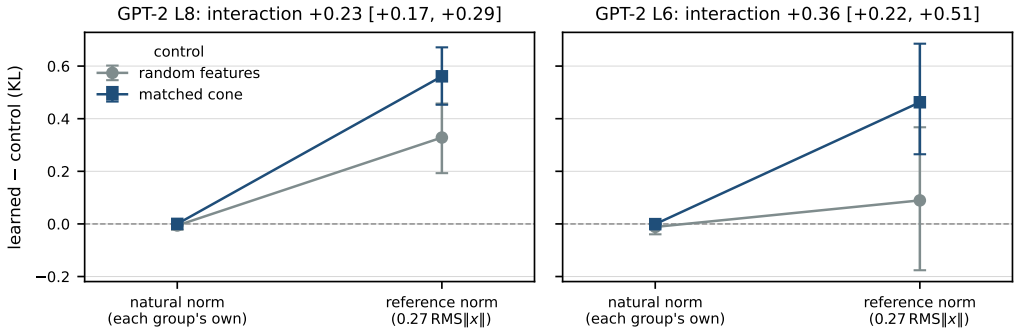


Figure 9: **Table 63, drawn.** At each group's natural norm neither control separates from the learned group; at the reference norm both do, and the matched cone more. The lines fan out rather than run parallel: that is the interaction. Intervals are bootstraps over groups, controls and corpus shards.

dominates instead), and `audit` says so: run on the reference-norm, matched-cone cell of either layer it returns INADMISSIBLE, because the response is not linear around that dose (super-linear at L8, saturating at L6, judged on both ratios touching the reference; PREREG §100). Two fixes made in this revision changed how it reaches that answer: its dose check had swept a fixed range of absolute norms, $1/78$ of the reference, instead of the dose supplied, and admitted the cell (PREREG §97.1); and it had judged linearity on the top ratio of its ladder alone, which a response with a bump at the reference passes (PREREG §100).

What the group definition does here, and it is not what we first reported. A separate sweep of group width at the reference norm against matched cones (PREREG §87, 24 window starts) resolves every width positive at L8, rising monotonically with width over $1, 16, 64, 256$ from $+0.28$ to $+0.72$; L6 follows the same curve from an unresolved $+0.17$ at width 1. It is measured at the same unvalidated dose as the reference-norm row of table 63, so it is a statement about that dose. This is the tabular aggregation finding — bundling makes a direction more sensitive at fixed variance share (appendix X.21) — replicating on a dictionary we did not train, with a resolved slope. An earlier version of this appendix, run at six groups, reported that the width sweep *reversed* the point estimate on both layers, and quoted a $16.5\times$ span built on that reversal. Both are withdrawn (PREREG §87.1). The six starts happened to be low-sensitivity ones, and we read the sign of an interval that covered zero, which is the thing this paper tells other people not to do.

What it does not show, at the strength we registered. No cell resolves negative, so the pre-registered condition for a sign reversal on a public dictionary (§83, §87) is not met, and we predicted twice that it would be. The paper's sign-level claim therefore rests on our tabular cell and on the atoms-versus-groups contrast of appendix X.37.1. What a public dictionary supports independently is the claim at the level of *magnitude*, plus one specific mechanism, aggregation, at the level of sign.

And one obstruction transfers by its absence. GPT-2's variance-share ceiling is 0.929 at L8 and 0.955 at L6, against 0.243 in the tabular setting. C2 is a statement about the spectrum of the activation

covariance. GPT-2's residual stream is *more* concentrated, not less — one outlier direction carries over 90% of the variance — but that raises the ceiling far above every learned group's variance share, so the geometric obstruction that makes two-thirds of our registered band unreachable does not bind here: at 100 cones, 0 of 24 learned groups lacked a caliper match. That is the mechanical reason C2 bites hardest on tabular activations, and the ladder confirms it rather than assuming it.

