SAME FEATURES, OPPOSITE CONCLUSIONS: CONCEPT ASSIGNMENT IN SAE EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) help researchers study the features learned by neural networks, but only a small subset of these features can usually be inspected. These features may correspond to concepts, meaning human-interpretable properties or categories. Choosing that subset of features requires a way to score features and evaluate the selections of concepts. We investigated whether the relative performance of two feature-selection methods (top-480 and uniform-480) varies depending on the concepts used for evaluation. Across 163 trained SAE checkpoints, we compared feature scores calculated using either the 480 most strongly activating examples or 480 randomly sampled active examples. We represented each feature using its decoder direction and measured how well it separated active from inactive examples. We assigned each selected feature the concept that maximized either binary F1 or absolute cosine similarity between its decoder direction and a known concept direction. We then evaluated concept detection on held-out examples using precision, recall, and F1 between feature and concept presence. Across matching-pursuit and standard SAEs, changing the concept assignment rule reversed which selection method achieved higher mean F1, even though the feature scores and selections were unchanged. The reversals recurred with fresh scoring, assignment, and evaluation data and new candidate pools, although statistical support was weaker for matching-pursuit SAEs under a common-seed analysis. The original test across 105 held-out checkpoints did not support the hypothesis that uniform sampling selects better concept detectors overall. In a separate experiment, training SAEs on data where concepts occurred independently greatly reduced disagreement between the assignment rules, even when the models were evaluated on the same data. These results show that comparisons between feature-selection methods can depend on how concepts are assigned to features. Although our experiments do not establish generalization beyond one synthetic generator or identify a universally better method, they show that changing the evaluation target can reverse conclusions while feature scores and selections remain fixed.

1 INTRODUCTION

Sparse autoencoders (SAEs) decompose a model's activations into thousands of features, far more than anyone can inspect. Investigators therefore rank features with a cheap score and read only the top few, and synthetic benchmarks with known generating concepts can test which ranking method finds better concept detectors. Such verdicts can guide which methods practitioners adopt.

Such a verdict can depend on which concept each selected feature is scored against. Behavioral assignment gives a feature the concept it detects best on separate data. Geometric assignment gives it the generating factor maximizing absolute decoder–concept cosine, the rule behind Synth-SAEBench's ground-truth F1 (GT-F1) (Chanin & Garriga-Alonso, 2026b). The two can disagree sharply. In one standard SAE (a ReLU SAE trained with an L1 penalty), a latent detects 617 of 618 occurrences of its behavioral concept but only 977 of 9,193 occurrences of that concept's grandparent, which geometric assignment selects (Table 1).

We ask whether this choice can change which ranking method looks better. Two selectors score each feature by decoder-readout AUROC against inactive examples, using either its 480 most strongly activating examples (top-480) or 480 randomly sampled active examples (uniform-480). We evaluate

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

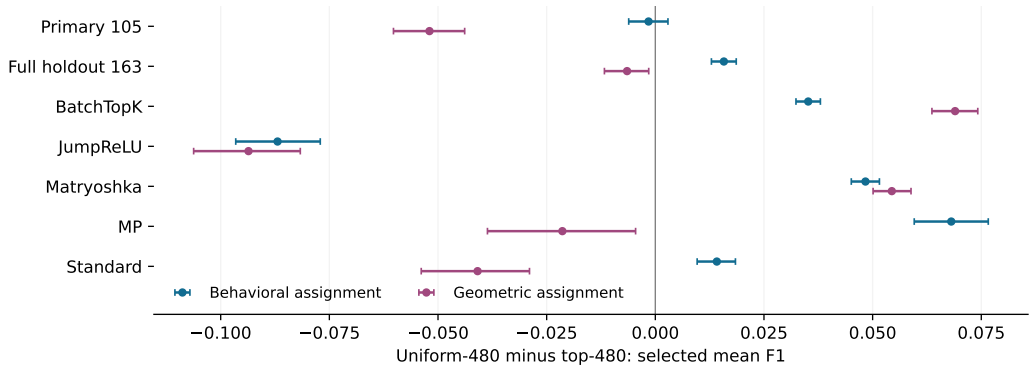


Figure 1: **Changing the assigned concept can reverse a selector comparison.** Uniform-minus-top F1 for the same two fixed selections under behavioral and geometric assignment; values right of zero favor uniform-480. The full 163-checkpoint panel, MP and standard reverse. The primary 105 checkpoints (JumpReLU, MP and standard) do not: their behavioral contrast is unresolved and their geometric contrast favors top-480. Bars are paired 95% intervals; only the primary behavioral comparison is confirmatory.

each selector's fixed selection under both assignment rules on held-out data, so that only the evaluation target changes.

Changing the assigned concept can reverse that preference. Across 163 released SynthSAEBench checkpoints, uniform-480's held-out F1 advantage changes from $+0.016$ under behavioral assignment to -0.006 under geometric assignment, with opposite preferences in matching-pursuit (MP) (Costa et al., 2025) and standard SAEs (Figure 1). These comparisons are secondary: our prespecified primary hypothesis, that uniform sampling improves selection on 105 checkpoints from families absent from calibration, is unsupported. A later replication, fixed before evaluation, redraws scoring, assignment and evaluation samples and candidates for the same models and generator. It reproduces the standard-SAE reversal under both uncertainty analyses, whereas MP support is lost when seeds are grouped across sparsities. Retraining studies leave the reversal unresolved, and on one language-model SAE both assignments favor the top-activation selector (Sections 3.4 and 4).

Prior work shows that example sampling and metric choice affect SAE evaluation (Oikarinen et al., 2025; Paulo & Belrose, 2026; Chanin, 2026), and feature dashboards and practical rankings rely on top-activating examples (Bricken et al., 2023; Templeton et al., 2024; Sharma, 2026). We instead hold the score, the selections and the F1 formula fixed and vary only the assigned concept. Our contributions are threefold. First, we show that concept assignment alone can reverse a selector comparison, and we map where the reversal holds and where it does not. Second, we measure the assignment effect descriptively: on redrawn samples, switching assignment moves the selector contrast by 4.66 F1 percentage points on average, compared with 2.91 points between seed checkpoints at fixed family and sparsity. Third, we provide audit code and a simple reporting rule: state what a selected feature is meant to detect, and report selector comparisons under each justified target specified before comparison.

2 Evaluation design

Figure 2 separates feature scoring and selection from concept assignment and evaluation, the two operations that we repeat while holding each selector's selected-feature weights fixed. Changing the sampling rule defines the separate comparison between selectors.

2.1 Pipeline

Score and select. Let $z_j(x) \geq 0$ be feature j's encoder activation and $r_j(x) = x^\top d_j / \|d_j\|$ its decoder readout. On scoring set S, let $F_j = \{x : z_j(x) > 0\}$ be the active examples. Top-480

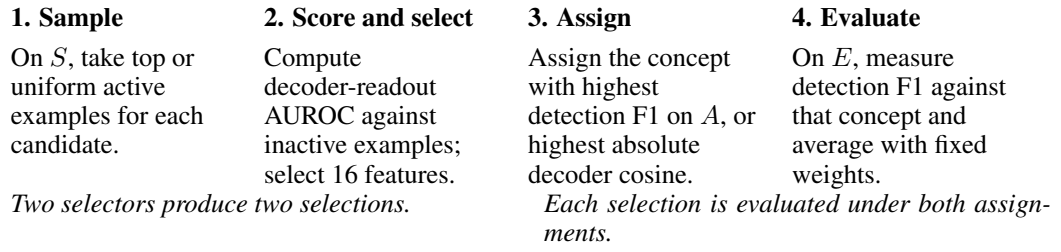


Figure 2: **Only the evaluation target changes.** The scoring set S supplies top or uniform examples for scoring and selection; A is the assignment set and E the held-out evaluation set. We assign concepts by detection F1 on A or by decoder cosine, then evaluate on E, with scores and selection weights unchanged.

takes the 480 largest activations in F_j, whereas uniform-480 samples 480 active examples without replacement. For either active subset P_j, the feature score compares decoder readouts against a shared, fixed sample Z_j of 4,000 inactive examples:

$$\mathrm{AUROC}_j(P_j) = \Pr(r_j(P_j) > r_j(Z_j)) + \tfrac{1}{2}\Pr(r_j(P_j) = r_j(Z_j)). \tag{1}$$

We rank 256 eligible candidates by this score and select 16. A feature is eligible if it is active on more than 0.1% and fewer than 10% of scoring examples, and on at least 480 of them. In these sets the 480-example floor raises the minimum to 0.2%.

