DOES THE ROUTE MATTER? FUNCTIONAL VALIDATION OF SHARED–PRIVATE SAES

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) are commonly evaluated by reconstruction and feature descriptions, neither of which establishes that a learned decomposition matters to downstream computation. We introduce an operator-matched test of *functional routing validity*: correct and wrong private routes replace only the component represented by a training-only low-dimensional transform while preserving the activation's orthogonal residual. We instantiate the test with a shared–private SAE estimator based on principal-vector contrasts and group-level horseshoe shrinkage. On repository-derived prose and code, correct routing lowers held-out next-token negative log likelihood at three GPT-2 and three Pythia-70M layers. On independent text, the effect replicates at all three Pythia layers but GPT-2 intervals include zero. A 10-seed Pythia experiment removes the algebraically forced shared-subspace intersection and retains positive route effects on repository (0.227 $[0.111, 0.342]$) and independent text (0.432 $[0.354, 0.510]$); 50%-randomized routing is intermediate. The same test transfers to RxRx1 cellular embeddings, where correct HUVEC/RPE routing lowers perturbation-label loss by 0.234 $[0.209, 0.259]$ and improves top-1 accuracy by 5.04 percentage points. Calibrated simulations return to zero under known nulls. Functional routing validity is thus measurable across modalities while remaining model- and corpus-dependent.

1 INTRODUCTION

Dense internal representations are now collected routinely: residual-stream activations of language models at every layer and token, and embeddings of cellular images from models trained on millions of wells. Sparse autoencoders (SAEs) are the customary tool for decomposing such activations into a dictionary of sparse directions (Huben et al., 2024; Gao et al., 2025): the dictionary is linear, the code is sparse, reconstruction improves predictably with width, and benchmark suites compare dictionaries on a common footing (Karvonen et al., 2025). When activations come from two observed domains, for example prose and code, or two cell types, a natural extension is a shared–private SAE, in which a shared dictionary reconstructs what the domains have in common and a private expert for each domain reconstructs what is specific to it. The observed domain label selects the private expert, and we call that selection the route.

Existing evaluations of such decompositions fall into three lines. The first scores reconstruction, sparsity, and feature descriptions (Huben et al., 2024; Gao et al., 2025; Karvonen et al., 2025). However, an activation error is not an output error, since a small error in a sensitive direction changes the next-token distribution while a large one elsewhere does not, and frozen or random components can approach trained SAEs on these scores (Korznikov et al., 2026; Chanin, 2026). The second line intervenes on individual features and measures circuits, concept removal, or behavior (Marks et al., 2025; Karvonen et al., 2024). However, individual atoms are seed dependent while their span is reproducible (Paulo & Belrose, 2025; Gerasimov et al., 2026), so a single-atom intervention tests a basis choice as much as a representation. The third line uses interchange interventions to ask whether a representation realizes a specified variable (Chaudhary & Geiger, 2024; Geiger et al., 2025), and routed architectures such as Switch SAE (Mudide et al., 2025) allocate examples to experts for capacity but are evaluated by reconstruction. However, none of these asks the question a routed decomposition raises: does the downstream computation use the distinction between the private

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.

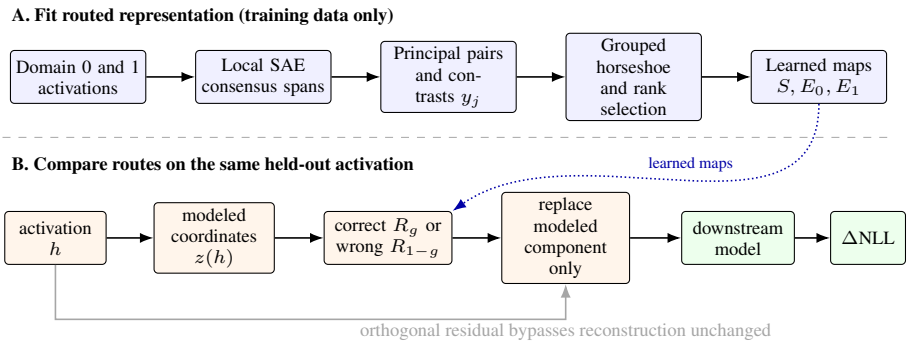

Figure 1: The two stages of the study. Panel A fits the shared map S and private experts E_0, E_1 from training activations only: local spans are paired through principal vectors, their contrasts y_j enter a grouped normal-means regression, and the grouped horseshoe with a step-up rule decides which paths are shared. Panel B compares routes on one held-out activation. The gray lane is the orthogonal residual, which bypasses reconstruction, so correct and wrong conditions differ only in the private expert applied to the modeled component.

experts, or does the matching expert merely fit the activation statistics better? A shared–private SAE can therefore pass every current score while its route is functionally inert.

Answering that question raises three statistical challenges. (1) Ownership is not identifiable from reconstruction and sparsity, because an always-on shared atom and the same atom copied into every private expert reconstruct identically; the decision must be made on cross-domain geometry. (2) The unit of that geometry is a direction in a q-dimensional coordinate space rather than a coordinate, so the shrinkage that decides shared against private must act on whole vectors and must not depend on the coordinate basis. (3) A functional test must change the route while holding fixed the operator, the represented rank, the held-out tokens, and the unmodeled part of the activation.

In this article we address these challenges with a grouped-horseshoe ownership estimator and an operator-matched routing test. Challenge (1) is answered by pairing the two local SAE spans through principal vectors and treating each pair's sign-aligned difference as one observation of a normal-means regression, so that ownership becomes the question of which coefficient vectors are zero (Section 3). Challenge (2) is answered by placing one half-Cauchy local scale on each whole contrast vector, the single-level grouped horseshoe of Xu et al. (2016), whose posterior shrinkage factor is rotation invariant and has a one-dimensional closed form that we integrate deterministically, and by a beta–Bernoulli step-up rule that turns the ranking into a shared rank. Challenge (3) is answered by a residual-preserving intervention that replaces only the modeled component of an activation and returns the orthogonal residual unchanged, so correct and wrong routes differ in the private expert alone (Section 4). Unlike feature-level interventions, our unit is a domain-indexed private reconstruction; unlike interchange interventions, we claim no variable correspondence, only that the estimated destination preserves more downstream information for its observed domain.

Our contributions are (1) the formulation of shared–private ownership as a grouped normal-means regression with a closed-form grouped-horseshoe posterior and an explicit rank decision, together with two qualifications that bound what the prior contributes; (2) the residual-preserving routing test and the identity that fixes its interpretation; and (3) evidence at three GPT-2 and three Pythia-70M layers, on two text corpora, in a floor-free geometry with ten refits, on RxRx1 cellular embeddings, and in calibrated null simulations, including a negative GPT-2 replication that bounds the language-model claim. Sections 2 to 4 build the estimand, the estimator, and the test; Sections 5 and 6 report the experiments and the limits. Code, configurations, and archived run artifacts are provided as supplementary material.

2 Data, estimand, and the statistical problem

Data and coordinates. Let $h \in \mathbb{R}^d$ be an activation at a fixed transformer hidden-state index, or a fixed-length image embedding, observed with a domain label $g \in \{0, 1\}$. In the language

experiments the domains are prose and code, d is the hidden width, and a token's activation is scored by the next-token loss of the same frozen model after the activation is edited. In the imaging experiments the domains are two cell types, $d = 128$, and a well's embedding is scored by a frozen readout that predicts which of 1,108 siRNA perturbations the well received. Training and evaluation observations are disjoint. A transform fitted on training activations only contains a mean $\mu \in \mathbb{R}^d$, orthonormal principal-component columns $C \in \mathbb{R}^{d\times q}$, and positive scales $s \in \mathbb{R}^q$; for row vectors,

$$z(h) = ((h - \mu)C) \oslash s, \qquad H(z) = \mu + (z \odot s)C^\top. \tag{1}$$

The modeled coordinates $z(h) \in \mathbb{R}^q$ are where all ownership estimation takes place; the residual $h - H(z(h))$ lies in the $(d - q)$-dimensional orthogonal complement of $\mathrm{span}(C)$ and is never modeled.

