DIFFERENT FEATURES, REPRODUCIBLE RESPONSES: SPARSE AUTOENCODERS ACROSS MODEL CHANGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) represent model activations using sparse combinations of learned features. Each learned feature has a latent activation; features are often compared across dictionaries by matching these latents. We ask whether this is sufficient to assess the reproducibility of an observed activation response. Across independently fitted sparse dictionaries before and after model adaptation, compact combinations recover responses that no affine single-latent predictor recovers to the same accuracy. The singleton comparator searches the entire target dictionary and optimizes directly on evaluation inputs; group memberships and coefficients are fixed beforehand. We compare three ways to select groups: unbalanced optimal transport (UOT), descriptor cost, and calibration-response correlation. Across TinyLlama and Qwen3, with five fits per endpoint, compact groups increase recovery across all target fits. Token responses and unchanged-model refits show the same qualitative separation, and simple cost-selected groups outperform UOT groups. These findings distinguish the recoverability of an observation from the reappearance of an individual latent. Cross-dictionary references can retain compact realizations and their measured accuracy, allowing response evidence to accumulate without a canonical feature vocabulary.

1 INTRODUCTION

SAE interpretability increasingly uses named features across analyses: dashboards and feature atlases label them (Lin, 2023; Templeton et al., 2024), model-diff reports compare them (Lindsey et al., 2024), persistence analyses track them, and causal studies ablate them (Marks et al., 2025). Those uses raise a reproducibility question: which properties of a feature remain recoverable across model adaptation and independent SAE fitting? SAE permutation symmetry permits arbitrary reindexing, so index agreement alone cannot answer this question. Independent fits can also learn different decompositions. Even after finding the best counterpart, a single target latent may fail to reproduce the response that motivated the original observation.

A feature observation concerns a measured response pattern: which inputs activate a latent and how its activation varies across them. The latent belongs to one fitted dictionary; the observation can be tested through latents learned in other fits. A cross-dictionary relation identifies candidate counterparts; their responses on new inputs test whether the pattern recurs. We call a calibrated predictor of the source response a realization of the observation. It may use one target latent or a compact combination, with coefficients fitted on separate inputs. This separates the response we want to study from the dictionary that first exposed it. Unbalanced optimal transport (UOT) supplies one construction of the candidate relation; we compare it with simpler alternatives.

We study fine-tuning trajectories from a shared pretrained base. At each checkpoint we fit a fresh SAE and compare groups selected by UOT, descriptor cost, and calibration-response correlation. UOT estimates a joint weighted relation while allowing the dictionaries to assign different activity masses to corresponding features. Same-checkpoint refits test whether singleton limitations also arise without model adaptation. For each of two models, we train five sparse dictionaries independently at each endpoint and test all 25 cross-checkpoint relations on documents excluded from SAE fitting, calibration, and matching.

To distinguish a matching failure from a limitation of single-latent references, we select the best of all 4096 target latents and optimize its scale and offset on evaluation inputs. This bounds the recovery

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.

attainable by any affine singleton matcher on those inputs. In TinyLlama, separately calibrated four-latent UOT groups exceed that ceiling, recovering document responses in all five target fits for 31.0% of eligible source coordinates, versus 19.3% for the oracle at normalized squared error at most .25. A fixed-protocol Qwen3 replication, token responses, and unchanged-model refits show the same qualitative separation. Some observations have adequate singleton realizations in some fits but need combinations in others. Requiring a single-latent successor in every fit would therefore miss responses that remain recoverable across all five fits. Additional correspondence studies cover four model families.

Contributions. We test response recovery by comparing compact, separately calibrated predictors with an evaluation-optimized affine-singleton ceiling. Across independently fitted TinyLlama and Qwen3 dictionaries, we measure recovery beyond that ceiling under adaptation and unchanged-model refitting, accounting for all eligible source responses. We compare three group selectors under one calibration procedure. Transport and coefficient comparisons distinguish candidate selection, activity-mass constraints, and response weighting. The resulting reference records the source observation, its realizations in other dictionaries, and their evaluation errors.

2 Related Work

SAEs decompose model activations into sparse, inspectable directions (Bricken et al., 2023; Huben et al., 2024) and support large feature atlases, released dictionaries, and benchmarks (Gao et al., 2025; Templeton et al., 2024; Lieberum et al., 2024; Karvonen et al., 2025). SAE objectives and operating points matter: JumpReLU, Top-k, and BatchTopK variants change reconstruction, sparsity, and interpretability tradeoffs (Rajamanoharan et al., 2024; Bussmann et al., 2024); feature splitting and absorption further shape which latents an SAE forms (Chanin et al., 2025). Sparse feature circuits provide complementary causal tools given a fixed SAE (Marks et al., 2025).

Leask et al. (2025) demonstrate incompleteness and non-atomicity across dictionary sizes, challenging canonical units of analysis. Paulo and Belrose (2026) study dependence on initialization: in their 131K-latent Llama 3 8B example, only $\sim$30% of latents share a counterpart under both encoder and decoder Hungarian matchings at cosine ≥ 0.7, with substantial orphan fractions over the tested seed counts. Our question is whether the source response remains recoverable when both the model and its independently fitted dictionary change, even if no single latent reproduces it. Gurnee et al. (2024) give the model-neuron analog (1–5% of GPT-2 neurons universal across seeds). Song et al. (2026) report high assignment-based decoder consistency with suitable SAE architectures. Their dictionary-level average does not determine whether a particular response has an adequate singleton realization.

Gerasimov et al. (2026) study feature reappearance and functional importance, and show that individually unstable features can span reproducible decoder subspaces. Their synthetic example explains how basis ambiguity can preserve shared structure despite unstable individual latents. Our response-recovery test asks a further question: can a specific input-dependent activation be reconstructed in another fit, at a stated accuracy, from coefficients fixed before evaluation? Decoder-subspace overlap alone does not determine this because latent activations also depend on the encoder. The affine-singleton oracle distinguishes a need for multiple latents from unsuccessful matching. A recent cross-seed benchmark reports comparisons of cosine and Sinkhorn matchers using ablation fingerprints and feature substitution across coverage levels (Anonymous Authors, 2026).

Lan et al. (2024) find similarities between SAE feature spaces after activation-based matching across models. Related alignment methods address permutation symmetries, linear mode connectivity, and orthogonal shape metrics (Ainsworth et al., 2023; Entezari et al., 2022; Williams et al., 2021). Khosla and Williams (2024) compare neural-unit responses by balanced OT with uniform masses; Kapoor et al. (2026) allow partial correspondence with capped marginals and a selected total matched mass. We use activity-dependent SAE masses with KL relaxation and test response corroboration across independently fitted checkpoint dictionaries.

Fine-tuning can enhance existing entity-tracking mechanisms (Prakash et al., 2024); sparse parameter grafting can localize task-specific updates (Panigrahi et al., 2023). Crosscoders jointly learn shared sparse feature coordinates across base and fine-tuned models, with model-specific decoder vectors (Lindsey et al., 2024). Their training objective can create apparent model-specific features (Minder

et al., 2025); Delta-Crosscoder targets narrow fine-tuning changes through a shared dictionary and a change-focused loss (Kassem et al., 2026). SAE-Track (Xu et al., 2025) instead initializes each checkpoint SAE from its predecessor. Reusing a base-model SAE can also reconstruct an adapted model well in some settings (Kissane et al., 2024). This retains one dictionary. We instead ask whether an observation can be corroborated after the dictionary is learned independently, so we train the dictionaries separately and recover their relation afterward. We then test what activation structure the inferred relations preserve across independent fits; same-seed transitions provide supporting comparisons. Cao et al. (2026) use Wasserstein distances between activation-weighted hidden-state distributions to compare individual SAE features across layers. Our transport instead couples activity references across the features of two dictionaries.

3 METHOD

3.1 RECOVERING A RESPONSE IN ANOTHER DICTIONARY

A response observed in one dictionary may lack a good individual counterpart in another. We test whether the response can nevertheless be recovered from a small group of target latents. Comparing that prediction with the best single latent separates a failure of individual matching from a failure to recover the observation.

Let $f_i(x)$ be source latent i's response to input x, and let $z_j^t(x)$ be target latent j's response to the same input. The response is either a document's mean activation over content tokens or token-level activation over aligned content-token positions; each convention receives a separate fit. Content tokens exclude padding and position zero. For a selected target set J_{ir} in dictionary r, we fit

$$\widehat{f}_{ir}(x) = b_{ir} + \sum_{j \in J_{ir}} w_{irj} z_j^{t,r}(x), \qquad w_{irj} \geq 0, \quad |J_{ir}| \leq k. \tag{1}$$

Descriptor inputs select the UOT and cost-based memberships. Separate calibration inputs fit the nonnegative coefficients and free intercept by least squares. We freeze these predictions before evaluation. The primary budget is $k = 4$; we also report $k = 1, 2, 8$. A fitted group may use fewer than k nonzero coefficients. Transport masses select candidates but are distinct from the response coefficients w_{irj}; Sections 3.2–3.3 specify these selections.

We divide the frozen prediction's mean squared error by the source response's evaluation variance, obtaining normalized squared error e_{ir}. An error at most $\tau = .25$ means evaluation $R^2 \geq .75$; we report the full tolerance curve. With $\bar{f}_i$ the evaluation mean, the singleton oracle is

$$e_{ir}^{\text{single}} = \min_{j,\alpha,\beta} \frac{\sum_x [f_i(x) - \alpha z_j^{t,r}(x) - \beta]^2}{\sum_x [f_i(x) - \bar{f}_i]^2} = 1 - \max_j \operatorname{corr}(f_i, z_j^{t,r})^2. \tag{2}$$

It considers every target latent, either slope sign, and evaluation-optimal scale and offset. For this formula we assign zero correlation to constant targets, which permit the constant predictor; constant source responses have undefined normalized error. Thus no affine singleton selection procedure can improve upon this oracle on the evaluated corpus. If the oracle exceeds τ but a frozen four-latent predictor meets it, the smallest adequate nonnegative affine support satisfies $2 \leq k_{ir}^{\star}(\tau) \leq 4$ on that corpus. This conclusion concerns the stated response and predictor class, not arbitrary nonlinear transformations or a population impossibility theorem.

For each source observation, we count the target dictionaries in which its frozen realization meets the tolerance. Recovery in all five fits requires five successes. We separately count observations for which groups succeed in every fit but the singleton oracle fails in one or more. These observations need multiple latents in at least one dictionary; they need not require groups in every dictionary. Failed and undefined outcomes remain in the denominator.

The five highest-response documents for Figure 1's source concern photographs or imaging, including a report on Fujifilm and Sprint's photo-printing service. In target fit B_3, the calibrated realization uses two nonzero terms, $\widehat{f} = -.000304 + .417088 z_{1571} + .435091 z_{317}$. Its error is .172, while the best affine singleton has error .351. In B_1 the singleton has error .104 and is more accurate than the group. The observation admits different adequate realizations across fits; it does not require every realization to contain several latents.

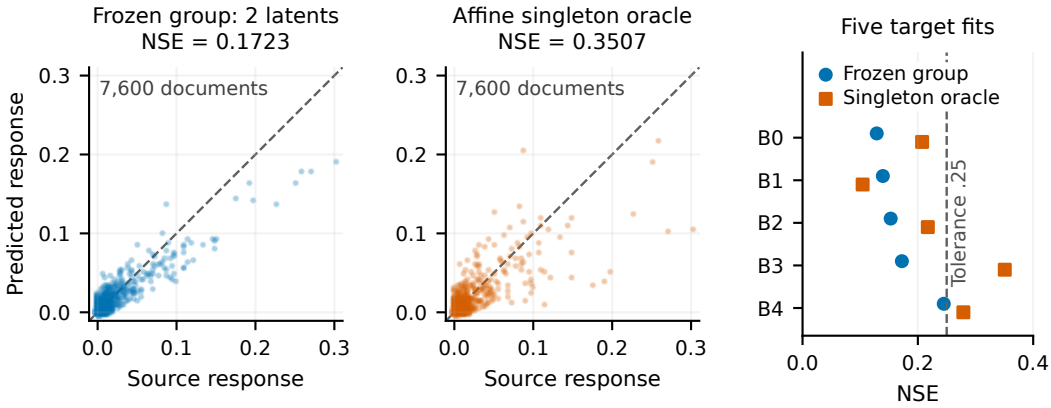


Figure 1: One response with different realizations across target dictionaries. The scatter plots show all 7,600 evaluation documents for TinyLlama source A_0 feature 8 and target fit B_3, on identical axes. NSE is normalized squared error. The right panel shows the same source in all five adapted fits. The source is the lowest index whose groups succeed in all five fits at error .25 while the singleton fails in at least two; B_3 is the lowest-index fit where the singleton fails.

3.2 Problem setup and descriptors

Each checkpoint receives its own SAE. Superscripts s and t identify source and target quantities; cross-checkpoint comparisons run from earlier to later checkpoints. Appendix D.1 collects notation and terms. Descriptors determine matches; local indices identify endpoints and define the index-based controls. For the primary TinyLlama and Qwen3 studies, decoder directions are compared in the native residual coordinates of each base–adapted model pair; no signed permutation or Procrustes transform is applied. Appendix F.3 reports separate alignment checks for an earlier TinyLlama L11 pair (Schönemann, 1966).

For source feature i and target feature j, the descriptor cost combines normalized decoder-direction similarity, a fixed-corpus activation profile, prevalence (the fraction of tokens on which a feature activates), and activation magnitudes. Write $z_i(n,u)$ for feature i's activation at token u in document n. Let $\bar{z}_i = \mathbb{E}[z_i]$ and $q_i = \mathbb{E}[z_i^2]$ denote its mean activation and raw second moment on the descriptor corpus. The raw second moment measures mean squared activation without subtracting the mean. A nearly constant but active feature can therefore receive substantial mass even when its variance is small.

Here $\hat{d}_i$ is the unit decoder direction: it specifies the direction, rather than the magnitude, of feature i's contribution. For ordered documents $n = 1, \ldots, N$, let T_n contain nonpadding token positions, excluding position zero in the primary study. Token conventions for the additional studies are specified in Appendices E.10 and C.13. Define document mean $\mu_{in} = |T_n|^{-1} \sum_{u \in T_n} z_i(n,u)$ and active fraction $r_{in} = |T_n|^{-1} \sum_{u \in T_n} \mathbf{1}[z_i(n,u) > 0]$. Prevalence $\pi_i = M^{-1} \sum_n |T_n| r_{in}$ pools all $M = \sum_n |T_n|$ tokens, as do the expectations defining $\bar{z}_i, q_i$. The profile concatenates $\tilde{p}_i = (\mu_{i1}, \ldots, \mu_{iN}, r_{i1}, \ldots, r_{iN})$. If $2N > 64$, project to 64 dimensions with one fixed Gaussian matrix Φ shared across checkpoints: $p_i = \Phi \tilde{p}_i$. Otherwise set $p_i = \tilde{p}_i$. Normalize nonzero p_i to $\hat{p}_i$. The primary study assigns cosine zero to a zero profile; conventions for the additional studies are detailed in Appendix C.2. The descriptor cost used by UOT and cost selection is

$$C_{ij} = w_{\text{dir}} \cdot \frac{1 - \left\langle \hat{d}_i^s, \hat{d}_j^t \right\rangle}{2} + w_{\text{prof}} \cdot \frac{1 - \left\langle \hat{p}_i^s, \hat{p}_j^t \right\rangle}{2} + w_\pi \left| \pi_i^s - \pi_j^t \right| + w_\mu \frac{\min\left(\left| \log(\bar{z}_i^s + \delta) - \log(\bar{z}_j^t + \delta) \right|, 5 \right)}{5} + w_q \frac{\min\left(\left| \log(q_i^s + \delta) - \log(q_j^t + \delta) \right|, 5 \right)}{5}. \tag{3}$$

The default weights are $w_{\mathrm{dir}} = 1.0$, $w_{\mathrm{prof}} = 0.5$, and $w_\pi = w_\mu = w_q = 0.25$, with log stabilizer $\delta = 10^{-9}$. We choose these weights heuristically, prioritizing decoder direction, then activation profile, with smaller weights on prevalence and the two magnitude summaries. Default UOT hyperparameters are candidate edge budget $k_{\mathrm{cand}} = 8$, entropic regularizer $\varepsilon = 0.01$, and unbalanced KL weights $\rho_s = \rho_t = 0.5$. The candidate budget limits alternatives, ε controls how concentrated the plan is, and equal ρ values penalize marginal deviations symmetrically.

3.3 Candidate selection and reference records

For each source, cost selection takes the k lowest-cost targets in the graph defined below, UOT selection takes the k targets with largest plan mass, and response selection takes the k largest absolute calibration correlations over the full target dictionary. Cost and UOT candidates require positive raw second moment on the descriptor corpus before the numerical mass floor is applied. All three receive the same coefficient fit and evaluation.

UOT construction. UOT supplies a joint candidate relation and activity-mass diagnostics. Balanced OT also permits splitting and merging; our reason for relaxing its marginals is that separately fitted dictionaries need not assign the same activity mass to corresponding features, and some features may lack a suitable counterpart. We test the effect of these constraints on response recovery separately in Section 5.5. For example, the constructed split $dz = d(z/2) + d(z/2)$ preserves reconstruction but halves the summed raw second moment (Appendix C.3). We use these moments as soft activity references. Let $a_i = q^s_i / \sum_{i'} q^s_{i'}$ and $b_j = q^t_j / \sum_{j'} q^t_{j'}$ be source and target feature masses, assuming positive total second moment in each dictionary. The primary study floors each raw moment at 10^{-12} before normalization; conventions for the additional studies are specified in Appendix C.2.

The primary Top-k fits and the six additional TinyLlama/Qwen3 endpoint pairs have unit-norm decoder columns; their moments are measured in that stored latent scale. Rescaling a decoder column and inversely rescaling its activation would preserve reconstruction but change q_i, so a correspondence record must specify the latent scale. The support $\mathcal{S}$ is the union of the eight lowest-cost targets per source and the eight lowest-cost sources per target. These neighbors need not admit a plan with the exact source and target marginals. The plan trades off descriptor agreement against discrepancies in the activity assigned by each dictionary. For these inputs, we solve

$$P^\star := \operatorname*{arg\,min}_{\substack{P \geq 0 \\ \mathrm{supp}(P) \subseteq \mathcal{S}}} \left(\langle C, P \rangle + \varepsilon \sum_{(i,j) \in \mathcal{S}} P_{ij}(\log P_{ij} - 1) + \rho_s \, \mathrm{KL}(P\mathbf{1} \,\|\, a) + \rho_t \, \mathrm{KL}(P^\top \mathbf{1} \,\|\, b) \right), \tag{4}$$

where $\langle C, P \rangle = \sum_{(i,j) \in \mathcal{S}} C_{ij} P_{ij}$ is total descriptor cost, $\mathbf{1}$ is the all-ones vector, and $\mathrm{KL}(x \| y) = \sum_\ell [x_\ell \log(x_\ell / y_\ell) - x_\ell + y_\ell]$ is generalized KL divergence. The objective is solved by generalized Sinkhorn updates with kernel $G_{ij} = \exp(-C_{ij}/\varepsilon)$ on $\mathcal{S}$, following the entropic-OT formulation of Cuturi (2013) and the unbalanced extensions of Chizat et al. (2018), Séjourné et al. (2019); Singh and Jaggi (2020) use OT to align and fuse independently trained networks. The row and column sums of $P^\star$ may differ from a and b. Multiple entries may connect one source to several targets or several sources to one target; interpreting these connections as feature splitting, merging, or disappearance requires further evidence. Appendix E.1 defines separate transition summaries using these activity references.

Response weights and reference records. On the UOT memberships we also compare uniform and transport-relative coefficients, fitting only a nonnegative global scale and an intercept. These comparisons separate support selection from response weighting.

A reference starts with a source dictionary, a local feature index, and a specified response observation. For each target fit it records the candidate relation, fitted support, coefficients, intercept, and evaluation error. Low accuracy, an undefined measurement, and no proposed counterpart are distinct outcomes. The UOT relation also retains source and target activity references and transported marginals; their discrepancies may have either sign. Changing the response convention or construction defines a new reference version. Adjacent relations need not compose into a semantic identity. Earlier selected-link experiments extract a partial one-to-one core and measure response correlations and activating-document overlap; their selector and denominators are given separately in Appendix B.1.

4 EXPERIMENTAL SETUP

Primary sparse studies. We study TinyLlama L11 and Qwen3-1.7B L14 before and after AG News classification fine-tuning, at steps 0 and 3000. For each model, we independently train five width-4096 Top-k SAEs per endpoint, with at most 64 active features per token. Within each model, base dictionaries $A_0, \ldots, A_4$ use seeds 101–105 and adapted dictionaries $B_0, \ldots, B_4$ use seeds 201–205. No fit is initialized from another. We evaluate all 25 base-to-adapted pairs and, at each checkpoint, each of the ten fit pairs with the lower-index fit as source. Each model thus contributes 45 comparisons among ten dictionaries on one adaptation.

The SAE-training, calibration, descriptor, and evaluation partitions contain 15,982, 998, 1,998, and 7,600 documents after exact-text deduplication and a fixed near-duplicate filter. Qwen3 repeats the fixed response protocol on the same documents, retokenized for that model. Sparsity and the common training budget are chosen using reconstruction calibration alone; every fit receives approximately 12 million token presentations. These partitions separate evaluation from new SAE fitting, calibration, and matching; they do not establish exclusion from language-model pretraining or the earlier fine-tuning run. Appendices A.1 and A.3 give training diagnostics, provenance, and measured residual change.

The 998 calibration documents also fit response coefficients. Source eligibility requires positive responses on at least twenty descriptor and ten calibration documents and calibration variance above 10^{-12}. It retains 20,474 source coordinates in TinyLlama and 19,972 in Qwen3, out of 20,480 for each model, without selecting on singleton success. All memberships and coefficients were locked before recovery errors were evaluated. The evaluation corpus had already been used for the earlier correspondence study (Appendix A.2).

Additional correspondence studies. These compare TinyLlama, Qwen3, Llama-3.2, and Gemma after low-rank adaptation, plus TinyLlama after full fine-tuning. Their denser dictionaries, response conventions, and unavailable SAE-training document identities require separate reporting (Appendices C.7 and C.8); five-checkpoint comparisons are in Appendix E.

5 RESULTS

We first compare compact response recovery with the affine-singleton ceiling, then test recurrence across fits and recovery in an unchanged model. We distinguish these findings from the contribution of the particular relation construction.

5.1 RECOVERY BEYOND THE AFFINE-SINGLETON CEILING

Compact realizations recover responses that are unavailable to every affine singleton at the same accuracy. At normalized squared error at most .25, four-latent UOT groups increase all-five document recovery over the singleton oracle in both models (Table 1). Token responses show the same separation. Each group's coefficients were fixed on separate calibration inputs; the oracle instead chooses and fits its best latent on evaluation.

In TinyLlama, 12.5% of eligible source coordinates are recovered by groups in all five target fits while the singleton oracle fails in at least one; 0.8% show the reverse all-five outcome. Among the document responses recovered by UOT groups in all five fits, the singleton oracle fails in at least one fit for about two-fifths in each model.

The improvement appears in every source fit of each model. Averaged over the 25 source–target pairs, document recovery is 48.4% for four-latent UOT groups versus 37.1% for the affine-singleton oracle in TinyLlama, and 37.4% versus 28.8% in Qwen3. Figure 2 shows the TinyLlama error-tolerance curves; Appendices A.2 and A.3 report per-fit counts and paired outcomes.

5.2 ACCURACY, GROUP SIZE, AND RECURRENCE

Among the TinyLlama records newly recovered in all five fits, median worst-fit error falls from approximately .33 for the singleton oracle to .20 for the groups. Appendix A.2 reports the paired error reductions within these newly complete records.

Table 1: Source responses recovered in all five adapted dictionaries at normalized squared error at most .25 (%). Each entry averages five source-fit fractions over their fitting-defined eligible populations. Groups contain at most four latents and are frozen before evaluation. The frozen singleton selects and fits one signed affine predictor on calibration inputs; the oracle instead selects and fits on evaluation. Document and token predictors are fitted separately. Random-group entries average 20 fixed document-only draws; token controls were not run (Appendix A.5).