We score and rank each of 32 uniform draws separately before averaging their selected outcomes, using the expected outcome over tied candidates at an exact score cutoff. These conventions define selection weights that sum to one and represent each feature's expected share of the selected mean. Integer midranks preserve AUROC ties, and Appendix A gives the exact tie formula and secondary sampling variants. The selection budget k denotes the number of features retained, with $k = 8, 32$ used as sensitivities.

Assign and evaluate. For detection F1, a feature predicts a concept whenever it is active. Using SynthSAEBench's known concept directions and presence labels (Chanin & Garriga-Alonso, 2026b), we assign each feature either the concept maximizing detection F1 on assignment set A or the concept maximizing absolute decoder cosine. The latter follows the benchmark's per-latent classification rule, which differs from its separate mean correlation coefficient based on Hungarian matching. We assign concepts independently to each latent and resolve tied assignments by choosing the smallest concept ID.

On independent evaluation set E, we measure each selected feature's precision, recall and F1 against its assigned concept, then average those per-feature outcomes with the selection weights rather than pooling confusion counts. Behavioral assignment optimizes F1 on A, whereas geometric assignment optimizes alignment, so their held-out scores answer different target-specific questions. Both rules impose one concept per latent and permit several latents to receive the same concept. Mean per-latent F1 measures detection, not semantic identity, causal usefulness or distinct-concept coverage.

A worked example. The example in the introduction is latent 341 in a standard SAE at target $L_0 = 30$, seed 0, active on 1,092 held-out examples. Uniform-480 selects this latent in every draw at budget 16; top-480 never does. We chose this illustrative case post hoc using the deterministic case-selection rule in Appendix G.

Table 1: **Same latent, different target.** Held-out scores for latent 341; concept 97 is the grandparent of concept 2207 in the generator tree.

Assigned target	Precision	Recall	F1
Behavioral: concept 2207	0.565	0.998	0.722
Geometric: ancestor 97	0.895	0.106	0.190

The assignment interaction. For fixed normalized selection weights w_U, w_T, define

$$\Delta_\ell = \sum_j (w_{Uj} - w_{Tj})\,\mathrm{F1}_{\ell j}, \qquad I = \Delta_D - \Delta_A. \tag{2}$$

Here $\ell \in \{A, D\}$ indexes behavioral and geometric rules, respectively, and $\mathrm{F1}_{\ell j}$ is feature j's held-out F1, with subscript A denoting the behavioral rule, subscript D the geometric rule, and standalone A its fitting data. Positive contrasts favor uniform and negative contrasts favor top, so opposite point signs constitute a point reversal. We require bounds that exclude zero in opposite directions to support a joint reversal, leaving comparisons unresolved when this criterion fails, even if the interaction I differs from zero. Because a feature has identical F1 when its assignments agree, only features assigned different concepts contribute to I, with each contribution determined by its selection weights and outcomes under the two targets.

2.2 Data, baselines and uncertainty

The original panel has 163 held-out checkpoints spanning five training families and seven target sparsities, each with 4,096 latents representing a generator with 16,384 true directions in 768 dimensions and hierarchical, correlated concepts. Each scoring, assignment, diagnostic and evaluation set contains 240,000 examples: scoring data S and assignment data A come from the twelve-checkpoint calibration stage, while we drew new data for exploratory diagnostics C and selection evaluation E. A fixed permutation chose 256 selection candidates and a disjoint diagnostic pool, and all 163 cases completed. We saved scoring and assignment decisions with hashes before evaluating C, then diagnostic decisions before evaluating E. "Frozen" and "sealed" refer to these local records, as we did not publicly preregister the studies. Appendix A specifies the data sets and generator, and Table 6 summarizes the evidence stages and their roles.

A calibration regression failed to predict uniform-minus-top F1 better than a constant, so its criterion for panel expansion failed (Appendix B). The later holdout used a separate calibration-informed protocol recorded in Supplement S8.

The sole original confirmatory test asks whether uniform-480 improves behaviorally assigned selected F1 on 105 checkpoints from the JumpReLU, MP and standard families, which were absent from calibration. Another 58 checkpoints extend the two calibration families, BatchTopK and Matryoshka. All original-panel means weight family and sparsity cells equally and then available seeds equally. Paired 95% percentile intervals combine shared evaluation-example draws with within-cell seed resampling: seeds are drawn separately within each family and sparsity cell. The example draws are shared across methods and assignments; Appendix A gives the resampling counts. All original family, assignment and diagnostic comparisons are secondary and uncorrected.

These intervals condition on the observed families, sparsities, generator, scoring draws, pools and fitted maps (feature-to-concept assignments), with example resampling representing variation in evaluation data for the existing models and generator. The supplement contains full records, including seed-only resampling that fixes evaluation examples (Supplement S1) and common-seed sensitivities that average checkpoints with the same seed label across family and sparsity cells, then use those groups as uncertainty units despite unverified sharing of training randomness. Tests with redrawn samples and candidates use their own corrected directional bounds, described below.

To assess absolute selection quality alongside relative contrasts, we report F1 against the exact candidate-pool expectation for random selection and a label-free mean-difference baseline M. Motivated by mean-difference selection (Kantamneni et al., 2025), M uses the standardized mean difference in decoder readout between active and inactive examples, as defined in Appendix A. Supplements S1–S2 retain the other secondary comparators. A reversal or relative gain alone does not establish useful selection: the selected features must also outperform the relevant random and mean-difference baselines.

3 SYNTHETIC RESULTS

3.1 ASSIGNMENT CAN REVERSE THE COMPARISON

On the primary 105 checkpoints, uniform-minus-top F1 is -0.002 $[-0.006, 0.003]$, leaving the positive-average hypothesis unsupported while permitting effects in either direction. A five-block common-seed sensitivity also spans zero, $[-0.012, 0.009]$ (Supplement S1). Table 7 also gives the secondary full-panel average, which includes the calibration families.

Changing only the assignment produces a different pooled comparison on the 163-checkpoint panel: uniform-minus-top F1 changes from $+0.016$ under behavioral assignment to -0.006 under geometric assignment, with neither paired interval containing zero (Figure 1). MP changes from $+0.068$ to -0.021 and standard from $+0.014$ to -0.041, while BatchTopK and Matryoshka remain positive and JumpReLU remains negative. In the primary 105, JumpReLU offsets MP and standard under behavioral assignment, leaving the pooled contrast unresolved. Under geometric assignment all three favor top, with a pooled contrast of -0.052. The reversal is therefore specific to the secondary full-panel, MP and standard comparisons.

The two family point reversals persist at budgets 8, 16 and 32 (Table 11), and positive behavioral and negative geometric point estimates occur in 21/35 MP and 22/35 standard checkpoints at budget 16. These counts describe checkpoint signs. Figure 3 shows cell-mean behavioral contrasts by family and sparsity. Family comparisons are observational, with eligibility and activation rates also differing across families. Although released BatchTopK (Bussmann et al., 2024), Matryoshka (Bussmann et al., 2025) and JumpReLU (Rajamanoharan et al., 2024) models all use JumpReLU inference, their sampling outcomes differ in sign.

Absolute selection quality. For MP, top and uniform F1 fall from 0.733 and 0.801 under behavioral assignment to 0.529 and 0.508 under geometric assignment, while the corresponding standard-SAE values fall from 0.638 and 0.652 to 0.585 and 0.544. Under geometric assignment, uniform exceeds random by 0.0841 in MP with an interval excluding zero, compared with only 0.0098 $[0.0002, 0.0188]$ in standard, although its point score is at or below random in 5/35 MP and 11/35 standard checkpoints. Each assignment evaluates performance against its own target, with assignment and random-baseline results reported in Tables 8 and 9.

The primary uniform-480 versus M contrast is -0.0007 $[-0.0052, 0.0037]$, leaving the comparison with the mean-difference baseline unresolved.

A separate allocation variant replaces the fixed 480 examples with a fixed fraction of each feature's active examples, yielding a positive primary sampling contrast of $+0.0102$ $[0.0055, 0.0149]$. This allocation change has family-specific effects, making the standard contrast negative and the MP contrast more positive while also changing how compute is allocated across features (Appendix B).