The routed object and the estimand. The estimator of Section 3 returns a shared reconstruction $S : \mathbb{R}^q \to \mathbb{R}^q$ and private experts E_0, E_1, each a sparse top-k autoencoder. Route $r \in \{0, 1\}$ reconstructs

$$R_r(z) = S(z) + E_r\{z - S(z)\}, \tag{2}$$

the shared part first and the private expert on what the shared part leaves. Routing is correct when $r = g$ and wrong when $r = 1 - g$. For domain g, evaluation block i, and position t, let $\ell_{git}(r)$ be the downstream loss after route r is applied through the intervention of Section 4. The primary estimand is the balanced wrong-minus-correct loss

$$\Delta_{\mathrm{bal}} = \tfrac{1}{2}\sum_{g=0}^{1} \mathbb{E}_{i,t}\left[\ell_{git}(1 - g) - \ell_{git}(g)\right], \tag{3}$$

paired on identical held-out observations. A positive value means the expert indexed by the observed domain preserves more downstream information than the other expert. Note that Δ_{bal} is a property of the fitted decomposition inside a fixed downstream model, not a claim about the domains themselves.

The statistical problem. Before Δ_{bal} can be measured, the estimator must decide which directions of $\mathbb{R}^q$ the domains share and what each owns privately. Reconstruction cannot decide this: a direction placed in S, or copied into both E_0 and E_1, reconstructs every training activation identically, so the shared rank is unidentified by any reconstruction objective and must be read from cross-domain geometry. Section 3 therefore estimates one local span per domain, pairs the spans, and turns each pair into one vector observation of a regression whose zero coefficient vectors are the shared directions. Each hyperparameter of that regression has one job, and the rule that extracts the shared rank is stated separately from the prior that ranks the candidates, following the decoupling of shrinkage from selection in Hahn & Carvalho (2015).

3 Ownership as a grouped normal-means regression

We build the estimator in the order in which it is computed: a local span per domain, a pairing of the two spans, a regression on the paired differences, a prior that ranks its rows, a rule that cuts the ranking, and sparse atoms fitted after the cut.

Local consensus spans. Within each domain, three top-k SAEs of width $r = 20$ and $k = 4$ are trained on bootstrap resamples of the training coordinates $z(h)$. The atoms of each replica are matched one-to-one to a medoid replica by absolute cosine similarity, sign aligned, and averaged over the replicas whose match exceeds 0.80. We use the numerical column span of this consensus dictionary, with orthonormal basis $Q_g \in \mathbb{R}^{q\times r}$, as the local span of domain g, and not its atoms, because atoms vary with the training seed while the subspace they span is reproducible (Paulo & Belrose, 2025; Gerasimov et al., 2026).

Principal pairs and the regression. The singular value decomposition of the $r \times r$ matrix $Q_0^\top Q_1 = U\,\mathrm{diag}(\rho_1, \ldots, \rho_m)V^\top$, with $m = r$ and $1 \geq \rho_1 \geq \cdots \geq \rho_m \geq 0$, pairs the principal vectors $a_j = Q_0 U_j$ and $b_j = Q_1 V_j$ of the two spans. After choosing the sign of b_j so that $a_j^\top b_j \geq 0$, the jth cross-domain contrast is

$$y_j = (a_j - b_j)/\sqrt{2} \in \mathbb{R}^q, \qquad \|y_j\|_2^2 = 1 - \rho_j. \tag{4}$$

If a direction lies in both spans, its aligned principal vectors coincide and $y_j = 0$; if the spans differ along path j, y_j is large. In the population idealization in which each local span is the orthogonal direct sum of a common subspace of rank r_0 and a private subspace, with largest canonical correlation between the two private subspaces $\rho < 1$, the matrix $Q_0^\top Q_1$ is block diagonal with an $r_0 \times r_0$ identity block, so exactly r_0 of the ρ_j equal one and the rest are at most ρ; ownership is then the multiplicity of unit principal similarities and no atom correspondence is needed. Orthogonality is required, since a private direction leaning on the common subspace can enlarge the intersection beyond r_0. Stacking the contrasts as rows gives the working regression

$$y_j = \theta_j + \varepsilon_j, \qquad \varepsilon_j \sim N_q(0, \sigma^2 I_q), \qquad j = 1, \dots, m, \tag{5}$$

a normal-means regression with identity design, m vector responses, and one coefficient vector $\theta_j \in \mathbb{R}^q$ per row. Ownership is now a row-sparsity question: path j is shared when $\theta_j = 0_q$ and private otherwise. The Gaussian error is a calibrated working model for estimation error in the aligned principal vectors, not a sampling law implied by SAE training, and no claim below depends on it holding exactly.

Grouped horseshoe prior. The expected pattern is sparse at the level of rows: most paths are nearly shared and a few differ strongly. We therefore place a global–local scale mixture on each whole row, with one local scale per row rather than per coordinate,

$$\theta_j \mid \lambda_j, \tau \sim N_q(0, \tau^2 \lambda_j^2 I_q), \qquad \lambda_j \overset{\text{iid}}{\sim} C^+(0, 1), \tag{6}$$

where $C^+(0, 1)$ is the standard half-Cauchy law. Equations 5–6 are the single-level grouped horseshoe of Xu et al. (2016) with an orthogonal within-group design; the marginal of θ_j/τ is the Grouped Horseshoe distribution of He & Wand (2025), and $q = 1$ recovers the horseshoe of Carvalho et al. (2010). Each hyperparameter has one job. The noise variance σ^2 sets the unit of contrast energy, $T_j = \|y_j\|_2^2/\sigma^2$. The global scale τ pulls every row toward zero, the shared end. The local scale λ_j lets one whole row escape that pull through the Cauchy tail. Their combination is the shrinkage weight

$$\kappa_j = \frac{\sigma^2}{\sigma^2 + \tau^2 \lambda_j^2} \in (0, 1), \qquad \mathbb{E}(\theta_j \mid y_j, \lambda_j) = (1 - \kappa_j)\, y_j, \tag{7}$$

which multiplies all q coordinates of row j by the same factor, so that κ_j near one means a shared path and near zero a private one. Grouping answers challenge (2): the prior covariance within a row is isotropic, so an orthogonal change of the q coordinates leaves T_j and every posterior summary unchanged, whereas separate local scales λ_{jr} would make the decision depend on the coordinate basis and could shrink fragments of one path. All fits use $\sigma^2 = 0.001$ and $\tau = 0.1$. Given the energy unit σ^2, the prior enters Equation 7 only through $\tau^2 \lambda_j^2/\sigma^2$, so the scale-free quantity to report is the ratio $\tau/\sigma = \sqrt{10}$ (Piironen & Vehtari, 2017a). We fix τ rather than estimate it because its sparse-normal-means calibration is tied to the prevalence of nonzero rows (van der Pas et al., 2014), which is the quantity being estimated; Lemma 1 shows that this costs nothing for the ranking.

Lemma 1 (Grouped shrinkage posterior)**.** *Under Equations 5–6, write $c = \sigma^2/\tau^2$. The posterior of κ_j depends on y_j only through T_j and has density*

$$p(\kappa_j \mid T_j) \propto \frac{\kappa_j^{(q-1)/2} (1 - \kappa_j)^{-1/2}}{c + (1 - c)\kappa_j}\, e^{-T_j \kappa_j / 2}, \qquad 0 < \kappa_j < 1, \tag{8}$$

and, whenever this posterior is nondegenerate, $\frac{d}{dT_j}\mathbb{E}(\kappa_j \mid T_j) = -\frac{1}{2}\operatorname{Var}(\kappa_j \mid T_j) < 0$.