	TinyLlama		Qwen3	
Method	Document	Token	Document	Token
Frozen singleton	17.6	7.5	12.5	4.5
Affine singleton oracle	19.3	7.6	15.0	4.7
UOT groups	31.0	15.1	23.3	9.2
Cost groups	33.3	15.7	25.2	9.7
Response-selected groups	32.4	16.5	23.4	10.1
Random groups	0.0	–	0.0	–

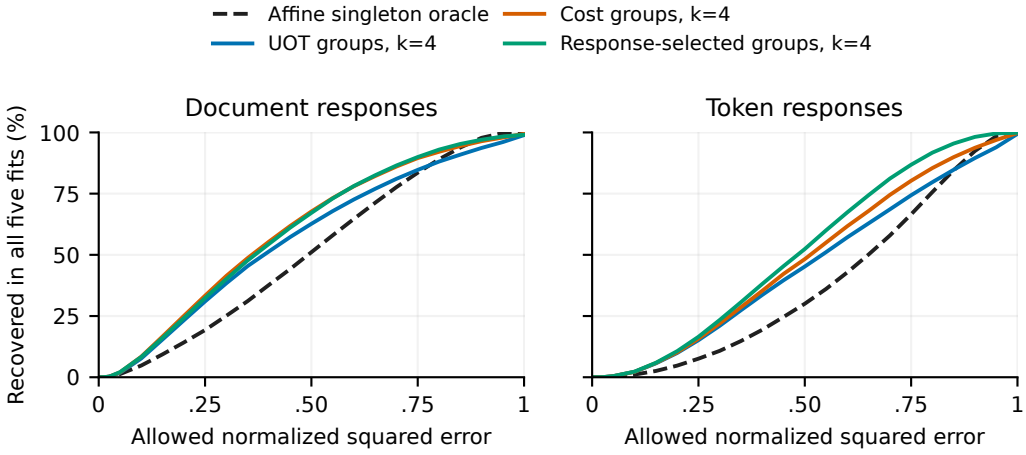


Figure 2: TinyLlama recovery across the complete declared error-tolerance grid. Each curve averages five source-fit fractions, requiring success in all five target dictionaries. Group predictions are frozen before evaluation, whereas the affine singleton is optimized on evaluation. Cost and calibration-response selection receive the same four-feature budget and coefficient fit as UOT. The curves describe one model adaptation, not uncertainty across independent adaptations.

Among the newly complete TinyLlama document records, the affine singleton exceeds error .25 in one target fit for 45.6%, in two to four for 46.4%, and in all five for 8.0%. Qwen3 has a similar distribution (Appendix A.3). All five target fits describe the same adapted model: whether an affine singleton meets the declared accuracy can depend on the SAE fit. In both models, two-latent UOT groups already exceed the singleton ceiling; increasing the budget to four and eight raises recovery further. Token responses show the same ordering.

5.3 Refitting alone can require compact realizations

The limitation also occurs when the transformer is unchanged. Across independent refits of both endpoints in each model, four-latent UOT groups increase per-pair recovery over the singleton oracle (Appendices A.2 and A.3). Even here, four-latent UOT predictors recover only about 38% of eligible token-response records in TinyLlama and 27–30% in Qwen3, averaged over refit pairs at the same error threshold. Thus adaptation is not necessary for a response to lack an adequate singleton realization. The cross-checkpoint results establish the additional recovery across a changed model. These cohorts reuse dictionaries and differ in their source populations; subtracting their means does not isolate a causal effect of adaptation.

5.4 SUPPORTING CORRESPONDENCE EVIDENCE

Individual correspondences also corroborate shared response patterns across fits. In the sparse TinyLlama study, selected links cover 55.4% of source coordinates per cross-checkpoint pair on average, with mean document-response correlation .813 on evaluation inputs. For sources linked in all five target fits, their counterparts have greater mean overlap in highly activating documents than re-paired controls (Appendices B.3 and B.4). Compact groups extend this evidence by recovering some responses in all five target fits when the affine-singleton oracle fails in at least one.

Additional correspondence studies cover four model families. In all six TinyLlama/Qwen3 comparisons with both endpoint SAEs independently reseeded, selected-pair overlap exceeds shuffled and frequency-matched controls. Their different dictionary densities, response conventions, and fitting-provenance limits are reported separately (Appendices F.4.3 and C.7).

5.5 WHAT THE RELATION CONSTRUCTION CONTRIBUTES

Cost- and calibration-response-selected groups both exceed UOT recovery in all four model–response settings in Table 1. Cost selection gives the highest document recovery, while calibration-response selection gives the highest token recovery. The response selector uses neither the descriptor cost in Eq. 3 nor transport, so the recovery beyond single latents also occurs without those weights or marginal constraints. In TinyLlama, cost selection recovers 97.4% of the complete UOT document records, as well as more records that UOT does not recover. The phenomenon is therefore shared by several constructions; these results do not attribute it specifically to joint transport allocation.

Uniformly sampling four descriptor-active latents with the same coefficient fit yields no all-five recovery of eligible document responses in either model across twenty fixed draws (Table 1). Extra coefficients alone do not reproduce the selected groups' recovery under this uniform baseline (Appendix A.5).

Table 2: Transport attribution on identical dense support. Entries are the percentage of A_0 responses recovered in all five adapted fits at error $\leq .25$, using at most four features. Unlike Table 1, these fractions concern one source fit. All three transports use the same cost, masses, entropy, and calibration.

Construction	Documents	Tokens
Affine singleton oracle	19.6	7.5
Dense UOT	30.0	14.5
Balanced OT	29.6	14.2
Rejection transport	33.9	15.8

To isolate the marginal constraints from candidate support, we also solve UOT, balanced OT, and transport with an explicit unmatched option on the same dense cost matrix. This comparison covers thirteen TinyLlama pairs: one source fit against all five adapted fits, plus four refits at each endpoint. All methods use the same activity masses, entropy, group budget, and coefficient fit. UOT and balanced OT give similar recovery; rejection transport recovers more responses (Table 2). All three exceed the singleton oracle in both response units. Thus the demonstrated benefit of compact realizations does not require KL marginal relaxation. Appendix A.4 gives the objectives, refit results, and numerical checks.

Activity references and transport constraints. The role of marginal relaxation is to accommodate different activity allocations. In six additional TinyLlama/Qwen3 comparisons, the candidate graphs cannot carry the exact activity marginals; one forces at least 13.07% of any balanced plan outside its graph (Appendix C.3). This explains a constraint of exact conservation on those graphs, not an advantage in primary response recovery. The accompanying partial-matching and constructed-correspondence comparisons include stronger rejection baselines (Appendices C.4 and C.9). Activity convention also matters: first-position contributions dominate some reference weights, motivating their exclusion in the primary study. Replacing second moment with variance in both cost and activity masses preserves above-shuffle correspondence in all six sensitivity pairs, whereas removing

decoder direction fails in three. Appendices C.13 and C.6 report the full mass and weight-sensitivity diagnostics.

Response calibration matters separately. With the same TinyLlama UOT memberships, uniform relative coefficients yield 4.1% all-five document recovery, transport-relative coefficients yield 19.1%, and freely calibrated nonnegative coefficients yield 31.0%. The first two variants each fit a global nonnegative scale and an intercept, so the difference concerns relative coefficients rather than an arbitrary overall scale. Qwen3 shows the same ordering (Appendix A.3). Transport-relative coefficients improve on uniform weighting, and fitting relative coefficients separately improves recovery further.

6 Discussion and Limitations

Implications for SAE research. Failure to find an adequate individual counterpart does not settle whether an observation is recoverable. The singleton oracle rules out better affine single-latent matching as an explanation for the rescued responses; compact predictions show how those responses can instead be measured in other dictionaries. A feature study can retain its source response, the fitted realizations at each checkpoint, and their errors. A realization may use one latent in one fit and several in another, so multiple counterparts do not by themselves show that adaptation split a representation. Refitting tests the response through another decomposition. Local indices remain addresses for inspection and intervention; the cross-dictionary reference connects the evidence.

This practice complements studies of individual-feature reappearance, which measure a different property. It also preserves the distinction between matching and validation: descriptor similarity proposes a relation, while evaluation responses test its usefulness. The successful calibrated predictions reproduce scalar responses; they do not establish semantic splitting, native decoder-contribution transfer, or interchangeable interventions. A separate activation-substitution study reports lower pooled spread of class-score margin changes across model settings and SAE fits for matched groups than for same-index, shuffled, and frequency-matched controls; mixed signs, perturbation size, and decoder geometry limit its interpretation (Appendix G).

Scope. The primary sparse evidence covers one layer and one adaptation in each of TinyLlama and Qwen3, with ten fitted dictionaries per model. Both use the same task and corpus; adaptation seeds and domains are not replicated. How this recovery gap changes with dictionary width, training budget, or reconstruction quality remains untested. Full-SFT evidence is limited to TinyLlama. Width or convention changes require a newly estimated relation; the separate Gemma Scope experiment does not validate transfer of existing families (Appendix F.5). Activation-overlap evidence is limited to the four layers studied here. The references are conditional on the source dictionary, response convention, and calibration distribution; arbitrary re-anchoring or path composition need not recover the same response. Calibration can absorb changes in scale and offset, so successful recovery does not itself quantify how the original feature's expression changed. Source coordinates are observations, not counts of distinct semantic concepts.

7 Conclusion

A response can remain recoverable across independently fitted dictionaries even when no individual target latent provides an adequate realization. Compact predictions calibrated separately recover responses beyond an evaluation-optimized affine singleton across adaptation and in unchanged-model refits. UOT supplies joint relations between dictionaries with different activity allocations; simpler selectors recover more responses overall. These findings support references that retain the observation, its compact realizations, and their measured accuracy. Researchers can then compare response evidence across decompositions without assuming a canonical feature vocabulary.

ETHICS STATEMENT

Feature tracking can support interpretability, debugging, and model comparison. Using inferred lineages to guide interventions requires attention to their assumptions and evidence; we report the fitted relations, activity-mass conventions, controls, and experimental coverage.

REPRODUCIBILITY STATEMENT

Section 4 and Appendices A.1–A.5 specify fitting, partitions, frozen predictions, evaluation, and controls. The supplement contains saved outcomes, example predictions, input hashes, and scripts reproducing the primary tables and numerical checks. Full training and inference require the external assets listed in Appendix H; Appendix I reports compute. Appendix C.8 documents provenance limits of the supporting studies, and Appendix C.3 gives the transport arguments and proofs.

AI USE STATEMENT

We used generative AI tools to help develop the conceptual framework and mathematical claims, draft and check proofs, refine hypotheses and experiments, implement methods and analyses, process data, prepare figures, review literature, interpret results, and draft and edit the manuscript. AI-assisted code generated the constructed correspondence controls algorithmically from feature descriptors. The authors are responsible for the final manuscript, claims, code, and other artifacts.

REFERENCES

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *ICLR*, 2023.

Anonymous Authors. Benchmarking cross-seed feature correspondence in sparse autoencoders. Anonymous manuscript, 2026. https://openreview.net/pdf?id=5cy6WtSC8f.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024.

Tue M. Cao, Nguyen Do, and My T. Thai. Semantic optimal transport for sparse autoencoder feature matching and circuit compression. *arXiv preprint arXiv:2605.28567*, 2026.

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 *NeurIPS*, 2025.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *ICLR*, 2022.

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 *ICLR*, 2025. arXiv:2406.04093.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal neurons in GPT2 language models. *Transactions on Machine Learning Research*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. arXiv:2106.09685.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *ICLR*, 2024.

Chaitanya Kapoor, Alex H. Williams, and Meenakshi Khosla. Partial Soft-Matching Distance for Neural Representational Comparison with Partial Unit Correspondence. In *ICLR*, 2026. https://proceedings.iclr.cc/paper_files/paper/2026/hash/f13d74c6666087aa6eea3d17820e6a23-Abstract-Conference.html.

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 *ICML*, 2025. arXiv:2503.09532.

Aly Kassem, Thomas Jiralerspong, Negar Rostamzadeh, and Golnoosh Farnadi. Delta-Crosscoder: Robust crosscoder model diffing in narrow fine-tuning regimes. *arXiv preprint arXiv:2603.04426*, 2026.

Meenakshi Khosla and Alex H. Williams. Soft Matching Distance: A metric on neural representations that captures single-neuron tuning. In *Proceedings of UniReps*, volume 243 of *Proceedings of Machine Learning Research*, pp. 326–341, 2024. https://proceedings.mlr.press/v243/khosla24a.html.

Connor Kissane, Robert Krzyzanowski, Arthur Conmy, and Neel Nanda. SAEs (usually) Transfer Between Base and Chat Models. *Alignment Forum*, 2024. https://www.alignmentforum.org/posts/fmwk6qxrpW8d4jvbd/saes-usually-transfer-between-base-and-chat-models.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Quantifying feature space universality across large language models via sparse autoencoders. *arXiv preprint arXiv:2410.06981*, 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 *ICLR*, 2025.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *BlackboxNLP Workshop*, 2024.

Johnny Lin. Neuronpedia: Interactive reference and tooling for analyzing neural networks. 2023. Software available from https://www.neuronpedia.org.

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.

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 *ICLR*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *ICLR*, 2017. arXiv:1609.07843.

Julian Minder, Clément Dumas, Caden Juang, Bilal Chugtai, and Neel Nanda. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. In *NeurIPS*, 2025. arXiv:2504.02922.

Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. Task-specific skill localization in fine-tuned language models. In *ICML*, 2023.

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

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *ICLR*, 2024.

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

Peter H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019.

Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. In *NeurIPS*, 2020.

Xiangchen Song, Aashiq Muhamed, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona T. Diab, Virginia Smith, and Kun Zhang. Mechanistic interpretability should prioritize feature consistency in sparse autoencoders. In *ACL*, pp. 2172–2210, 2026. https://aclanthology.org/2026.acl-long.99/.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Christopher Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, Anthropic, 2024.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*, 2020.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations. In *NeurIPS*, 2021.

Yang Xu, Yi Wang, Hengguan Huang, and Hao Wang. Tracking the feature dynamics in LLM training: A mechanistic study. *arXiv preprint arXiv:2412.17626*, version 3, 2025. https://arxiv.org/abs/2412.17626.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *ICLR*, 2024.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. TinyLlama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NeurIPS*, 2015. arXiv:1509.01626.

A PRIMARY EXPERIMENTS AND RESPONSE RECOVERY

Appendices A.1–A.5 specify TinyLlama training, response evaluation, the Qwen3 replication, and the transport and random-group comparisons. Earlier selected-link measurements form the separate study in Appendix B. Appendix C gives transport foundations, mass-compatibility proofs, and correspondence controls. Transition studies begin in Appendix E; margin interventions are in Appendix G. The separate head-level study is retained in the Supplementary analyses, Section H3. These retain their original cohorts and conventions, with terminology defined in Appendix D.1.

A.1 TINYLLAMA SPARSE DICTIONARIES: FITTING AND PROVENANCE

Model and corpus. The base model is TinyLlama-1.1B-intermediate-step-1431k-3T, revision `59f6f375b26bde864a6ca194a9a3044570490064`; the adapted endpoint is the stored step-3000 AG News classification adapter in the accompanying model manifest, with LoRA (Hu et al., 2022) rank 16 and $\alpha = 32$. We record the input to zero-based transformer block 11, using right padding and a 256-token length cap. The two checkpoints use identical token IDs and document offsets. Fitting, descriptors, and evaluation exclude padding and original position zero. Writing H_t for the stacked residual vectors at checkpoint t, direct float64 accumulation over the 472,566 evaluation content positions gives $\|H_{3000} - H_0\|_F^2/\|H_0\|_F^2 = .1666$, or $\|H_{3000} - H_0\|_F/\|H_0\|_F = .408$. Mean token-level residual cosine is .915. Centering the base and change fields separately gives a squared-energy ratio .1885; the mean change accounts for 2.91% of uncentered change energy. These descriptive measurements quantify residual change, not an isolated adaptation effect on the selected-link scores. They were computed from existing caches after correspondence evaluation; the released record includes cache hashes and the calculation. Float16 residual caches have relative storage error below 5×10^{-4}; one-document independent CPU full-forward checks corroborate the layer and checkpoint identities.

The corpus uses AG News train/test snapshots with saved hashes and row IDs. NFKC normalization, case folding, and whitespace collapse define exact-text deduplication. A fixed class-stratified selection is followed by 64-component MinHash candidate retrieval and exact word-five-shingle Jaccard at least .8. Connected components spanning partitions retain only the highest-priority partition: evaluation, descriptors, calibration, then training. There are no replacements after filtering. Final partition counts are 15,982/998/1,998/7,600; all official test rows remain. Labels balance the partitions but do not enter matching. The rule establishes the stated document separation, not complete semantic deduplication or independent news events. The dataset was familiar from earlier analyses; language-model pretraining and historical adaptation exposure are not established by these new partitions.

Architecture and training. The Top-k SAE uses $z = \mathrm{TopK}_{64}(\mathrm{ReLU}(W_e(h - b_d) + b_e))$ and reconstruction $Dz + b_d$, with trainable biases and unit-norm decoder columns. Encoder and decoder parameters share a transpose initialization within a fit, then optimize independently; no parameters are shared between fits. A training-only scalar, the square root of mean squared distance from the training mean, normalizes optimization inputs. Native-unit exports rescale both biases and codes while retaining unit decoder directions. All exports and their hashes are recorded.

Training uses Adam with betas $(.9, .999)$, batch size 1024, learning rate 4×10^{-4}, 500-step linear warmup, gradient-norm cap 1, projected decoder gradients, and decoder renormalization after each step. The decoder bias starts at the geometric median of a fixed training-only sample. Auxiliary reconstruction uses coefficient $1/32$ and at most $d_{\mathrm{model}}/2$ features inactive for the preceding million training presentations. Uniform sampling with replacement supplies 12,000,256 presentations per fit from 1,005,631 content-token positions; fitted dictionaries see 1,005,621–1,005,629 distinct positions.

We measure centered FVU as $\sum_u \|h_u - \hat{h}_u\|^2 / \sum_u \|h_u - \bar{h}_{\mathrm{cal}}\|^2$, where $\bar{h}_{\mathrm{cal}}$ is the mean of all 62,520 retained calibration tokens at that checkpoint and reconstructions are in native units. Before fitting the remaining seeds or constructing relations, the first source and target fits determine the common budget through reconstruction calibration alone. Starting at 12 million presentations, the rule adds four-million blocks, up to 24 million, if either calibration FVU improves by more than 3% relatively over the preceding four million presentations. The observed improvements are 1.28% and 1.20%, so all fits use the initial budget. The declared $k = 64$ quality rule requires calibration FVU at most .40 and at most 10% of features inactive over the last million training presentations; both initial fits pass, so $k = 128$ is not used. All ten fits are retained. Independent arithmetic over all 62,520 calibration tokens per fit reproduces centered FVU .266712–.270187, exactly 64 strictly positive codes per token, and zero never-active calibration features. Calibration participates in this operating-point choice; it is not the final response-evaluation partition.

A.2 TinyLlama response recovery: protocol and reproducibility

The TinyLlama experiment uses the ten saved sparse dictionaries and 45-pair schedule in Section 4, with the document partitions and native-unit activation convention in Appendix A.1. It was specified after the earlier correspondence results and historical group results were known. The evaluation corpus is reused; it is not a previously untouched replication. The protocol, source code, input hashes, and all group memberships and coefficients were saved before computing the new evaluation errors. No group budget, coefficient rule, or tolerance was changed in response to these errors.

Source accounting and fitting. The measurable source population is defined on fitting inputs: positive document response on at least twenty descriptor and ten calibration documents, and calibration response variance above 10^{-12} in the relevant response unit. This retains 4096, 4094, 4096, 4092, and 4096 coordinates in source fits $A_0, \ldots, A_4$, respectively. The same source population is used for every target fit and construction. Undefined evaluation errors, including evaluation variance at most 10^{-12}, remain recorded and cannot count as successful recovery. Scores against the full width are also recoverable from the released counts. Eligibility is not conditioned on the selected-link core or on singleton failure.

UOT memberships are the largest plan entries in each source row of the saved union-top-eight graph; cost memberships are the lowest costs in the identical graph. Targets must be descriptor-active. Ties use ascending target indices. Calibration-response selection ranks absolute calibration correlations over all 4096 target features. Membership budgets are 1, 2, 4, and 8. Constant calibration columns remain in the nominal membership but receive zero coefficient. Nonnegative least squares is solved in standardized coordinates with a free intercept, then exported to native response units. There is no decoder reconstruction constraint or unit-sum coefficient constraint. The saved records distinguish nominal and nonzero support sizes.

Uniform and transport-relative weights are normalized within each UOT group and then receive a single nonnegative calibration gain and an intercept. Cost groups additionally retain an analogous $\exp(-C_{ij}/\varepsilon)$ weighting comparison. The unrestricted nonnegative fit uses the same target membership. The deployable singleton comparator selects by absolute calibration correlation over the entire target dictionary and freezes its signed slope and intercept. It differs from the stronger evaluation oracle in Eq. 2.

Evaluation. We compute error from actual frozen predictions, including the calibration intercept, and divide by evaluation source variance. We do not replace this error with one minus squared correlation: that would recalibrate on evaluation. Errors may exceed one. The oracle alone receives evaluation-optimal target selection and affine coefficients, with a constant-target fallback. The fixed tolerance grid is $0, .05, \ldots, 1$, supplemented by .01 and .025; $.10, .25, .50$ are descriptive reference points. At .25, root-mean-square prediction error is at most half the source response's evaluation standard deviation.

All-five recovery requires success in each of the five adapted dictionaries for the same source coordinate. Fractions first divide by the eligible count in each source fit and then average over source fits. Source-fit records are not distinct semantic concepts. The saved arrays retain success in zero through five target fits, group-only and oracle-only outcomes, and the number of target fits requiring a group. Same-checkpoint results average per-pair recovery over the ten directed comparisons induced

Table 3: TinyLlama: Source responses recovered in all five adapted dictionaries (%). Entries average the five source-fit fractions over their fitting-defined eligible populations. Group size is at most four; coefficients are fixed before evaluation. The singleton oracle instead selects and calibrates on evaluation. Document and token predictors are fitted separately.

	Document responses			Token responses		
Allowed error	.10	.25	.50	.10	.25	.50
Affine singleton oracle	4.8	19.3	51.1	1.1	7.6	30.0
UOT groups	7.7	31.0	62.7	2.2	15.1	45.2
Cost groups	8.4	33.3	67.7	2.2	15.7	48.3
Response-selected groups	8.3	32.4	67.1	2.3	16.5	52.4

by each endpoint's unordered pair schedule, with the lower seed index as source. These are not independent model adaptations.

Additional recovery at error .25. Across the five source fits, UOT groups complete 520, 516, 518, 504, and 510 document-response records unavailable as complete singleton-oracle records. The corresponding token counts are 311, 327, 319, 293, and 303. Requiring a group in every target fit gives document counts 33, 45, 42, 47, and 39, and token counts 18, 33, 24, 22, and 16. Conditional on the newly complete document records, 73.1–78.1% reduce the worst target-fit error by at least .10. These conditional distributions describe the recovered subset; they are not an unconditional treatment effect or an independent confirmation. The paired error summaries and full recurrence arrays are included in the machine-readable release.

Of the 2,568 newly complete document records, the affine singleton exceeds error .25 in exactly one target fit for 1,171 (45.6%), in two to four fits for 1,191 (46.4%), and in all five for 206 (8.0%). Across source fits, the median worst singleton errors are .324–.333 and group errors .196–.204. With UOT group budgets 1, 2, 4, and 8, all-five document recovery is 14.1%, 23.9%, 31.0%, and 37.1%; token recovery is 7.2%, 12.2%, 15.1%, and 17.4%. The frozen one-feature predictor is not the evaluation oracle. Among all complete document records, 6,177 are recovered by both UOT and cost groups, 164 by UOT alone, and 644 by cost alone. This overlap analysis does not define a combined selector using evaluation outcomes.

Table 4: Mean per-pair recovery at normalized squared error .25 (%). UOT and cost groups use at most four features and separately calibrated coefficients. The 25 transitions and ten refits at each endpoint reuse ten dictionaries.

	Document responses			Token responses		
Comparison	Oracle	UOT	Cost	Oracle	UOT	Cost
Across checkpoints	37.1	48.4	50.1	19.6	27.3	27.9
Base-model refits	43.2	55.9	57.8	27.5	37.8	38.6
Adapted-model refits	44.2	56.9	58.9	27.8	37.9	38.7

Illustrative record. Figure 1 selects the smallest eligible A_0 index with group success in all five adapted fits and singleton-oracle failure in at least two. This rule selects source 8 from 283 qualifying sources, and the first failing target is B_3. The example illustrates the already evaluated population. Across $B_0, \ldots, B_4$, its group errors are $.129, .139, .153, .172, .245$ and singleton errors $.207, .104, .217, .351, .279$. The nominal four-feature B_3 support has two nonzero coefficients. Explicit predictions over all 7,600 documents reproduce the saved group and oracle errors within 2.5×10^{-15}. The source response is zero on 5,928 documents; these points remain in the scatter plots. Full-precision coefficients, all-document predictions, hashes, and the figure builder are supplied in `data/response_example/`.

Numerical checks. The oracle was checked against direct affine least squares, including anticorrelated and constant responses. Tests also cover shifted evaluation means and fitted groups that reduce

to one feature. Independent raw-design fits for 675 calibration cases agree with the moment-based predictions to a maximum absolute difference 5.27×10^{-15}. Direct evaluation of 210 frozen predictors agrees with the moment-based mean squared errors to 2.34×10^{-16}. These numerical checks verify the reported arithmetic; they do not constitute a fresh replication of training or generalization to another setting.