Exposure to disagreeing features. Uniform assigns 60.7% of its weight to disagreeing MP features compared with top's 39.5%, while the corresponding standard-SAE values are 42.4% and 32.5%. We call this weight disagreement exposure and define its gap as uniform minus top. MP and standard have the largest gaps among the five families and are the two whose contrasts change sign. JumpReLU also has a positive gap but keeps the same selector preference, with disagreements typically joining different trees rather than the mainly ancestor–descendant pairs in MP and standard. The decomposition in Appendix B (Eq. 6) expresses the interaction through both exposure and feature-level F1 changes. These family associations are descriptive.

With trained models and evaluation data E fixed, two independent assignment refits preserve the reversals, and features retaining their labels across refits carry more than 99% of standard's interaction. Appendices A and B report the refits and post-hoc decomposition.

Appendix B defines the saturation and activation-frequency diagnostics evaluated on C; Supplement S2 retains the failed diagnostic selector rules and complete controls.

3.2 EFFECT SIZE RELATIVE TO CHECKPOINT VARIATION

On the fresh-sample replication (Section 3.3), we compare assignment and checkpoint changes in uniform-minus-top F1 within each family and sparsity cell. We measure assignment change as the mean absolute difference between targets for a fixed checkpoint, and seed-checkpoint change as the mean absolute difference between distinct checkpoints at a fixed target, averaged over both targets. We then weight cells equally and take the ratio of the two means, as detailed in Appendix G.

The assignment change is 4.66 F1 percentage points versus 2.91 between seed checkpoints, a pooled ratio of $1.60\times$ (Table 2). The assignment change exceeds the seed-checkpoint change in 19/35 cells. Family-level ratios exceed one for MP, standard and BatchTopK and fall below one for JumpReLU and Matryoshka. Leaving out each seed label from this replication gives a budget-16 sensitivity range of 1.53–1.73. Appendix G reports the separate budget and original-evaluation comparisons.

Table 2: **On average, assignment moves the selector contrast more than changing checkpoints does.** Mean absolute change in uniform-minus-top F1 on the fresh-sample replication (percentage points, budget 16), with cells weighted equally.

Panel	Assignment	Seed checkpoint	Ratio
All 163	4.66	2.91	$1.60\times$
BatchTopK	2.96	1.57	$1.88\times$
JumpReLU	3.17	4.43	$0.71\times$
Matryoshka	0.77	1.55	$0.50\times$
MP	11.81	4.58	$2.58\times$
Standard	4.61	2.42	$1.91\times$

This post-hoc comparison concerns two fixed assignments and the observed checkpoint grid, where candidate pools, inactive references Z_j and uniform draws also differ between checkpoints, so the denominator does not isolate training randomness. Regenerating the evaluation pipeline for the same checkpoints changes the contrast by 2.94 points on average across targets and cells, a change of similar magnitude that we report separately. The observed assignment changes are comparable in magnitude to checkpoint differences, with the ratio serving as a descriptive scale comparison rather than a variance decomposition.

3.3 REPLICATION WITH FRESH SAMPLES

Under a separately frozen protocol, we redrew scoring, assignment and evaluation samples and candidate pools for all 163 models while retaining their weights and generator. The prespecified budget-16 criterion requires positive behavioral and negative geometric contrasts in both MP and standard, and all four Bonferroni directional bounds pass under within-cell resampling. Table 3 compares this result with the common-seed analysis.

Table 3: **Fresh samples: standard reverses under both uncertainty schemes, MP only within cells.** Uniform-minus-top F1 at budget 16 on the same models with new samples and candidates. Parentheses give Bonferroni-corrected one-sided bounds: the behavioral lower bound must exceed zero and the geometric upper bound must fall below zero. Within-cell bounds resample seeds and examples; common-seed bounds use t_4 on five seed blocks, conditional on E.

Family	Seed grouping	Behavioral (lower)	Geometric (upper)	Joint reversal
MP	Within cells	$+0.0713$ $(+0.0614)$	-0.0361 (-0.0168)	Supported
MP	Common-seed	$+0.0713$ $(+0.0587)$	-0.0361 $(+0.0024)$	Unresolved
Standard	Within cells	$+0.0156$ $(+0.0108)$	-0.0258 (-0.0143)	Supported
Standard	Common-seed	$+0.0156$ $(+0.0073)$	-0.0258 (-0.0130)	Supported

Each common-seed block groups one of five seed labels across seven sparsity cells within the same generator, and all five blocks have opposite point signs. MP's geometric effects nevertheless vary enough that the corrected upper bound crosses zero, leaving its joint reversal unresolved under this

analysis. Opposite point signs also recur at budgets 8 and 32, although the corrected replication criterion concerns budget 16. Appendix B shows the block results, and Appendix H gives the audit code.

3.4 NATIVE SCORES AND RETRAINING

Table 4 summarizes the extension studies. Two reuse the 70 released MP and standard checkpoints: the native-score study replaces AUROC with the mean of the 50 largest or 50 randomly sampled positive encoder activations (top-50, uniform-50), while the hierarchy-cap study replaces depth-3 labels with their depth-2 ancestors. Neither resolves the joint reversal (Appendix B). Table 12 gives estimates and uncertainty conventions; exploratory fixed-selection dependence analyses are in Appendix D.

Table 4: **What the extension studies support.** Each row is a separate study. "Unresolved" means at least one required directional interval includes zero or supports the other direction. Native-score, hierarchy-cap and paired-width rows use two-sided 98.75% t_4 intervals.

Change tested	Models and comparison	Supported conclusion
Native activation score	Same 70 released MP/standard models; top-50/uniform-50	Joint reversal unresolved in both families
Hierarchy-cap	Same models; replace depth-3 labels by depth-2 ancestors	Joint reversal unresolved in MP and standard
Paired-width training	New MP/standard, widths 4k/16k; original generator	No panel resolves the original joint reversal
Generator/training draws	Two historical and two new standard-SAE blocks, kept separate	Neither two-block panel resolves joint reversal
Natural SAE and concept bank	Native top-50/uniform-50; two assignment fits	Top preferred under both assignments

A paired-width study retrained five seeds each of MP and standard SAEs at widths 4k and 16k on the original generator. Standard retrained at the released width (4k) retains positive behavioral and negative geometric point estimates, but both intervals span zero: $+0.0031$ $[-0.0441, 0.0503]$ and -0.0267 $[-0.1387, 0.0852]$. At width 16k, MP favors top under behavioral assignment, -0.0282 $[-0.0562, -0.0003]$, reversing the original behavioral preference, although unequal achieved sparsity prevents attributing this change solely to width. Appendix C gives paired-width-study quality, baselines and uncertainty; Supplement S4 retains those records for the exploratory historical generator pilot, which crossed two generator and training draws with seven sparsities.

New generator draws. Under an exploratory protocol fixed beforehand, we trained 14 standard SAEs: two generator and training blocks with separate random draws, each with all seven sparsities and about 200M examples per model. At budget 16, the first block reverses in point estimates, with behavioral and geometric contrasts of $+0.0034$ and -0.0021, while the second favors top under both targets, with contrasts of -0.0039 and -0.0392. These 14 models supply only two independent block units, and the descriptive equal-block intervals span zero for both assignments (Appendix C). We retain every sparsity cell and keep this extension separate from the historical pilot.

3.5 TRAINING STRUCTURE CHANGES THE LABELS

To separate training effects from measurement effects, we vary the training generator and fit labels on either of two data populations: the original dependent concepts or independently sampled concepts. We then evaluate each fitted map on both populations. This extends a twelve-model dependence intervention with a sixteen-model ablation that varies hierarchy and an additional correlation component (Appendix D). On common original assignment data, disagreement is 38.45% with both components, 37.82% with hierarchy alone, 0.14% with correlation alone and 0.13% with neither: four paired models per configuration, all 4,096 features (Table 18). Hierarchy itself induces dependence, so these conditions do not equate all forms or strengths of correlation.

The contrast persists at target sparsities 15 and 45 and two new hierarchy shapes: disagreement ranges from about 4% to 68% with hierarchical training, versus below 0.3% with independent training (Figure 5). These exploratory extensions cover one SAE family and generator framework, with each hierarchy change affecting several properties together.