In other words, the group dimension enters only through the exponent of κ_j, and larger contrast energy always receives less shrinkage. Three things follow. First, Equation 8 is a one-parameter exponential family in the natural parameter $-T_j/2$, so the posterior mean $\bar{\kappa}_j$, variance, and interval are obtained by deterministic quadrature on 2,001 equally spaced values of $\log \lambda_j \in [-12, 12]$, at a cost of $m \times 2{,}001$ density evaluations, and no sampler is needed. Second, at $\tau = \sigma$ the induced prior on κ_j is $\mathrm{Beta}(1/2, 1/2)$ for every q, the U-shaped density that a binary ownership decision calls for. Third, with common q, σ, τ the order of the rows by $\bar{\kappa}_j$ does not depend on τ, so τ is a reporting scale for κ_j and not a decision scale. The proof is in Appendix B.

Shared-rank decision. Continuous shrinkage ranks rows but sets none of them to zero, so the discrete decision is made by a separate rule (Hahn & Carvalho, 2015). Order the rows $j_{(1)}, \ldots, j_{(m)}$ from smallest to largest $\bar{\kappa}_j$, least shrunk first, and compare each row's zero model $H_0 : \theta_j = 0$ with a Gaussian slab $H_1 : \theta_j \sim N_q(0, vI_q)$,

$$b(T_j) = \log BF_{10,j} = -\frac{q}{2}\log\left(1 + \frac{v}{\sigma^2}\right) + \frac{T_j}{2}\,\frac{v}{v+\sigma^2}. \tag{9}$$

The nested model whose private set is the first k rows scores

$$S_k = \sum_{i=1}^{k} b(T_{j_{(i)}}) + \log \frac{B(a+k,\, b+m-k)}{B(a,b)}, \qquad k = 0, \ldots, m, \tag{10}$$

where the beta–Bernoulli term is the multiplicity adjustment of Scott & Berger (2010) with $a = b = 1$, a uniform prior on the number of private rows. The maximizer $k^\star$ is the estimated number of private paths and $m - k^\star$ is the shared rank. We use $v = 0.01 = \tau^2$, so the slab and the ranking prior agree on how large a private contrast is; v is the finite slab variance of the regularized horseshoe (Piironen & Vehtari, 2017b), not a horseshoe parameter. We avoid the canonical half-threshold $\bar{\kappa}_j < 1/2$ because it inherits the calibration of τ, which we fixed for a reason that forbids tuning it to the answer, and because grouped half-thresholding classifies substantially worse than a spike-and-slab benchmark in He & Wand (2025). Lemma 2 in Appendix B shows that the slab score is the degenerate member of a family of strictly increasing evidences that includes the fully integrated horseshoe evidence, so the nested path in Equation 10 is the same for every member and the slab only fixes the scale at which the multiplicity penalty is paid.

Sparse atoms after the decision. The $m - k^\star$ selected pairs are averaged, $(a_j + b_j)/\|a_j + b_j\|_2$, and re-orthogonalized into a shared basis $B \in \mathbb{R}^{q\times(m-k^\star)}$. The shared map S is a top-k SAE of width $m - k^\star$ and $k = 3$ trained on the projected coordinates $zBB^\top$, initialized by FastICA inside the selected span and anchored to that initialization with strength 1.0 so that its atoms stay inside the span the decision chose. Each private expert E_g is a top-k SAE of width 10 and $k = 3$ trained on the residual $z - S(z)$ of its own domain with an incoherence penalty of 0.01 against the shared dictionary, which stops an expert from relearning shared atoms and reintroducing the ambiguity of challenge (1). Local replicas train for 400 Adam steps and the shared and private maps for 300, with batch size 128 and learning rate 0.003; this training dominates the runtime, since the statistical layer costs only one $r \times r$ decomposition and $m \times 2{,}001$ scalar density evaluations.

Two qualifications. First, because Lemma 1 holds under common q, σ, τ, ranking by $\bar{\kappa}_j$ is the same as ranking by raw energy $1 - \rho_j$. The grouped prior therefore contributes a rotation-invariant shrinkage scale and row-level uncertainty summaries, not a demonstrated ranking advantage, and Section 5 audits this equivalence rather than claiming estimator superiority. Second, two rank-r subspaces of $\mathbb{R}^q$ intersect in at least $2r - q$ dimensions. In the primary design, $r = 20$ and $q = 32$, so at least eight of the ρ_j equal one whatever the data; across 101 archived fits, 100 have exactly eight similarities above 0.999999 and the selected shared rank is 8 to 10. The fitted shared rank in that design is a forced intersection plus at most two paths, not a measurement of what the domains share. A floor-free design requires $2r \le q$; the $q = 48$ experiment of Section 5 satisfies it.

4 THE FUNCTIONAL ROUTING TEST

Residual-preserving intervention. A full low-rank replacement would set h to $H(R_r(z(h)))$ and discard the $(d-q)$-dimensional component of h outside $\mathrm{span}(C)$, confounding the route effect with truncation damage that has nothing to do with the route. We instead intervene with

$$I_r(h) = h + H(R_r(z(h))) - H(z(h)), \tag{11}$$

which adds to h the difference between the routed and the identity reconstruction of its modeled component.

Proposition 1 (Residual preservation)**.** *Let $P_C = CC^\top$. For every route r and every h, $\{I_r(h) - h\}(I - P_C) = 0$.*

In other words, correct and wrong routes alter only the PCA span fitted on training activations and return exactly the same unmodeled residual. This is the identification property of the experiment and the answer to challenge (3): the two conditions share the operator, the represented rank, the coordinate system, the held-out tokens, and the residual, and differ only in which private expert acts on $z - S(z)$. The proof and a comparison with full replacement are in Appendix B.

Conditions. Every held-out block is scored under identical token identities in each condition. The unmodified model and a no-op hook, which must reproduce baseline logits within 10^{-6} and does so with maximum difference exactly zero in every completed run, anchor the scale. Correct and wrong routing apply Equation 11 with R_g and R_{1-g}. Shared-only reconstruction replaces R_r by S and asks whether the routed private term adds information at all. A deterministic 50%-randomized route assigns exactly half of the scored positions in each block to the wrong expert under a fixed control seed; it is a route-intensity control, not a new estimator, and should fall between correct and wrong if the route carries information. Direct private-subspace ablation against a per-block rank- and energy-matched random subspace asks the different, destructive question of whether the selected private span is unusually necessary; we report it separately.

Metrics. For language models the primary loss is next-token negative log likelihood (NLL), the mean of $-\log p(\text{next token})$ over scored positions, lower being better; it is the model's own training objective, so a change in it is a change in what the model computes. When the next token is drawn from the unedited model, as in the simulations, its expected difference between two edits equals the difference in $\mathrm{KL}(p_{\mathrm{base}} \| p_{\mathrm{edit}})$ (Appendix B); on real text the target is the observed token, so we record that label-free divergence separately, together with the root-mean-square size of the edit, so that every effect is reported beside its magnitude. For RxRx1 the loss is the perturbation NLL of a frozen cosine nearest-centroid readout with a softmax temperature calibrated on training wells, and we also report top-1 accuracy, the proportion of wells whose true condition ranks first, in percentage points. Intervals are 95% Student-t intervals over independently refitted estimators after averaging evaluation blocks within seed and domain, five refits for the primary grid and ten for the floor-free experiment; they quantify refit variability, not population uncertainty over models or corpora.