Reproduction. The source package includes the frozen aggregate results, descriptive analysis, and a script that regenerates Table 1 and Figure 2. The experiment records retain per-source prediction errors, oracle targets, coefficient arrays, membership identities, eligibility masks, and input/code hashes. The compact supplement provides the retained primary-method outcomes and example predictions; full coefficient arrays, model weights, and activation caches are separate large inputs. Appendix A.4 gives the completed dense common-support comparison separately.

A.3 QWEN3 REPLICATION

Model and fitting. We repeat the fixed response-recovery protocol in Qwen3-1.7B, using the input to zero-based block 14 before and after step-3000 AG News adaptation. The stored adapter has rank 16 and $\alpha = 32$. The same four document partitions are retokenized with the pinned Qwen tokenizer and a 256-token cap; removing padding and position zero leaves 51,348 calibration and 388,186 evaluation tokens. Checkpoint and full-forward checks verify the recorded layer and adapter. The evaluation residuals have relative Frobenius displacement .379 and mean token cosine .928, computed under the same content-token convention as the TinyLlama measurement. Input hashes identify the exact caches.

Five independent TopK64 dictionaries per endpoint use width 4096 and the architecture, optimizer, seed schedule, and reconstruction-based training rule in Appendix A.1. All ten receive 12,000,256 token presentations. Calibration FVU ranges from .1824 to .2272, mean active-feature count from 63.99988 to 64, and the fraction never active on calibration from .0046 to .0183. The training-last-million inactivity diagnostic is a different quantity and ranges from .00049 to .00146. Qwen3 is a model replication on the shared task and corpus; it does not supply a new domain or repeated adaptation seeds.

Response accounting. Selectors, coefficient fits, group budgets, the affine-singleton oracle, tolerance grid, and the 25 cross-checkpoint plus 20 same-checkpoint pair schedule are unchanged. Eligibility retains 4005, 3982, 4000, 3990, and 3995 coordinates in source fits $A_0, \ldots, A_4$ for both response units. No eligible source has undefined evaluation error; undefined outcomes remain recorded for the full source dictionaries. At error .25, UOT groups recover in all five target fits for 22.9–23.9% of eligible document responses across source fits and 8.9–9.3% of token responses. They complete 1906 document-response records and 933 token-response records for which the singleton oracle fails in at least one fit. Among the newly complete document records, the affine singleton exceeds error .25 in one target fit for 846 (44.4%), in two to four for 903 (47.4%), and in all five for 157 (8.2%). For token records, 69 require groups in every fit. Complete per-fit counts, all declared tolerances, and both group-only and oracle-only outcomes are released.

Per-pair recovery also increases in both unchanged-model refit cohorts (Table 6). Cost selection outperforms UOT at the primary four-feature budget, as in TinyLlama. Uniform, transport-relative, and freely calibrated weights on the same UOT memberships give all-five document recovery of 5.1%, 13.4%, and 23.3%, respectively. The first two still calibrate an overall nonnegative gain and intercept. No dense-transport comparison is added for Qwen3; the common-support attribution experiment remains the TinyLlama comparison in Appendix A.4.

Numerics and reproduction. All 90 pair/unit coefficient files are locked before evaluation. Fits are certified against the original moments at KKT residual below 10^{-7}. When the spectral solver fails that check, direct fits on centered calibration responses enumerate the at most 256 supports and are checked against an independent bounded least-squares solve at the same objective and tolerance. This changes neither memberships nor scientific criteria. Independent raw-design checks of calibration fits and explicit evaluation predictions agree with the recorded calculations. The experiment records retain numerical fixtures, fit and evaluation locks, per-source outcomes, and frozen primary predictors. The compact supplement includes the fixtures, locks, and outcomes; full coefficient arrays are separate

Table 5: Qwen3: Source responses recovered in all five adapted dictionaries (%). Entries average the five source-fit fractions over their fitting-defined eligible populations. Group size is at most four; coefficients are fixed before evaluation. The singleton oracle instead selects and calibrates on evaluation. Document and token predictors are fitted separately.

	Document responses			Token responses		
Allowed error	.10	.25	.50	.10	.25	.50
Affine singleton oracle	3.7	15.0	42.4	0.8	4.7	22.0
UOT groups	5.0	23.3	52.6	1.3	9.2	34.1
Cost groups	5.6	25.2	57.5	1.4	9.7	37.3
Response-selected groups	5.0	23.4	55.0	1.4	10.1	39.9

Table 6: Qwen3 per-pair recovery at normalized squared error at most .25 (%). Entries average 25 cross-checkpoint pairs or ten recorded-direction unordered pairs for each unchanged-model refit cohort. These are per-pair fractions, not all-five-fit fractions or independent adaptation replicates. Group size is the stated maximum; group predictors are frozen before evaluation.

Response	Cohort	Oracle	UOT k=2	UOT k=4	UOT k=8	Cost k=4
Document	Cross-checkpoint	28.8	32.5	37.4	41.2	38.9
Document	Base refits	34.2	39.1	44.7	48.7	46.1
Document	Adapted refits	36.5	42.2	48.1	52.6	49.9
Token	Cross-checkpoint	12.8	15.7	17.9	19.7	18.4
Token	Base refits	19.3	23.8	27.1	29.5	27.7
Token	Adapted refits	21.4	26.0	29.5	32.4	30.4

large artifacts. The source archive contains the aggregate results and a builder for the two-model main table and these replication tables; the supplementary code verifies saved outcomes without repeating model inference.

A.4 Transport attribution on common dense support

This comparison uses the saved TinyLlama dictionaries and response partitions. Its thirteen directed pairs are $A_0 \to B_0, \ldots, B_4$, $A_0 \to A_1, \ldots, A_4$, and $B_0 \to B_1, \ldots, B_4$. Every real source–target edge is available to each transport method. The cost is Eq. 3; activity references are the same normalized, floored descriptor second moments used in the sparse experiment. Entropy is fixed at $\varepsilon = .01$. Dense UOT uses Eq. 4 with $\rho_s = \rho_t = .5$. Balanced OT replaces the KL penalties with exact constraints $P\mathbf{1} = a$ and $P^\top \mathbf{1} = b$.

Rejection transport adds one dummy feature to each dictionary and solves balanced transport with augmented marginals $[a, 1]$ and $[b, 1]$. Both real–dummy costs are .125 and the dummy–dummy cost is zero. Only the real–real block supplies group members. Its largest row entries select up to k descriptor-active targets, exactly as for the other two transports. There is no rejection-flow threshold that excludes difficult sources. Budgets, coefficient fits, source eligibility, and evaluation tolerances are unchanged. Complete recurrence concerns five target dictionaries for transitions and four for each refit cohort (Table 7).

All costs and protocol choices were locked before solving, and all 39 relations and both units' coefficients were locked before recovery evaluation. The already known sparse results and the reused evaluation corpus are recorded in the protocol. For the same A_0 transitions, dense comparisons use 4096 eligible sources; they are not pooled with the five-source-fit averages in Table 1.

The initial balanced Sinkhorn solve for $A_0 \to B_1$ stopped at its iteration cap with marginal error 5.4×10^{-8}, above the specified 10^{-8} requirement. Before any predictor fitting or evaluation, we added Newton–conjugate-gradient acceleration of the same convex semi-dual and continued from the saved potentials. The objective, marginals, entropy, and convergence requirement remained fixed. All completed balanced and rejection plans satisfy both marginal constraints to L_1 error at most

Table 7: Dense-support recovery at error $\leq .25$ (%). Groups contain at most four features. Complete recovery requires success in five target fits for transitions and four for each refit cohort; per-pair rates average the same directed comparisons.

		Documents		Tokens	
Cohort	Construction	Per pair	Complete	Per pair	Complete
A_0 to adapted fits	Affine singleton oracle	37.7	19.6	19.8	7.5
A_0 to adapted fits	Dense UOT	47.7	30.0	26.8	14.5
A_0 to adapted fits	Balanced OT	47.5	29.6	26.7	14.2
A_0 to adapted fits	Rejection transport	50.6	33.9	28.0	15.8
A_0 to base refits	Affine singleton oracle	43.6	25.5	27.9	12.4
A_0 to base refits	Dense UOT	54.8	37.5	37.5	23.0
A_0 to base refits	Balanced OT	55.0	37.3	37.4	22.5
A_0 to base refits	Rejection transport	58.1	41.7	39.1	24.5
B_0 to adapted refits	Affine singleton oracle	43.7	25.7	27.7	12.5
B_0 to adapted refits	Dense UOT	55.3	39.0	36.9	23.1
B_0 to adapted refits	Balanced OT	55.9	39.1	36.9	22.9
B_0 to adapted refits	Rejection transport	58.3	42.2	38.4	24.3

7.87×10^{-9}; UOT stationarity residuals are at most 4.93×10^{-13}. Reconstructing every full plan from saved duals reproduces all top-eight memberships. Independent raw-design checks of 234 calibration fits agree to 1.1×10^{-12}, and direct predictions for 234 evaluation cases reproduce mean squared error to 6.0×10^{-17}.

The source package retains the aggregate curves, objectives, convergence records, and protocol amendments in `data/dense_transport/`; `scripts/build_dense_tables.py` regenerates the two reported tables. Full experiment records retain duals, memberships, per-source errors, and the scientific implementation. The compact supplement includes saved outcomes and code, but omits the duals and reference masses needed to reconstruct the full plans. These comparisons test how marginal constraints select useful response supports; they do not assign semantic meaning to transport deficits.

A.5 Uniform random-group control

The random-group control was specified after the primary study. We draw four distinct descriptor-active target latents uniformly for each source coordinate, with twenty fixed random draws for every pair in the original 25 cross-checkpoint and 20 unchanged-checkpoint comparison schedule, separately in TinyLlama and Qwen3. We use the original document responses, source eligibility, nonnegative coefficient fit and free intercept; calibration-constant selected columns receive zero slopes without resampling. All memberships and coefficients across both models are frozen before evaluating any random-group outcome. We report every draw at all original error tolerances; draw variation describes the random selector, not independent model replications. The control tests uniform uninformed membership selection at the same nominal coefficient budget; it does not test activity-matched groups, establish semantic splitting, or resolve dictionary-width and training-budget sensitivity.

Table 8: Document-response recovery at normalized squared error .25 (%). Panel (a) uses the all-five metric of Table 1. Panel (b) reports mean pairwise random-group recovery: 25 cross-checkpoint pairs and ten refit pairs at each unchanged checkpoint. Random entries are mean [minimum, maximum] across 20 fixed draws; these quantify selector variation, not model replications.

(a) All-five recovery

Model	Oracle	UOT	Cost	Response	Random
TinyLlama	19.28	30.97	33.32	32.37	0.0 [0.0, 0.0]
Qwen3	15.02	23.30	25.23	23.35	0.0 [0.0, 0.0]

(b) Random-group pairwise recovery

Model	Cross-checkpoint	Base refits	Adapted refits
TinyLlama	0.076 [0.054, 0.095]	0.082 [0.066, 0.127]	0.093 [0.061, 0.115]
Qwen3	0.087 [0.073, 0.100]	0.104 [0.080, 0.133]	0.093 [0.058, 0.128]

B SELECTED-LINK CORRESPONDENCE

This study evaluates individual selected links by response correlation and activating-example overlap, rather than the frozen compact predictors in Appendix A. Its methods, coverage, cross-fit coherence and alternative matchers are reported together below. These results do not enlarge the primary response-recovery study's model or task coverage.

B.1 LINK SELECTION AND EVALUATION

Selected links. A source–target edge is selected when it has maximal $P^{\star}_{ij}$ in both its row and column, has mass at least 10^{-6}, and accounts for at least 0.7 of both its row's and column's transported mass. Neither endpoint may have another incident edge of mass at least 10^{-6} carrying at least 0.2 of its transported mass. This selects a partial one-to-one core from the full many-to-many plan, called *canonical* in the figures and earlier tables. Historical event tables also apply birth/death exclusions before assigning *persistent* labels (Appendix E.1).

Testing selected links. For a selected pair (i, j), let A_i and B_j contain the up to twenty evaluation documents with the largest strictly positive responses. We measure activating-document overlap by Jaccard similarity, $J_{ij} = |A_i \cap B_j|/|A_i \cup B_j|$ when both sets are nonempty; otherwise the primary score is undefined. For these selected-link experiments the response is the mean latent activation over a document's nonpadding tokens, excluding position zero. We also measure centered correlation between document-response vectors and cosine and centered correlation between aligned token-response vectors. Descriptor documents determine all links before evaluation responses are opened.

The selected-pair mean averages the defined scores for selected counterparts. The all-width Jaccard yield sums these overlaps and divides by all 4096 source features, including those without a selected counterpart. An undefined score contributes no demonstrated overlap to this yield, but is excluded from the defined-score mean and counted separately. No selected score is undefined in the primary study. A missing selected link does not establish that the model lacks the corresponding property.

Comparisons. Re-pairing the same selected source and target populations tests whether their pairing carries response information; the TinyLlama selected-link study computes the exact expected pair score over all such re-pairings. Strong alternatives use the same descriptor cost for nearest-neighbor, mutual-nearest-neighbor, and Hungarian matching, with decoder-only nearest-neighbor as a separate descriptor comparison. We report both equal selected counts, which may use different sources, and exactly the UOT-selected source queries. Across target fits, we measure overlap between counterparts of the same source and retain missing fits in the query denominator. Appendix B.2 gives the ranking, coherence, and numerical rules.

B.2 TinyLlama measurement protocol

The fitting protocol is in Appendix A.1. This section records the earlier selected-link measurements, whose denominators differ from the compact-response experiment.

Relations and ranking. All descriptors and selections use only the 1,998 descriptor documents. One shared Gaussian projection, seed 20260916, maps concatenated document means and active fractions to 64 dimensions. Eq. 3 uses the stated weights, zero-profile cosine zero, and raw-second-moment references floored at 10^{-12} before separate normalization. All 45 union-top-eight plans use $\varepsilon = .01$, $\rho_s = \rho_t = .5$, and the selected-link rule in Appendix B.1. Maximum saved stationarity residual is 4.92×10^{-13}, below the declared 10^{-8} requirement.

Nearest-neighbor, mutual-nearest-neighbor, and Hungarian links are ranked by their common descriptor cost, with index tie-breaking. Decoder-only nearest-neighbor uses decoder cost for both matching and ranking. Selected UOT links are ranked by the lesser of their row and column transported-mass fractions, then edge mass and indices. Its reported curve consists of ranked subsets of the declared selected core, not a threshold sweep extending to every source. Saved counts are 10, 25, 50, 100, 250, 500, the UOT-selected count, and each method's available count. Fixed-query comparisons evaluate NN and Hungarian on exactly the UOT source queries.

Outcomes and denominators. Document responses average retained tokens. Positive top sets contain at most twenty documents, with row-index tie-breaking; no zero responses fill missing positions. A pair with an empty top set has undefined Jaccard. Undefined centered correlations require a zero centered norm. In the completed study every feature has at least twenty positive evaluation documents and no selected metric is undefined. Token cosine and centered correlation use the 472,566 aligned evaluation positions. For K selected sources and targets with defined pair scores, the exact re-pairing expectation is $K^{-2} \sum_{i,j} s(i, j)$. It describes random pairing of those endpoints, not uncertainty over model adaptations.

For each source dictionary, complete families contain selected counterparts in all five target fits. Mean and weakest target–target Jaccards summarize the ten target pairs, then average over source features. The re-pairing reference independently permutes target populations, with 1000 draws. These are conditional pairing controls, not tests against homogeneous feature matchability. Full-query reporting first retains the union of sources with any selected link, then additionally reports all-width yields with denominator 4096×10. The union omits 324–379 sources with no selected counterpart; the all-width result includes them. Missing or undefined slots contribute no demonstrated overlap to yields; they are not assigned semantic absence.

On identical UOT-realized slots and the union-query denominator, UOT coherence yields are .2036–.2081, NN .2049–.2087, and Hungarian .2033–.2078. Allowing all ten target-pair slots on the same union queries gives NN .3669–.3693 and Hungarian .3263–.3319. The resulting increase includes their greater resolution coverage. None of these values is a conditional mean over realized slots alone.

Dense selected-link comparisons. Dense balanced and rejection solves cover the predeclared thirteen pairs $A_0 \to B_0, \ldots, B_4$, $A_0 \to A_1, \ldots, A_4$, and $B_0 \to B_1, \ldots, B_4$. Rejection adds a dummy node of mass one on each side with rejection cost .125. Both methods use the identical complete cost and activity references; the maximum row/column L_1 residual must not exceed 10^{-5}. Twelve balanced solves and all thirteen rejection solves pass. Balanced $B_0 \to B_2$ reaches the iteration limit with row residual 1.10045×10^{-5} and is excluded from scoring. This failure remains in the released record.

Ranking row-maximum edges by the smaller marginal fraction supplies equal-count comparisons separately from the unchanged selected-link rule. Relative to UOT, mean matched-count Jaccard differences for balanced transport are $-.00710/-.00489/-.00896$ across transitions/base refits/adapted refits, using 5/4/3 eligible pairs. Rejection differences are $+.02297/+.02386/+.02303$, using 5/4/4 pairs. Each difference uses the same eligible pairs. The unchanged balanced and rejection selectors return fewer than the requested UOT count in every usable solve; those smaller populations are not relabeled equal-count results.

Verification and uncertainty. The release includes document manifests, model/SAE identities, protocol and implementation hashes, pair and coherence summaries, and code to regenerate the main numerical tables from these saved records. Independent raw-code checks reproduce document Jaccard and correlation for $A_0 \to B_0$ and $A_0 \to A_1$, and A_0 coherence including fixed-query baselines. All 45 pairs and five coherence populations also pass saved-score consistency and coverage checks. These checks have different scopes: independent raw-response recomputation was not performed for every pair and metric. Pair ranges describe fitting variation with shared dictionaries and documents; they are not confidence intervals across independent adaptations. No edge-based standard error is reported.

Table 9: Response distributions in the earlier TinyLlama selected-link study. Each entry is the equal-pair mean of a within-pair median or tenth percentile among selected links, not a pooled-edge quantile. All selected values are defined. Pair ranges and the full ranked comparison curves are in the bundled numerical record.

Population	Metric	Median	10th percentile
Across checkpoints	Document Jaccard	.440	.110
Across checkpoints	Document correlation	.875	.566
Across checkpoints	Token correlation	.802	.463
Base refits	Document Jaccard	.510	.130
Base refits	Document correlation	.900	.593
Base refits	Token correlation	.842	.499
Adapted refits	Document Jaccard	.538	.137
Adapted refits	Document correlation	.902	.614
Adapted refits	Token correlation	.843	.515

B.3 TinyLlama response correspondence

Across all 25 endpoint pairs, 2,180–2,323 selected links recover related held-out document and token responses (Table 10), covering 55.4% of source coordinates on average and 60.8% of source activity-reference mass. Mean activating-document Jaccard is .444, compared with .002 under re-pairing of the same endpoints. Centered document-response correlation is .813 and centered token-response correlation is .745. Aligned-token correlations retain both within-document and between-document response structure.

The all-width yield is .246 rather than the selected-link mean .444: queries without selected links remain in the denominator. Every feature has at least twenty positive-response evaluation documents, and all selected scores are defined. Same-checkpoint refits have greater mean overlap under their own selections. Appendix B.2 supplies the full distributions and pair records.

B.4 Corroboration across TinyLlama target dictionaries

For a source feature linked to several target fits, we test whether its target counterparts activate on the same evaluation documents. Links to each target fit are inferred separately from descriptor data; this agreement is measured afterward. For each source dictionary, 683–745 of its 4096 features have selected counterparts in all five target fits. Their mean target–target Jaccard is .580–.606, compared with .0027–.0028 under re-pairing of the selected target populations. Even the weakest target pair per complete family has mean overlap .412–.436 (Table 11). The agreement is therefore not confined to a single favorable source–target pairing.

Complete families comprise 16.7–18.2% of each source dictionary; between 324 and 379 sources have zero selected counterparts across all five target fits. Summing demonstrated overlap across all source/target-pair slots gives all-width yields .185–.190. These yields retain missing evidence without interpreting an unresolved link as absence of the property. On the same source features and pairs of target fits scored for UOT, nearest-neighbor and Hungarian give similar target–target overlap; their additional counterparts increase the overlap summed over all sources and pairs of target fits (Appendix B.2).

Table 10: Earlier TinyLlama selected-link study: ten width-4096 dictionaries with $k = 64$ from one adaptation. Entries average pairs equally; link counts give ranges. Refits hold the model fixed but select their own links, so column differences do not isolate adaptation effects. Jaccard is intersection-over-union of sets of up to 20 highest-activating documents. Re-pairing randomly pairs the same selected features; correlations are centered. Yield divides summed selected-link Jaccard by all 4096 source queries.

Measurement	Across checkpoints 25 pairs	Base refits 10 pairs	Adapted refits 10 pairs
Selected links	2180–2323	2300–2433	2349–2442
Source-count coverage (%)	55.39	58.06	58.35
Document Jaccard	.444	.495	.498
Re-paired document Jaccard	.0021	.0021	.0021
Document correlation	.813	.835	.840
Token correlation	.745	.778	.781
All-width Jaccard yield	.246	.287	.291

Table 11: Earlier TinyLlama selected-link study: five-target-fit corroboration, separately for each source dictionary. A complete family has one selected counterpart in every target fit. Mean and weakest Jaccards summarize the ten target pairs within each complete family, then average families. All-width yield instead retains all 4096×10 source/target-pair slots, including sources without any selected counterpart. These are descriptive results from one adaptation.

Source	Complete families of 4096	Mean Jaccard	Mean weakest Jaccard	All-width yield
A_0	741	.588	.417	.189
A_1	707	.591	.418	.185
A_2	745	.580	.412	.189
A_3	738	.589	.416	.190
A_4	683	.606	.436	.188

Independently trained target dictionaries thus corroborate these activating-document patterns through several decompositions of the adapted model, without requiring a canonical source dictionary.

B.5 TinyLlama correspondence quality and coverage

At the UOT-selected count, same-cost nearest-neighbor, mutual-nearest-neighbor, Hungarian, and decoder-only nearest-neighbor matching all have higher mean activating-document overlap in every endpoint pair. These rankings select different source populations. On exactly the UOT-selected source queries, Jaccards are .4443 for UOT, .4472 for nearest-neighbor, and .4442 for Hungarian. Response correspondence is therefore recoverable under several matching procedures; these scores do not establish a UOT retrieval advantage at this coverage.

The saved coverage curves also test smaller, higher-ranked subsets. Table 12 reports activating-document overlap at common counts using each method's descriptor-only ranking. UOT's concentration ranking yields higher overlap at 250 and 500 links, whereas the simple matchers lead at the full selected-core count. These rankings trade coverage for activating-document overlap and can select different source features at the same count.

In the earlier selected-link analysis, dense balanced and rejection transport use the same costs and activity references on thirteen predeclared primary-study pairs. On the five $A_0 \to B_r$ endpoint pairs, ranking each source's maximum-flow edge gives matched-count Jaccard .0071 below sparse UOT for balanced transport and .0230 above for rejection. This comparison changes support as well as marginal constraints. Across the full thirteen-pair schedule, 12 balanced and all 13 rejection solves pass the numerical tolerance; the failed balanced refit is excluded from scoring. Appendix B.2 reports these rankings and the separate unchanged-selector results. The completed response-recovery comparison uses separately solved plans (Appendix A.4).

Table 12: TinyLlama activating-document Jaccard at equal source-query counts, averaged over the 25 cross-checkpoint pairs. The first two counts cover 6.1% and 12.2% of source coordinates; K is the UOT-selected count in each pair (2180–2323, mean coverage 55.4%). UOT ranks its selected core by marginal concentration; other methods rank by their defining cost. Source populations can differ, and NN can reuse targets. Full curves and correlation metrics accompany Appendix B.2.

Method	250 links	500 links	K links
UOT	.645	.621	.444
Same-cost NN	.601	.595	.460
Mutual NN	.601	.595	.461
Hungarian	.601	.595	.461
Decoder NN	.604	.606	.464

C TRANSPORT FOUNDATIONS, CORRESPONDENCE CONTROLS, AND REPRODUCIBILITY

This section gives the transport arguments and supporting correspondence studies. The multi-model endpoint panel is in Appendix C.7, with its training and evaluation provenance in Appendix C.8. Appendix C.3 gives activity-mass compatibility arguments, and Appendix C.9 specifies the constructed correspondence controls. Appendix C.11 reports source-query accounting; Appendix C.13 separates descriptor and evaluation document groups and examines position-weighted activity. Each study below states its response convention and relation to the primary experiment.

C.1 THE LINEAGE OBJECT AND DEPENDENCE ON COORDINATES

Let D_s, D_t denote two fitted dictionaries, and let a lineage record contain their identifiers, descriptor specification, candidate graph $\mathcal{S}$, reference masses a, b, fitted plan P, and its row and column sums. Weighted edges express the proposed relation; local feature numbers address its endpoints. Selecting the one-to-one core is an optional summary of this record.