The F1 consequences depend, in a post-hoc grouping, on which labels change: parent-label losses and descendant-label gains nearly cancel in the twelve-model intervention, while new hierarchy-only models give behavioral-minus-geometric F1 on disagreements of about $+0.44$ at target 15 and -0.03 at 45 as the label mix changes. These gaps use independent-domain E and the opposite subtraction order to I; unlike I, they average disagreements without selector weights. A census of 175 released checkpoints (163 holdout plus 12 calibration) further shows that label relations and consequences vary by family (Appendix E). Although the twelve-model study's selector-interaction criterion fails, the studies show that both training configuration and the population used for measurement affect concept labels. This conclusion is bounded by only approximately matched concept rates, differing reconstruction quality and a second factorial replicate that was outcome-informed (Appendix D).

4 A NATURAL-MODEL AUDIT

On a Gemma-2-2B Gemma Scope layer 12 residual SAE of width 16k (Lieberum et al., 2024), evaluated on held-out RAVEL cities (Huang et al., 2024), top-50 is preferred under both assignment rules, each fitted on two disjoint samples (Table 5). All four corrected intervals favor top-50 despite substantial assignment disagreement.

The concept bank contains eight concepts—five continents, China, Russia and the United States—evaluated on 771 held-out RAVEL cities, with scoring based on eight contexts for each of 778 other cities. Each selector chooses 16 of 256 eligible candidates using top-50 or uniform-50 *native activation means*, and we rank 32 uniform draws separately before averaging their selected outcomes. We hold these selections fixed across maps fitted on two disjoint assignment samples of 515 and 524 cities, maximizing detection F1 for behavioral assignment and *signed* decoder cosine for geometric assignment. The latter uses label-derived residual mean-difference directions as proxies for concepts, in contrast to the known directions in the synthetic generator.

Table 5: **Top-50 stays preferred under both natural assignment rules.** Uniform-minus-top F1; negative values favor top-50. Conditional 98.75% intervals use 20,000 paired, country-stratified evaluation-entity resamples, corrected for four contrasts; A_{fit} and A_{refit} are disjoint fitting samples.

Map fit	Assignment	Δ	98.75% interval
A_{fit}	Behavioral	-0.0274	$[-0.0431, -0.0115]$
A_{fit}	Geometric	-0.0408	$[-0.0552, -0.0260]$
A_{refit}	Behavioral	-0.0326	$[-0.0464, -0.0190]$
A_{refit}	Geometric	-0.0385	$[-0.0517, -0.0250]$

Absolute F1 ranges from 0.277 to 0.297 for top and from 0.237 to 0.269 for uniform, descriptively above random's 0.113–0.126. The post-hoc paired interactions are -0.013 and -0.006, with 95% intervals spanning zero, even though on A_{fit} the assignment rules disagree on 62.9% of candidates and on features carrying 52.3% of uniform and 43.8% of top selection weight (Supplement S6).

A post-hoc top-50 decoder-AUROC adaptation retains all four negative contrasts (Appendix F), supporting the same preference under both tested scores here. Model, domain, bank, assignment geometry and sampling design all differ from the synthetic setting, preventing attribution to one difference. Cross-sample selection-weighted label agreement ranges from 0.703 to 0.813, and both assignment refits retain negative contrasts; these refits measure label sensitivity conditional on the model and concept bank.

Supplement S8 records code-freeze timing and the analysis-code hash check added after outcomes, while Appendix F summarizes the separate country-detector study using fitted detectors. Its uniform-versus-top detector-F1 comparison remains unresolved, with full results in Supplement S6.

5 Related work

SAEBench and SynthSAEBench provide broad evaluation frameworks (Karvonen et al., 2025; Chanin & Garriga-Alonso, 2026b). Board-game models offer board-state labels for evaluating features inside trained transformers (Karvonen et al., 2024). On the synthetic substrate, Chanin (2026) shows that sparse probing tracks ground-truth mean correlation coefficient (GT-MCC) more closely than GT-F1, with the GT-F1 correlation turning negative in some settings. The same study shows that changing how many features a metric selects can reverse the sign of its correlation with GT-MCC (spurious-correlation removal, SCR, moves from $\rho = +0.65$ to -0.64 as top-N grows). Our comparison holds the score, selection and F1 formula fixed and changes the assigned concept. Chanin (2026) also measures seed and snapshot variance; van der Ben et al. (2026) study sampling and pipeline variance. Our scale comparison places a specified target reassignment alongside observed checkpoint changes.

F1-based concept–unit round trips with multi-neuron units (Zhang et al., 2026) and coalition matching (Klotz et al., 2026) address ambiguity in relating concepts to units, while ablation position changes unit-level conclusions (Noël, 2026). Our comparison instead asks how the assigned concepts of individual selected latents affect selector preference. Feature hedging in SAEs narrower than the true feature set (Chanin et al., 2025a) and sparsity misspecification (Chanin & Garriga-Alonso, 2026a) limit geometric recovery, whose relationship to causal utility is itself imperfect (Bal, 2026). Feature absorption (Chanin et al., 2025b) and SynthSAEBench tie hierarchy to feature mixing, while Grandien et al. (2026) study learned hierarchy and absorption, providing context for our training studies of competing labels. Downstream feature effects distinguish detection, causal relevance and effect geometry (Hoang et al., 2026), of which our F1 audit measures detection (Appendix I).

6 Discussion and limitations

What the reversal reflects. In a post-hoc decomposition on the original released MP and standard panels, latents whose decoder points to an ancestor of the concept they detect, as with latent 341, account for most of the net change in the selector contrast (Supplement S8). Uniform-480 places more weight on disagreeing latents, and its lower geometric scores are consistent with descendant detectors recalling only part of broader ancestor concepts. In this sense, a benchmark's matching rule is itself a choice of target. Disagreement alone does not decide the preference: in the natural audit, the rules disagree on most candidates, yet both prefer top-50.

Limitations. The resolved synthetic reversal is concentrated in particular released families and AUROC selections, and the fresh-sample study retains those models and the generator. Recurrence remains unsupported in wider models, including the earlier development study of three TopK SAEs with 16,384 latents (Supplement S2). The released panel's single generator, limited seeds, 4,096 latents for 16,384 true features and absence of vanilla TopK restrict generalization, while unequal achieved sparsity limits attribution to capacity. The natural result likewise concerns one SAE and a fixed bank with proxy directions, leaving generality across layers and tasks to further study.

Recommendations. Use the behavioral target to evaluate detectors of the concept each feature detects best, and the geometric target to evaluate detection of the generating factors assigned by decoder directions. On standard checkpoints evaluated on fresh samples, between the two selectors, these targets favor uniform and top, respectively: choosing uniform costs 2.58 F1 points under the geometric target, while choosing top gives up 1.56 under the behavioral target (Table 3). These differences quantify the cost of choosing the selector preferred under the other target.

An independently specified task fixes its target before selector comparison. Broader concept-detector comparisons can evaluate several justified targets specified beforehand, reporting absolute F1, random and simple baselines, and paired contrasts for each. The natural preference is stable only over the tested targets, while recurrence across new generator draws remains unresolved. Probing and steering address other notions of utility (Kantamneni et al., 2025; Wu et al., 2025; Arad et al., 2025; Wang et al., 2026). Appendix H provides the audit and replay code.

Reproducibility statement

Download and extract the supplementary code archive, install NumPy and SciPy, and run the commands in Appendix H to replay saved results. The archive includes the Gemma audit arrays: fitted assignments, selection weights, per-feature outcomes and conditional bootstrap draws. The audit uses these saved arrays to recompute contrasts and intervals; it does not regenerate the original activation caches, refit assignments or rerun neural inference. The archive also provides code, frozen protocols, unsuccessful analyses and source hashes.

AI use statement

In this work, we used generative AI tools to aid and polish the writing, to draw figures, and to help with coding. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Dana Arad, Aaron Mueller, and Yonatan Belinkov. SAEs are good for steering – if you select the right features. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10241–10259, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.519. URL https://aclanthology.org/2025.emnlp-main.519/.

Mohamed Abdessalem Bal. From geometric recovery to causal validation: A reproducible audit of sparse autoencoder features, from superposition geometry to causal inertness. *arXiv preprint arXiv:2607.12166*, 2026. doi: 10.48550/arXiv.2607.12166. URL https://arxiv.org/abs/2607.12166.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. https://transformer-circuits.pub/2023/monosemantic-features.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024. doi: 10.48550/arXiv.2412.06410. URL https://arxiv.org/abs/2412.06410. NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6077–6101. PMLR, 2025. doi: 10.48550/arXiv.2503.17547. URL https://proceedings.mlr.press/v267/bussmann25a.html.