5 Experiments

Design. We evaluate GPT-2 at hidden-state indices 3, 7, and 11 and Pythia-70M-deduped at indices 2, 4, and 6 (Biderman et al., 2023). For each model–layer and estimator seed, 3,800 prose and 200 code tokens fit the transform and the estimator, and eight disjoint blocks of 100 tokens per domain are held out; the imbalance is deliberate, being the regime in which the rare domain's expert is hardest to estimate. Indices 3 and 11 were added after the design was frozen and differ only in intervention site. The repository corpus takes prose and code from one scientific-software repository, so that topic is approximately controlled while modality changes; because that source changed during development, primary results use one recorded version. The independent corpus has no shared provenance: prose is the complete mypyc documentation directory and code is the complete CPython json package, both copied with per-file hashes before any independent-corpus run began. Complete configurations and seed-level outcomes are in Appendix C.

5.1 Correct routing improves prediction at every primary cell

Correct routing beats wrong routing at all six cells on the repository corpus (Table 1), and all 30 seed-level contrasts are positive. A random-effects summary of the six cell means is $0.426\ [0.284, 0.568]$, but layers share model weights and text and are not independent replications. Correct routing also beats shared-only reconstruction in every cell, by 0.289 to 0.730 NLL with all 30 seed contrasts positive, so the routed private term adds information rather than merely avoiding the wrong expert.

5.2 The effect survives floor-free geometry, and randomized routing is intermediate

The primary design forces an eight-dimensional intersection between the local spans (Section 3). We therefore refit Pythia index 4 with $q = 48$ and $r = 20$, for which the forced floor is zero, on both corpora with ten estimator seeds. No fit has a principal similarity above 0.999999, and the selected

Table 1: Balanced wrong-minus-correct next-token NLL, Equation 3. Brackets are seed-level 95% intervals. Independent-text replication is family dependent.

Cell	Repository corpus	Independent corpus
GPT-2 index 3	$0.812\ [0.551, 1.073]$	$0.157\ [-0.129, 0.443]$
GPT-2 index 7	$0.367\ [0.166, 0.568]$	$0.229\ [-0.058, 0.516]$
GPT-2 index 11	$0.618\ [0.074, 1.163]$	$0.175\ [-0.244, 0.594]$
Pythia index 2	$0.305\ [0.190, 0.420]$	$0.258\ [0.197, 0.319]$
Pythia index 4	$0.345\ [0.210, 0.480]$	$0.316\ [0.198, 0.434]$
Pythia index 6	$0.362\ [0.275, 0.450]$	$0.278\ [0.235, 0.320]$

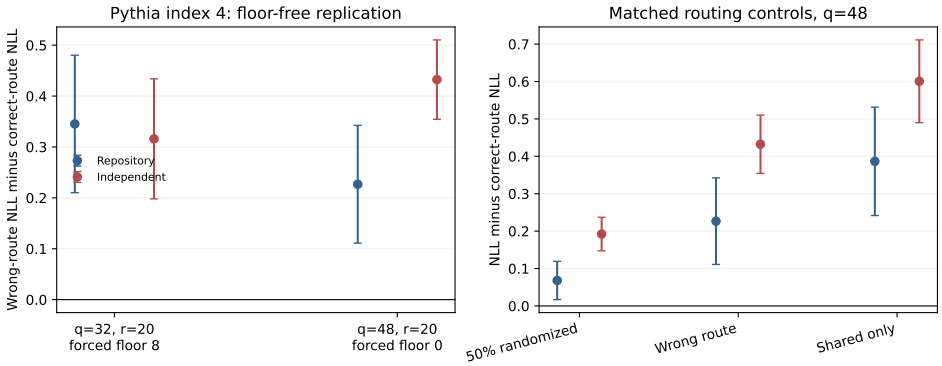

Figure 2: Left: Pythia index-4 routing survives removal of the forced shared-subspace floor; primary $q = 32$ intervals use five refits and floor-free $q = 48$ intervals use ten. Right: in both floor-free fits the randomized route lies between correct and wrong, and shared-only reconstruction loses still more information. Bars are seed-level 95% intervals.

shared rank falls to 2–4 on repository text and 3–5 on independent text. Wrong-minus-correct NLL remains positive, $0.227\ [0.111, 0.342]$ with 9 of 10 seeds positive and $0.432\ [0.354, 0.510]$ with 10 of 10 (Figure 2); shared-only-minus-correct is $0.387\ [0.242, 0.531]$ and $0.601\ [0.490, 0.711]$. The 50%-randomized route is intermediate, exceeding correct by $0.068\ [0.017, 0.119]$ and $0.192\ [0.147, 0.237]$, while fully wrong routing exceeds randomized by $0.159\ [0.089, 0.228]$ and $0.240\ [0.195, 0.285]$, a dose response under one operator. The raw-energy audit of Section 3 is exact here: across all 20 fits, ranking by increasing $1 - \rho_j$ matches ranking by decreasing $\bar{\kappa}_j$ with Spearman correlation 1.0 and the evidence-selected supports coincide, so holding the rank rule fixed gives the same routes and outcomes.

5.3 Independent text replicates for Pythia and not for GPT-2

The independent corpus sharpens the conclusion in a way that pooling would hide (Figure 3). Pythia replicates at every layer: 15 of 15 seed contrasts are positive, the family summary is $0.275\ [0.223, 0.326]$, and every seed-paired change from the repository corpus includes zero. GPT-2 does not clear the same gate: 11 of 15 seeds are positive, each layer interval includes zero, the family summary is $0.189\ [-0.093, 0.472]$, the paired drops at indices 3 and 11 exclude zero, and shared-only reconstruction reverses at index 3. A plausible mechanism is generic reconstruction cost: at GPT-2 index 3 correct routing itself sits about 2.47 NLL above the unmodified model on this corpus, so the route contrast is measured on top of a large and seed-variable baseline shift; we have not yet tested this. A six-cell inverse-variance summary is dominated by the precise Pythia cells and is not used. The defensible language-model claim is Pythia replication across two corpora, with corpus-dependent GPT-2 evidence.

5.4 The same test transfers to cellular embeddings

RxRx1 provides a modality transfer with independent experimental batches (Sypetkowski et al., 2023). We average two image sites per well, fit on the official training portion of 25,441 wells

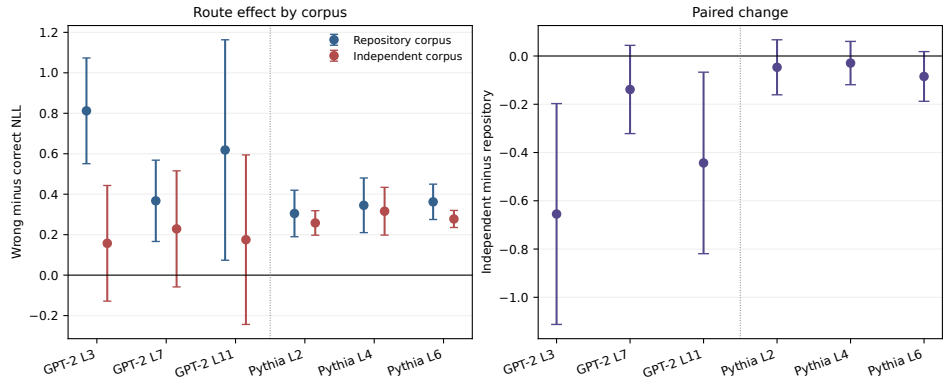

Figure 3: Route effect by corpus (left) and estimator-seed-paired change from repository to independent text (right). Pythia replicates tightly; GPT-2 becomes smaller and uncertain.