X.46 The most diverse matched control: proof and every cell

Proposition 1. Let $a_1 \geq \cdots \geq a_d$ be the eigenvalues of $A = \Sigma / \operatorname{tr}\Sigma$ with orthonormal eigenvectors v_i, and for a unit q let $w_i = (q^\top v_i)^2$, $t = q^\top A q = \sum_i a_i w_i$ and $\mathrm{PR}(q) = 1/\sum_i w_i^2$. Then $\mathrm{PR}(q) \leq \mathrm{PR}_{\max}(t) := 1/\min\{\sum_i w_i^2 : w \geq 0,\ \sum_i w_i = 1,\ \sum_i a_i w_i = t\}$, with equality for $q^* = \sum_i \sqrt{w_i^*}\, v_i$; and every such q has $w_1 \geq (t - a_2)/(a_1 - a_2)$ when $a_1 > a_2$.

Proof. As q ranges over the unit sphere, w ranges over exactly the probability simplex: any w in the simplex is realised by $q = \sum_i \pm\sqrt{w_i}\, v_i$, and $\sum_i w_i = \|q\|^2 = 1$. The constraint $t = \sum_i a_i w_i$ is linear, so the minimum of the strictly convex $\sum_i w_i^2$ over the resulting polytope is attained, unique, and realised by q^*. For the second claim, $t = a_1 w_1 + \sum_{i \geq 2} a_i w_i \leq a_1 w_1 + a_2(1 - w_1)$. □

The optimum has the form $w_i = \max(0, \lambda + \mu a_i)$ (KKT), which `tabdiff.diagnostics.max_participation_ratio` finds by nested bisection; a test checks that the constructed direction attains the target share and the optimum, and that no random direction at the same share exceeds it. The inequality is elementary; what it adds is a decision made *before* sampling: whether the comparison at the learned arm's share can have a population control at all.

Table 64: The most diverse control any generator could build at the learned arm's variance share t (PREREG §102). *Recorded*: what the cone generator produced; *closed form*: the bound implied by $w_1 \geq (t - a_2)/(a_1 - a_2)$. Generated by `exp_core_tables.py`.

cell	t	a_1	$\mathrm{PR}_{\max}(t)$	closed form	recorded	classification
L0↔L4 v2	0.332	0.379	2.21	3.70	degenerate	necessarily degenerate
L0↔L4 v25	0.327	0.369	1.40	1.45	degenerate	necessarily degenerate
L10↔L17 v2	0.391	0.433	1.29	1.31	degenerate	necessarily degenerate
L10↔L17 v25	0.003	0.388	190.72	192.00	non-degenerate	non-degenerate, as recorded
L2↔L1 v2	0.296	0.333	2.38	3.23	degenerate	necessarily degenerate
L2↔L1 v25	0.215	0.243	1.76	2.11	degenerate	necessarily degenerate
L7↔L11 v2	0.520	0.524	1.02	1.02	degenerate	necessarily degenerate
L7↔L11 v25	0.095	0.316	14.70	192.00	non-degenerate	non-degenerate, as recorded
L9↔L16 v2	0.332	0.338	1.05	1.05	degenerate	necessarily degenerate
L9↔L16 v25	0.019	0.233	135.13	192.00	non-degenerate	non-degenerate, as recorded
GPT-2 L8 (24 groups)	0.0005–0.91	—	1.0–768	—	—	23 of 24 admit PR ≥ 8
GPT-2 L6 (24 groups)	0.0003–0.94	—	1.0–768	—	—	23 of 24 admit PR ≥ 8

What it decides. All seven tabular cells whose cone control we had recorded as degenerate are *necessarily* degenerate — $\mathrm{PR}_{\max}(t)$ between 1.02 and 2.38, far under the threshold of 8 — so the failure belongs to the geometry, not to our generator, and no amount of engineering would have produced a population control there (PREREG §102; predicted for at least five). The three non-degenerate cells have learned arms at small shares, where a PR of 15–190 is available. On GPT-2, 23 of 24 groups per layer admit one; the exception is the group lying on the dominant direction.