David Chanin. Are sparse autoencoder benchmarks reliable? *arXiv preprint arXiv:2605.18229*, 2026. doi: 10.48550/arXiv.2605.18229.

David Chanin and Adrià Garriga-Alonso. Sparse but wrong: Incorrect L0 leads to incorrect features in sparse autoencoders. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pp. 12889–12915. PMLR, 2026a. URL https://openreview.net/forum?id=aYqF2Cv6Cr. arXiv:2508.16560.

David Chanin and Adrià Garriga-Alonso. SynthSAEBench: Evaluating sparse autoencoders on scalable realistic synthetic data. *arXiv preprint arXiv:2602.14687*, 2026b. doi: 10.48550/arXiv.2602.14687. URL https://arxiv.org/abs/2602.14687.

David Chanin, Tomáš Dulka, and Adrià Garriga-Alonso. Feature hedging: Correlated features break narrow sparse autoencoders. *arXiv preprint arXiv:2505.11756*, 2025a. doi: 10.48550/arXiv.2505.11756. NeurIPS 2025 Mechanistic Interpretability Workshop.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems*, volume 38, pp. 82318–82355, 2025b. doi: 10.52202/085713-2755. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/764ff7477b8e24dbe01531f6791e8bdf-Abstract-Conference.html. arXiv:2409.14507.

Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *Advances in Neural Information Processing Systems*, volume 38, pp. 34381–34424. Curran Associates, Inc., 2025. doi: 10.52202/085713-1154. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/311bd32bcd952f93cf820bacd6955f0a-Abstract-Conference.html. arXiv:2506.03093.

Katarzyna Filus and Sebastian Pokuciński. Measuring monosemanticity in sparse autoencoders via latent activation coherence. *arXiv preprint arXiv:2607.17770*, 2026. doi: 10.48550/arXiv.2607.17770.

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. doi: 10.48550/arXiv.2606.12138.

Nils Grandien, David Steinmann, Felix Friedrich, and Kristian Kersting. Do sparse autoencoders learn meaningful concept hierarchies? *arXiv preprint arXiv:2606.22994*, 2026. doi: 10.48550/arXiv.2606.22994. URL https://arxiv.org/abs/2606.22994.

Phu Gia Hoang, Anwoy Chatterjee, Tanmoy Chakraborty, Iryna Gurevych, and Subhabrata Dutta. Sparse autoencoders encode both concepts and functions: The downstream geometry of feature effects. *arXiv preprint arXiv:2607.24645*, 2026. doi: 10.48550/arXiv.2607.24645. URL https://arxiv.org/abs/2607.24645.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, 2024. doi: 10.18653/v1/2024.acl-long.470. URL https://aclanthology.org/2024.acl-long.470/.

Mikkel Godsk Jørgensen and Lars Kai Hansen. Steering LLMs? Actually, sparse autoencoders can outperform simple baselines. *arXiv preprint arXiv:2605.31183*, 2026. doi: 10.48550/arXiv.2605.31183. URL https://arxiv.org/abs/2605.31183.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? A case study in sparse probing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29018–29049. PMLR, 2025. doi: 10.48550/arXiv.2502.16681. URL https://proceedings.mlr.press/v267/kantamneni25a.html.

Adam Karvonen, Benjamin Wright, Can Rager, Rico Angell, Jannik Brinkmann, Logan Smith, Claudio Mayrink Verdun, David Bau, and Samuel Marks. Measuring progress in dictionary learning for language model interpretability with board game models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 83091–83118. Curran Associates, Inc., 2024. doi: 10.52202/079017-2644. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9736acf007760cc2b47948ae3cf06274-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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning*

Research, pp. 29223–29264. PMLR, 2025. URL https://proceedings.mlr.press/v267/karvonen25a.html. arXiv:2503.09532.

Jonas Klotz, Cassio F. Dantas, Pallavi Jain, Diego Marcos, and Begüm Demir. Evaluating the interpretability of sparse autoencoders with concept annotations. In *Computer Vision – ECCV 2026*, Lecture Notes in Computer Science. Springer, 2026. doi: 10.48550/arXiv.2606.24716. URL https://arxiv.org/abs/2606.24716. arXiv:2606.24716.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Sanity checks for sparse autoencoders: Do SAEs beat random baselines? *arXiv preprint arXiv:2602.14111*, 2026. doi: 10.48550/arXiv.2602.14111.

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 *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.19.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. A sober look at the unsupervised learning of disentangled representations and their evaluation. *Journal of Machine Learning Research*, 21(209):1–62, 2020. URL https://www.jmlr.org/papers/v21/19-976.html. arXiv:2010.14766.

Valentin Noël. Where you measure decides what you measure: Position selection in ablation-based SAE evaluation. *arXiv preprint arXiv:2608.13337*, 2026. doi: 10.48550/arXiv.2608.13337.

Tuomas Oikarinen, Ge Yan, and Tsui-Wei Weng. Evaluating neuron explanations: A unified framework with sanity checks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47090–47131. PMLR, 2025. doi: 10.48550/arXiv.2506.05774. URL https://proceedings.mlr.press/v267/oikarinen25a.html.

Gonçalo Paulo and Nora Belrose. Evaluating SAE interpretability without generating explanations. In *The Fourteenth International Conference on Learning Representations*, 2026. doi: 10.48550/arXiv.2507.08473. URL https://openreview.net/forum?id=kHhMs642rR.

Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. Better than average: Paired evaluation of NLP systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2301–2315, 2021. doi: 10.18653/v1/2021.acl-long.179. URL https://aclanthology.org/2021.acl-long.179/. arXiv:2110.10746.

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. doi: 10.48550/arXiv.2407.14435. URL https://arxiv.org/abs/2407.14435.

Deepshik Sharma. Analyzing visual aircraft representations with sparse autoencoders. *arXiv preprint arXiv:2606.15468*, 2026. doi: 10.48550/arXiv.2606.15468.

Kaden Stillwagon, Alexandra D. VandeLoo, and Craig R. Forest. Maximum Matching Accuracy: An instance segmentation evaluation metric utilizing globally optimal matching. *arXiv preprint arXiv:2606.10107*, 2026. doi: 10.48550/arXiv.2606.10107. URL https://arxiv.org/abs/2606.10107.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread, https://transformer-circuits.pub/2024/scaling-monosemanticity/, 2024. Posted to arXiv as arXiv:2605.29358 in May 2026.

Sinie van der Ben, Neele Roch, Anna Hedström, and Mennatallah El-Assady. Building fast, evaluating slow: Pipeline choices dominate autointerpretability score variance. *arXiv preprint arXiv:2607.19386*, 2026. doi: 10.48550/arXiv.2607.19386. URL https://arxiv.org/abs/2607.19386.

Xu Wang, Yan Hu, Benyou Wang, and Difan Zou. Does higher interpretability imply better utility? A pairwise analysis on sparse autoencoders. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Q4ooLNOFeR. arXiv:2510.03659.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67035–67080. PMLR, 2025. URL https://proceedings.mlr.press/v267/wu25a.html.

Chenhao Zhang, Chris Lin, and Su-In Lee. A geometric view for understanding concept learning and neuron interpretation in sparse autoencoders. *arXiv preprint arXiv:2606.07007*, 2026. doi: 10.48550/arXiv.2606.07007. URL https://arxiv.org/abs/2606.07007.

GUIDE TO THE SUPPORTING EVIDENCE

The core appendix gives the methods, central evidence and limits needed to assess the main claims, while the separate *Supplementary evidence* PDF contains full study records. We interpret each study using its stated population and uncertainty unit, without pooling studies or combining their intervals. Label transfer means evaluating labels fitted on one domain in another domain.