Proposition 1 (Relabeling of a fixed transport problem)**.** *For finite costs, nonnegative finite reference masses, and $\varepsilon, \rho_s, \rho_t > 0$, Eq. 4 has a unique minimizer on its fixed support. If the source and target features, their costs, masses, and candidate graph are relabeled by permutation matrices R, T, the minimizer becomes $RP^\star T^\top$.*

Proof. The zero plan is feasible with finite objective, using generalized KL conventions. Any plan with finite objective has zero entries on edges incident to a zero reference mass. On the remaining finite-dimensional feasible set, the entropy term is strictly convex and coercive; the cost is linear and marginal KL terms are convex. A unique minimizer therefore exists. Permuting entries preserves every term of the objective and bijectively maps the feasible sets, proving the claimed transformation. □

This standard relabeling property lets a plan refer to local coordinates without treating their numbers as cross-dictionary identities; it does not establish semantic identifiability. The candidate graph must be relabeled too: index-based tie-breaking at a top-k cutoff need not regenerate the same graph from tied costs. The archived support is retained for the sparse feasibility certificates.

For normalized a, b, the zero plan has objective $\rho_s + \rho_t$, so UOT remains well posed even when exact-marginal transport on the sparse graph is infeasible. This does not mean the optimizer returns literal zero mass on every unsuitable edge. Entropy and finite KL penalties generally give soft attenuation, and marginal sums can exceed their references. We report both reference and transported masses rather than interpreting normalized edge weights alone as survival probabilities.

C.2 NUMERICAL CONVENTIONS OF THE ARCHIVED PLANS

The six archived TinyLlama/Qwen3 sparse plans are numerically consistent with effective solver references $a_i^{\mathrm{num}} = \max(q_i^s, 10^{-12}) / \sum_{i'} \max(q_{i'}^s, 10^{-12})$, and analogously for b^{num}. They are also numerically consistent with zero cosine similarity for a zero-norm activation profile (dissimilarity

$1/2$). These conventions are identified from the retained numerical artifacts; the original solver source was not recovered. They reconstruct every archived union-top-8 edge in all six pairs and give maximum absolute objective-gradient residual below 2×10^{-9} on positive plan entries. Equation 4 applies to this numerical reconstruction with the effective references and costs. With the nominal zero references instead, the tiny positive archived weights incident to those nodes have infinite generalized-KL cost. No zero-norm decoder occurs in these six pairs, so their archives do not identify a zero-decoder convention.

The historical six-pair dense comparisons use nominal raw-second-moment references; both their dense and direct descriptor comparisons use maximum zero-profile dissimilarity. The support differences arise from these conventions rather than numerical ties. Re-solving under both the archived and later numerical conventions, including rebuilding support under the later convention, leaves all six original selected cores unchanged. On fixed archived support, the convention-only plan L_1 change is below 2.2×10^{-12}; the larger replay-to-archive discrepancy, below 5.0×10^{-10}, is consistent with finite solver convergence. All six capacity witnesses below survive the reference floor: the worst-case perturbation bound for any witness deficit is below 3×10^{-12}. Reference-mass coverage and certificate values retain the nominal-mass convention. The selected-link results are unchanged under either numerical specification.

C.3 Mass compatibility and the cost of enlarging support

Activity-mass compatibility places a separate constraint on correspondence: candidate coverage does not guarantee balanced feasibility. For any source subset I, a balanced plan supported on $\mathcal{S}$ must satisfy

$$\sum_{i \in I} a_i \;\le\; \sum_{j \in N_{\mathcal{S}}(I)} b_j, \tag{5}$$

where $N_{\mathcal{S}}(I)$ is the set of reachable target features. The mass from I cannot exceed its candidate neighbors' total capacity. All six TinyLlama/Qwen3 endpoint graphs violate this condition. The largest violation, TinyLlama seed 3, identifies a source subset with 34.0334% mass but only 20.9629% reachable target capacity. Thus exact balanced OT cannot retain both this support and these masses. On full support, the same certificate forces at least 13.0706% of mass outside the local candidate relation (Eq. 6); the certificate does not determine whether those edges are valid correspondences. The violations vary substantially in size: one Qwen3 violation is only 3.43×10^{-8} (Appendix C).

The capacity condition also gives a quantitative statement when the graph is enlarged. For any exact-marginal plan P and source subset I, at most $b(N_{\mathcal{S}}(I))$ of the mass $a(I)$ can reach its original candidate neighbors. Hence

$$\sum_{(i,j) \notin \mathcal{S}} P_{ij} \;\ge\; \big[a(I) - b(N_{\mathcal{S}}(I))\big]_+. \tag{6}$$

In particular, the TinyLlama-3 certificate forces at least .1307055542 of unit mass outside the archived candidate graph in any dense balanced solution. This is an off-support bound, not a false-correspondence bound: the sparse graph is a modeling restriction and its excluded edges are not known to be semantically false.

More generally, if $\mathcal{R}$ is a stipulated valid relation and $M_{\mathcal{R}}$ is the maximum mass of a nonnegative coupling supported on $\mathcal{R}$ with row sums at most a and column sums at most b, then every unit-mass balanced plan places at least $1 - M_{\mathcal{R}}$ outside $\mathcal{R}$. Indeed, its restriction to $\mathcal{R}$ is such a subcoupling. If the only valid edges form a permutation π, then $M_{\mathcal{R}} = \sum_i \min(a_i, b_{\pi(i)})$ and this bound becomes $\frac{1}{2}\|a - b_\pi\|_1$. These bounds follow directly from the marginal constraints.

Squared activity need not be conserved by a feature split. Consider unit decoder directions d, e with source reconstruction $dz + ey$, where $\mathbb{E}[z^2] = \mathbb{E}[y^2] = 1$. Replacing dz by $d(z/2) + d(z/2)$ preserves the reconstruction exactly, but changes the unnormalized squared-activity masses from $(1, 1)$ to $(1/4, 1/4, 1)$. Normalization gives source masses $(1/2, 1/2)$ and target masses $(1/6, 1/6, 2/3)$. The stipulated relation connects the first source to its two copies and the second source to itself. Its first source has mass $1/2$ but corresponding target capacity only $1/3$, forcing at least $1/6$ of any balanced coupling outside that relation. An inactive source feature embeds this example at equal

width. The cross term between the copies explains why reconstruction preservation does not imply additivity of per-feature second moments.

This example identifies a mismatch between the chosen activity measure and exact conservation even when every active source has a counterpart. It does not show that the archived SAEs underwent this split. Changing the reference measure, using a partial/rejection coupling, or relaxing marginal penalties as in KL-UOT can avoid the forced cross-relation flow. A marginal deficit alone cannot distinguish this reparameterization from a missing counterpart.

Activity measures under KL relaxation. Nearest-neighbor lookup selects candidates independently; a fixed-cardinality partial assignment selects a disjoint list of pairs. A transport reference additionally couples the decisions through source and target activity measures. Exact balanced transport fixes both measures; rejection uses them as upper bounds on the retained submarginals; Eq. 4 instead penalizes relative deviations and allows either decrease or increase. The latter is a tractable, regularized way to treat activity moments as references whose compatibility can be inspected. It does not calibrate a probability of semantic identity. The constructed controls also support rejection as an alternative (§C.9).

An isolated-edge example. For a single edge with positive reference masses a, b and cost c, the derivative of Eq. 4 is $c + \varepsilon \log p + \rho_s \log(p/a) + \rho_t \log(p/b)$. Writing $D = \varepsilon + \rho_s + \rho_t$, its optimum is

$$p^\star = e^{-c/D} a^{\rho_s/D} b^{\rho_t/D}.$$

If $c = 0$ and $a = b = m \in (0, 1)$, then $p^\star = m^{(\rho_s+\rho_t)/D} > m$. Thus transported mass can exceed its activity reference even for a perfect zero-cost match under the stated entropy convention. In a larger graph, incident edges are coupled through the marginal sums. Raw expansion or shortfall is a property of the fitted transport problem, not a direct measurement of semantic birth or death.

C.4 Exact-marginal controls

The six comparisons reuse frozen descriptors and saved 7600×4096 endpoint activation tensors. Costs follow Eq. 3, with equivalent coefficients .05 for each unscaled capped-log term. The solves use normalized raw-second-moment marginals and $\varepsilon = .01$. The balanced solver uses exact-marginal Sinkhorn exponent 1; UOT uses $.5/(.5 + .01)$. All dense solves converged, with maximum balanced marginal L_1 residual below 10^{-8}. Small-problem solver checks reproduce the known zero-cost balanced outer-product solution and agree with the existing full-support log-domain UOT implementation.

Under the original selected-pair rule, dense UOT retains 764 pairs in total versus 528 for balanced OT, with row-average overlap .10284 versus .10278. At equal selected count, UOT has higher overlap in five of six rows (Table 13); this secondary selector ranks mutual-best edges by the smaller marginal fraction without the original .7 threshold or branch exclusions. The direct descriptor matchers below nevertheless exceed the original sparse-UOT core in every row. These comparisons separate the weighted relation from selected-pair retrieval quality.

The historical six-pair comparison controls support by running both methods densely. Separately, each original sparse graph violates Eq. 5. Singleton capacity checks and an integer-capacity maximum-flow computation identify witness subsets; the certificates below recompute their deficits using the original float64 masses. The zero-profile convention explains the reconstructed-support differences (Appendix C.2); the archived and reconstructed graphs both have the stated obstruction. Positive neighbor coverage alone does not establish feasibility. Adding edges or rejection nodes, or using partial transport, changes the feasible set and can remove the obstruction.

Table 13: Dense transport with identical costs and reference masses. K/J gives selected count/mean activating-example Jaccard under the original selector. Equal-K uses the sparse-UOT count, ranking mutual-best edges by their smaller row or column fraction. Q3/TL: Qwen3/TinyLlama; suffix: target seed. †Balanced TL-3 evaluates 112/113 pairs; scoring the excluded pair as zero gives .09070. All other shown transport selections evaluate every selected pair. Six fixed SAE pairs, without confidence intervals across SAE solutions.

Pair	UOT K/J	Balanced K/J	Equal K	UOT J	Balanced J
Q3-1	142 / 0.103	104 / 0.114	149	0.1043	0.1080
Q3-2	176 / 0.110	131 / 0.107	185	0.1107	0.1059
Q3-3	158 / 0.105	102 / 0.112	155	0.1123	0.1117
TL-1	87 / 0.099	77 / 0.098	91	0.1034	0.1009
TL-2	90 / 0.097	74 / 0.084	94	0.0979	0.0882
TL-3	111 / 0.104	40 / 0.102	113	0.1078	0.0915†

Table 14: Validated capacity deficit on the archived candidate graph. The Q3-2 obstruction is numerically tiny; the TL-2 and TL-3 deficits are substantial fractions of total reference mass.

Pair	Positive capacity deficit
Q3-1	0.005902180968
Q3-2	3.42952566e-08
Q3-3	0.007736269904
TL-1	0.001502075665
TL-2	0.07481098752
TL-3	0.1307055542

Table 15: Reference-mass coverage of each method's selected endpoints (source / target, percent). This sums reference masses over selected endpoints, not transported mass or causal importance. More selected coordinates can cover less reference mass.

Pair	Sparse UOT	Dense UOT	Dense balanced
Q3-1	18.75 / 20.36	18.74 / 20.34	18.66 / 20.18
Q3-2	13.92 / 15.09	13.90 / 15.07	13.76 / 14.83
Q3-3	13.75 / 17.64	13.77 / 17.66	13.63 / 17.43
TL-1	1.94 / 1.56	1.90 / 1.54	1.62 / 1.53
TL-2	1.99 / 1.58	2.01 / 1.56	33.11 / 26.48
TL-3	2.55 / 1.64	2.50 / 1.63	0.79 / 0.67

C.5 DIRECT PARTIAL ASSIGNMENTS AT THE SAME SELECTED COUNT

We evaluate descriptor-only matcher selections on the full saved activation tensors, with selections fixed before overlap evaluation. Exact-K matching minimizes total descriptor cost over partial assignments of cardinality K; the Hungarian comparator solves a full assignment and retains its K lowest-cost edges. Both use the original sparse-UOT count and are distinct from post-processing an existing UOT plan. The selected-count comparison was added after the original experiments; pairs were selected without using evaluation overlap.

Both-empty activation sets are excluded by the original evaluator, so equal selected count need not mean equal evaluated count. Table 16 gives both counts and a sensitivity analysis that scores both-empty pairs as zero. Both competitors have higher mean overlap than the original sparse-UOT core in every row under either denominator. The smallest exact-K advantage under the all-selected denominator is .000708 (TL-3), and no solution-level uncertainty estimate establishes that this small difference is meaningful. The selected endpoint populations differ; this metric cannot establish a globally best matcher or validate the full plan's branching/unmatched structure.

Table 16: Fixed selected count K. N is the evaluated count; each two-part J entry gives the mean over pairs with nonempty activation-set union / the mean when both-empty pairs score zero. Every original sparse-UOT pair is evaluated.

Pair	K	UOT J	Exact N	Exact J	Hung. N	Hung. J
Q3-1	149	0.0978	145	0.1326/0.1291	146	0.1301/0.1275
Q3-2	185	0.1071	178	0.1332/0.1282	179	0.1332/0.1289
Q3-3	155	0.1037	149	0.1290/0.1240	150	0.1283/0.1242
TL-1	91	0.0906	89	0.1348/0.1318	89	0.1371/0.1341
TL-2	94	0.0822	92	0.1106/0.1082	92	0.1173/0.1148
TL-3	113	0.1010	111	0.1036/0.1017	111	0.1068/0.1049

C.6 DESCRIPTOR AND SOLVER SENSITIVITY

We evaluate the ten configurations below on the same six endpoint pairs, with thirty shuffled repairings per configuration and pair. Configuration values and activation hashes are preserved in the evidence package. The variance condition uses $\max(q - \bar{z}^2, 10^{-12})$ in both the reference masses and the second-moment cost term; it is not an isolated mass-only ablation. Weight-removal conditions rescale the remaining weights as specified in their saved configurations. The magnitude-heavy condition uses equal coefficients .37 on the five raw cost components, whose two log differences are capped at five without division by five. In Eq. 3's normalized-component convention its vector is $(.37, .37, .37, 1.85, 1.85)$: each magnitude term has five times the coefficient of each of the other three components. This is the archived condition named `equal_weights_norm`.

Table 17: All ten saved evaluation-set sensitivity configurations. Final column counts fixed endpoint pairs whose selected mean exceeds all thirty shuffled means; it is descriptive, not a multiplicity-adjusted significance statement.

Configuration	Selected K	Mean J range	Above all shuffles
Default	91–185	0.082–0.107	6/6
Decoder only (normalized)	170–221	0.053–0.104	6/6
Magnitude-heavy	52–158	0.030–0.124	6/6
Remove profile (normalized)	136–208	0.073–0.100	6/6
Remove direction (normalized)	22–133	0.005–0.133	3/6
$\varepsilon = .005$	203–273	0.052–0.092	6/6
$\varepsilon = .05$	9–40	0.081–0.202	6/6
$\rho = .25$	93–193	0.079–0.106	6/6
$\rho = 1$	87–168	0.083–0.106	6/6
Variance mass and cost	81–185	0.078–0.105	6/6

For $\rho = .25, .5, 1$, mean source reference mass labeled death changes from .278 to .0667 to zero, although selected overlap remains above every shuffled mean in all six pairs. These labels depend on the mass model and thresholds; they do not independently establish feature disappearance.

Index and exact-link sensitivity. A broader plan audit contains fourteen configurations on each of the same six fixed pairs: the default, decoder-only and magnitude-heavy costs, five leave-one-component-out costs, $\varepsilon = .005, .02, .05$, $\rho = .25, 1$, and variance in both mass and cost. No selected edge has equal indices in 83 of the 84 setting–pair combinations. The exception removes decoder direction on TinyLlama seed 3: one of 133 links (.75%) has equal indices. The four settings without activation-overlap evaluation are removal of prevalence, mean or second moment, and $\varepsilon = .02$. The 84 plan settings reuse the same fitted SAE pairs; activation-overlap evaluations cover a subset of those 84 settings.

Exact selected-edge sets are more sensitive to the matching specification than the above-shuffle activation-overlap result: their intersection-over-union with the stored default is 8–22% under magnitude-heavy reweighting, 38–63% under decoder-only cost, and 62–81% under variance mass and cost. These compare *edge identities*, not sets of activating examples. Above-shuffle activation overlap therefore does not make individual links insensitive to the matching specification. These diagnostics reuse the saved plans and fitted SAEs.

Maximum-mass assignment and descending-mass greedy selection necessarily retain the default selected core; their additional edges have lower overlap but exceed shuffled controls on the six rows with complete saved activation tensors (Appendix F.4). At $T_{\text{persist}} = .85$, canonical/same-index overlap ratios are 42–111$\times$ on the six historical LoRA rows (Appendix G.2). The gauge ablation preserves the dissociation on one TinyLlama pair (Appendix F.3).

C.7 Multi-model endpoint correspondence

Multi-model endpoint panel. The endpoint design fixes a step-0 SAE and retrains three step-3000 SAEs in each of five settings: TinyLlama L11, Qwen3 L14, Llama-3.2-1B L8, and Gemma-2-2B L13 with low-rank adaptation (LoRA), plus TinyLlama L11 with full supervised fine-tuning (SFT). LoRA uses rank 16 and $\alpha = 32$; full-SFT updates all backbone parameters. SAEs have width 4096 and train for eight epochs at learning rate 3×10^{-4}. The six recovered TinyLlama/Qwen3 pairs use activation-penalty coefficient $L_1 = .2$, with mean active counts L_0 of 1588–1611 and 743–750 during target-SAE training. Their response statistic uses the final nonpadding token. Historical fitting-document identities were not recovered, so their separation from evaluation cannot be rechecked (Appendix C.8).

The supporting TinyLlama and Qwen3 transition study compares five checkpoints; Appendix E specifies the WikiText-2 and AG News grids, shared-seed protocol, and alternating-seed control.

Local indices and cross-checkpoint links. Among selected pairs, the three designated same-seed reference rows have equal-index incidence 95–100%; asymmetric endpoint SAEs have incidence $\leq 1.2\%$ (Table 18). Selected target-index sets are 93.5–100% disjoint across the five independently seeded target-SAE triads (Appendix F.2). These target-ID overlaps show no detectable enrichment over uniform random subset sampling at $K \ll N = 4096$ (upper-tail hypergeometric $p \in [.161, 1]$; Appendix F.2).

Activation correspondence across independent fits. Across all 15 endpoint comparisons, selected pairs have greater activation overlap than the same-index and shuffled controls. Selected UOT pairs have mean intersection-over-union 0.053–0.107, compared with 0.0007–0.0044 for same-index pairs and 0.0050–0.0134 for shuffled pairs. Each row exceeds all of its shuffled means (seven rows use 30 draws, eight use 1000; descriptive comparisons without adjustment across rows). The $K = 67$–197 selected pairs cover 1.6–4.8% of each width-4096 dictionary by count. Gemma-2-2B's three rows have the smallest absolute selected-pair Jaccards (0.053–0.076), still 12–18$\times$ their same-index controls. Table 38 reports same-seed descriptor-matched references.

Table 18: Endpoint activation correspondence. K is selected-pair count; Id% is the fraction with equal source/target indices. J is mean intersection-over-union of sets of up to twenty strongest positively activating AG News test examples. Selected coverage is 1.6–4.8% by count. Same-index control indices are sampled outside the selected sets. Same-seed controls are in Table 38.

Condition	K	Id%	Selected J	Shuffled J	Same-index J
Asymmetric-seed audit: five cells × three endpoint seeds					
TinyLlama LoRA asym (0,1)	91	0.0	0.091	0.007	0.002
TinyLlama LoRA asym (0,2)	94	0.0	0.082	0.007	0.002
TinyLlama LoRA asym (0,3)	113	0.0	0.101	0.006	0.002
Qwen3 LoRA asym (0,1)	149	0.0	0.098	0.012	0.002
Qwen3 LoRA asym (0,2)	185	0.0	0.107	0.012	0.001
Qwen3 LoRA asym (0,3)	155	0.0	0.104	0.013	0.001
Llama-3.2-1B LoRA asym (0,1)	182	0.0	0.088	0.005	0.001
Llama-3.2-1B LoRA asym (0,2)	192	0.0	0.095	0.006	0.001
Llama-3.2-1B LoRA asym (0,3)	197	0.0	0.090	0.005	0.001
Gemma-2-2B LoRA asym (0,1)	67	0.0	0.073	0.009	0.004
Gemma-2-2B LoRA asym (0,2)	70	0.0	0.053	0.009	0.004
Gemma-2-2B LoRA asym (0,3)	76	0.0	0.076	0.009	0.004
TinyLlama full-SFT asym (0,1)	86	0.0	0.083	0.008	0.003
TinyLlama full-SFT asym (0,2)	81	0.0	0.096	0.008	0.002
TinyLlama full-SFT asym (0,3)	83	1.2	0.101	0.009	0.003

C.8 Training, evaluation, and scale records for the supporting studies

The six TinyLlama/Qwen3 comparisons have saved full source and target SAE weights, descriptors, evaluation configurations, and full activation tensors. Metadata inside all twelve SAE files record width 4096, eight epochs, learning rate .0003, an L_1 penalty coefficient of .2, and the training-space label `normalized_unit_rms` (RMS: root mean square). The three source copies per model have identical hashes and record seed 0; targets have distinct hashes, record seeds 1, 2, and 3, and all files record no warm start. Their hashes agree with the training manifest. The metadata record separate fits; complete training random-state histories are unavailable. Source descriptor arrays and reference masses are identical across the three TinyLlama relations, whereas Qwen3 has two descriptor/mass versions despite identical source SAE parameters. The cause is unresolved; the Qwen3 query aggregation retains these versions rather than isolating target-seed variation.

Every decoder column in these twelve files has unit Euclidean norm (maximum measured deviation 3.22×10^{-15}). The second-moment masses therefore refer to this fixed latent scaling convention. A transformation $d_i \mapsto \alpha d_i$, $z_i \mapsto z_i/\alpha$ leaves reconstruction unchanged but changes q_i by α^{-2}; the activity measure therefore depends on the latent scaling convention. Unit decoder norms also do not equalize intervention sizes when the latent coefficient differences change.

Table 19: Training diagnostics stored with the recovered SAE weight hashes. The scores were recorded during training rather than measured on the evaluation set. Appendix E.2 reports calibration results for separate cohorts.

Artifact	Training L_0	Training fraction of variance unexplained
Qwen3 source	664.396	.001705
Qwen3 target 1	749.774	.001493
Qwen3 target 2	743.003	.001504
Qwen3 target 3	749.527	.001489
TinyLlama source	1716.729	.002459
TinyLlama target 1	1605.692	.002016
TinyLlama target 2	1611.420	.001965
TinyLlama target 3	1587.623	.002050

Artifact availability differs across the broader fifteen-row panel. All fifteen compact result files are retained. Full weights and full evaluation tensors were checked for the six TinyLlama/Qwen3 rows. The three Llama-3.2 and three TinyLlama full-SFT rows retain selected-column evaluation tensors but lack a complete provenance record linking target weights to their generating configurations in the inspected release. The three original Gemma rows retain compact summaries; their selected-column tensors and complete generating chain were not recovered. Public Gemma Scope artifacts describe a separate experiment and do not fill this gap. These differences limit how fully each condition can be reproduced.

The six-row evaluation reconstruction loads the AG News test split explicitly, uses right padding and maximum length 256, records the residual layer and adapter, and pools each feature at the final non-padding token. It produces 7600×4096 float32 arrays at each endpoint. The six original selected-pair means are reproduced with a maximum absolute difference of approximately 2.8×10^{-17}, supporting the chosen artifacts and evaluation convention. This agreement of six means does not establish byte-identical tensors or control samples.

The historical SAE-training and descriptor-building text files were not recovered; projected profiles do not retain document identities. Consequently, disjointness of the SAE-training and descriptor-building documents from AG News test examples cannot be rechecked from the current archive. The original protocol specifies separate fitting and evaluation corpora, but their document-level disjointness could not be verified. A fully reproducible generalization test requires archived fitting and evaluation texts or hashes, explicit disjointness, and the generating implementation. The controlled benchmark's calibration/evaluation split applies only to that benchmark.

Shuffle-based ranks and p-values are reported per audited condition ($n = 30$ for seven original rows and $n = 1000$ for eight; $n = 30$ in the new six-row comparisons) and are not family-wise or across-condition multiple-testing corrected unless explicitly noted.

The historical class screen retains the selected-feature counts and four-transition persistence rates, but not the screening-score formula or branch-tracing implementation. The feature-removal LOO records omit the mass-renormalization and support-update rules. The head-level sweep retains outcome summaries, but not the ordering of canonical subsets or handling of requests larger than the available set. These archived diagnostics are separate from the primary sparse study, whose fitting protocol and selection/evaluation implementations are included in the evidence release.