X.47 Dose and control as a curve

table 63 compares two doses. Here the same groups, controls and read-out are swept over multiples of each group's natural RP norm, with the clean log-probabilities kept in FP32 after a floor check showed the FP16 storage of table 63 to be four orders of magnitude below the smallest effect it measured (PREREG §101, §103).

Three readings. First, learned minus matched cone is zero to within ± 0.005 at and below the natural norm, and resolves only at the reference norm, on both layers; learned minus random features is unresolved at every multiple up to $16\times$ and resolves at the reference on L8 only. Second, Γ, the

Table 65: Dose as a multiple of each group's natural RP norm, and the fixed reference norm, on GPT-2 with the 2×2's groups and controls (PREREG §103). Brackets are simultaneous 95% bands over the dose grid. KL no-op floor $2.4e-07$ with FP16 clean storage and 0 with FP32, against a median learned effect at the natural norm of $6.2e-04$ (L8). Injecting each group's own per-token offsets instead of their broadcast mean changes the natural-norm effect by a median factor of 8.3 (L8) and 13.1 (L6); coherence 0.21, 0.22. Generated by `exp_core_tables.py`.

layer	dose	Δ_R	Δ_M	$\Gamma = \Delta_M - \Delta_R$
L8	0.25×	-0.0004 $[-0.001, +0.001]$	-0.0000 $[-0.000, +0.000]$	$+0.0004$ $[-0.000, +0.001]$
L8	0.5×	-0.0015 $[-0.005, +0.002]$	-0.0000 $[-0.001, +0.000]$	$+0.0015$ $[-0.002, +0.005]$
L8	1×	-0.0060 $[-0.021, +0.009]$	$+0.0001$ $[-0.002, +0.002]$	$+0.0061$ $[-0.008, +0.021]$
L8	2×	-0.0187 $[-0.068, +0.030]$	$+0.0033$ $[-0.001, +0.008]$	$+0.0220$ $[-0.030, +0.074]$
L8	4×	-0.0449 $[-0.170, +0.080]$	$+0.0124$ $[-0.001, +0.025]$	$+0.0574$ $[-0.075, +0.189]$
L8	16×	-0.0824 $[-0.493, +0.329]$	$+0.1043$ $[-0.009, +0.218]$	$+0.1867$ $[-0.177, +0.551]$
L8	ref.	$+0.3284$ $[+0.172, +0.484]$	$+0.5611$ $[+0.417, +0.705]$	$+0.2327$ $[+0.152, +0.314]$
L6	0.25×	-0.0006 $[-0.002, +0.001]$	-0.0001 $[-0.000, +0.000]$	$+0.0005$ $[-0.001, +0.002]$
L6	0.5×	-0.0024 $[-0.008, +0.004]$	-0.0003 $[-0.001, +0.001]$	$+0.0021$ $[-0.003, +0.007]$
L6	1×	-0.0115 $[-0.040, +0.017]$	-0.0008 $[-0.004, +0.003]$	$+0.0107$ $[-0.016, +0.037]$
L6	2×	-0.0379 $[-0.135, +0.060]$	$+0.0040$ $[-0.005, +0.012]$	$+0.0418$ $[-0.061, +0.144]$
L6	4×	-0.0793 $[-0.285, +0.127]$	$+0.0096$ $[-0.004, +0.023]$	$+0.0889$ $[-0.125, +0.303]$
L6	16×	-0.1693 $[-0.705, +0.366]$	$+0.0523$ $[-0.034, +0.138]$	$+0.2216$ $[-0.272, +0.715]$
L6	ref.	$+0.0894$ $[-0.245, +0.424]$	$+0.4625$ $[+0.174, +0.751]$	$+0.3731$ $[+0.174, +0.572]$