Core evidence	Complete records
A–B: Protocol, selection and boundaries	S1–S3; S8 execution
C: Retraining and model quality	S4
D–E: Dependence, label transfer and persistence	S5
F: Natural audit and secondary detectors	S6
G–H: Worked example, scale and replay	S1; S8
I: Closest prior-work distinctions	S7

A PROTOCOL, DATA SETS AND UNCERTAINTY

We fit behavioral labels by maximizing detection F1 on A and assign geometric labels using the factor of maximum absolute normalized decoder cosine, then evaluate the same fixed binary activity on E. The archive calls these rules activation and decoder assignment, respectively, while plain A denotes assignment data and subscript A in Δ_A denotes the behavioral rule. Each map attaches one concept to a feature, using fitted relationships between feature activity and concepts for behavioral labels and the benchmark's per-latent decoder-cosine classification rule for synthetic geometric labels. Dictionary recovery and causal utility require separate measurements, and an independently specified task retains its own concept as the evaluation target. Supplements S1 and S8 provide full methods and execution records.

Table 6 distinguishes the development, calibration and evaluation studies. Development evaluation first used decoder-based assignment; behavioral assignment was added as an exploratory control before the holdout protocol fixed both. The main released panel's fixed seed-42 generator has 16,384 unit directions in 768 dimensions, 128 root concepts, four children per internal node and maximum depth three. A child can occur only with its parent and siblings cannot co-occur, with probabilities compensating for these constraints and concepts outside the trees treated as unstructured labels.

The released grid contains five families, seven target sparsities (15–45), five seeds and width 4,096, leaving $175-12 = 163 = 18+40+105$ checkpoints after calibration exclusions whose seed, sparsity or family, respectively, was absent from calibration. We retain the original 105-model test as primary and use 240,000 examples for each data set: original scoring and assignment seeds are 917101 and 917102, and diagnostic and evaluation seeds are 918103 and 918104. For the fresh replication we redrew scoring, assignment and evaluation data with seeds 920101, 920102 and 920104, together with new candidates on the same models. Eligibility requires $0.001 < p(z_j > 0) < 0.10$ and at least 480 active scoring examples, giving an effective minimum rate of 0.002, and a seeded permutation supplies the first 256 eligible IDs to selection and the next 256 to the original diagnostic pool. These constants were fixed in the holdout protocol before C and E were generated. We resolve assignment ties by the lowest concept ID and seal maps and selections before held-out evaluation under the freeze convention in Section 2.2, using "in-house" for runs performed by us.

Sampling and scoring. We compute decoder-readout AUROC for top-480 and uniform-480 from 480 active examples and 4,000 inactive examples Z_j (Eq. 1), ranking each of uniform's 32 draws separately before averaging selected outcomes. The matched-fraction sampling variants use

$$q = \min\{1, 480/\operatorname{median}_j n_j\}, \qquad n_{q,j} = \lceil q n_j \rceil, \quad n_j = |F_j|. \tag{3}$$

This allocation matches counts within a latent and approximately matches fractions across latents while allowing total compute to vary. Additional conditions are decile-480, with 48 examples per activation decile; all/4k, with all active examples against the same 4,000 references; and T_{all}, with all active and inactive examples. Uniform AUROC is conditionally unbiased for all/4k AUROC, although ranking by estimated scores can bias the F1 of selected features. Writing N_j for the inactive

examples, the label-free baseline is

$$M_j = \frac{\mathbb{E}[r_j \mid F_j] - \mathbb{E}[r_j \mid N_j]}{\mathrm{SD}(r_j)}. \tag{4}$$

We use the exact candidate-pool mean as the random-selection baseline and average cutoff ties analytically: for candidates $\mathcal{J}_>$ above the cutoff, tied candidates $\mathcal{J}_=$, counts $n_>, n_=$ and budget k, the mean under rule ℓ is

$$\overline{\mathrm{F1}}_\ell = k^{-1}\left[\sum_{j\in\mathcal{J}_>} \mathrm{F1}_{\ell j} + \frac{k-n_>}{n_=}\sum_{j\in\mathcal{J}_=} \mathrm{F1}_{\ell j}\right]. \tag{5}$$

We preserve AUROC ties through integer midrank numerators, so selection averages the entire saturated set whenever at least k candidates score one.

Uncertainty and execution. We construct the original 95% intervals by crossing 500 shared example bootstraps with 10,000 within-cell seed resamples and weighting all 35 cells equally, including the six cells with three seeds and the remaining cells with five. These intervals condition on the realized generator, cells, candidates and scoring, assignment and diagnostic data, with checkpoint seeds and evaluation samples as the uncertainty units. Fresh primary tests use four one-sided Bonferroni bounds at $\alpha = 0.0125$, while common-seed t_4 bounds provide a separate sensitivity conditional on fixed E (Appendix B). Supplement S8 records caching amendments, evaluation-time completion of the binning helper and verification scope.

Eligibility and assignment refits. Family pool composition qualifies between-family comparisons because ancestor–descendant disagreements account for 25.2% of MP pools and 33.9% of standard pools, close to their total disagreement rates of 25.9% and 34.1%. We fit two independent assignment maps with seeds 919202 and 919203, retaining all MP selection weight and 95.12% of standard top weight and 94.84% of standard uniform weight on features whose labels stay unchanged in both refits. Those standard features contribute -0.05489 to I without renormalization, compared with totals of -0.05517 and -0.05512 in the two refits, showing sample stability within the fixed trained basis. These refits do not test whether that learned basis is unique across training runs.

Table 6: **Evidence stages and their roles.** We report each stage with its own uncertainty convention: the original holdout reuses calibration scoring and assignment data with new diagnostic and evaluation samples, while the replication redraws scoring, assignment and evaluation data and candidates on the same generator and checkpoints. Assignment refits and baseline sensitivities reuse observed outcomes exploratorily, and the direct natural study has a separate execution record (Appendix F).

Stage	Models or task	Role
Development	Three 16k TopK; two 4k BatchTopK and two 4k Matryoshka	Fresh frozen selections, followed by exploratory sampling, baseline, and assignment controls
Calibration	Twelve released 4k BatchTopK/Matryoshka checkpoints	Sampling comparison and failed prospective predictor; informs new protocol
Holdout	163 remaining released checkpoints; 105 from new families	Primary confirmatory test on 105; family, assignment, and diagnostic sensitivities
Fresh replication	Same 163 checkpoints, new samples and candidates	Four directional tests fixed before sampling
Dependence intervention	12 initial, 16 factorial, 16 extension models; census of 175 released models	Separate training and assignment effects; test sparsity, shape and family boundaries
Natural assignment	One Gemma-2-2B SAE; eight concepts, 771 evaluation cities	Fixed native selections, two map fits; same-direction preference
Natural detectors	Same SAE; twelve countries, 329 test cities	Label-screened selection and detector comparison

B SELECTION RESULTS AND BOUNDARIES

The original 105-model criterion fails despite positive secondary results on the full 163-model panel. In the earlier twelve-checkpoint calibration, mean F1 is 0.804 for top and 0.842 for uniform. We fitted its frozen ridge regression to predict per-checkpoint uniform-minus-top F1 from the top-480 score-one fraction and the log ratio of tie-set to pool median numbers of active examples. Expansion required complete schema-valid data, lower cross-validated MSE than a training-fold constant, fraction matching that reduced the selected log-rate gap in at least four of six cells, and absolute F1 effects of at least 0.01 in at least two cells. The predictor's MSE of 0.000850 exceeds the constant's 0.000575, so the expansion criterion fails. The later holdout used a separate calibration-informed protocol whose freeze chronology appears in Supplement S8. Earlier development comparisons use ranking adaptations of the source statistics, as defined in Supplement S2.

Table 7: **The primary 105-checkpoint comparison does not support a uniform-480 advantage, and the family means run in opposite directions.** The table reports behaviorally assigned mean F1 for 16 of 256 candidates, with each row weighting sparsity cells equally and uniform means averaging 32 separately ranked draws. Positive uniform-minus-top differences favor uniform-480, all variants average cutoff ties, and paired 95% intervals cross example and seed resampling. Primary 105 is the sole confirmatory population.