Packaged records. The compact supplement supplies the primary response outcomes, source tables, supporting code, and aggregate summaries; it does not include every historical per-edge record or diagnostic input. The reported comparisons distinguish original measurements from later analyses using the fitted SAEs. The selected-count comparison was specified after the original experiments and has no solution-level confidence interval.

Scope of the supporting studies. The historical endpoint panel covers five model/training settings, with additional comparisons that reseed both endpoints on TinyLlama L11 and Qwen3 L14. These response-overlap studies use denser dictionaries, different response conventions and incompletely recovered fitting provenance. They cannot be pooled with primary response recovery or used to attribute score differences to the transport method. Margin substitutions are separate descriptive intervention measurements; the head-level study is retained in the Supplementary analyses, Section H3. The primary study's fitting and evaluation protocol is specified in Appendix A.1.

C.9 Controlled benchmark construction and calibration

We use the step-0 descriptors from Q3-1 and TL-1, which represent two distinct decoder dictionaries. Source files shared by multiple endpoint comparisons represent the same dictionary. Each dictionary supplies three calibration pools and nine evaluation pools, each with 64 source atoms and 64 reserve atoms. Pools are globally disjoint within the dictionary. All atoms used in an earlier partition that treated duplicated source files as distinct dictionaries are excluded from calibration and evaluation. The partition seed is 313272028. Eligible atoms have positive second moment and nonzero decoder and profile vectors. Marginals are recomputed and normalized after subsampling. Applying the six transformation cases to every pool produces 36 calibration and 108 evaluation problems.

Complete overlap copies and permutes the source. In the mass-shift case, each target atom receives an independently drawn factor $f_j = \exp(\xi_j)$, with $\xi_j \sim N(0, .5^2)$; its mean is multiplied by f_j and its second moment by f_j^2. Replacement removes 16 or 32 counterparts and substitutes uniformly chosen reserve atoms. The 25% replacement case with drift adds an isotropic vector of norm .35 to each retained decoder direction before renormalization, and applies lognormal amplitude drift with log standard deviation .25. The split case duplicates 16 atoms; each child receives half its parent's second moment and its parent's mean divided by $\sqrt{2}$. This conserves activity mass rather than reconstructed activations. Target indices are permuted in every case. These constructed perturbations set the control conditions; their magnitudes are not estimates of training drift, and their correspondences do not establish semantic lineage.

Known copied/parent–child identifiers define valid edges. The reference G transports the maximum valid mass subject to source and target capacity. Let V_i be the set of valid target children or copies of source i, and let $B_i = \sum_{j \in V_i} b_j$. For $B_i > 0$, set $G_{ij} = b_j \min(1, a_i/B_i)$ on valid edges and zero elsewhere; set the entire row to zero if $B_i = 0$. Thus a one-to-one copy receives $\min(a_i, b_j)$, and a split distributes $\min(a_i, B_i)$ in proportion to its children's target reference masses. In particular, mass shifts can leave reference mass unmatched even when every feature has a counterpart. Unmatched activity is therefore not a feature-death label.

For each plan, false mass is $\sum_{G_{ij}=0} P_{ij}$, missing true mass is $\sum_{G_{ij}>0}(G_{ij} - P_{ij})_+$, and excess true mass is $\sum_{G_{ij}>0}(P_{ij} - G_{ij})_+$. Their sum is $\|P - G\|_1$. True-mass recall is $\sum_{G_{ij}>0} \min(P_{ij}, G_{ij}) / \sum_{ij} G_{ij}$. Reporting all components prevents a zero plan or arbitrary mass expansion from appearing successful. This reference favors a conservative subcoupling; UOT permits mass above these reference caps, so the component errors matter as well as their sum.

All methods use the original cost and complete support between real atoms, with $\varepsilon = .01$. Balanced rejection adds one dummy node of mass 1 to each side, costs λ on real-to-dummy and dummy-to-real edges, zero cost on the dummy-to-dummy edge, and entropy on the expanded coupling. Its real block has submarginals bounded by a, b; for transported real mass T, the linear rejection cost is $2\lambda(1 - T)$. This expanded-coupling regularizer is not identical to KL-UOT. Calibration minimizes the same mean $\|P - G\|_1$ over both dictionaries, three calibration pools and six cases, using $\rho \in \{.03, .06, .125, .25, .5, 1, 2\}$ and $\lambda \in \{.015, .03, .06, .125, .25, .5, 1\}$. Calibration selects $\rho = .125$ for UOT and $\lambda = .125$ for balanced rejection; evaluation examples do not select either parameter. We also report original $\rho = .5$ and $\varepsilon = .005, .02$ sensitivities without retuning.

All 2,808 final solves converged, with maximum balanced marginal L_1 residual 9.993×10^{-10} (rounded upward). Stiff balanced/rejection cases use damped Newton refinement of the same scaling objective after ordinary Sinkhorn iterations. Every saved plan and error decomposition was recomputed from the 144 raw problem files. The compact supplement includes calibration-grid and per-dictionary summaries, input and implementation hashes, partition checks, and recorded residuals; the full raw problem arrays are separate inputs. Pool averages describe subsamples of two dictionaries and do not support population inference.

The original $\rho = .5$ has false mass 15.41% and 28.70% at 25% and 50% replacement, respectively; its total L_1 errors are 21.35% and 33.86%. Calibrated UOT errors are 15.78% and 16.08%. Across all six cases, calibrated UOT has mean L_1 error 10.05%, 13.58%, and 22.20% at $\varepsilon = .005, .01, .02$. UOT places excess mass on complete overlap, and the rejection baseline has lower errors. Every retained atom has a correct nearest cost neighbor in these controls. They isolate the consequences of mass constraints when descriptors are informative; they do not establish performance on difficult semantic matching or calibrate birth/death in real trajectories. Normalization within these small dictionaries can also make activity mass highly concentrated, so these effect sizes should not be extrapolated to width-4096 dictionaries.

Results and interpretation. The known correspondences define a reference coupling capped by source and target activity masses. At 25% and 50% replacement, balanced OT sends 32.48% and 55.80% of unit reference mass outside the known relation; calibrated UOT reduces these to 3.24% and 6.50% with 100% recall after capping each valid edge at its reference value. UOT also places 12.53% and 9.58% excess mass above those caps. Balanced transport augmented with explicit rejection recovers the reference more accurately (Table 20). Balanced OT also solves the complete-overlap

and split controls. UOT's total mass on complete overlap is 111.16% at calibrated $\rho = .125$, versus 102.75% at the original $\rho = .5$; the original endpoint experiments retain $\rho = .5$.

These controls support relaxing or rejecting incompatible mass rather than forcing complete correspondence. The rejection comparator adds dummy nodes to absorb unmatched mass; transport between the original features can then fall below their reference marginals. It is a strong alternative when the desired reference is a subcoupling bounded above by both activity measures. KL-UOT instead penalizes relative marginal discrepancies smoothly and permits increases as well as decreases; these are different mass models, and the constructed subcoupling reference favors rejection. Every retained source has a correct target as its nearest cost neighbor, so rejection can exploit that separation. The split preserves activity mass, not SAE reconstruction. The controls test mass handling under constructed correspondences; they do not establish semantic lineage.

Table 20: Constructed controls, averaging 18 evaluation pools per case. All entries are percentages; mass uses the unit reference scale. UOT and rejection penalties are selected on calibration pools disjoint from evaluation pools. False mass lies outside the known relation; excess mass exceeds the reference on valid edges. Recall caps each valid edge at its reference mass. Rejection L_1 error sums false, excess, and missing reference mass. Values rounded to 0.00 may be nonzero.

Case	Balanced false mass	UOT false mass	UOT excess mass	UOT recall	Rejection L_1 error
Complete overlap	0.00	0.00	11.16	100.00	0.00
Mass shift	33.08	0.00	15.13	97.04	0.03
25% replacement	32.48	3.24	12.53	100.00	0.00
50% replacement	55.80	6.50	9.58	100.00	0.00
25% + drift	39.31	3.12	5.36	97.08	0.00
Mass-conserving split	0.00	0.00	8.61	98.74	0.01

C.10 Multi-fit descendants and conditional controls

The descendant analysis holds fixed the source features with a selected-core edge in all three target fits. Each resulting family contains one descendant per target dictionary. For each source, we evaluate the mean and minimum of its three target-target Jaccards, then average over sources. The comparator holds all three descendant populations fixed and independently permutes the second and third relative to the first, using 10,000 draws. It tests conditional pairing structure, not excess source recurrence or full-dictionary coverage. Results describe four studied cells; neither the source features nor the permutation draws provide independent population replications.

Table 21: Three-fit descendant coherence on selected source anchors. Entries are observed / re-pairing mean. The weakest-pair statistic takes the minimum within each source family before averaging.

Cell	Sources	Mean pairwise J	Weakest-pair J
Qwen3 LoRA	20	0.2319 / 0.0304	0.1670 / 0.0049
TinyLlama LoRA	9	0.1773 / 0.0247	0.1277 / 0.0022
TinyLlama full-SFT	9	0.1101 / 0.0182	0.0607 / 0.0017
Llama-3.2 LoRA	25	0.2566 / 0.0172	0.1683 / 0.0009

A class signature is the vector of a feature's mean activations in the four AG News classes, centered across classes and normalized to unit norm. We compare signatures by cosine across disjoint class-stratified example halves, averaging both half assignments over 200 splits and scoring a comparison as zero if either centered norm is at most 10^{-8}. On the triple families, these scores also exceed exact-set re-pairing. This is coarse class correspondence, not unique semantic identity. A separate hard-negative analysis uses sources selected in both members of each target-fit pair, including sources absent from the third fit. The three pairwise intersection sizes are 53/51/59 (Qwen3), 18/28/32 (TinyLlama LoRA), 15/27/18 (TinyLlama full-SFT), and 70/60/71 (Llama-3.2), totaling 502 records, each representing a source and pair of target fits. For each fit pair, a fixed class-stratified split supplies selection and evaluation halves. On the selection half, we rank all descendants separately

by Jaccard and class-signature cosine, then choose the other descendant with the smallest sum of these descending ranks; rank-sum ties are broken by the candidate's local feature index. On the evaluation half, the margin is the source-linked descendant's score minus the alternative's score. Signature comparisons use disjoint stratified subhalves within each outer half. Each record averages both assignments of selection/evaluation half and both query directions. Jaccard margins are positive in 12/12 fit-pair summaries, but class-signature cosine margins are positive in 0/12. Averaging the three fit-pair means within each cell and then weighting the four cells equally gives margins .0976 and $-.4596$, respectively. Only 83/502 records have positive margins for both outcomes. The comparison remains conditional on the selected descendants.

Direct target-target Hungarian matching uses the Qwen3 and TinyLlama LoRA pairwise-common anchors just defined: their six intersection sizes sum to 241. On those fixed target endpoint sets it recovers 241/241 source-mediated edges using full descriptor cost and 239/241 using decoder cost alone. The comparison uses these pairwise intersections, which differ from the 87 links in the 20 Qwen3 and 9 TinyLlama families selected in all three fits. Thus the source-mediated correspondences agree with a separately computed direct assignment on these endpoint sets. This is agreement between matching routes on selected populations, not path independence of the full weighted plan. The source bridge does not add pairing information once these nodes are fixed; source-based node selection and query-level reporting are different questions. Recomputing the observed Qwen3/TinyLlama triple Jaccards from example sets extracted from the saved full activation tensors reproduced the values exactly, using the same observations.

C.11 Source-query population and conditioning

This section describes the historical resolver and its selected-link evidence. The calibrated response realizations in the primary study additionally retain the coefficients, intercepts, and prediction errors defined in §3.1; they do not inherit the historical selected-core restriction.

Reference resolution. The historical resolver uses six stored sparse plans and their source and target artifact identifiers. The compact supplement includes its source code and aggregate query summaries, but omits the plan data required to run queries. A query specifies a source dictionary and local index; each target fit returns the incident edge masses, selected-core membership, reference masses, and transported marginals. Candidate fractions P_{ij}/u_i are defined only for $u_i > 0$ and are not semantic probabilities. A row with no positive transported edge has no transported candidate; every source row in the six archived plans has a positive edge. Dictionary identifiers remain fixed while a changed descriptor or solver specification defines a new relation version. The operation organizes the fitted components into source-anchored references; it does not collapse connected components into equivalence classes.

The minimum scientific record specifies the observation, its input population and measurement rule, and the corroborating, conflicting, or unresolved outcomes in each assessed fit. It also records source and target SAE identifiers, checkpoint/layer/width/seed metadata, descriptor version, OT hyperparameters and persistence threshold, weighted source–target edges, reference and transported marginals, selected count K, selected-endpoint count and reference-mass coverage, and evaluation/control summaries. Support attaches to the measured property: shared activating examples do not by themselves establish conserved magnitude, semantic identity, or interchangeable interventions.

The illustrative sources were fixed before inspecting their evaluation overlap. For each model, plan-only rules choose the smallest source index with the most core-linked fits, then one branching proposal and one low-relative-mass proposal. All six examples, including their weak candidates, are retained in the separate full historical archive. TinyLlama source 784 is the first of these choices. Its target endpoints in fits 1/2/3 are 2241/2591/3409, with Jaccards .1429/.0811/.1765 and transported/reference source-mass ratios .678/.649/.564. Decoder-nearest selects the same endpoints.

The all-source query analysis uses the six original archived sparse plans, their selected cores, and saved 7600×4096 evaluation tensors. Every source feature is queried against all three target fits. Sources are grouped into strata by the number of fits with a selected-core edge, irrespective of exclusive node-label precedence; this number is not a validated semantic-resolution count. The diagnostic compares the strongest transport edge with the decoder-cosine nearest neighbor. Scores use at most 20 strictly positive activating examples, ordered by decreasing activation then example

ID; both-empty sets score zero. No ambiguous cutoff occurred. We average three fit scores within each source, then average sources.

Table 22: All-query accounting. r counts fits with a selected-core edge. Mass is the sum of source reference masses in the stratum, in percent; Qwen3 ranges retain its two source descriptor/mass versions. Scores are unweighted source means across all three fits.

Cell	r	Sources	Ref. mass (%)	Top-plan J	Decoder NN J
Q3	0	3750	67.3606–67.3617	0.0064	0.0083
Q3	1	223	18.9035–18.9041	0.0658	0.0755
Q3	2	103	13.6789–13.6792	0.1016	0.1059
Q3	3	20	0.0559–0.0560	0.1425	0.1425
TL	0	3867	94.9918	0.0100	0.0129
TL	1	169	3.7391	0.0624	0.0644
TL	2	51	1.0716	0.0984	0.0963
TL	3	9	0.1975	0.1688	0.1729
Q3	All	4096	100	.0127	.0150
TL	All	4096	100	.0136	.0164

A source selected in one target fit need not have a selected descendant in another, so cross-fit reproducibility must be assessed over all target fits. Among the 346 Qwen3 and 229 TinyLlama sources selected in at least one fit, averaging only selected fits gives mean Jaccard .09601 and .08204; evaluating the strongest plan candidate in every fit gives .08087 and .07457 for the same sources. Only 20 and 9 sources, respectively, have selected links in every fit.

Averaging only selected fits changes the within-query denominator, not the source population. There are 2249 Qwen3 and 666 TinyLlama source features with empty evaluation example sets, all retained. Original selected-core means reproduce, and no selected-core edge in these six rows conflicts with the source death threshold. Core membership and transported/reference marginal ratios remain separate reported fields.

Source encoder/decoder weights and biases are identical across the three relations within each model. Qwen3 nevertheless has two source descriptor and reference-mass versions; TinyLlama has identical source descriptors and masses. We retain relation-specific masses and do not attribute Qwen3 differences solely to target-SAE seed. These descriptive comparisons reuse existing trained artifacts, have no solution-level uncertainty estimates, and do not evaluate the full weighted activation family. The historical document-provenance limitation in Appendix C.8 also applies here.

C.12 Class-conditioned endpoint shuffles

For the six recovered endpoint rows, we apply class-conditioned shuffles to pairs already selected by the default and normalized decoder-only costs. We hold target endpoints fixed and permute entire source top-20 positive-activation sets, first without restriction, then within source majority-class signatures (ties retained), and finally within identical four-class count vectors. The last control preserves each edge's selected source and target class histograms, the selected endpoint populations, and each example's total occurrence count on each side. It does not fix pairwise activation prevalence or pairwise example-popularity similarity. Class labels are used to form shuffle groups, without refitting the relations; decoder-only cost still uses activation-derived reference masses.

The observed movable-subset mean exceeds its null expectation in all 12 comparisons, but randomization power is uneven. Default TinyLlama permits only 6, 10, and 17 endpoints to move, with exact upper-tail fractions $1/8$, $1/32$, and $1/768$. Singleton contributions are fixed: the full-K score is a constant plus K_m/K times the movable-subset score, so these are not independent tests. The majority-class control leaves 90–182 default endpoints movable and every observed mean exceeds all 10,000 sampled null means; majority class alone does not control class purity.

All twelve canonical means reproduce the saved evaluation, with no both-empty canonical pairs or ambiguous positive cutoffs. Analytical expectations were independently checked on five conditions, with complete enumeration of four small permutation spaces. The better-powered conditions show

Table 23: Exact-class-histogram control on the movable subset. K_m counts source endpoints in groups of at least two endpoints with identical four-class histograms. Null means are analytical expectations; tails are exact where marked E and otherwise from 10,000 permutations with the $+1$ correction (M), unadjusted across rows.

Cost	Row	K	K_m	Observed J	Null mean	Upper tail
Default	Q3-1	149	32	0.07880	0.04508	0.000500 M
Default	Q3-2	185	49	0.09311	0.04843	$1/10001$ M
Default	Q3-3	155	30	0.07421	0.03913	$1/10001$ M
Default	TL-1	91	6	0.07261	0.03631	0.125000 E
Default	TL-2	94	10	0.12832	0.06416	0.031250 E
Default	TL-3	113	17	0.12925	0.06308	0.001302 E
Decoder only	Q3-1	178	48	0.07756	0.04480	$1/10001$ M
Decoder only	Q3-2	221	72	0.09958	0.05302	$1/10001$ M
Decoder only	Q3-3	196	39	0.09618	0.04772	$1/10001$ M
Decoder only	TL-1	181	30	0.06003	0.03346	0.002075 E
Decoder only	TL-2	170	34	0.06311	0.03194	$1/10001$ M
Decoder only	TL-3	199	49	0.08769	0.04363	$1/10001$ M

pairing-specific same-example overlap beyond the class composition preserved by this control. This control tests activation overlap conditional on class histograms; it reuses evaluation documents and source dictionaries and does not compare alternative matchers.

C.13 Document-split correspondence and position-weighted activity mass

The document-split experiment separates descriptor construction from evaluation by document group and measures how token position affects the activity reference. It uses saved TinyLlama L11 SAEs and residual caches from the AG News LoRA trajectory without additional model or SAE training. It is separate from the fifteen final-token Jaccard comparisons in Table 18.

Fits, documents, and relation construction. The dictionary symbols in this subsection are local to this historical cohort and do not denote the newly fitted dictionaries in Section 4. Let $A_0, \ldots, A_4$ denote the five step-0 restart SAEs and $B_0, \ldots, B_4$ the five step-3000 restart SAEs. A_r and B_r share initialization; the twenty $A_r \to B_s$ pairs with $r \neq s$ use independent seeds. Ten unordered refit pairs at each checkpoint supply twenty same-checkpoint comparisons. Three further target SAEs, C_1, C_2, C_3, are the original TinyLlama asymmetric-audit targets, distinct from $B_1, \ldots, B_4$; the shared source A_0 is linked to the original source weights by its recorded hash. These three extensions were added after the initial protocol of twenty-five transition comparisons and twenty refit comparisons was specified.

The two endpoint caches contain 688,198 rows in 4,500 groups with matching group IDs and lengths. A check of 256 rows supports the row alignment; token IDs and document strings were not recovered. Even group IDs supply 2,250 descriptor groups and odd IDs supply 2,250 evaluation groups. After excluding each group's first stored position, these contain 339,663 and 344,035 rows, respectively. The exported encoders are applied directly in their stored scale. Descriptors, costs, supports, and selections use only descriptor groups, with the default cost weights, union top-eight support, $\varepsilon = .01$, and $\rho_s = \rho_t = .5$. A shared Gaussian projection maps each profile to 64 dimensions. Squared-activity references are floored at 10^{-12} before normalization. The positional concentration was inspected before this protocol was specified; evaluation responses were not used to fit the relations.

Correspondence on separate groups. For each selected link, evaluation computes the cosine between its two vectors of latent responses on nonfirst evaluation rows, and the corresponding centered correlation. Cosine is zero when either feature is inactive. Centered correlation can be undefined for inactive pairs; these values are omitted from the all-index mean. Selected-link and selected-endpoint re-pairing correlations are finite in all 48 comparisons. The re-pairing control is the exact mean over the selected-source $\times$ selected-target score matrix. Same-index cosine averages all 4,096 indices, unlike the original final-token control drawn outside selected endpoint sets. Hungarian

assignments and mutual nearest neighbors use the same descriptor cost and retain their K lowest-cost pairs, where K is the selected UOT link count. Table 24 averages each statistic equally across comparisons; it does not treat tokens or reused dictionaries as independent replications.

Table 24: Document-split TinyLlama correspondence. Entries are mean cosines between latent-response vectors over evaluation rows; K is selected UOT link count per comparison. Hungarian and mutual-NN columns use the same K. The three original-audit targets are an extension of the initial protocol.

Pairs	n	K	UOT	Re-paired	Same index	Hungarian	Mutual NN
Independent endpoints	20	89–127	.765	.393	.128	.807	.808
Original-audit targets	3	99–122	.764	.407	.128	.810	.809
Shared-seed endpoints	5	1592–1697	.892	.554	.491	.802	.802
Same-checkpoint refits	20	89–214	.756	.393	.128	.806	.806

Selected cosine and centered correlation exceed their respective re-pairing controls in all 48 comparisons. In the twenty independent endpoint pairs, selected cosine ranges from .733 to .782 versus .343–.422 for re-pairing; centered correlation averages .591 versus .026, with ranges .564–.629 and .020–.034. Hungarian and mutual-NN centered correlations average .690 each. Both hard matchers have higher cosine than the UOT core in every independent endpoint and original-audit-target comparison. For shared-seed pairs, same-index cosine on the selected source population averages .885, close to the selected .892; the all-index control in the table uses a different population. Selected links thus show activation correspondence under the descriptor/evaluation split; separation from the SAE-training data remains unverified.

What pooled activity mass weights. Across all 688,198 cached rows, the 4,500 first stored positions constitute .654% of rows but contribute 58.65–69.53% of summed squared latent activity in $A_0, \ldots, A_4$ and 75.49–83.30% in $B_0, \ldots, B_4$. Each dictionary has 6–14 features receiving more than half of their own squared activity at these positions. These diagnostic percentages pool both document halves; descriptor-half percentages differ slightly. The caches do not identify these positions as BOS tokens or establish an attention-sink mechanism.

For the descriptor-half $A_0 \to B_0$ reanalysis, the all-token plan has stable-core mass .209 (the average of source and target reference mass assigned persistent event labels; Appendix D.1). Keeping the all-token plan's state assignments fixed and reweighting by nonfirst-position activity raises its stable-core mass to .844. Rebuilding the relation using nonfirst positions gives stable-core mass .881, with 1,577 selected links shared between the all-token plan's 1,579 and the rebuilt plan's 1,592 links. That rebuild changes profiles, prevalence, activation moments, costs, support, and references together; it is not an isolated mass-only ablation. The fixed-state comparison demonstrates why a large shift in mass-weighted persistence need not represent a corresponding increase in recovered links. Both position conventions are well-defined measurements. Neither measures the fraction of model knowledge, and these endpoint findings do not supply a positional decomposition for every archived trajectory table.

Response combinations beyond the selected core. For source features active on evaluation rows (second moment above 10^{-12}) without a selected link, we test whether several target responses agree more closely with the source than one target alone. Let B_i contain targets whose edges have mass at least 10^{-6} and carry at least .2 of the source row's transported mass. There are 563–737 sources with $|B_i| \geq 2$ per independent-seed pair, each with two to four branches. The predictor $\hat{z}_i(x) = \sum_{j \in B_i} P_{ij} z_j^t(x)$ uses descriptor-fitted weights and is scored on the 344,035 nonfirst evaluation rows. Cosine is unchanged by normalizing each row's weights. Table 25 averages equally over sources, then over the twenty pairs. For each dictionary pair, averaging over eligible source features gives higher cosine for the weighted sum than for either single-target comparison. Centered correlation also improves over the target with the largest transported mass. Equal-weight groups do better still. Combining target latents therefore recovers the source response more closely in these historical fits. The improvement could reflect complementary responses or averaging over noisy counterparts; it does not establish semantic splitting or show that the transport weights can reconstruct model activations.

Table 25: Response agreement outside the selected one-to-one core. The best single target is chosen among all 4096 targets using evaluation responses; the other nonrandom choices use descriptor data. Groups contain the same number of targets as the qualifying UOT branches. Low-cost groups use the lowest-cost candidates within the same row support. Random groups sample descriptor-active targets. Dashes denote metrics not evaluated for that comparator.