Table 2: RxRx1 held-out perturbation prediction. Positive values favor correct over wrong routing.

Domain	Wrong−correct NLL	Correct−wrong top-1
HUVEC	$0.286\ [0.214, 0.358]$	$6.01\ [4.91, 7.12]$ pp
RPE	$0.182\ [0.139, 0.226]$	$4.07\ [3.37, 4.76]$ pp
Balanced	$0.234\ [0.209, 0.259]$	$5.04\ [4.61, 5.47]$ pp

from 23 experiments, and evaluate 13,263 wells from 12 unseen experiments. HUVEC and RPE are the observed domains, the 32 PCA coordinates retain 71.1% of training variance, and the estimator, centroids, and softmax temperature are fitted on training wells only. Wrong routing raises balanced perturbation NLL by $0.234\ [0.209, 0.259]$ and correct routing improves top-1 accuracy by 5.04 points $[4.61, 5.47]$ (Table 2); both contrasts favor the correct route in every estimator fit and every held-out experiment. Shared-only-minus-correct NLL is $0.345\ [0.330, 0.359]$ and the randomized route is again intermediate at $0.113\ [0.095, 0.130]$ (Figure 4). Correct reconstruction does not improve on the original embedding, whose top-1 accuracy is 97.6% for HUVEC and 96.6% for RPE against 85.0% and 88.7% after routing; the claim is comparative phenotype preservation under a matched reconstruction, not representation improvement or gene-specific causal identification.

5.5 Simulations return to zero under known nulls

The simulation program contains 2,450 spectral cells and 40 fits of the complete consensus, horseshoe, step-up, and conditional-SAE pipeline (Appendix F). Synthetic activations contain shared sparse factors, domain-private factors, and orthogonal noise; a fixed categorical output head controls whether private factors affect prediction, and expected NLL under the unedited distribution removes label-sampling noise. The effect increases with private output strength, declines as the private subspaces become similar, and vanishes under high noise or severe rank misspecification. Under known nulls it is centered at zero: with no private-output dependence the mean is $0.000007\ [-0.000001, 0.000016]$, and under 200 identical-domain replicates it is -0.000004 $[-0.000011, 0.000004]$ with 47% positive replicates. Full low-rank replacement incurs far more generic output damage than residual preservation, which is why Equation 11 is the primary operator, and full-estimator fits remain positive at 40, 80, 200, and 800 rare-domain observations, so the instability seen in small real code samples is activation and corpus specific. Direct ablation is the one heterogeneous quantity: private-minus-random NLL is positive at Pythia indices 2 and 6 and negative at Pythia index 4 and GPT-2 index 7, because destroying a span and rerouting its reconstruction are different interventions, so necessity is kept separate from routing validity.

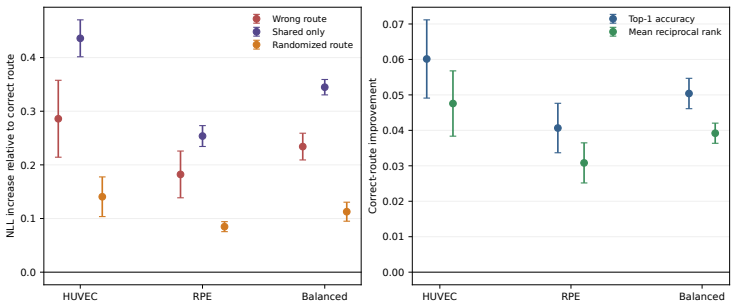

Figure 4: RxRx1 controls relative to correct routing (left) and gains from correct instead of wrong routing (right). Bars are seed-level 95% intervals.

6 Related work and limitations

Activation patching shows that conclusions depend on the replacement operator and the metric (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024); Proposition 1 and pairing on identical observations are our answer. Feature steering varies across contexts and perturbs unrelated features (Duan, 2026), so every route effect is reported beside its magnitude and divergence. Single-token feature effects depend strongly on layer and SAE family (Cho et al., 2026), consistent with the heterogeneous ablation results we keep apart from routing. On the statistical side, the grouped horseshoe is used here as a ranking and uncertainty device under fixed scales; group selection consistency for one-group global–local priors (Paul et al., 2025) assumes a fixed design and half-thresholding, and neither premise holds for data-dependent principal vectors decided by Equation 10, so we claim no asymptotic selection guarantee. Horseshoe credible sets are not honest over the whole parameter space (van der Pas et al., 2017), and the row-level intervals we store describe ambiguity within the contrast model, not uncertainty from SAE training, which the refit intervals carry.

Four limitations bound the claims. First, the raw-energy audit reproduces the grouped-horseshoe ordering and support exactly under the fitted common scales, so the experiments are not evidence of estimator superiority; fixed-rank, independent-domain, random-shared, and coordinatewise-shrinkage comparisons would be required for that, and the floor-free result makes selected rank algebraically unforced without making it semantically identifiable. Second, five refits are a small uncertainty sample for the primary grid, the intervals average blocks before comparing seeds, and the all-positive repository result is descriptive rather than 30 independent replications; a hierarchical block–seed–layer analysis would separate evaluation from refit uncertainty. Third, aggregate NLL does not say whether gains concentrate on syntax, identifiers, or domain-diagnostic tokens, so no semantic-localization claim is made. Fourth, GPT-2 and Pythia-70M are small, their independent-corpus behavior differs, and the imaging result rests on one encoder and two cell types; a larger model, diverse corpus families, learned routing, mixed-domain inputs, further cell types, and an independent encoder are needed to establish scope.

7 Conclusion

We have formulated shared–private ownership as a grouped normal-means regression on principal-vector contrasts, ranked its rows with a grouped horseshoe whose posterior is one-dimensional and rotation invariant, cut the ranking with a multiplicity-adjusted step-up rule, and tested the resulting route with an intervention that edits the modeled component of an activation and returns its residual untouched. Correct routing preserves more downstream information in repository-corpus transformer activations and RxRx1 cellular embeddings, and calibrated simulations recover zero under known nulls. Independent text sharpens the result: Pythia replicates at all tested layers while GPT-2 becomes uncertain and corpus dependent. At Pythia index 4 the effect survives floor-free geometry, and randomized routing produces intermediate damage. Functional routing validity is therefore measurable and transferable, but its scope must be established by replication rather than inferred from reconstruction. The natural next step is a block-within-seed hierarchical model of the route effect that separates evaluation from refit uncertainty.

REPRODUCIBILITY STATEMENT

All preprocessing and estimator fitting use training observations only. Runs store resolved configurations, model and tokenizer revisions, package versions, corpus and token hashes, evaluation masks, fitted transforms, estimator identifiers, block-level outcomes, and summaries in immutable directories. Five estimator seeds for the primary grid, ten for the floor-free experiment, and paired control seeds are explicit. Figure and analysis scripts regenerate the reported summaries from archived cell tables, while a manuscript audit checks numerical claims, citation keys, figure pairs, no-op gates, and artifact hashes. The supplementary material specifies the estimator, configurations, seed-level effects, corpus provenance, simulations, and RxRx1 evaluation protocol.

AI USE STATEMENT

Generative AI tools were used in accordance with the conference guidelines to refine the grammar and tone of the paper, format BibTeX citations, and refine LaTeX displays while editing the manuscript. The authors determined the research questions, experimental design, estimands, interpretation, and all scientific conclusions. All AI-assisted text was reviewed by the authors; numerical claims were checked against archived run artifacts and citations against the cited sources. The authors take responsibility for the complete submission.

REFERENCES

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

David Chanin. Are sparse autoencoder benchmarks reliable?, 2026.

Maheep Chaudhary and Atticus Geiger. Evaluating open-source sparse autoencoders on disentangling factual knowledge in gpt-2 small, 2024.