difference between the two controls, contains no learned term, rises monotonically with dose on both layers and is resolved only at the reference: the choice of control matters only where the dose is large. We had registered that all three quantities would be equivalent to zero within ± 0.02 at small dose; that holds for Δ_M and fails for Δ_R and Γ, whose bands are wide because the random-feature pool is heterogeneous. Third, and not predicted: at the natural norm, injecting a group's own per-token offsets rather than their mean direction broadcast to every token changes the effect by a median factor of 8–13, because the offsets point in different directions at different tokens (coherence 0.21–0.22). A broadcast mean direction is therefore not the group's ablation, and every GPT-2 number in this paper — the reversal, the 2×2, this curve — is a statement about the norm-fixed sensitivity of a direction, not about the effect of removing the features.

X.48 The 2 × 2 as the group's actual removal

Because a broadcast mean direction is not the group's ablation (appendix X.47), we reran table 63 with every arm injected per token (PREREG §106): the learned arm is the group's own offsets u_i; the random-feature arm is a random group's own per-token offset directions; the cone arm is its fixed direction; and all three share the learned group's per-token norm profile, so only the direction at each token differs.

Table 66: The 2×2 of table 63 with every arm injected per token, as the group's actual removal (PREREG §106). All arms share the learned group's per-token norm profile; only the direction at each token differs. Learned minus control, KL. Generated by `exp_core_tables.py`.

cell	L8	L6
natural norm, random features	-0.054 $[-0.159, -0.005]$	-0.129 $[-0.393, -0.010]$
natural norm, matched cone	-0.002 $[-0.024, +0.007]$	-0.013 $[-0.067, +0.009]$
reference norm, random features	-0.214 $[-0.368, -0.083]$	-0.355 $[-0.657, -0.104]$
reference norm, matched cone	$+0.426$ $[+0.343, +0.510]$	$+0.270$ $[+0.094, +0.454]$
dose main effect	$+0.134$ $[+0.056, +0.216]$	$+0.029$ $[-0.133, +0.202]$
control main effect	$+0.346$ $[+0.275, +0.447]$	$+0.370$ $[+0.245, +0.579]$
interaction	$+0.587$ $[+0.509, +0.670]$	$+0.509$ $[+0.395, +0.627]$

The control now decides the sign. At the reference norm the learned groups are *less* disruptive than random feature groups (-0.214 at L8, -0.355 at L6) and *more* disruptive than matched cones ($+0.426$, $+0.270$), both resolved: the same intervention at the same dose gives opposite answers depending on which control is called "random". Against random feature groups the deficit is already

resolved at the natural norm (-0.054, -0.129) — the direction of our tabular result — while against the cone it is zero there under both injection forms. The control main effect ($+0.35$, $+0.37$) now exceeds the dose main effect ($+0.13$ at L8, unresolved at L6), the reverse of the broadcast 2×2. We had predicted a positive natural-norm gap against both controls; both predictions failed, one of them with the opposite sign.

X.49 One dose scale, and no dose in this paper passes it

Three quantities had shared one name. Earlier versions wrote every dose as a multiple of α^*, defined in the text as the norm fraction at which a label-independent fallback reaches half its full-scale effect. A review found that the numbers carrying the name were in three different coordinates: on TabPFN, 0.1011 was an *output* change (the cost of the full overwrite) and 0.3 a multiple of the shared group's own vector; on GPT-2 the "full-scale fallback" was normalised to $\alpha\|x_i\|$ and averaged over $\pm$, so $\alpha = 1$ was not an overwrite at all. The statements built on them — "$2.67\alpha^*$" for the tabular gaps, "$0.44\alpha^*$, admissible" for GPT-2, the bracket $[1.25, 3.92]\alpha^*$ — are withdrawn (table 10). Where earlier appendices (appendix X.30) still quote α^*, it is that older, model-specific coordinate, and no cross-model comparison should be read from it.