Target response	Cosine	Centered correlation
Single target with largest edge mass	.651	.271
Best single target (evaluation cosine)	.675	—
UOT branches, transport weights	.729	.341
Same branches, equal weights	.731	—
Low-cost group, cost-kernel weights	.719	.348
Low-cost group, equal weights	.737	—
Random group, equal weights	.433	—

Scope and reproducibility. The experiment covers one model, layer, task, and adaptation trajectory. To assess the relaxed marginals, we compare transport shortfalls with activity changes measured using each fixed SAE at both checkpoints. Among features with second moment above 10^{-12} at the fitting checkpoint, the pre-specified broad change label marks a second-moment ratio (other checkpoint / fitting checkpoint) outside $[.5, 2]$ or response cosine below $.8$. Nearest-neighbor cost ranks this label better by AUROC than UOT shortfall, $1 - u_i/a_i$ for sources or $1 - v_j/b_j$ for targets. The narrower target label, a base/current second-moment ratio below $.5$ for a fixed target SAE feature, favored UOT shortfall over nearest-neighbor cost by mean AUROC. This directional comparison was not specified in the initial protocol. Fitting-document provenance remains unavailable for the original fifteen-row analysis (Appendix C.8). Accompanying numerical records retain all per-comparison cosine and correlation summaries, the positional shares, and the aggregation used for Table 24; rerunning encodings and transport also requires the residual caches and SAE artifacts.

D Notation and diagnostic conventions

Appendices A.1 and A.2 specify primary fitting and response evaluation. The definitions below interpret the separate correspondence, transition, and intervention diagnostics. Their cohort-specific records remain distinct from the primary recovery experiment. Supplementary analyses, Sections H1–H6, retain the exploratory studies and detailed supporting grids.

In retained table and figure labels, "canonical" or "can" means descriptor-selected, "cell" means a model/layer/task/training setting, and "asym" means different SAE initialization seeds. "Same-solution" means separately fitted SAEs with shared initialization, not fixed SAE weights.

D.1 Notation and terminology

Source and target features have local indices i and j; superscripts s, t distinguish their dictionaries. Section 3.2 defines document and token indices, and Eq. 3 defines the descriptor cost. The terms below define the diagnostic conventions and translate the labels retained in supporting figures and tables.

Sparse autoencoder (SAE). A model that reconstructs activations using a sparse combination of learned decoder directions; z_i is a feature's latent activation and d_i its decoder direction.

Response realization. The predictor in Eq. 1: at most k target activation streams, nonnegative coefficients w_{irj}, and a free intercept b_{ir} fitted on calibration inputs. Its coefficients are distinct from transport masses.

Normalized squared error. Mean squared error of a frozen response prediction divided by source variance on evaluation inputs. The affine singleton oracle minimizes this quantity over all target latents, slopes, and intercepts on evaluation itself.

All-five recovery. The same source response meets the stated error tolerance in each of five independently fitted target dictionaries. It need not require more than one latent in every fit.

Activation profile. The vector of document-wise mean activations and active-token fractions, in a common document order. Writing $\tilde{p}_i = (\mu_{i1}, \ldots, \mu_{iN}, r_{i1}, \ldots, r_{iN})$, set $p_i = \Phi\tilde{p}_i$ if $2N > 64$, otherwise $p_i = \tilde{p}_i$. The Gaussian matrix $\Phi \in \mathbb{R}^{64\times 2N}$ is fixed across checkpoints, and $\hat{p}_i = p_i/\|p_i\|$ for nonzero p_i; zero-profile conventions for archived plans and later comparisons are distinguished in Appendix C.2.

Prevalence. π_i, the fraction of descriptor-corpus tokens for which $z_i > 0$. It pools tokens, rather than weighting documents equally.

Second moment. $q_i = \mathbb{E}[z_i^2]$, including the squared mean: $q_i = \mathrm{Var}(z_i) + \bar{z}_i^2$. It supplies the default activity reference mass and one cost component.

Probe documents. Documents used to construct matching descriptors. The primary study records 1,998 descriptor documents separate from fitting and evaluation. Historical descriptor-document identities were not recovered (Appendix C.8).

Evaluation corpus. The 7,600 AG News test documents. The primary study verifies their separation from its new SAE fitting, calibration, and descriptor partitions; historical fitting separation remains unverified.

Transport masses. a, b are normalized activity *references*; $u = P\mathbf{1}$ and $v = P^\top\mathbf{1}$ are the actual transported marginals. Under UOT, u and v may fall below or exceed their references.

Support. $\mathcal{S}$, the allowed source–target edges: the union of the lowest-cost k_{cand} neighbors in each direction. Candidate coverage does not imply balanced feasibility or selection of any particular link.

Entropy parameter. ε controls concentration of transport mass relative to the descriptor cost; changing it changes the count and quality of selected links.

Marginal penalty. ρ_s, ρ_t weight generalized KL deviations of transported marginals from a, b. Both increases and decreases are permitted at finite penalty.

Selected pair. An edge in the partial one-to-one core extracted from $P^\star$, using the full rule in Appendix B.1. This core does not restrict the primary compact-response study.

Selection rule. Mutual-best edge, mass at least 10^{-6}, at least .7 of both transported marginals, and at most one qualifying branch at each endpoint. A qualifying branch has mass at least 10^{-6} and at least .2 of its endpoint's transported marginal.

Selected pair in archived labels. "Canonical and "can are legacy names for a descriptor-selected pair. They do not denote an independently established canonical semantic feature.

Branch. An edge meeting the branch mass/fraction cut. Multiple branches are possible in the full plan but excluded at selected-core endpoints. Plan-defined split/merge labels do not themselves establish semantic splitting or merging.

Scaffold mass. In the archived transition analyses, reference mass assigned a persistent event label, with source and target sides reported separately. The stable-core statistic averages the two sides. Node labels follow Appendix E.1's precedence rules; selected-edge membership alone does not override a birth/death label.

Same-seed trajectory. Each checkpoint SAE is fitted separately from the same initialization seed. Archived "same-solution" tables refer to this protocol; SAE weights are not held fixed across checkpoints.

Different-seed endpoint audit. Archived labels use "asymmetric or "asym for these comparisons. Endpoint comparisons fixing the step-0 SAE at seed 0 and fitting the step-3000 SAE with another seed. This varies dictionary initialization along a fixed model trajectory.

Off-diagonal pairs. Comparisons using seed pairs $(1, 2), (2, 3), (3, 1)$, so neither endpoint uses seed 0. Both endpoints vary; this does not isolate a source-only seed effect.

Model/training setting. Called a "cell in archived figures: a model, layer, task and training-protocol setting; the historical endpoint panel has five such settings, each with three target-SAE seeds.

Jaccard. Intersection-over-union $|A \cap B|/|A \cup B|$. Unless explicitly called an edge-set comparison, A, B are sets of up to twenty strongest positively activating examples. The primary study ranks document-mean responses and treats any empty endpoint as undefined. Both-empty pairs are excluded in the historical evaluator and scored as zero when evaluating all source queries; other comparisons report zero-scored sensitivity where specified.

Equal-index control. For activation overlap, pairs (k, k) use indices sampled without replacement outside both selected index sets. For margin patching, target indices stay fixed and the source coefficient uses the same local index. These test literal index reuse, not optimized descriptor assignments.

Shuffled control. Random re-pairing of the fixed selected source and target sets, testing the particular correspondence while keeping endpoint membership fixed.

Can/Sidx. Mean activating-example Jaccard for selected pairs divided by that for equal-index controls. Can/Shuf replaces the denominator by shuffled controls; these ratios are not absolute overlaps or intervention-effect ratios.

Evaluation. Normalized squared error of frozen compact response predictions, with the affine-singleton oracle and recurrence across fits. Supporting selected-link studies measure response overlap, document and token correlation, and cross-fit coherence. Historical intervention analyses additionally assess joint latent substitutions.

Margin patching. Jointly substitute source latent activations through the target decoder, then measure the change in true-class logit margin. Appendix G defines the perturbation and the patched-minus-original sign convention.

D.2 Bounded event statistics

The lineage matcher is built from a relaxed unbalanced transport plan. Raw plan mass is useful diagnostically, but it is not a probability distribution over event types. In particular, the total matched mass can exceed 1 under the relaxed objective. We therefore report bounded quantities derived from the exclusive event partition:

- source-side {persistent, branch/split, pruning/death, ambiguity} mass shares, partitioning total source mass and summing to 1;
- target-side {persistent, branch/merge, novelty/birth, ambiguity} mass shares, partitioning total target mass and summing to 1;
- persistent-edge count fractions: the number of archived persistent pairs divided by dictionary width (4096);
- *stable-core*, defined as the symmetric average of source and target persistent shares; this is a bipartite diagnostic and is therefore not a member of either side's exclusive partition.

The four shares per side lie in $[0, 1]$ by construction. Stable-core also lies in $[0, 1]$ but is reported separately from the per-side partitions; tables that show stable-core alongside source-only branch and ambiguity values are therefore not constrained to sum to one.

D.3 Within-checkpoint baselines and uncertainty

The restart floor is estimated by training multiple fresh SAEs on the *same* checkpoint and computing the resulting within-checkpoint lineage agreement. It provides a within-checkpoint baseline against which to compare temporal agreement. A temporal similarity score below the restart-floor score does not by itself establish statistical indistinguishability from that baseline.

Threshold and descriptor sweeps test event rules, descriptor weights, entropic regularization, and candidate-top-k pruning. The Supplementary analyses, Section H1, retains the feature-removal LOO summaries and their incomplete-method status (Appendix C.8). A naive bootstrap that duplicates source features changes the candidate graph and transport problem; it does not directly give uncertainty for the original fixed-dictionary statistic. A justified resampling analysis must specify its estimand and how fitting or matching is repeated. This objection does not apply to every bootstrap design.

E TRANSITION DIAGNOSTICS

E.1 EVENT RULES AND SENSITIVITY

Event rules. Let $u_i = \sum_j P^\star_{ij}$, $v_j = \sum_i P^\star_{ij}$, $f^r_{ij} = P^\star_{ij}/u_i$, and $f^c_{ij} = P^\star_{ij}/v_j$, taking a fraction as zero when its denominator is zero. An eligible edge has $P^\star_{ij} \geq 10^{-6}$. Let B^r_i count eligible row edges with $f^r_{ij} \geq .2$ and B^c_j count eligible column edges with $f^c_{ij} \geq .2$. A selected edge must be eligible and mutual-best in plan mass, satisfy $f^r_{ij}, f^c_{ij} \geq .7$, and have $B^r_i, B^c_j \leq 1$. This selected-pair rule is used for the common-input analysis; the archived temporal and restart tables use the additional exclusions described below.

Source labels are assigned in the priority order death ($u_i < .5a_i$), persistent (incident to a selected edge), split ($B^r_i \geq 2$), then ambiguous. Target labels use birth ($v_j < .5b_j$), persistent, merge ($B^c_j \geq 2$), then ambiguous. Node labels and selected-edge membership are distinct because death/birth has priority. Each event share is the sum of the reference masses of nodes with that label on the corresponding side. Stable-core is the symmetric average of source and target persistent shares. These labels summarize the fitted plan; the ρ sensitivity in Appendix C precludes interpreting them as calibrated semantic birth/death rates.

Legacy temporal and restart summaries. In the recovered TinyLlama CLS temporal and endpoint-restart records, low transport is tested as $(a_i - u_i)/\max(a_i, 10^{-12}) > .5$, with the target expression mirrored. Selected candidates incident to a death-labeled source or birth-labeled target are removed before assigning persistent labels on either side. This rule reproduces all saved edge sets and node labels in the 24 records with complete descriptors. The current extraction above keeps edge membership separate from each node's low-transport priority; the common-input results in Table 35 explicitly use that version.

Numerical conventions. The descriptor replay assigns maximum cosine dissimilarity to zero-norm vectors; the archived zero-profile convention is distinguished in Appendix C.2. The descriptor replay uses the same capped log convention as Eq. 3. Minimum edge mass is 10^{-6}; branch and death/birth thresholds are as above. The reproducibility record includes selected edges, reference and transported masses, convergence residuals, and hashes of imported implementations and inputs.

Threshold and descriptor sweep results. Threshold sweeps over the persistent / branch / birth event thresholds yield 28 of 36 transition pairs with stable-core range below 0.1 across all nine conditions. Descriptor-sweep robustness varies across eight conditions; TinyLlama L5 is handled separately as a layer-specificity control. Candidate-top-k pruning meets the robustness criterion in every pair (32/32 robust), descriptor-weight perturbations are robust in 22/32 pairs, and entropic regularization ε is the most parameter-sensitive sweep axis (16/32 robust). Sensitivity concentrates in the CLS settings (Qwen3 CLS especially) and in TinyLlama Top-k. The incompletely reconstructed LOO diagnostic is retained separately in the Supplementary analyses, Section H1.

Table 26: Robustness counts using a stable-core range threshold of 0.1 across descriptor-weight, ε, and candidate-top-k sweep grids. Candidate pruning meets this criterion in every pair; ε is the most parameter-sensitive axis; sensitivity concentrates in the CLS and Top-k conditions. TinyLlama L5 is handled separately as a layer-specificity control.

Condition	weights robust	ε robust	top-k robust
Qwen3 Top-k L14	4 / 4	3 / 4	4 / 4
Qwen3 L14 ReLU (wiki)	3 / 4	3 / 4	4 / 4
Qwen3 L21 ReLU (wiki)	3 / 4	1 / 4	4 / 4
Qwen3 L14 CLS ReLU	0 / 4	0 / 4	4 / 4
TinyLlama L11 ReLU (wiki)	4 / 4	3 / 4	4 / 4
TinyLlama L16 ReLU (wiki)	3 / 4	3 / 4	4 / 4
TinyLlama Top-k L11	4 / 4	2 / 4	4 / 4
TinyLlama CLS L11	1 / 4	1 / 4	4 / 4
All 8 conditions	**22 / 32**	**16 / 32**	**32 / 32**

E.2 Operating points of the supporting fits

These historical operating points concern different cohorts from the six endpoint pairs with recorded $L_1 = .2$ in Appendix C. Equal L_1 does not imply equal L_0 across corpora, activation conventions, or fitted SAEs; these values do not describe the primary $k = 64$ study.

Table 27: TinyLlama calibration points.

TinyLlama setting	Approx. L_1	Approx. L_0
historical dense ReLU regime	0.005	2507–2660
sparse ReLU reference regime	0.2	839
second sparse ReLU regime	0.3	521
over-sparse ReLU calibration point	0.5	235
Top-k regime	fixed active count	839

Table 28: Qwen3-1.7B calibration points.

Qwen3-1.7B setting	Approx. L_1	Approx. L_0
historical dense ReLU regime	0.005	1861–1955
denser sparse ReLU regime	0.1	839
middle ReLU regime	0.2	450
sparser ReLU regime	0.3	289
Top-k regime	fixed active count	450

E.3 Same-seed transition summaries

The supporting transition study uses TinyLlama-1.1B (Zhang et al., 2024) and Qwen3-1.7B (Yang et al., 2025) with checkpoints at steps 0, 100, 300, 1000, 3000. We study wiki-style language-model fine-tuning on WikiText-2 (Merity et al., 2017) and AG News classification fine-tuning (Zhang et al., 2015), ReLU and Top-k SAEs, and two-to-three layers per model. For the shared-seed transition tables, we train a separate SAE at each checkpoint with the same initialization seed; restart floors and asymmetric audits deliberately vary the SAE seed. The shared-seed transition grid contains nine LoRA conditions; the SFT extension adds three full-SFT trajectories. The asymmetric audit adds Llama-3.2-1B from the Llama family (Grattafiori et al., 2024)[1] as a third architecture and Gemma-2-2B (Gemma Team et al., 2024) as a fourth, using endpoint step $0 \to 3000$ pairs. A separate TinyLlama L11 trajectory alternates seeds $[0, 1, 0, 1, 0]$ across all five checkpoints to test intermediate transitions.

Same-seed scaffold statistics summarize activity mass under a fixed SAE training convention. Independently seeded comparisons test the dependence on that convention. Fixed L_1 does not induce comparable L_0 across TinyLlama and Qwen3; the calibration-sensitivity summary below links to the full sweep in Supplementary analyses, Section H4. These historical summaries use all-token raw second moments. The first-position audit in Appendix C.13 shows that such weights can be dominated by a few positional responses; it does not decompose every trajectory, and scaffold mass is not a fraction of semantic knowledge.

Table 29 summarizes the same-seed scaffold result. The lenient test asks whether persistent source mass exceeds death source mass; it passes 35/36 LoRA transitions and 12/12 full-SFT transitions. The stringent test asks whether persistent source mass exceeds total source reorganization (death + split + ambiguous mass); it passes 24/36 LoRA transitions and 6/12 source-side full-SFT transitions. Thus the chosen event rule assigns little source mass to death in the same-seed lineage, and wiki-style regimes often have persistent mass dominating total plan-defined reorganization. Classification regimes and the early TinyLlama L5 layer-specificity control are more reorganized.

For the six endpoint pairs evaluated with bounded asymmetric scaffold events, the lenient endpoint test passes 6/6 because death mass is near zero, but the stringent mass-dominance test is met in only

[1] Llama-3.2-1B model card: https://huggingface.co/meta-llama/Llama-3.2-1B.

1/6 (Appendix E.9). Same-seed scaffold statistics therefore describe dynamics conditional on the SAE training convention; they do not establish that literal SAE feature IDs persist across independently trained dictionaries.

Intermediate-transition asymmetric audit. The intermediate asymmetric trajectory shows the same separation across all four adjacent TinyLlama L11 transitions. We alternate SAE seeds $[0, 1, 0, 1, 0]$ across checkpoints $\{0, 100, 300, 1000, 3000\}$, making every adjacent pair asymmetric. The persistent/reorganization source-mass shares, in chronological order, are $0.05/0.95$, $0.18/0.82$, $0.21/0.79$, and $0.02/0.98$, with selected-edge counts $K = 85, 98, 117, 131$ (lenient passes 2/4, at transitions where death mass is near zero). Stringent persistence under shared seeds therefore depends on the SAE solution (Appendix E.10). Under the asymmetric protocol, non-persistent mass routes mostly into split/merge/ambiguous events rather than death — a plan-defined pattern motivated by feature splitting and absorption (Chanin et al., 2025).

Table 29: Same-seed scaffold statistics. The lenient test compares plan-defined persistence and death; the stringent criterion is met in a majority of transitions in six of nine LoRA conditions, primarily wiki-style regimes. SAEs at successive checkpoints share the initialization seed. Target-side and full-SFT analogues are in Appendix E.6 and Appendix E.11. The source-side ratios are $r_s^{\text{lenient}} = p_s/(p_s + d_s)$ and $R_s^{\text{strict}} = p_s/(p_s + \text{reorg}_s)$, where p_s is persistent source mass, d_s is death source mass, and reorg_s is total non-persistent source mass. Superscripts L and S abbreviate lenient and stringent. Per-condition medians are taken over transitions with defined ratios.

	Lenient	Stringent	Medians	
Condition	$\text{pers}_s > \text{death}_s$	$\text{pers}_s > \text{all reorg}_s$	$\text{med}(r_s^{\text{lenient}})$	$\text{med}(R_s^{\text{strict}})$
TinyLlama L5 wiki ReLU (layer control)	4 / 4	0 / 4	1.000	0.051
TinyLlama L11 wiki ReLU	4 / 4	4 / 4	0.997	0.667
TinyLlama L11 wiki Top-k	3 / 4	3 / 4	0.983	0.622
TinyLlama L11 CLS ReLU	4 / 4	1 / 4	0.763	0.367
TinyLlama L16 wiki ReLU	4 / 4	4 / 4	0.998	0.654
Qwen3 L14 wiki ReLU	4 / 4	3 / 4	0.999	0.526
Qwen3 L14 wiki Top-k	4 / 4	4 / 4	0.993	0.985
Qwen3 L14 CLS ReLU	4 / 4	1 / 4	1.000	0.472
Qwen3 L21 wiki ReLU	4 / 4	4 / 4	1.000	0.810
All 9 LoRA conditions	**35 / 36**	**24 / 36**	**0.997**	**0.622**
Condition-level majority	**9 / 9**	**6 / 9**	—	—

E.4 Expanded transition summary

The table below expands the same-seed scaffold result of §E.3. Appendix E.7 gives the within-checkpoint restart comparisons.

Detailed transition records. The seven per-transition grids, including persistent-edge fractions and branch/ambiguity mass, are retained in Supplementary analyses, Section H5. TinyLlama L5 and L16 remain in the summary, restart, and stringent-ratio tables here.

Table 30: Transition and restart-floor summaries for the historical transition conditions. SC denotes temporal stable-core; RF denotes the same-checkpoint restart floor. TinyLlama layer 5 is included as a layer-specificity control: its temporal stable-core range lies below its same-checkpoint restart-floor range; this early-layer setting is reported separately from the middle-layer settings.

Setting	Task	SAE	Layer	SC range	RF range
TinyLlama L11	wiki	ReLU	11	0.630–0.710	0.001–0.004
TinyLlama L11	wiki	Top-k	11	0.582–0.696	0.013–0.088
TinyLlama L11	AG News CLS	ReLU	11	0.193–0.536	0.092–0.106
TinyLlama L16	wiki	ReLU	16	0.582–0.722	0.019–0.062
TinyLlama L5	wiki	ReLU	5	0.044–0.094	0.115–0.381
Qwen3 L14	AG News CLS	ReLU	14	0.158–0.747	0.196–0.249
Qwen3 L14	wiki	ReLU	14	0.430–0.592	0.015–0.052
Qwen3 L14	wiki	Top-k	14	0.929–0.989	0.010–0.014
Qwen3 L21	wiki	ReLU	21	0.674–0.870	0.055–0.078

E.5 Scope of the early-layer diagnostics

The TinyLlama L5 temporal and same-checkpoint restart statistics in Appendices E, E.4, and E.7 measure transport rather than endpoint activation overlap. The separate L5 endpoint overlap summary lacks a verifiable evaluation record and is excluded. The historical endpoint activation-overlap evidence concerns the four layers listed in Appendix C.7.

E.6 Stringent scaffold ratios

The strict denominator includes all plan-defined reorganization: death + split + ambiguity on the source side, and birth + merge + ambiguity on the target side. Table 31 gives both sides per transition.

Table 31: Stringent scaffold ratios R_s (source-side) and R_t (target-side). Each ratio is persistent reference mass divided by persistent plus total-reorganization reference mass on the corresponding side. Values follow the transitions $0 \to 100$, $100 \to 300$, $300 \to 1000$, and $1000 \to 3000$.

Condition	R_s per transition	R_t per transition	src pass $R > 0.5$	tgt pass $R > 0.5$
TinyLlama L5 wiki	0.074, 0.053, 0.050, 0.040	0.114, 0.049, 0.043, 0.048	0 / 4	0 / 4
TinyLlama L11 wiki ReLU	0.617, 0.702, 0.692, 0.641	0.643, 0.717, 0.703, 0.622	4 / 4	4 / 4
TinyLlama L11 wiki Top-k	0.567, 0.678, 0.701, 0.431	0.654, 0.713, 0.464, 0.855	3 / 4	3 / 4
TinyLlama L11 CLS	0.426, 0.558, 0.308, 0.231	0.483, 0.513, 0.235, 0.156	1 / 4	1 / 4
TinyLlama L16 wiki	0.640, 0.701, 0.668, 0.595	0.665, 0.743, 0.646, 0.569	4 / 4	4 / 4
Qwen3 L14 wiki ReLU	0.311, 0.656, 0.517, 0.535	0.550, 0.528, 0.481, 0.509	3 / 4	3 / 4
Qwen3 L14 wiki Top-k	0.921, 0.984, 0.986, 0.990	0.936, 0.986, 0.988, 0.989	4 / 4	4 / 4
Qwen3 L14 CLS	0.204, 0.808, 0.468, 0.476	0.113, 0.685, 0.450, 0.585	1 / 4	2 / 4
Qwen3 L21 wiki	0.662, 0.875, 0.815, 0.806	0.685, 0.864, 0.805, 0.819	4 / 4	4 / 4
Aggregate	**24 / 36** src > 0.5	**25 / 36** tgt > 0.5	**6 / 9** maj.	**6 / 9** maj.

A majority of transitions has $R > 0.5$ in $6/9$ settings; this is descriptive stratification, not a standalone sign test. TinyLlama L5 and both AG News CLS settings have low strict ratios, with split-labeled

mass exceeding persistence in the more reorganized settings. The 35/36 persistence-versus-death count uses the weaker denominator.

E.7 WITHIN-CHECKPOINT RESTART COMPARISONS

These three-checkpoint stable-core values are descriptive within-model baselines; comparisons reuse the same trajectories.

Table 32: Restart-floor stable-core mass fractions for all restart-calibrated conditions in the historical transition study.