Seonglae Cho, Zekun Wu, Kleyton Da Costa, Rishi Kalra, Ilham Wicaksono, and Adriano Koshiyama. Are single-token sparse autoencoder features causally necessary? layer-depth and sae-family effects, 2026.

Evan Duan. Pre-intervention prediction of sparse autoencoder steering side effects, 2026.

Leo Gao, Tom Dupre 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*, volume 2025, pp. 26721–26754, 2025.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

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

P. Richard Hahn and Carlos M. Carvalho. Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *Journal of the American Statistical Association*, 110 (509):435–448, 2015. doi: 10.1080/01621459.2014.993077.

Virginia X He and Matt P Wand. The grouped horseshoe distribution and its statistical properties. *Stat*, 14(1):e70050, 2025.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Adam Karvonen, Can Rager, Samuel Marks, and Neel Nanda. Evaluating sparse autoencoders on targeted concept erasure tasks, 2024.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, et al. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability. 2025.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Sanity checks for sparse autoencoders: Do saes beat random baselines?, 2026.

Samuel Marks, Can Rager, Eric 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*, volume 2025, pp. 23888–23923, 2025.

Anish Mudide, Josh Engels, Eric Michaud, Max Tegmark, and Christian Schroeder de Witt. Efficient dictionary learning with switch sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 101830–101844, 2025.

Sayantan Paul, Prasenjit Ghosh, and Arijit Chakrabarti. Consistent group selection using global–local shrinkage priors in sparse normal linear regression. *Annals of the Institute of Statistical Mathematics*, 2025. doi: 10.1007/s10463-025-00950-z. arXiv:2302.04715.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. 2025.

Juho Piironen and Aki Vehtari. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 905–913. PMLR, 2017a. URL `https://proceedings.mlr.press/v54/piironen17a.html`.

Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017b. doi: 10.1214/17-EJS1337SI.

James G Scott and James O Berger. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pp. 2587–2619, 2010.

Maciej Sypetkowski, Morteza Rezanejad, Saber Saberian, Oren Kraus, John Urbanik, James Taylor, Ben Mabey, Mason Victors, Jason Yosinski, Alborz Rezazadeh Sereshkeh, et al. Rxrx1: A dataset for evaluating experimental batch correction methods. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4285–4294. IEEE, 2023.

Stéphanie van der Pas, Botond Szabó, and Aad van der Vaart. Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis*, 12(4):1221–1274, 2017. doi: 10.1214/17-BA1065.

Stéphanie L. van der Pas, Bas J. K. Kleijn, and Aad W. van der Vaart. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2):2585–2618, 2014. doi: 10.1214/14-EJS962.

Zemei Xu, Daniel F Schmidt, Enes Makalic, Guoqi Qian, and John L Hopper. Bayesian grouped horseshoe regression with application to additive models. In *Australasian Joint Conference on Artificial Intelligence*, pp. 229–240. Springer, 2016.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

A Full estimator specification

A.1 Posterior computation and selection

The use of a common local shrinkage scale for every coefficient in a predefined group follows the grouped horseshoe construction of Xu et al. (2016); our groups are principal-vector contrast paths rather than additive-model basis expansions, the design is the identity, and the groups do not overlap. Three departures from that regression model are deliberate. Xu et al. (2016) write the group prior variance as $\sigma^2\tau^2\delta_g^2$ and place half-Cauchy and Jeffreys priors on τ and σ^2; we write it as $\tau^2\lambda_j^2$, so their global scale corresponds to our $\tau/\sigma = \sqrt{10}$, and we fix both τ and σ^2 for the reason given in Section 3. Their coefficients are retained and sampled by the auxiliary-variable Gibbs sampler of Makalic and Schmidt; here θ_j is integrated out analytically and the one-dimensional posterior of Lemma 1 is integrated on a grid. We verified that the grid posterior agrees with that Gibbs sampler, run with the same fixed τ and σ^2, to Monte Carlo error in mean, variance, and interval. The implementation integrates Equation 8 on 2,001 equally spaced values of $\log\lambda_j \in [-12, 12]$, with the half-Cauchy density in $\log\lambda_j$ proportional to $\lambda_j/(1+\lambda_j^2)$, and stores the posterior mean, variance, and 95% equal-tailed interval of κ_j for every path. The step-up rule in Equation 10 is evaluated for $k = 0, \ldots, m$ and the maximizer is stored with the log Bayes factors of Equation 9. All fits use $\sigma^2 = 0.001$, $\tau = 0.1$, $v = 0.01$, and $a = b = 1$. The canonical half-threshold rule $\bar{\kappa}_j < 1/2$ is retained in the implementation as an option and is not used in any reported fit. Selected shared principal vectors are averaged and re-orthogonalized, a shared sparse dictionary is initialized with FastICA, and each private expert is trained on the residual left by shared reconstruction.

A.2 Dimensional floor audit

Two rank-r subspaces of $\mathbb{R}^q$ intersect in dimension at least $2r - q$. In the primary design $r = 20, q = 32$, so eight shared directions are forced. Of 101 archived fits, 100 have exactly eight principal similarities above 0.999999. The shared rank is 8 in 52 fits, 9 in 44, and 10 in 5. The estimator selects the forced paths and sometimes one or two additional paths with similarity 0.92–1; the largest nonselected similarity is 0.916. A floor-free design instead requires $2r \leq q$.

We implemented the first design at Pythia index 4 on both text corpora. With $q = 48, r = 20$, the lower bound is zero. None of the 20 estimator fits has a principal similarity above 0.999999; the largest observed similarity is 0.996675. The selected rank is 2–4 (mean 3.1) on repository text and 3–5 (mean 4.4) on independent text. Thus the robustness result does not inherit the forced eight-dimensional intersection of the primary design.

A.3 Raw contrast-energy audit

Because a_j and b_j are sign-aligned unit vectors,

$$\|y_j\|_2^2 = \tfrac{1}{2}\|a_j - b_j\|_2^2 = 1 - a_j^\top b_j = 1 - \rho_j.$$

For each of the 20 floor-free estimators we recomputed the stable ordering by increasing $1 - \rho_j$ and compared it with decreasing posterior mean shrinkage. Every complete ordering is identical (Spearman correlation 1.0 in every fit). At the evidence-selected rank, every selected support is also identical. Replacing horseshoe ordering by raw energy while holding the rank rule fixed therefore leaves the shared basis, private destinations, and all downstream results unchanged. This is an equivalence finding, not estimator-superiority evidence.

B Proofs

B.1 Proof of Lemma 1

Marginalizing θ_j in Equations 5–6 gives $y_j \mid \lambda_j \sim N_q(0, (\sigma^2+\tau^2\lambda_j^2)I_q)$, whose density depends on y_j only through $\|y_j\|_2^2$ and on λ_j only through $\kappa_j = \sigma^2/(\sigma^2+\tau^2\lambda_j^2)$; it equals $(\kappa_j/\sigma^2)^{q/2}e^{-T_j\kappa_j/2}$ up to a constant. The change of variables $\lambda_j \mapsto \kappa_j$ has $\lambda_j^2 = c(1-\kappa_j)/\kappa_j$ and Jacobian proportional to $\kappa_j^{-3/2}(1-\kappa_j)^{-1/2}$, and the half-Cauchy density $\propto (1+\lambda_j^2)^{-1} = \kappa_j/\{c+(1-c)\kappa_j\}$. Multiplying

the three factors gives Equation 8. Since T_j is unchanged by $y_j \mapsto Oy_j$ for orthogonal O, the posterior is rotation invariant. Equation 8 is an exponential family in the natural parameter $-T_j/2$ with sufficient statistic κ_j, so $\frac{d}{dT_j}\mathbb{E}(\kappa_j \mid T_j) = -\frac{1}{2}\operatorname{Var}(\kappa_j \mid T_j)$, which is negative whenever the posterior is nondegenerate. At $\tau = \sigma$ we have $c = 1$ and the denominator is constant; the prior factor, which is Equation 8 without the likelihood contribution $\kappa_j^{q/2} e^{-T_j\kappa_j/2}$, is then $\kappa_j^{-1/2}(1-\kappa_j)^{-1/2}$, the $\mathrm{Beta}(1/2, 1/2)$ density, for every q.