The definition (PREREG §96), identical on both models. A dose is the relative norm $\rho = \overline{\|\delta_i\|/\|x_i\|}$. It is calibrated by the mean overwrite $\delta_i(\lambda) = \lambda b_i$, $b_i = \bar{x} - x_i$ with $\bar{x}$ the context mean (train label tokens for TabPFN, the corpus for GPT-2), applied one-sided, so $\lambda = 1$ is exactly the overwrite. c^* is its effect at $\lambda = 1$, λ^* the fraction at which it reaches $c^*/2$, and $\rho^* = \lambda^* \overline{\|b_i\|/\|x_i\|}$ the half-effect norm. The estimators broadcast one direction to every position and scale the matrix to $0.27\|X\|_F$, so the reference is $\rho_{\text{ref}} = \overline{0.27\,\mathrm{RMS}_j\|x_j\|/\|x_i\|}$. The rule, fixed before the run: a dose is *validated* iff $\rho_{\text{ref}} \le \rho^*$ *and* the direction's own response is linear, at the metric's order, over 0.5–$2\rho_{\text{ref}}$ ($|\text{ratio} - 1| \le 0.25$). It is an operating rule, not a sufficient condition for a causal reading, and we use it that way. Failing it does not mean the intervention has no causal effect: $F(x+\delta) - F(x)$ is exactly defined at any dose. It means the number cannot be read as the direction's *locally proportional* sensitivity, the quantity a "this group matters k times more than random" sentence implicitly reports. The three layers are distinct — the intervention is implemented as intended (identity checks); the comparison is interpretable at this operating point (this rule, support, geometry); the intervention changes the concept it is said to change (a behavioural test we do not run) — and ADMISSIBLE certifies only the first two.

Table 67: One dose scale on both models (PREREG §96). c^*: the effect of an exact mean overwrite ($\lambda = 1$); λ^*: the overwrite fraction that reaches $c^*/2$; $\rho^* = \lambda^* \overline{\|b_i\|/\|x_i\|}$; ρ_{ref}: the reference norm as the estimators apply it. Each direction's own response over 0.5–$2\rho_{\text{ref}}$, as the ratio to the metric's order (1 is linear, tolerance 0.25): TabPFN 1.06 below and 1.29 above the reference (medians; 1 of 15 directions within tolerance); GPT-2 0.95 and 0.51 (0 of 48). Generated by `exp_dose_table.py`.

model, data	c^*	λ^*	ρ^*	ρ_{ref}	ρ_{ref}/ρ^*
TabPFN, `california_housing`	0.069	0.57	0.361	0.270	0.75
TabPFN, `diabetes`	0.032	0.64	0.403	0.270	0.67
TabPFN, `abalone`	0.045	0.53	0.354	0.270	0.76
TabPFN, `bike_sharing`	0.223	0.54	0.357	0.270	0.76
TabPFN, `concrete`	0.059	0.58	0.363	0.270	0.74
GPT-2 L8 (KL)	5.13	0.75	0.573	0.768	1.34
GPT-2 L6 (KL)	5.50	0.72	0.555	0.943	1.70

What it finds (table 67). Neither model passes. On TabPFN the reference norm is below the half-effect norm (0.67–$0.76\rho^*$), which we predicted it would not be; but around it the headline's own directions respond super-linearly, and 14 of 15 fall outside the tolerance on the upper step. On GPT-2 the reference is past the half-effect norm ($1.34\rho^*$ at L8, $1.70\rho^*$ at L6) because a few very-high-norm tokens dominate the RMS it is set from, and all 48 learned directions saturate around it. We had predicted GPT-2 would pass. Every *dose validated* entry in table 47 is therefore *no* or n/a. The practical lesson is the one appendix X.45 shows directly: a reference norm set from a global RMS can be tens of times the natural size of the perturbations being compared.

The positive control, in the headline's estimand. appendix X.30 reported a pre-registered positive control: rescale the label-independent fallback to the intervention's norm and require it to produce at least $3\times$ the largest other arm. It returned $1.64\times$. Because that was a raw $|\Delta\mathrm{pred}|$ on an α grid at one pair, not the headline's norm-fixed, sign-symmetrised, fifteen-dictionary estimand, we ran it again through that estimand with nothing else changed (PREREG §81).