Setting	step 0	step 300	step 3000
TinyLlama L5 wiki ReLU	0.381	0.207	0.115
TinyLlama L11 wiki ReLU	0.001	0.004	0.001
TinyLlama L11 wiki Top-k	0.042	0.088	0.013
TinyLlama L11 AG News CLS	0.092	0.106	0.092
TinyLlama L16 wiki ReLU	0.062	0.038	0.019
Qwen3 L14 wiki ReLU	0.052	0.039	0.015
Qwen3 L14 wiki Top-k	0.010	0.014	0.013
Qwen3 L14 AG News CLS	0.249	0.243	0.196
Qwen3 L21 wiki ReLU	0.073	0.078	0.055

E.8 SENSITIVITY TO THE SAE OPERATING POINT

The supporting sweep adds eight operating-point settings to the nine LoRA conditions, giving 68 adjacent transitions. Its four lenient tests compare persistent source mass with death mass, persistent target mass with birth mass, their conjunction, and stable-core mass with combined birth/death mass. The respective aggregate pass counts are 65/68, 67/68, 64/68, and 66/68. These activity-weighted event counts do not replicate the primary compact-response experiment. The complete settings, formulas, and table are retained in Supplementary analyses, Section H4.

E.9 ENDPOINT COMPARISONS WITH INDEPENDENT SAE SEEDS

Scaffold preservation under shared SAE seeds need not extend to independently seeded SAEs. Table 33 reports the asymmetric endpoint comparison. For the six endpoint pairs evaluated with bounded scaffold events, the lenient test passes because death mass is near zero; under high cross-solution uncertainty, non-persistent mass is routed mostly to split or ambiguous events. The stringent mass-dominance test is met in $1/6$ endpoint pairs.

Table 33: Endpoint scaffold metrics under asymmetric SAE seeds, computed from the per-transition pairwise OT outputs. Lenient means persistent source mass exceeds death source mass; stringent means persistent source mass exceeds death+split+ambiguous source mass. The same-seed reference combines 36 LoRA and 12 full-SFT transitions; only that row also reports the target-side stringent count.

Condition	persistent src mass	total src reorg.	lenient	stringent
Llama-3.2-1B asym seed 1	0.0042	0.9958	yes	no
Llama-3.2-1B asym seed 2	0.0046	0.9954	yes	no
Llama-3.2-1B asym seed 3	0.9337	0.0663	yes	yes
TinyLlama full-SFT asym seed 1	0.0366	0.9634	yes	no
TinyLlama full-SFT asym seed 2	0.3353	0.6647	yes	no
TinyLlama full-SFT asym seed 3	0.0186	0.9814	yes	no
Asymmetric endpoint aggregate	–	–	6/6	1/6
Same-seed reference	–	–	47/48	30/48 src, 32/48 tgt

E.10 Alternating-seed transitions and common-input checks

The intermediate asymmetric audit uses a TinyLlama L11 LoRA trajectory in which fresh seed-1 SAEs are used at checkpoints $\{100, 1000\}$ and the canonical seed-0 SAEs are reused at $\{0, 300, 3000\}$, yielding an alternating $[0, 1, 0, 1, 0]$ pattern across the five checkpoints. Table 34 reports the four resulting transitions. In this TinyLlama L11 trajectory, using independently seeded SAEs at adjacent checkpoints often reassigns mass that was persistent under shared seeds to split or ambiguous categories. The lenient criterion can still pass, but no transition meets the stringent criterion.

Table 34: Intermediate-transition asymmetric audit, TinyLlama L11. Persist/death/reorg are source-mass fractions; reorg = death + split + ambiguous source. The stringent criterion is met in $0/4$ transitions, while the lenient criterion is met in $2/4$ transitions where persistent source mass exceeds death source mass. Adjacent checkpoints use independently seeded SAEs.

Transition (asym SAE pair)	K	persist mass	death mass	total reorg	lenient	stringent
$0 \to 100$	85	0.048	0.173	0.952	no	no
$100 \to 300$	98	0.185	0.0003	0.815	yes	no
$300 \to 1000$	117	0.207	0.210	0.793	no	no
$1000 \to 3000$	131	0.023	0.016	0.977	yes	no

The source data consist of the four per-transition pairwise OT outputs and a derived scaffold-summary file. The latter is included as `track_b_intermediate/scaffold_summary.json` in the original supplement.

Mass fractions in these trajectory summaries use all-token raw second moments in the stored latent scale. The position audit in Appendix C.13 demonstrates concentration in this weighting for a TinyLlama endpoint/restart cohort; it does not decompose every condition or intermediate transition by position.

Common-input remeasurement. We re-encode 15 saved TinyLlama CLS SAEs on the five canonical residual caches, each containing 688,198 token rows in the same 4,500 ordered document groups. One fixed Gaussian projection is shared across all re-encoded profiles. The exported encoder is applied directly in its stored scale; measured second moments reproduce the canonical and endpoint-restart archives to relative L_2 error below 1.4×10^{-6}. The archived restart profiles

have different cross-fit geometry despite near-identical aggregate statistics; the missing historical projection and document identities prevent assigning that discrepancy to a unique cause. We solve all eight shared/alternating-seed transitions, 22 same-checkpoint refit pairs, and one direct endpoint pair with the stated UOT objective and current event rule. Table 35 shows that stable-core mass remains dependent on SAE initialization under common descriptor measurement. The maximum objective-gradient residual over these 31 solves is 5.0×10^{-13}; every supported edge and both marginals are retained.

Table 35: Common-input UOT, TinyLlama CLS L11. Stable core is the symmetric average of source and target persistent reference-mass shares; distributed source is the split-labeled source reference-mass share. Columns compare shared seed 0 with the alternating $[0, 1, 0, 1, 0]$ sequence.

	Stable core		Distributed source	
Transition	Shared	Alternating	Shared	Alternating
$0 \to 100$	.455	.052	.400	.553
$100 \to 300$	.536	.183	.274	.556
$300 \to 1000$	.271	.215	.472	.334
$1000 \to 3000$	.194	.016	.601	.586

Recovering a positional response. We select the split-labeled source with the highest reference mass in the first TinyLlama asymmetric endpoint plan before inspecting its examples. Source 120 carries 31.74% of descriptor activity mass and has no selected-core counterpart in any of the three target fits, but the full UOT relation has two, two, and three qualifying source-row branches: targets $\{1190, 776\}$, $\{806, 3360\}$, and $\{861, 3479, 1754\}$. Direct encoding of the cached token stream shows that the source and all seven branch endpoints activate at the first stored token in all 4,500 groups; each places more than 99.99999% of its squared activation there and none activates at the last stored token. In contrast, the final-token AG News test readout sees source 120 on only two examples and gives zero activating-example overlap with every candidate. Thus that readout misses the dominant positional activity linked by this relation. The full relation retrieves candidates for a positional response that the final-token measurement misses. Individual branch latents already recover this dominant response; the analysis retains all 29 candidates, their masses, and their position traces. This example measures positional correspondence on the descriptor corpus; held-out semantic validity and additive reconstruction are not evaluated.

E.11 Full fine-tuning correspondence

The full-SFT extension uses three TinyLlama-1.1B AG News trajectories (lr $1e{-5}$, 3000 steps, seeds $42, 43, 44$), with checkpoints $\{0, 100, 300, 1000, 3000\}$ and L11 SAEs. It evaluates scaffold summaries and independently seeded endpoint correspondence, not the primary compact-response protocol.

Same-seed scaffold extension to full SFT. Applying the four lenient scaffold tests from Appendix E.8 to the 12 consecutive SFT transitions gives 12/12 lenient source, 12/12 lenient target, 12/12 joint, and 12/12 stable-core dominance; the stringent test passes 6/12 source and 7/12 target. Aggregating these with the 9 historical LoRA conditions gives 47/48 lenient source, 47/48 lenient target, 46/48 joint, and 48/48 stable-core dominance. These remain plan-defined activity-mass summaries.

Asymmetric endpoint correspondence under full SFT. We also retrained the step-3000 SAE from independent seeds while keeping the step-0 SAE fixed at seed 0 on the SFT seed-42 trajectory. These are the three TinyLlama full-SFT asymmetric rows in Table 18.

Same-seed SFT endpoint controls. Both SAEs use seed 0, giving high selected-pair identity $a(p) = p$. The same-index Jaccard control instead samples outside the selected set; its scores describe a different pair population even when selected pairs share indices.

Table 36: Full-SFT asymmetric endpoint audit. The representational pattern matches the LoRA asymmetric audit; the canonical-minus-same-index margin contrast varies in sign across endpoint seeds. Same-seed controls are reported separately in Table 37. The table notation Δ_m denotes the condition-level mean of Δm.

Endpoint SAE seed	baseline acc	K	Id%	can J	sidx J	$\Delta_m^{\text{can}} - \Delta_m^{\text{sidx}}$	paired CI_{95}	
1	0.943	86	0.0	0.083	0.003	$+0.006$	$[-0.002, +0.014]$	$\approx$
2	0.943	81	0.0	0.096	0.002	$+0.050$	$[+0.040, +0.060]$	$\succ$
3	0.943	83	1.2	0.101	0.003	-0.063	$[-0.070, -0.056]$	$\prec$

Table 37: Same-seed full-SFT endpoint positive controls. High identity makes canonical and same-index nearly the same intervention; the canonical-minus-same-index margin contrasts have no consistent sign across seeds. The table notation Δ_m denotes the condition-level mean of Δm.

SFT seed	baseline acc	K	Id%	can J	sidx J	shuf J	$\Delta_m^{\text{can}} - \Delta_m^{\text{sidx}}$	paired CI_{95}
42	0.943	1758	98.2	0.118	0.010	0.010	-0.021	$[-0.024, -0.018]$
43	0.947	1776	98.9	0.127	0.012	0.010	$+0.007$	$[+0.005, +0.009]$
44	0.946	1735	98.3	0.123	0.013	0.011	-0.014	$[-0.017, -0.011]$

F CORRESPONDENCE CONTROLS

These checks separate initialization, raw index labels, edge selection, residual coordinates, and SAE conventions. Their response measures and fitted cohorts are specified separately from the primary recovery protocol.

F.1 SHARED-INITIALIZATION AND SAME-CHECKPOINT CONTROLS

Table 38 compares descriptor-selected pairs with shared or different SAE seeds. Shared-seed fits have high equal-index incidence and $K \in [41, 2102]$, versus $K \in [67, 197]$ in the different-seed endpoint pairs. This is an initialization control, not evidence that raw indices have semantic identity.

Table 38: Same-seed reference and matched-control comparisons. These diagnostic rows are excluded from the independently seeded endpoint aggregate.

Condition	K	Id%	Selected J	Shuffled J	Same-index J
Same-seed reference: canonical/same-index ratios 1.0–2.6× versus 12–137× in the original asymmetric comparisons					
Qwen3 same-seed	41	95	0.114	0.012	0.044
TinyLlama same-seed	74	100	0.060	0.004	0.059
Gemma-2-2B same-seed	2102	98.9	0.092	0.009	–
Diagnostic same-seed matched controls (excluded from the asymmetric aggregate)					
TinyLlama LoRA same-seed matched control	1697	98	0.106	0.008	0.037
Llama-3.2-1B same-seed matched control	701	90	0.102	0.006	0.034

The historical Gemma same-seed summary reports same-index Jaccard .036, omitted from this table because its generating control could not be reconciled. At width 4096 and $K = 2102$, excluding even one selected set leaves at most 1994 indices, fewer than the preserved sampler requires without replacement. No fallback occurs in that sampler, and the row-specific generating record was not recovered. This control is not used to support the asymmetric comparison.

The following table gives the separate same-checkpoint, different-seed transport summaries.

Table 39: Same-step cross-run stable-core mass fractions. The step-0 baseline is excluded from inference.

Same-step cross-run atlas	step 100	step 300	step 1000	step 3000
TinyLlama, $L_0 \approx 839$	0.439	0.628	0.360	0.545
Qwen3-1.7B, $L_0 \approx 839$	0.899	0.092	0.207	0.210

F.2 Raw index overlap and fixed-source recurrence

Independent target dictionaries permit arbitrary index permutations. The reported target sets are 93.5–100% disjoint across the five target-SAE triads; this is a raw-label check, not evidence that matched activation responses fail to recur. Source indices instead belong to one fixed dictionary, so recurrence of selected source features is a different question.

Target-index null. None of the 15 target-set comparisons shows detectable enrichment over independent uniform subsets at these set sizes (upper-tail hypergeometric probabilities .161–1). This does not rule out weaker sharing below the test's power. Full overlap grids and the null calculation are retained in Supplementary analyses, Section H6.

Source-matchability nulls. With four restart fits available for calibration, the stratified restart-count/prevalence null gives observed/expected source overlaps 1.269 (Qwen3) and 1.298 (TinyLlama), with upper-tail Monte Carlo probabilities approximately 5×10^{-6} and .00115. A separate restart-calibrated conditional-Bernoulli null gives overlaps .882 and .877 and probabilities .987 and .932. Both results are retained: neither null was selected as primary after observing the outcomes. The contrast shows sensitivity to the matchability model rather than establishing a stable source-feature population. Supplementary analyses, Section H6, preserves both constructions, calibration rules, draw counts, and formulas. The primary response-recovery claim does not depend on raw-index overlap or source-set recurrence.

F.3 Residual-coordinate alignment check

We re-solve the full lineage at TinyLlama L11 (asymmetric-seed configuration: step-0 SAE at seed 0, step-3000 SAE at seed 1, matching the TinyLlama seed-1 rerun row of Appendix G.2) under three gauge methods, holding the SAE pair fixed across all three. The three methods are: (i) *signed-permutation* (the canonical method, a discrete Hungarian assignment with sign flips); (ii) *identity* (no alignment is applied between the source and target residual basis); (iii) *Procrustes* (Schönemann, 1966) (a continuous orthogonal rotation R obtained by SVD of the cross-correlation $X_s^\top X_t$). With the SAE pair held fixed, only the gauge transformation differs across conditions, so any difference in the OT correspondence is attributable to the gauge.

Table 40: Gauge-method ablation at TinyLlama L11 (asymmetric-seed configuration, step-0 SAE seed 0, step-3000 SAE seed 1). All three gauge methods produce the same canonical-vs-same-index dissociation. The signed-permutation gauge matches identity (Hungarian assignment margin median 2,239), so signed and identity rows coincide exactly; Procrustes finds a non-trivial rotation $\|R - I\|_F / \|I\|_F \approx 0.10$ and yields the same conclusion. The table notation Δ_m denotes the condition-level mean of Δm.

Gauge	baseline acc	K	Id%	canonical Jaccard	same-index Jaccard	shuffled Jaccard	paired CI_{95} on $\Delta_m^{\mathrm{can}} - \Delta_m^{\mathrm{sidx}}$
signed-permutation	0.949	94	0	0.094	0.002	0.008	$[+0.004, +0.009]$
identity (no alignment)	0.949	94	0	0.094	0.002	0.008	$[+0.004, +0.009]$
Procrustes	0.949	98	0	0.089	0.002	0.008	$[+0.006, +0.011]$

F.4 EDGE SELECTION AND INDEPENDENTLY SEEDED COMPARISONS

To separate the persistence rule from the edge selector, we post-process each saved unbalanced-OT plan with Hungarian assignment and greedy mass ranking, then apply the same row-fraction ≥ 0.7 and column-fraction ≥ 0.7 persistence rule. This describes the effect of additional selection rules on an already fitted plan; retaining its concentrated core is structurally guaranteed for these selectors.

For each of the 15 asymmetric conditions, the saved results compare Hungarian assignment on plan mass, greedy selection on plan mass, and the unchanged unbalanced-OT plan as the canonical baseline. The guarantee below applies to maximum-mass Hungarian assignment and global descending-mass greedy selection. We then count how many edges pass the persistence rule under each matcher and check whether the unbalanced-OT persistent set is contained in the Hungarian and greedy persistent sets.

Table 41: Post-processing of saved UOT plans across the 15 asymmetric endpoint rows. The unbalanced-OT persistent set is fully contained in both the Hungarian and greedy persistent sets in $15/15$ rows. Hungarian and greedy give identical post-filtered sets because each nonzero edge satisfying both transported-mass fractions ≥ 0.7 must survive either selector. The UOT event rule has additional branch/precedence restrictions; its smaller count is not evidence that the alternative selectors independently rediscover the same correspondence. The post-filtered counts average about $2.1\times$ the UOT count (per-row ratios 1.79–2.72$\times$). The Jaccard separation in Table 18 is computed on the unbalanced-OT subset, which is recovered as a subset of either post-processed persistent set. The additional Hungarian-extra and greedy-extra edges are evaluated in §F.4.1; they have lower absolute canonical Jaccard than UOT-persistent but row-wise top-rank $0/30$ against the shuffled null in each of the $6/6$ TinyLlama and Qwen3 LoRA rows (descriptive checks, not adjusted across rows).

Condition	UOT K	Hungarian K	Greedy K	UOT $\subseteq$ Hung	UOT $\subseteq$ Greedy
TinyLlama LoRA asym (0,1)	91	192	192	yes	yes
TinyLlama LoRA asym (0,2)	94	193	193	yes	yes
TinyLlama LoRA asym (0,3)	113	216	216	yes	yes
Qwen3 LoRA asym (0,1)	149	384	384	yes	yes
Qwen3 LoRA asym (0,2)	185	418	418	yes	yes
Qwen3 LoRA asym (0,3)	155	384	384	yes	yes
Llama-3.2-1B asym (0,1)	182	325	325	yes	yes
Llama-3.2-1B asym (0,2)	192	344	344	yes	yes
Llama-3.2-1B asym (0,3)	197	366	366	yes	yes
Gemma-2-2B asym (0,1)	67	182	182	yes	yes
Gemma-2-2B asym (0,2)	70	158	158	yes	yes
Gemma-2-2B asym (0,3)	76	178	178	yes	yes
TinyLlama full-SFT asym (0,1)	86	189	189	yes	yes
TinyLlama full-SFT asym (0,2)	81	159	159	yes	yes
TinyLlama full-SFT asym (0,3)	83	202	202	yes	yes
Mean (asym only, $n = 15$)	121.4	259.3	259.3	15/15	15/15

The smaller UOT event set reflects its additional selection rules. For a nonnegative saved plan, an edge with both row and column fractions at least .7 has total competing incident mass at most $(6/7)P_{ij}$. Any complete maximum-mass assignment excluding it can be improved by swapping the two assigned edges incident to its endpoints to include it; the other new edge has nonnegative mass. Greedy descending-mass selection also retains it, since every competing incident edge is smaller. Filtering either assignment by the original fractions therefore retains the same concentrated edges. This guarantee applies to the stated maximum-mass and global descending-mass procedures. The saved results contain counts and intersections, but their generating implementation was not recovered. These checks concern post-processing of UOT and are distinct from the direct cost-based matchers in Appendix C. The full transport plan represents multi-edge correspondence that these one-to-one

summaries discard. This comparison does not isolate marginal relaxation, because balanced transport also allows multi-edge couplings; the matched-support balanced comparison is in Appendix C. The extra Hungarian/greedy edges are evaluated in §F.4.1: they have lower absolute canonical Jaccard than UOT-persistent but row-wise top-rank $0/30$ against the shuffled null in each of $6/6$ TinyLlama and Qwen3 LoRA conditions (descriptive checks, not adjusted across rows).

F.4.1 Jaccard on Hungarian and greedy persistent/extra edge sets

This subsection evaluates the additional edges retained by the post-processing described in Appendix F.4. For the six TinyLlama/Qwen3 LoRA endpoint conditions with saved full z_0, z_{3000} tensors, we computed top-20 activating-example Jaccard on the alternative matcher edge sets. **Hungarian-persistent** and **Greedy-persistent** apply the row- and column-fraction ≥ 0.7 persistence rule after the matcher's full-plan assignment. **Hungarian-extra** and **Greedy-extra** are the edges admitted by the alternative matcher but absent from UOT-persistent.

Interpretation. In these six measured LoRA conditions, the surplus edges remain above shuffled baselines despite lower absolute canonical Jaccard. Hungarian-extra and greedy-extra edges have lower absolute canonical Jaccard than UOT-persistent (0.021 vs. 0.097 pooled) but higher Can/Shuf ratios because their shuffled baselines are smaller: mean per-condition Can/Shuf ratios $23.80\times$ (Hungarian-extra) and $24.39\times$ (greedy-extra) versus $10.85\times$ (UOT-persistent), with row-wise top-rank $0/30$ against 30 shuffles in each of the $6/6$ conditions for both extras (descriptive checks, not adjusted across rows).

Table 42: Hungarian and greedy Jaccard on the six TinyLlama and Qwen3 LoRA asymmetric conditions. Pooled rows show the mean of per-condition Jaccards; the Can/Shuf column is the mean of per-condition Can/Shuf ratios (mean-of-ratios), not the ratio of pooled means. For UOT-persistent the ratio of pooled means is $0.0971/0.0097 \approx 10.0\times$, while the mean of per-condition ratios is $10.85\times$.

Edge set	K	Canonical J	Shuf J	Same-idx J	Can/Shuf
TinyLlama LoRA L11 (0,1)					
UOT-persistent	91	0.0906	0.0071	0.0026	12.85×
Hungarian-persistent	192	0.0529	0.0029	0.0028	18.28×
Greedy-persistent	192	0.0529	0.0026	0.0023	20.53×
Hungarian-extra	101	0.0167	0.0016	0.0025	10.32×
Greedy-extra	101	0.0167	0.0015	0.0031	10.98×
TinyLlama LoRA L11 (0,2)					
UOT-persistent	94	0.0822	0.0068	0.0027	12.05×
Hungarian-persistent	193	0.0518	0.0028	0.0033	18.29×
Greedy-persistent	193	0.0518	0.0028	0.0008	18.48×
Hungarian-extra	99	0.0201	0.0011	0.0014	17.73×
Greedy-extra	99	0.0201	0.0019	0.0026	10.36×
TinyLlama LoRA L11 (0,3)					
UOT-persistent	113	0.1010	0.0064	0.0024	15.67×
Hungarian-persistent	216	0.0633	0.0031	0.0016	20.34×
Greedy-persistent	216	0.0633	0.0031	0.0025	20.23×
Hungarian-extra	103	0.0203	0.0008	0.0013	24.95×
Greedy-extra	103	0.0203	0.0010	0.0011	21.22×
Qwen3 LoRA L14 (0,1)					
UOT-persistent	149	0.0978	0.0120	0.0007	8.13×
Hungarian-persistent	384	0.0582	0.0033	0.0018	17.55×
Greedy-persistent	384	0.0582	0.0031	0.0009	18.65×
Hungarian-extra	235	0.0235	0.0008	0.0017	31.32×
Greedy-extra	235	0.0235	0.0008	0.0014	30.44×
Qwen3 LoRA L14 (0,2)					
UOT-persistent	185	0.1071	0.0120	0.0008	8.91×
Hungarian-persistent	418	0.0655	0.0039	0.0026	16.60×
Greedy-persistent	418	0.0655	0.0034	0.0009	19.34×
Hungarian-extra	233	0.0196	0.0008	0.0013	24.52×
Greedy-extra	233	0.0196	0.0004	0.0007	46.41×
Qwen3 LoRA L14 (0,3)					
UOT-persistent	155	0.1037	0.0138	0.0004	7.52×
Hungarian-persistent	384	0.0629	0.0036	0.0007	17.55×
Greedy-persistent	384	0.0629	0.0037	0.0015	17.06×
Hungarian-extra	229	0.0241	0.0007	0.0007	33.95×
Greedy-extra	229	0.0241	0.0009	0.0011	26.94×
Pooled across 6 conditions:					
UOT-persistent (mean of 6)		0.0971	0.0097	—	10.85×
Hungarian-persistent (mean of 6)		0.0591	0.0033	—	18.10×
Greedy-persistent (mean of 6)		0.0591	0.0031	—	19.05×
Hungarian-extra (mean of 6)		0.0207	0.0010	—	23.80×
Greedy-extra (mean of 6)		0.0207	0.0011	—	24.39×

F.4.2 HISTORICAL TOP-k AUDIT WITH FIXED ACTIVATION BUDGETS

The six ReLU endpoint targets in Appendix C.7 have recorded training L_0 1588–1611 (TinyLlama) and 743–750 (Qwen3). The Top-k conditions use a different activation regime and do not match these L_0 values exactly. We repeat the asymmetric audit in TinyLlama L11 and Qwen3 L14 using Top-k activation, which limits each token to at most k_{act} active features. We use the same Top-k calibration as the same-seed Top-k rows in Table 29: $k_{\text{act}} = 839$ for TinyLlama L11 (matching the existing `tinyllama_topk_same_base_study` configuration) and $k_{\text{act}} = 450$ for Qwen3 L14 (matching the existing `qwen3_17b_topk_same_base_study` configuration). Both endpoint SAEs are trained from scratch with `activation: topk`, source seed 0, target seed $N \in \{1, 2, 3\}$.