Population	Top-480	Uniform-480	M	Random	Uniform − top [95% CI]
Primary 105	0.653	0.652	0.652	0.505	-0.002 $[-0.006, +0.003]$
Full holdout 163	0.714	0.730	0.727	0.624	$+0.016$ $[+0.013, +0.019]$
BatchTopK	0.769	0.804	0.804	0.738	$+0.035$ $[+0.032, +0.038]$
JumpReLU	0.589	0.502	0.536	0.389	-0.087 $[-0.097, -0.077]$
Matryoshka	0.842	0.890	0.872	0.867	$+0.048$ $[+0.045, +0.052]$
MP	0.733	0.801	0.763	0.533	$+0.068$ $[+0.060, +0.077]$
Standard	0.638	0.652	0.657	0.593	$+0.014$ $[+0.010, +0.018]$

For fixed selection weights and F1 outcomes, only latents with disagreeing labels $\mathcal{J}_{\neq}$ contribute to the assignment interaction:

$$I = \Delta_D - \Delta_A = \sum_{j \in \mathcal{J}_{\neq}} (w_{Uj} - w_{Tj})(\mathrm{F1}_{Dj} - \mathrm{F1}_{Aj}). \tag{6}$$

The identity holds for all 163 checkpoints, and its post-hoc decomposition does not choose a uniquely correct target. We report the headline and baseline populations in Tables 8 and 9. Supplement S1 contains all seven baseline rows and the precision and recall decompositions.

Exposure is descriptive. Across 163 checkpoints, we regress interaction on the disagreement-exposure gap without checkpoint weights, obtaining slope -0.356, intercept -0.0026 and $r = -0.872$, or slope -0.420 and $r = -0.918$ after excluding JumpReLU ($n = 128$). These in-sample, non-causal fits weight checkpoints equally rather than matching the equal-cell means, and their association combines exposure with the feature-level F1 changes in the identity above. JumpReLU illustrates why both matter: its contrast moves from $\Delta_A = -0.0869$ to $\Delta_D = -0.0936$ while preserving the same selector preference. Exposure alone does not determine reversal.

Fresh samples and common-seed sensitivity. All 163 fresh-replication cases completed, and MP and standard pass the four within-cell directional bounds in Table 3. Standard also passes the common-seed sensitivity, whereas MP's geometric upper bound of $+0.0024$ leaves its geometric direction unresolved. Separately, the fresh 105-model behavioral contrast is unresolved at $+0.0029$ $[-0.0022, 0.0081]$. For five seed-label block means d_b, we calculate common-seed bounds as $\bar{d} \pm t_{4,0.9875} s_d/\sqrt{5}$, using the lower bound for behavioral and the upper bound for geometric assignment conditional on E. This grouping measures sensitivity to dependence among existing checkpoints, because released seed labels alone leave shared initialization and training-stream coupling across sparsities unverified.

Table 8: Headline assignment contrasts, exposure and per-checkpoint counts. Uniform-minus-top selected F1 at budget 16 under both assignments, with paired crossed example and seed 95% intervals. The Primary 105 Δ_A entry repeats the confirmatory test in Table 7; all other entries are secondary and unadjusted. "Disagree" gives the shares of uniform and top selection weight on latents whose two assigned concepts differ; "geom. anc." restricts it to latents whose geometrically assigned concept is a generator-tree ancestor of the behaviorally assigned one. "Counts" gives checkpoints with $\Delta_A > 0$ and $\Delta_D < 0$, then checkpoints where geometrically assigned uniform-480 exceeds random. These counts summarize point signs, not per-checkpoint significance tests. The lower panel includes all five families; Gap is uniform minus top disagreement exposure, in percentage points.

Population	Δ_A [95% CI]	Δ_D [95% CI]	Disagree U/T (%)
Primary 105	-0.0016 $[-0.0061, +0.0029]$	-0.0520 $[-0.0602, -0.0439]$	42.9/30.2
Full holdout 163	$+0.0158$ $[+0.0129, +0.0186]$	-0.0065 $[-0.0117, -0.0015]$	26.1/20.4
MP (35)	$+0.0681$ $[+0.0596, +0.0766]$	-0.0214 $[-0.0386, -0.0045]$	60.7/39.5
Standard (35)	$+0.0142$ $[+0.0096, +0.0184]$	-0.0409 $[-0.0539, -0.0290]$	42.4/32.5

Family	Disagree U/T (%)	Gap (pp)	Geom. anc. U/T (%)	Counts
BatchTopK	1.0/9.7	-8.69	—	—
JumpReLU	25.6/18.6	+7.03	—	—
Matryoshka	0.6/2.0	-1.32	—	—
MP	60.7/39.5	+21.19	52.8/35.0	21/35; 30/35
Standard	42.4/32.5	+9.90	30.4/18.2	22/35; 24/35

Table 9: Geometric-assignment baselines at budget 16 for the populations in the headline panel. We report secondary, unadjusted paired 95% intervals for differences against random selection, providing absolute-quality context for the selector comparison. These intervals do not establish equivalence or a validated selection policy.

Population	Random F1	Top-480 − random [95% CI]	Uniform-480 − random [95% CI]
Primary 105	0.4370	$+0.1193$ [+0.1120, +0.1268]	$+0.0674$ [+0.0615, +0.0733]
Full holdout 163	0.5755	$+0.0712$ [+0.0668, +0.0758]	$+0.0647$ [+0.0610, +0.0686]
MP	0.4236	$+0.1055$ [+0.0891, +0.1220]	$+0.0841$ [+0.0724, +0.0967]
Standard	0.5343	$+0.0507$ [+0.0435, +0.0583]	$+0.0098$ [+0.0002, +0.0188]

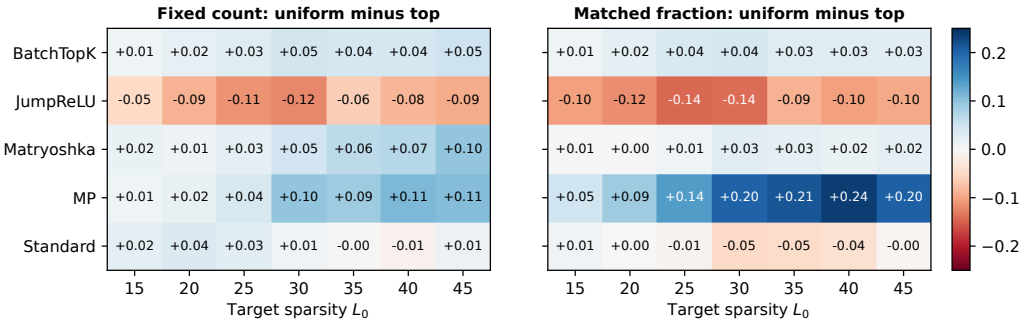


Figure 3: **Cell means expose the family and sparsity structure that the aggregate hides.** Positive cells favor uniform-480 and negative cells favor top-480. The panels show cell-mean behaviorally assigned F1 differences at budget 16, using five seeds per cell except for the six BatchTopK and Matryoshka cells that retain three after calibration exclusions. The fixed-480 panel on the left and matched-fraction panel on the right (Eq. 3) share a color scale, with uncertainty for aggregate comparisons in Table 7.

Table 10: Fresh replication by common-seed block. Each entry averages seven fixed sparsity cells, with no newly trained model. All point signs reverse. Under five-block uncertainty, standard passes both bounds while MP's geometric bound crosses zero (Table 3).

	Standard		Matching pursuit	
Seed block	Δ_A	Δ_D	Δ_A	Δ_D
0	$+0.0067$	-0.0147	$+0.0593$	-0.0185
1	$+0.0150$	-0.0358	$+0.0685$	-0.0617
2	$+0.0195$	-0.0261	$+0.0716$	-0.0036
3	$+0.0173$	-0.0309	$+0.0801$	-0.0411
4	$+0.0193$	-0.0215	$+0.0767$	-0.0557

Table 11: Original and fresh secondary budgets from saved checkpoint outcomes. Uniform-minus-top F1, equal means over seven sparsities and five seeds per family. These descriptive point contrasts do not replace the corrected budget-16 criterion.

Panel	Family	Budget	Behavioral Δ_A	Geometric Δ_D
Original	Standard	8	$+0.0235$	-0.0448
Original	Standard	32	$+0.0012$	-0.0314
Original	MP	8	$+0.0871$	-0.0226
Original	MP	32	$+0.0448$	-0.0391
Fresh	Standard	8	$+0.0245$	-0.0290
Fresh	Standard	32	$+0.0049$	-0.0186
Fresh	MP	8	$+0.0906$	-0.0369
Fresh	MP	32	$+0.0501$	-0.0404

Scorer and hierarchy boundaries. For the native-score comparison, we average 50 positive encoder activations without scale normalization on the same 70 MP and standard checkpoints and candidate pools, fixing selections before drawing E with seed 921104. The hierarchy-cap extension replaces depth-3 targets with depth-2 ancestors while retaining shallower and unstructured labels and all selections, then draws evaluation data with seed 921204. We test each extension with four two-sided 98.75% common-seed t_4 intervals that permit either reversal orientation, and neither meets its joint reversal criterion after multiplicity correction. Supplement S3 reports absolute scores, unchanged-target depth-cap-3 controls and sensitivities.