Table 68: The four §34.1 directions put through the headline's estimand at the headline's norm. The pre-committed thresholds are carried over unchanged.

arm	norm-fixed output sensitivity	n
label-independent fallback (positive control)	$\mathbf{1.547}$	5 datasets
isotropic	0.896	10
random alive group	0.671	15
shared group (the treatment)	0.220	15

The fallback is clearly above every other arm — the two-way cluster differences are $+1.327$ $[+0.972, +1.658]$ against the shared group, $+0.876$ $[+0.476, +1.227]$ against the random one and $+0.651$ $[+0.116, +1.116]$ against isotropic — but not by enough: $\mathbf{1.726}\times$ against the largest other arm, against a pre-committed $3\times$ ($2.272\times$ against the headline's own control arm, and §81 fixed that the registered denominator decides). No dose rescues it: across the α grid the ratio is 1.97 at 0.03, 1.87 at 0.1, 1.64 at 0.3, and at 1.0 the shared group overtakes the fallback. §81 recorded, before the run, that this was the expected verdict.

What survives, as a description. The learned arm occupies 0.215–0.231 in variance share and 0.220 in sensitivity against the control's 0.681, with complete separation ($\max_{\text{learned}} = 0.446 < \min_{\text{control}} = 0.498$), Cohen's $d = 3.96$, and a negative gap in five of five dataset clusters and in six of six unseen ones (appendix X.24).

History of the p-value, kept so the change is visible. After the positive control failed, an earlier version applied PREREG §34.1 — "report descriptively and drop the p-value" — to the headline and removed its p and Holm rejection. A later review pointed out that withholding a p neither restores a metric's validity nor is required by its absence, and that we had withheld it for one contrast while keeping it for another at the same norm. Both points are right, and the current version separates the two questions: every test in table 47 carries its p, which tests whether the arms differ in norm-fixed sensitivity at the norm used, and the *dose validated* column — now set by the rule above — says whether a causal reading is licensed. It is not, for any dose-dependent row.

X.50 The feature taxonomy we described is not the one we ran

appendix K originally said that a feature active on neither single-sided input is dead. The classifier does not say that. It decides death from the *joint* activation frequency, and then sweeps every live feature that is not one-sided into `shared`. A feature that fires only when both inputs are present — below τ on $(a_{\text{v2}}, \mathbf{0})$ and below τ on $(\mathbf{0}, a_{\text{v2.5}})$, above τ on the pair — is therefore in the shared group, not in the dead set.

This is not a corner case. Counting over all fifteen dictionaries that carry the headline: 399 of 3617 shared features, $\mathbf{11.0}\%$, reach the group that way. The rate is 8.0–14.5% and no dictionary is free of it. The group is what -0.461, the $3.1\times$, table 54's -0.544 and the withdrawn P1 are all measured on, so a wording fix is not enough.

We re-ran the headline with the group rebuilt under the rule the appendix stated — dropping the 399, changing nothing else, not the control arm, not the estimator, not the seed. The registration (§80) fixed the decision rule first: a shift under 0.10 with an interval still excluding zero keeps the number and corrects the prose; anything larger replaces the number.

The shift is a fiftieth of the threshold, so the numbers stand and the definition is restated. We record the direction anyway because we predicted the other one: the registration argued that features firing only on the joint input are *more* squarely "shared", so removing them should sharpen the contrast. It softened it, by an amount too small to interpret. What the check actually establishes is narrower than either prediction — the deficit is a property of the live non-exclusive population and not of the 11% that entered it by fall-through.