Pre-specified criteria. Four criteria for support in each condition were specified before any Top-k evaluation result (`scripts/TOPK_ASYM_PROTOCOL.md`): (i) raw Id% $\leq 5\%$, (ii) canonical $J \geq 10\times$ same-index J, (iii) canonical $J >$ shuffled J within each tested row, (iv) $K \geq 30$.

Table 43: Historical Top-k audit with fixed activation budgets. The six tested conditions pass the per-condition criteria: raw Id% $\leq 0.5\%$, canonical J/same-index J in 22.4–35.1×, canonical $>$ shuffled in 6/6, and $K \in [174, 436]$. This is descriptive across tested conditions, not a family-wise significance statement. The dissociation pattern — raw target-ID collapse together with descriptor-matched aggregate correspondence — also holds at these fixed activation budgets.

Condition	K	Id%	Can J	Shuf J	Sidx J	Can/Shuf	Can/Sidx
TinyLlama L11 (k_{act}=839) seed 1	394	0.0	0.0761	0.0037	0.0034	20.30×	22.38×
TinyLlama L11 (k_{act}=839) seed 2	183	0.5	0.0915	0.0036	0.0030	25.13×	30.68×
TinyLlama L11 (k_{act}=839) seed 3	174	0.0	0.1131	0.0037	0.0032	30.71×	35.12×
Qwen3 L14 (k_{act}=450) seed 1	416	0.0	0.1157	0.0049	0.0043	23.64×	27.21×
Qwen3 L14 (k_{act}=450) seed 2	436	0.0	0.1117	0.0047	0.0047	23.64×	23.88×
Qwen3 L14 (k_{act}=450) seed 3	435	0.0	0.1160	0.0047	0.0046	24.50×	25.42×

Interpretation. All 6 tested Top-k asymmetric conditions pass the per-condition criteria. Equal-index incidence among selected pairs is $\leq 0.5\%$, canonical/same-index Jaccard ratios are 22–35×, and persistent-set sizes (174–436) are larger than the corresponding ReLU rows (TinyLlama ReLU asym $K \in [91, 113]$; Qwen3 ReLU asym $K \in [149, 185]$; Table 18). Absolute canonical Jaccard is comparable to ReLU asym (0.076–0.116 here vs. 0.082–0.107 ReLU), and canonical-vs-same-index ratios are within the ReLU range. The dissociation therefore persists under a fixed Top-k activation budget as well as the high-L_0 ReLU regime.

F.4.3 OFF-DIAGONAL SOURCE/TARGET SAE SEED AUDIT

The asymmetric audit in §C.7 (Appendix C.7) fixes the source SAE at seed 0 and resamples the target SAE seed. We complement it with an off-diagonal audit in which the source step-0 SAE is trained at seed s and the target step-3000 SAE is trained at seed t, with $s \neq t$. We run three off-diagonal pairs $(s, t) \in \{(1, 2), (2, 3), (3, 1)\}$ for both TinyLlama L11 and Qwen3 L14 LoRA cells—a total of 6 conditions in which neither endpoint SAE uses seed 0.

Frequency-matched controls give mean Jaccard .0062–.0122 across the six pairs, with selected/control ratios 8.5–15.8×. Each control mean averages thirty re-pairings.

Interpretation. All 6 off-diagonal conditions yield 0% equal-index incidence and Can/Sidx ratios of 28–123×, comparable to the asymmetric audit's range for the same two architecture/adapter cells. The separation between index agreement and activation correspondence therefore persists when both endpoint SAEs are independently seeded.

Table 44: Off-diagonal audit. Each row pairs a source step-0 SAE at seed s with a target step-3000 SAE at seed t ($s \neq t$); neither endpoint uses seed 0. The off-diagonal values are comparable to the same-cell asymmetric rows: TL $K = 91$–105, Can $J = 0.080$–0.107 versus TL asymmetric $K = 91$–113, Can $J = 0.082$–0.101; Qwen3 $K = 146$–160, Can $J = 0.104$–0.115 versus Qwen3 asymmetric $K = 149$–185, Can $J = 0.098$–0.107.

Condition	K	Id%	Can J	Shuf J	Sidx J	Can/Sidx
TinyLlama LoRA L11 ($1 \to 2$)	105	0.0	0.107	0.0080	0.0022	48.97×
TinyLlama LoRA L11 ($2 \to 3$)	100	0.0	0.082	0.0066	0.0029	28.44×
TinyLlama LoRA L11 ($3 \to 1$)	91	0.0	0.080	0.0079	0.0023	34.78×
Qwen3 LoRA L14 ($1 \to 2$)	151	0.0	0.108	0.0139	0.0009	122.75×
Qwen3 LoRA L14 ($2 \to 3$)	146	0.0	0.104	0.0137	0.0012	85.69×
Qwen3 LoRA L14 ($3 \to 1$)	160	0.0	0.115	0.0127	0.0012	95.21×

F.5 Public-SAE convention check

The 15 within-convention asymmetric conditions in Table 18 pair SAEs of the same architecture and width fitted at different model checkpoints, with different SAE seeds. We additionally evaluate a *cross-convention* pair, where the source SAE comes from a different family entirely.

Setup. We use the public Gemma Scope L13 width-16384 JumpReLU SAE (Lieberum et al., 2024) as the source and our trained ReLU step-3000 SAE (width 4096) as the target. Both SAEs encode the *same* probe activations, taken at `hook_resid_post` on Gemma-2-2B + LoRA at the AG News step-3000 endpoint (68,377 probe tokens). We compute descriptors using `compute_feature_descriptors` (decoder direction 1.0, profile 0.5, prevalence 0.25, mean-activation 0.25, second-moment 0.25) and fit sparse UOT with Sinkhorn iterations (top-k candidate edges 8, $\varepsilon = 0.01$, $\rho_s = \rho_t = 0.5$). Gemma Scope uses JumpReLU activations, so we encoded the public SAE with the corresponding activation rule; the existing ReLU and Top-k protocol is otherwise unchanged.

Result. Using only row/col-frac ≥ 0.7, without the selected-core minimum-edge-mass and branch-exclusion requirements, $K = 530$ edges have mean canonical Jaccard 0.0118 versus uniform-target-null mean 0.0009 (uniform-target draws over integer indices $\{0, \ldots, 4095\}$). The canonical/uniform-target-null ratio is $13.7\times$, with canonical at a reported rank of $0/30$ against this null (row-wise rank check over 30 null draws; not adjusted for multiple row-wise tests). Both fraction thresholds already imply mutual-best, even for rectangular plans; unequal width does not invalidate that rule. Recounting the saved plan gives 366 pairs after the minimum-mass gate and 314 after branch exclusion, versus 530 under fractions alone. These narrower sets have no activation-overlap evaluation here. The archived event list additionally excludes source-death/target-birth states and retains one edge; it is a different selector. We therefore report the 530-edge diagnostic separately. The original cross-convention generating evaluator was not recovered; these distinctions follow from retained summaries and the saved plan.

The release used is the Gemma Scope 2B residual-post layer-13 width-16k SAE at the median-L_0 release.

G Intervention diagnostics

G.1 Margin-patching protocol and spread

For class logits ℓ_c and true label y, the logit margin is $m = \ell_y - \max_{c \neq y} \ell_c$. We report $\Delta m = m_{\text{patched}} - m_{\text{original}}$; a positive change increases the true class's margin. Margin patching substitutes source latent activations into the target SAE reconstruction, jointly through the target decoder as specified below.

The margin-patching evaluation uses $n_{\text{margin_shuffle}} = n_{\text{margin_freq-matched}} = 30$ for the 12 TinyLlama, Qwen3, Llama-3.2-1B, and TinyLlama-SFT conditions and $n_{\text{margin_shuffle}} =$

$n_{\text{margin_freq-matched}} = 5$ for the 3 Gemma-2-2B conditions; the Gemma rows contribute to both the activation-overlap table and the margin-spread summary, with only five margin-null draws per row.

We compare variation in descriptor-selected activation substitutions with same-index, shuffled, and frequency-matched controls. Each observation is a fitted endpoint-pair mean effect.

Across five model/training settings with three endpoint seeds each, the pooled SD is lower for selected patches by factors $4.31\times$, $4.72\times$, and $4.29\times$ relative to same-index, shuffled, and frequency-matched controls (Figure 3). Selected margin changes have mean -0.0229 and SD 0.0159; same-index changes have mean -0.0391 and SD 0.0685. Shuffled and frequency-matched SDs are 0.0751 and 0.0681. Leaving out one setting at a time gives null/selected SD ratios in $[3.61, 5.03]\times$. The cached Llama-3.2/full-SFT subset gives $2.08\times$ after dividing by mean absolute latent substitution s (defined below).

Pooled SD combines within-setting variation with differences between setting means and scales. Leaving one setting out checks dependence on a single setting, but does not separate those components. Frequency matching and the amplitude proxy do not equalize joint residual perturbation size or decoder geometry. These are descriptive spread comparisons, not evidence of interchangeable interventions or causal portability.

Representational correspondence does not imply consistent intervention signs; the reported margin signs appear below. The separate head-level study is retained in the Supplementary analyses, Section H3.3.

The margin probe uses activation patching (Vig et al., 2020; Meng et al., 2022; Zhang and Nanda, 2024): it patches $z_{3000}[p] \leftarrow z_0[a(p)]$ and compares against same-index, shuffled, and frequency-matched patches; paired bootstrap intervals are over the 7,600 test examples, conditioning on the trained SAE pair and OT plan. The probe measures classifier responses to latent substitutions but does not identify causal effects of the corresponding feature families. Let Δm denote the true-class logit-margin change. For the selected target set $\mathcal{P}$, substitutions are applied jointly through the target decoder: $\Delta h = \sum_{p \in \mathcal{P}} (z_0[a(p)] - z_{3000}[p])\, W_{\text{dec},3000}[:, p]$. Unit decoder columns fix the feature-scale convention, but the joint perturbation also depends on latent differences and decoder geometry. Across the asymmetric endpoint rows, source and target activation magnitudes on persistent indices match closely on active entries (pooled $\bar{z}_0^{+} = 0.080$ vs. $\bar{z}_{3000}^{+} = 0.073$; per-condition active-only mean ratios in $[0.80, 1.04]$); a latent-amplitude sensitivity on the cached-tensor subset is reported below.

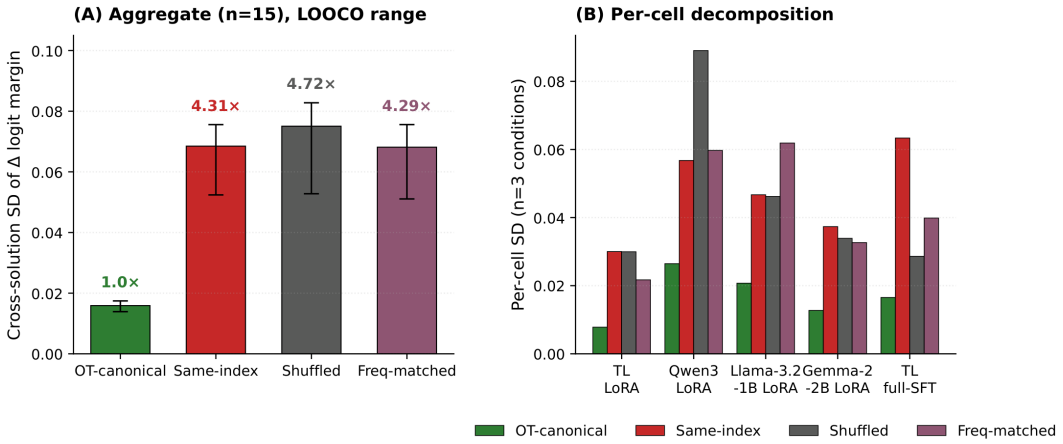


Figure 3: Intervention spread across model/training settings. Archived figure labels use "canonical" for descriptor-selected patches and "cell" for a setting. (A) Pooled SD of Δm over five settings $\times$ three endpoint seeds, with leave-one-setting-out ranges: selected SD is lower by $4.31\times$, $4.72\times$, and $4.29\times$ than same-index, shuffled, and frequency-matched controls. (B) Within-setting comparisons. These displays distinguish pooled and within-setting variation; neither matches joint perturbation size. On the cached subset, dividing by mean absolute latent substitution gives a shuffled/selected SD ratio of $2.08\times$.

Table 45 omits one setting at a time: TinyLlama LoRA, Qwen3 LoRA, Llama-3.2-1B LoRA, Gemma-2-2B LoRA, or TinyLlama full-SFT.

Table 45: Leave-one-cell-out intervention spread controls (n=15 asymmetric conditions). The frequency-matched null tests whether canonical low variance is explained by matching activation frequency; a latent-amplitude sensitivity is reported on the cached-tensor subset (see following table). The frequency-matched null has an SD 3.61–4.51× as large as canonical SD after any one architecture/adapter cell is removed. Holding out the Gemma-2-2B cell gives the frequency-matched $n = 12$ value of $4.30\times$; the $n = 15$ value is $4.29\times$ with Gemma-2-2B included. In the column labels, `can` denotes canonical SD.

Held-out cell	canonical SD	same-index SD / can	shuffled SD / can	frequency-matched SD / can
None (all 15 rows)	0.0159	4.31	4.72	4.29
Gemma-2-2B LoRA	0.0169	4.42	4.81	4.30
Llama-3.2 LoRA	0.0142	4.11	4.42	3.61
Qwen3 LoRA	0.0139	3.77	3.80	4.03
TinyLlama LoRA	0.0174	4.33	4.75	4.33
TinyLlama full-SFT	0.0164	4.34	5.03	4.51

In the Llama-3.2-1B rows, canonical pairs exceed the same-index activation-overlap null by 99.1×, 136.6×, and 92.5× across the three endpoint SAE seeds. Same-index and frequency-matched margin substitutions can nevertheless be near-neutral or positive while canonical substitutions are negative. Activation-pattern correspondence therefore does not determine the direction of classifier-head effects. The spread analysis measures observed variation in these patches, not causal portability.

Latent-amplitude sensitivity. We examine the $4.7\times$ cross-solution SD compression of Δm on OT-canonical patches relative to shuffled patches (§G.1) using mean absolute latent substitution, $s = \mathrm{mean}_{n,p\in\mathcal{P}} \left| z_0^{(n)}[a(p)] - z_{3000}^{(n)}[p] \right|$. The mean is over test examples indexed by n and selected target features $p \in \mathcal{P}$. On the $n = 6$ panel conditions with cached z_0, z_{3000} tensors (3 TinyLlama full-SFT and 3 Llama-3.2 LoRA seeds), pooled s is 0.070 for canonical and 0.096 for shuffled (shuf/canon $= 1.37\times$), against an observed Δm SD ratio of $2.90\times$ on this subset. Dividing each condition's mean Δm by its s yields a shuffled/canonical SD ratio of $2.08\times$. Shuffled draws use the identical column pool as canonical. The normalization accounts for latent substitution amplitudes; it does not measure the joint residual perturbation norm or causally decompose the spread difference.

Table 46: Mean absolute latent substitution s on the $n = 6$ subset of §G.1 with cached selected activations. The pooled shuf/canon s ratio is $1.37\times$ and the raw margin-change SD ratio is $2.90\times$. After dividing each condition's mean margin change by its s, the SD ratio is $2.08\times$. This proxy does not measure the joint residual norm. The table notation Δ_m denotes the condition-level mean of Δm.

Cell	Condition	s^{can}	s^{shuf}	Δ_m^{can}	Δ_m^{shuf}
Llama-3.2 LoRA	seed1	0.0714	0.1019	-0.0527	$+0.0441$
	seed2	0.0761	0.0949	-0.0120	-0.0028
	seed3	0.0721	0.0981	-0.0395	$+0.0896$
TinyLlama full-SFT	seed1	0.0635	0.0857	-0.0028	-0.0309
	seed2	0.0632	0.1010	-0.0357	-0.0608
	seed3	0.0599	0.0832	-0.0225	-0.0035
Pooled mean (entry-weighted)		0.0698	0.0957	—	—
Cross-condition SD of Δ_m		—	—	0.0185	0.0537

Subset ratio shuf/canon: $s = 1.37\times$, raw SD $= 2.90\times$, SD after division by $s = 2.08\times$ (full-panel unnormalized SD ratio: $4.72\times$).

G.2 THRESHOLD SENSITIVITY OF THE INTERVENTION AND OVERLAP OUTCOMES

The fixed threshold grid exposes two sensitive conditions: the selected-minus-same-index margin contrast changes sign for TinyLlama seed 1, and Qwen3 seed 2 reaches a high-mass selection with stronger same-index performance. The tables retain both the margin intervals and the separate activation-overlap outcome.

Table 47: Threshold-sensitivity sweep of the original six asymmetric LoRA conditions. Entries give K and a 95% paired example-bootstrap interval for the mean true-class margin-change difference $\Delta m_{\text{can}} - \Delta m_{\text{sidx}}$; intervals condition on the fitted SAEs and selected plan. T denotes T_{persist}. Leading zeros are suppressed in CI bounds. Symbols indicate whether the paired CI_{95} favors canonical ($\succ$), same-index ($\prec$), or spans zero ($\approx$). Canonical advantage persists across all thresholds for TinyLlama (TL) seeds 2,3 and Qwen3 (Q3) seeds 1,3. The counts describe the fixed 4-threshold grid without a multiple-testing correction.

Condition	$T = 0.70$ $K : \text{CI}_{95}$	$T = 0.75$ $K : \text{CI}_{95}$
TL seed 1	$91 : [-.010, -.005]$ $\prec$	$73 : [-.004, +.000]$ $\approx$
TL seed 1 rerun	$94 : [+.004, +.009]$ $\succ$	$74 : [-.004, +.002]$ $\approx$
TL seed 2	$94 : [+.019, +.026]$ $\succ$	$81 : [+.013, +.019]$ $\succ$
TL seed 3	$113 : [+.034, +.041]$ $\succ$	$95 : [+.042, +.048]$ $\succ$
Q3 seed 1	$149 : [+.125, +.145]$ $\succ$	$132 : [+.127, +.147]$ $\succ$
Q3 seed 2	$185 : [+.131, +.159]$ $\succ$	$156 : [+.002, +.016]$ $\succ$
Q3 seed 3	$155 : [+.049, +.061]$ $\succ$	$134 : [+.056, +.067]$ $\succ$
LoRA seed 43 matched	$1697 : [-.003, -.000]$ $\approx$	$1687 : [-.005, -.002]$ $\prec$
asymmetric $\succ$ count	5/6	5/6

Condition	$T = 0.80$ $K : \text{CI}_{95}$	$T = 0.85$ $K : \text{CI}_{95}$
TL seed 1	$60 : [-.002, +.002]$ $\approx$	$45 : [+.001, +.005]$ $\succ$
TL seed 1 rerun	$61 : [+.004, +.008]$ $\succ$	$46 : [-.002, +.002]$ $\approx$
TL seed 2	$59 : [+.027, +.031]$ $\succ$	$41 : [+.010, +.013]$ $\succ$
TL seed 3	$80 : [+.023, +.027]$ $\succ$	$54 : [+.017, +.020]$ $\succ$
Q3 seed 1	$110 : [+.107, +.122]$ $\succ$	$76 : [+.093, +.105]$ $\succ$
Q3 seed 2	$139 : [-.005, +.008]$ $\approx$	$112 : [-.025, -.013]$ $\prec$
Q3 seed 3	$117 : [+.048, +.058]$ $\succ$	$97 : [+.033, +.042]$ $\succ$
LoRA seed 43 matched	$1673 : [-.007, -.005]$ $\prec$	$1623 : [-.004, -.002]$ $\prec$
asymmetric $\succ$ count	4/6	5/6

Table 48: Activation overlap at $T_{\text{persist}} = .85$ on the six historical LoRA rows. These are the Jaccard values underlying the 42–111$\times$ canonical/same-index ratio; the preceding table evaluates a different, margin-patching outcome.

Row	K	Canonical J	Same-index J	Shuffled J
TL-1	45	.11099	.00194	.00758
TL-2	41	.10056	.00211	.00803
TL-3	54	.10483	.00249	.00803
Q3-1	76	.12876	.00221	.01233
Q3-2	112	.12925	.00116	.01300
Q3-3	97	.10664	.00121	.01469

H ASSET AND RELEASE NOTES

We use public model checkpoints (Zhang et al., 2024; Yang et al., 2025; Grattafiori et al., 2024; Gemma Team et al., 2024) and public text data (Merity et al., 2017; Zhang et al., 2015); license/terms and gated-status metadata are summarized in Table 49, with revision metadata pinned in released configs where applicable.

Table 49: Asset and release inventory. Gated model assets require the relevant HF account approval and acceptance of upstream terms; these terms should be checked for current access requirements. The submission bundle does not redistribute raw model weights or raw public datasets.

Asset	Type	Source / version	License / terms
TinyLlama-1.1B base	model	`HF TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T`	Apache-2.0
Qwen3-1.7B	model	`HF Qwen/Qwen3-1.7B`	Apache-2.0
Llama-3.2-1B	model	`HF meta-llama/Llama-3.2-1B` (gated)	Llama 3.2 Community License
Gemma-2-2B	model	`HF google/gemma-2-2b` (gated)	Gemma Terms of Use
WikiText-2	dataset	`HF wikitext` (`wikitext-2-raw-v1`)	HF metadata: CC BY-SA 3.0 / GFDL
AG News	dataset	`HF fancyzhx/ag_news`	HF metadata: unknown; corpus described for research/non-commercial use
Gemma Scope L13 (App. F.5)	SAE	`HF model google/gemma-scope-2b-pt-res`	CC-BY-4.0

The WikiText card lists CC BY-SA 3.0/GFDL in its metadata but CC BY-SA 4.0 in its licensing prose; the table reports the metadata. Historical model, tokenizer, and dataset revision identifiers are incomplete for some conditions (Appendix C.8).

The submitted archive `sae_iclr2027_supplement.zip` contains the manuscript source, *Supplementary analyses* (`historical_companion.tex` and its PDF), primary saved outcomes, example predictions, verification scripts, and supporting code and aggregate summaries. Its manifest records file hashes; its README specifies the omitted full coefficient arrays and historical per-edge records. Compact threshold codes, the protocol, and verification for the random-group control are included under `source/data/random_group_control`. Additional diagnostics and their input and implementation hashes are listed in Appendix C. The historical archive analyzed for the supporting studies has SHA-256 `b287d1b10412f960394020dac56014356b241834decad74c867185072ca8f118`.

I COMPUTE NOTES

TinyLlama sparse study. The ten Top-k fits used PyTorch MPS on a local Apple M5 GPU with 16 GB unified memory. Each fit received 12,000,256 token presentations. Logged training loops totaled approximately 2.4 hours, excluding setup, final calibration/export, activation encoding, and transport evaluation.

Random-group control. We recomputed the control locally on CPU using existing document responses: 20 random draws for each of 45 pairs in each of two models. This required no model or SAE training and no GPU computation.

Historical hardware: the original training experiments were reported on a single NVIDIA RTX 4090 GPU with 24 GB VRAM; no multi-GPU training was used. Evaluation-tensor reconstruction used Apple MPS, and the additional transport comparisons ran on CPU.

Per-stage cost (single condition, RTX 4090):

- LoRA fine-tuning to step 3000: $\approx$ 25 min (TinyLlama 1.1B), $\approx$ 45 min (Qwen3 1.7B); batch size 8, lr 2×10^{-4}, rank 16.
- Residual activation extraction over the SAE training set: ≈ 5 min per checkpoint.
- SAE training (width 4096, ReLU $L_1 = 0.005$ for the original calibration regime, lr 3×10^{-4}, 8 epochs): ≈ 15 min per checkpoint, ≈ 30 min for the Top-k variant.
- Restart-floor anchors: 5 fresh SAEs per anchor checkpoint $\times$ 3 anchor checkpoints per condition $= 15$ retrains ≈ 4 GPU-hours per condition.
- Descriptor construction and unbalanced OT: ≈ 2 min per transition, CPU-only.
- Activation-overlap evaluation and margin patching: $\approx$ 8–15 min per condition for the historical evaluation runs, including matched controls ($K \in [81, 1697]$ in this timing summary, 7600 AG News test examples, 5–1000 shuffle/frequency-matched draws; Gemma margin rows use the low end). The historical fifteen-row asymmetric panel has $K \in [67, 197]$.
- Head-level intervention sweep: ≈ 16 GPU-hours total for the 15-condition study, summed over independent single-GPU RTX 4090 jobs; the reported strict cross-cell statistic uses the $n = 6$ subset whose canonical sets match Table 18.

Aggregate: the original real-model runs were reported to total approximately 110–130 GPU-hours, including the main study, asymmetric-seed audit, gauge ablation, L_0 calibration-stress runs, sensitivity sweeps, exact LOO, and head-level intervention. This historical estimate excludes the later local MPS and CPU reanalyses.

Reproducibility: the compact supplement supports reanalysis of the included summaries; rerunning diagnostics that use full plans or activations requires additional inputs. Complete historical training and document-level reconstruction is unavailable; the retained records and limitations are given in Appendix C.8. Numerical agreement can also depend on hardware and library versions.