B.2 Evidence monotonicity

Lemma 2 (Evidence monotonicity)**.** *For any mixing distribution π on λ_j^2 in Equation 6 that is not concentrated at $\lambda_j = 0$, the marginal log-evidence $b_\pi(T)$ of row j against $\theta_j = 0$ satisfies $b'_\pi(T) = \{1 - \mathbb{E}_\pi(\kappa_j \mid T)\}/2 > 0$ and $b''_\pi(T) = \operatorname{Var}_\pi(\kappa_j \mid T)/4$, where the expectation and variance are under the posterior of κ_j given T induced by π. The slab score in Equation 9 is the member with π a point mass at $\lambda_j^2 = v/\tau^2$.*

In other words, every evidence in the family, including the fully integrated horseshoe evidence, is strictly increasing and convex in contrast energy, and the Gaussian slab is its degenerate member. What this buys is legitimacy for a slab-based decision inside a horseshoe estimator: the nested path in Equation 10 is the same for every member of the family, so the slab fixes only the scale at which the multiplicity penalty is paid, and any finite-sample statement about the step-up rule made for one strictly increasing score holds for all of them.

Proof. For a mixing distribution π on λ_j^2, the marginal density of y_j under $\theta_j \neq 0$ is $\int (\kappa/\sigma^2)^{q/2}(2\pi)^{-q/2} e^{-T_j\kappa/2}\, d\pi(\kappa)$, where π is transported to κ through Equation 7. Under $\theta_j = 0$ the density is the same expression with $\kappa = 1$. Hence $b_\pi(T) = \log \int \kappa^{q/2} e^{-T\kappa/2}\, d\pi(\kappa) + T/2$. The integral is a Laplace transform in $T/2$ with respect to the measure $\kappa^{q/2} d\pi(\kappa)$, so its log-derivative is $-\mathbb{E}_\pi(\kappa \mid T)/2$ and its second log-derivative is $\operatorname{Var}_\pi(\kappa \mid T)/4$, where the expectation is under the posterior of κ given T induced by π. Adding $T/2$ gives $b'_\pi(T) = \{1 - \mathbb{E}_\pi(\kappa \mid T)\}/2$, which is positive because $\kappa < 1$ on the support, and $b''_\pi(T) = \operatorname{Var}_\pi(\kappa \mid T)/4$. Taking π to be the point mass at $\lambda_j^2 = v/\tau^2$, so that $\kappa = \sigma^2/(\sigma^2 + v)$ with probability one, the integral collapses and $b_\pi(T) = \frac{q}{2}\log\{\sigma^2/(\sigma^2+v)\} + \frac{T}{2}\{1 - \sigma^2/(\sigma^2+v)\}$, which is Equation 9. Because every b_π is strictly increasing in T, the ordering of rows by $b_\pi(T_j)$ coincides with the ordering by T_j, hence by $\bar{\kappa}_j$ through Lemma 1, so the nested path in Equation 10 does not depend on the member of the family.

B.3 Proof of Proposition 1

Both $H(R_r(z(h))) - \mu$ and $H(z(h)) - \mu$ lie in the row span of $C^\top$. Their difference vanishes after right multiplication by $I - P_C$. Therefore

$$\{I_r(h) - h\}(I - P_C) = \{H(R_r(z(h))) - H(z(h))\}(I - P_C) = 0.$$

Full replacement instead gives $\{H(R_r(z(h))) - h\}(I - P_C) = -(h - \mu)(I - P_C)$, which deletes the residual regardless of the route.

B.4 Expected NLL and KL

For reference output $p(\cdot \mid h)$ and edited output $q_r(\cdot \mid h)$,

$$\mathbb{E}_{Y\sim p}\{-\log q_r(Y \mid h)\} = H\{p(\cdot \mid h)\} + \mathrm{KL}\{p(\cdot \mid h) \| q_r(\cdot \mid h)\}.$$

Subtracting correct from wrong cancels entropy. The synthetic study evaluates this expectation directly.

C Complete configurations

The floor-free robustness runs change PCA dimension to 48 and retain local width/top-k 20/4 and private width/top-k 10/3. They use estimator seeds 2301, 2303, 2309, 2311, 2333, 2339, 2341,

Table 3: Frozen language-model configuration.

Quantity	Value
Maximum tokens per domain	6,000
Training tokens, prose/code	3,800 / 200
Evaluation tokens per domain	800
Evaluation blocks	8 per domain, 100 tokens each
PCA dimension	32
Local replicas; width / top-k	3; 20 / 4
Private width / top-k	10 / 3
Estimator seeds	2301, 2303, 2309, 2311, 2333
Control seeds	9101, 9103, 9109, 9111, 9133
Model dtype/device	float32 / CPU

2347, 2351, and 2357, with paired control seeds 9101, 9103, 9109, 9111, 9133, 9137, 9151, 9157, 9161, and 9173. All other fitting and evaluation sizes remain unchanged. The randomized-route mask is deterministic given the control seed and assigns exactly half of scored positions, up to the unavoidable one-position parity difference within a 99-target block, to the wrong expert. The final token is always assigned correctly because it has no next-token target.

GPT-2 revision is `607a30d783dfa663caf39e06633721c8d4cfcd7e`; Pythia-70M-deduped revision is `e93a9faa9c77e5d09219f6c868bfc7a1bd65593c`. FastICA uses at most 1,000 iterations. Shared anchoring strength is 1.0. SAE fitting uses batch size 128 and learning rate 0.003, with 400 local and 300 conditional steps.

RxRx1 uses 25,441 training wells from 23 experiments and 13,263 evaluation wells from 12 separate experiments. Two sites are averaged per well. The 128-dimensional embeddings are mapped to 32 PCA coordinates, retaining 71.1% of training variance. A softmax temperature is calibrated on 8,000 training wells by leaving one training experiment out at a time. All five fits select shared rank eight.

D Additional language-model results

D.1 Seed-level repository-corpus effects

Table 4: Balanced wrong-minus-correct NLL by estimator seed on the repository corpus.

Cell	2301	2303	2309	2311	2333
GPT-2 index 3	.970	.541	.843	1.046	.660
GPT-2 index 7	.201	.419	.621	.279	.316
GPT-2 index 11	1.167	.146	.361	.995	.424
Pythia index 2	.207	.242	.349	.440	.286
Pythia index 4	.227	.323	.393	.277	.506
Pythia index 6	.347	.352	.471	.366	.275

D.2 Two versions of the repository corpus

Four cells were measured on both versions with identical model, layer, hyperparameters, and estimator seeds. All 20 effects are positive on each version. Pythia index 2 changes by $+0.002$ $[-0.192, 0.197]$, index 4 by -0.048 $[-0.124, 0.027]$, and index 6 by -0.019 $[-0.153, 0.116]$. GPT-2 index 7 changes by -0.364 $[-0.554, -0.175]$, from 0.731 to 0.367. These are versions of one source, not independent corpora.