Table 69: The headline under the rule we ran and under the rule we described. Only the learned arm's group definition changes.

estimator	as shipped (shared = complement)	strict (both sides active)	shift
support-matched	$\mathbf{-0.461}$ $[-0.581, -0.340]$	-0.459 $[-0.580, -0.339]$	$+0.002$
nearest-5 matched	-0.452 $[-0.505, -0.392]$	-0.452 $[-0.505, -0.392]$	0.000
covariate-adjusted	-0.438 $[-0.510, -0.367]$	-0.438 $[-0.510, -0.367]$	0.000
ratio	$3.09\times$	$3.07\times$	-0.02

Why this was not caught by a gate. Every gate we run compares a printed number to a stored number. This defect is a disagreement between a printed *sentence* and executed code, with no number in between, and nothing we have checks that. It was found by reading the classifier against the appendix. The same reading found the dead-feature range in §6.2: we printed 34.9–45.4%, which is the range in `results/exp_dictionary_quality`, a different set of dictionaries; the fifteen that carry the headline are 65.3–93.1% dead, mean 82.4%. An earlier audit (appendix X.43) had already "corrected" that slot once, from 36–45% to 34.9–45.4% — it checked that the printed value existed somewhere in `results/` and not what that value described. Value matching is not key matching, and key matching is not claim matching. This correction runs against us: the dictionaries are sparser and less healthy than we said.

X.51 Off-manifold encoding on a second crosscoder

Failure mode E1 is definitionally crosscoder-specific — it is about $\mathrm{enc}(a, \mathbf{0})$ — and every language-model dictionary elsewhere in this paper is a single-model SAE. So E1 rested on one crosscoder, one model pair, one layer pair, in the setting we ourselves call unusual. We trained a second crosscoder to find out whether the mode travels (PREREG §78, registered before the run).

The pair is `pythia-70m` against `pythia-70m-deduped` (Biderman et al., 2023): same architecture, same d_{model}, different training data, which is what makes it a diffing pair rather than two runs of one model. The crosscoder is the same recipe as appendix X.33's dictionaries — shared encoder over the concatenated activations, one decoder per model, TopK $k = 32$, $N = 4096$, 65k training positions — and the decomposition is the one section 5.2 applies to the tabular pair, with the effect measured as the KL divergence from splicing the modified activation back at the layer. We report one quantity per cell and attempt no diff.

Table 70: The interaction share on a language-model crosscoder, on the single-side code $\mathrm{enc}(a_A, \mathbf{0})$ and the joint code $\mathrm{enc}(a_A, a_B)$. All four dictionaries clear the $R^2 \geq 0.5$ gate on both sides.

layer	seed	side recon R^2 (single)	(joint)	$\lvert I\rvert/\lvert$replace$\rvert$ (single)	(joint)
3	42	0.863	0.939	0.299	$\mathbf{0.011}$
3	123	0.860	0.938	0.339	$\mathbf{0.016}$
1	42	0.564	0.854	0.331	$\mathbf{0.075}$
1	123	0.572	0.853	0.426	$\mathbf{0.069}$
mean				0.349	$\mathbf{0.043}$
tabular reference (section 5.2)				0.97	0.17–0.51

The direction replicates; the size does not. Single-side encoding inflates the interaction share in four of four dictionaries, by 4.4 to $27\times$ within one and $8\times$ on the mean, and it does so through the mechanism the tabular analysis identifies: zeroing the counterpart pushes the code off its training distribution, the side reconstruction falls ($0.86 \to 0.56$ at layer 1), and the residual it leaves is what the interaction term measures. But the levels are far lower than ours — 0.349 against 0.97, and 0.043 against 0.17–0.51. Two Pythia checkpoints are closer together than TabPFN v2 and v2.5, so zeroing one side costs less.

We pre-registered H53a, which required the single-side share to clear 0.5; at 0.349 it does not, so we do not claim it. What we claim is narrower and is what the table supports: *the mechanism behind E1 is a property of crosscoder analysis, established now on two crosscoders, two model pairs and two modalities; the magnitude is a property of the pair being diffed.* We also predicted the single-side

share would land at 0.6–0.8 and the joint share near 0.2. Both were wrong, and in the same direction — we overestimated how much the mode would travel.

It survives a first-order read-out. A reviewer noted that KL is quadratic at the clean point, so an interaction term $2ab$ appears in it even when the logit response is exactly additive, and that the tabular arm's cross-difference is taken on a signed, first-order output. We therefore re-measured the same four dictionaries on the signed change in log-probability of the realised next token, which is linear in the logits to first order (PREREG §92). The single-side share is 0.379 against 0.036 on the joint code — a gap of 0.34, marginally *larger* than KL's 0.31, with every dictionary between 0.32 and 0.46 single-side and between 0.01 and 0.06 joint. So KL's curvature does not produce the single-side excess: it persists on a read-out without that curvature. This rules out one alternative; it does not show which property of the off-manifold activation carries the interaction, and the log-probability of one token is itself only locally linear. What we still do not claim is that this is the *same mechanism* as TabPFN's: that arm's read-out is $|\Delta\mathrm{pred}|$, and a signed decomposition of it is in appendix E.1.

X.52 A THIRD SPARSITY OBJECTIVE, AND A DECISION RULE THAT COULD NOT SEE IT

"The dictionary family decides" has so far been a contrast between two points, L1+ReLU and TopK. That cannot separate "the sparsity objective is a deciding variable" from "these two dictionaries differ," so we added a third: JumpReLU (Rajamanoharan et al., 2024), a per-feature learned threshold with a straight-through gradient, whose sparsity comes from an L_0 penalty rather than from a global k or an L_1 on magnitudes (PREREG §79, registered before the run). Same model, width, budget, seeds, layers, matching and inference as table 51's arms.

Table 71: Three sparsity objectives on Pythia-70m-deduped at $N = 4096$. "Resolved" counts cells whose 95% interval excludes zero on the positive side; "sign" counts positive point estimates. Three of the four JumpReLU dictionaries miss the $[24, 40]$ band and are flagged rather than dropped, per §70's rule.

objective	resolved +	mean gap	sign +	mean L_0	mean recon R^2
L1+ReLU	$\mathbf{4/4}$	$+0.212$	$4/4$	44.4	0.783
JumpReLU	$1/4$	$+0.037$	$4/4$	29.0	0.779
TopK	$0/4$	-0.016	$2/4$	32.0	0.918

The axis has an interior point. JumpReLU sits between the other two on the mean gap and on how often it resolves, and its point estimates carry L1's sign in all four cells while TopK's carry it in two. It does not behave like TopK; it behaves like a weak L1. So the objective moves the estimate monotonically across the three points we have, and none of them is privileged — which is a weaker claim than a dichotomy and the one the data support.

Our decision rule could not have told us that. We pre-registered H54a/b/c *by counting resolutions only*: at most one of four resolving meant "with TopK." One of four resolved, so the rule returns that verdict — and it is wrong, because the rule puts "resolves as rarely as TopK" and "is a fifth the size of L1" in the same box. We should have written the rule on the effect size or the sign. We report the rule's output, and we follow the conclusion the evidence supports instead, which is the more conservative of the two. This is again a measurement choice of our own deciding a conclusion, and the first time the choice was a *decision rule* rather than an estimator: pre-registration constrains when you choose, not how well you choose.

Three grid extensions, recorded. Placing the penalty grid so that the realised L_0 lands in a pre-set band defeated us three times (§72.0, §75.0c, §79.0b); each time we saw only the L_0–λ curve, which is not an outcome variable, and widened the grid without touching the band or the verdict rule. JumpReLU's L_0 is especially steep in λ — $45 \to 22$ between $\lambda = 0.3$ and 1.0 — which is why three of its four dictionaries miss the band. A fourth defect was inert rather than adaptive: the first JumpReLU run applied the L_0 penalty to a boolean comparison, which carries no gradient, so the threshold never moved and every λ returned the same dictionary. We caught it because the sweep was suspiciously constant, before any gap was computed (§79.0).