D.3 INDEPENDENT-CORPUS SEED COUNTS AND SHARED-ONLY CONTROL

On independent mypyc/CPython text, Pythia indices 2, 4, and 6 each have 5/5 positive seed contrasts; GPT-2 indices 3, 7, and 11 have 4/5, 4/5, and 3/5. Shared-only-minus-correct remains positive at all Pythia layers. It is positive at GPT-2 index 7, uncertain at index 11, and negative in all five seeds at index 3. These outcomes prevent a model-family-wide GPT-2 claim.

Selection chronology and file rule. The independent corpus was materialized at 14:23 on August 19, 2026, and all six resolved configurations were written at 14:24, before the first model result completed at 14:25. The source rule was specified by directory rather than by individual file: all reStructuredText and Markdown files in the installed mypyc documentation directory and all Python files in the CPython `json` package were copied. The manifest stores source paths, byte counts, and SHA-256 hashes. Thus the corpus was immutable before evaluation and its membership was not selected using route outcomes. The experiment establishes provenance independence and technical prose/code transfer; it does not establish representativeness of prose and code in general.

D.4 FLOOR-FREE SEED-LEVEL EFFECTS

Table 5: Pythia index-4 floor-free ($q = 48, r = 20$) outcomes. WC is wrong-minus-correct NLL; RC is 50%-randomized-minus-correct NLL.

Corpus / seed	2301	2303	2309	2311	2333	2339	2341	2347	2351	2357
Repository WC	.287	.247	.528	.338	.261	.087	-.075	.112	.282	.201
Repository RC	.124	.049	.196	.147	.049	.002	-.047	.026	.069	.066
Independent WC	.291	.425	.363	.423	.530	.241	.446	.582	.533	.489
Independent RC	.143	.172	.151	.225	.220	.102	.132	.297	.270	.210

The corresponding repository shared ranks are 2, 4, 3, 3, 3, 2, 4, 2, 4, and 4; independent-corpus ranks are 4, 5, 4, 5, 5, 4, 5, 3, 4, and 5. Completed-run hashes, all block-level cells, spectra, posterior summaries, and seed-level contrasts are retained in the machine-readable artifacts.

E CONTROL COMPARISONS

Correct-versus-wrong asks whether route identity matters. Shared-only-versus-correct asks whether the routed private reconstruction adds predictive information. Correct-versus-unmodified quantifies generic reconstruction damage and is not ownership evidence. The 50%-randomized condition assigns half the scored positions to each of the correct and wrong experts under the same residual-preserving operator; it is a route-intensity control, not a new estimator. Direct private ablation versus a per-block rank- and energy-matched random subspace asks whether the selected private span is unusually necessary under a destructive intervention. The latter is heterogeneous: balanced private-minus-random NLL is 0.667 at Pythia index 2, -1.751 at index 4, 4.419 at index 6, and -1.692 at GPT-2 index 7. This is why the paper does not combine ablation with operator-matched reconstruction routing.

F SYNTHETIC EXPERIMENTS

The spectral robustness grid has 1,370 cells covering rare-sample mass, imbalance, downstream private strength, private-subspace similarity, noise, shared-rank misspecification, PCA rank, and identical-domain nulls. The interaction grid has 1,080 cells covering rare-sample-by-noise and rare-sample-by-similarity interactions, latent sparsity, local rank, and output vocabulary. Most settings use 30 independent replicates; the identical-domain null uses 200. Each cell has 256 evaluation observations per domain, and expected NLL is computed under the unedited categorical distribution.

The 40-cell full-estimator study fits the complete consensus, grouped horseshoe, nested selection, shared refinement, and private SAEs ten times at each of four training allocations. It stays positive with 40, 80, 200, and balanced 800 rare-domain observations. The two larger grids use covariance

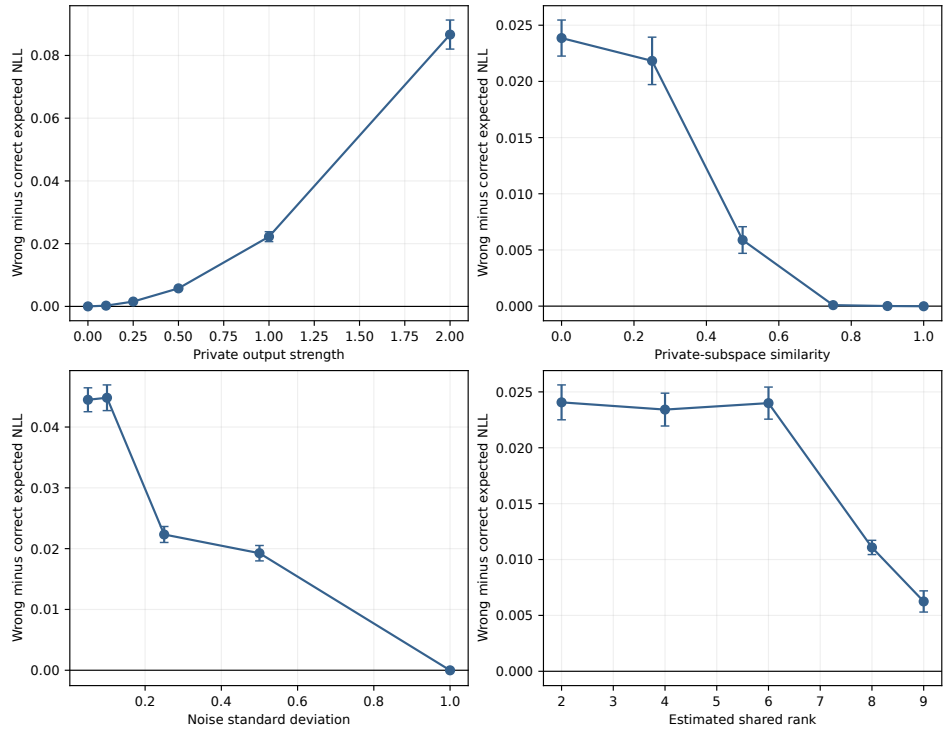

Figure 5: Synthetic routing effects across output relevance, private-subspace similarity, noise, and rank specification. Constructed nulls return effects centered at zero.

eigenspaces as a spectral proxy, permitting broad calibration while the smaller study verifies that the complete estimator preserves the mechanism.

G RxRx1 Supplementary Results

The endpoint is the experimental siRNA condition, not an isolated gene effect. After averaging over estimators, wrong-minus-correct NLL is positive in every held-out experiment: HUVEC batch effects range from 0.177 to 0.442 and RPE effects from 0.136 to 0.253. Correct-minus-wrong top-1 is also positive in all batches, ranging from 4.22 to 8.74 percentage points for HUVEC and 3.43 to 5.45 for RPE. Mean reciprocal rank improves by 3.92 points $[3.64, 4.20]$.

H Reproducibility and Artifact Structure

Every completed model run contains its resolved configuration, exact model and tokenizer revision, package versions, device and dtype, corpus source and token hashes, fitted activation transform, serialized estimator weights and identifier, token IDs and masks, block-level cell table, and summary. Existing completed directories are immutable. Analysis scripts recompute cell summaries, verify manifest hashes and no-op gates, and write machine-readable JSON. The floor-free analysis additionally verifies all 20 energy/shrinkage orderings and selected supports and exports the seed-level audit table. Figure scripts read those analyses. A manuscript audit checks numerical claims, citation keys, and figure pairs. The independent corpus is stored with source paths and per-file hashes rather than read live.

All language runs used 32-bit CPU inference and estimation; RxRx1 used the archived 128-dimensional embeddings rather than image-model training. The current manifests do not contain trustworthy wall-clock start and stop fields, and atomic staging makes filesystem timestamps unsuitable for reconstructing runtime after the fact. We therefore report the compute mode but do not

fabricate elapsed-time estimates. Future reruns should record wall-clock duration and peak memory in the manifest.