Tables 10 and 11 report common-seed means and secondary budgets, while Table 12 retains each principal panel's uncertainty convention. The original row is secondary and unadjusted, and fresh rows give corrected directional bounds within cells and across common seeds in parentheses, using behavioral lower bounds and geometric upper bounds. "Part" denotes an MP fresh-sample result that passes within cells but remains unresolved under common-seed inference, and "no" denotes failure of the specified joint reversal criterion. The natural row instead supports top-50 under both assignments and shows only the A fit, with both fits in Table 5. Exploratory fixed-selection dependence analyses appear separately in Appendix D.

Alternative allocation and mean-difference baselines. The primary 105-model uniform-minus-M contrast is unresolved at -0.0007 $[-0.0052, 0.0037]$, although M trails a selector in some families: JumpReLU top-480 gives 0.5891 versus 0.5363 for M, and MP uniform exceeds M by $+0.0375$ $[0.0284, 0.0465]$. On the original holdout at budget 16, uniform-480 shows no resolved advantage over M in standard under behavioral assignment (-0.0053 $[-0.0105, +0.00003]$). Under geometric assignment, the original-holdout uniform-minus-M contrasts are -0.0283 $[-0.0476, -0.0093]$ in MP and -0.0163 $[-0.0237, -0.0084]$ in standard. These comparisons do not evaluate top against M or establish baseline performance on fresh samples. Matched fractions yield a positive primary sampling contrast of $+0.0102$ $[0.0055, 0.0149]$. Standard changes direction from $+0.0142$ to -0.0195 $[-0.0235, -0.0154]$, MP increases to $+0.1623$ $[0.1533, 0.1713]$, and JumpReLU remains

Table 12: Detailed principal recurrence contrasts. Positive contrasts favor uniform and negative contrasts favor top. Natural uses 50 active examples, whereas the synthetic AUROC rows use 480. Native-score, hierarchy-cap and paired-width rows use two-sided 98.75% t_4 intervals across five common-seed blocks or training seeds. Fresh rows use one-sided 98.75% bounds for within-cell and common-seed analyses. The natural row uses two-sided 98.75% percentile bands and the historical generator pilot uses two-sided 95% t_1 intervals. †: original full-panel 95% crossed bootstrap intervals are secondary and unadjusted. Primary width-16k rows are in Table 15. We interpret each row under its own uncertainty convention.

Panel	Scorer	Δ_A [interval]	Δ_D [interval]	Joint
Released checkpoints reused				
Full holdout 163†	AUROC	$+0.0158$ $[+0.0129, +0.0186]$	-0.0065 $[-0.0117, -0.0015]$	yes
Fresh MP	AUROC	$+0.0713$ $(+0.0614/+0.0587)$	-0.0361 $(-0.0168/+0.0024)$	part
Fresh std	AUROC	$+0.0156$ $(+0.0108/+0.0073)$	-0.0258 $(-0.0143/-0.0130)$	yes
Hierarchy-cap MP	AUROC	$+0.0714$ $[+0.0542, +0.0886]$	-0.0344 $[-0.0789, +0.0102]$	no
Hierarchy-cap std	AUROC	$+0.0030$ $[-0.0119, +0.0179]$	-0.0362 $[-0.0566, -0.0158]$	no
Native MP	native	$+0.0387$ $[+0.0175, +0.0599]$	-0.0065 $[-0.0306, +0.0175]$	no
Native std	native	$+0.0656$ $[+0.0565, +0.0746]$	$+0.0017$ $[-0.0156, +0.0190]$	no
Natural	native	-0.0274 $[-0.0431, -0.0115]$	-0.0408 $[-0.0552, -0.0260]$	no
Models retrained				
Paired-width 4k std	AUROC	$+0.0031$ $[-0.0441, +0.0503]$	-0.0267 $[-0.1387, +0.0852]$	no
Paired-width 4k MP	AUROC	$+0.0774$ $[+0.0403, +0.1145]$	-0.0211 $[-0.0976, +0.0553]$	no
Pilot std	AUROC	$+0.0059$ $[-0.0557, +0.0676]$	-0.0103 $[-0.0278, +0.0071]$	no

negative. This separate allocation comparison changes compute as well as sampling and leaves the failed fixed-count primary result intact.

Selection diagnostics. We define top saturation as the fraction of candidates whose top-480 AUROC T equals one, write n_j for the number of active examples on S and restrict $\mathrm{median}_{T=1}$ to saturated candidates. The diagnostic pool is disjoint from the selection pool, with C-F1 denoting F1 on the diagnostic population, H3 and H4 denoting secondary diagnostic hypotheses, H1 the sampling test and H2 the saturation description. Only 53/163 holdout checkpoints exceed the descriptive 0.30 saturation threshold, all BatchTopK or Matryoshka, and all 53 have positive sampling gains. Positive gains also occur in 66 of the other 110 checkpoints, including 33/35 MP cases, while top exceeds random by more than 0.02 F1 in 128 checkpoints, falls within the descriptive ± 0.02 band in 18 and falls below it in 17.

H3 relates the diagnostic pool's scoring-count versus C-F1 Spearman correlation to selection-pool E top-minus-random F1. The association, weighting family and sparsity cells equally and seeds equally within cells, is 0.676 $[0.636, 0.715]$ across 163 and 0.253 $[0.142, 0.362]$ on the primary 105. Primary geometric assignment gives -0.095 $[-0.187, -0.003]$. H4 uses diagnostic F1 SD, adjusting for diagnostic mean, family and sparsity. Its full-panel rank association is 0.160 $[0.103, 0.218]$ and adjusted slope 0.842 $[0.643, 1.068]$. Resampling holds C fixed, omitting its estimation uncertainty. When selection and diagnostic features must have disjoint generator-tree roots, only four behavioral and five geometric cases meet the 64-candidate minimum. These label-aware statistics explain observed selection outcomes; because they require concept labels, they are not label-free selectors.

For the rate control, we sort the 256 candidates by number of active examples and then ID, divide them into equal bins and redistribute top-selection weight uniformly within each bin. The resulting top-minus-control contrast is 0.0363 $[0.0340, 0.0387]$ with 16 bins and 0.0298 $[0.0276, 0.0320]$ with 32, describing the residual contrast after this particular redistribution of weight. The partition is descriptive, not a causal decomposition. The separate post-hoc analysis relates each unsaturated pool's correlation between active-example count and F1 to its uniform-minus-top gain ($\rho = 0.423$). It reuses E on both axes, unlike H3's disjoint-pool design. Supplement S2 retains the sensitivities and diagnostic plot, with prospective selector prediction remaining an open question.

C RETRAINING AND MODEL QUALITY

The retraining studies leave the joint reversal unresolved at the primary budget, while differences in achieved sparsity and reconstruction quality, together with unverified convergence, limit attribution to model width.

Training recipes and conditioning. We report reconstruction quality as variance explained (VE), also denoted R^2 in the model records, and distinguish paired changes in model width from generator-block studies that change the generator draw at fixed width. The standard 200M-example studies use width 4,096, 195,312 batches of 1,024 (199,999,488 examples), a 65,104-step sparsity warmup, Adam learning rate 3×10^{-4}, betas $(0.9, 0.999)$, epsilon 10^{-8}, gradient clipping 1, decoder initialization norm 0.1 and initial L1 coefficient 2. We use BF16 autocast in both the historical and new generator-block studies and retain their fixed-budget final checkpoints without outcome-guided retuning. Their SAELens 6.51.0 differs from version 6.36.1 in the released standard metadata. Nominal recipe matches do not imply equal convergence, given the software differences and fixed training budgets. Supplement S4 records software pins, pairing, per-study precision and operational deviations.

Two additional independent generator and training blocks. We pair two new generator draws with new training seed ranges, each spanning target sparsities 15–45, using generator, initialization and training bases of 971000, 972000 and 973000 plus the block index. Separate scoring, assignment and evaluation populations contain 240,000 examples each, and we sealed all fourteen maps and selections before evaluation. Eligibility, candidate choice, top-480 and uniform-480 AUROC, 32 uniform draws, inactive references and tie handling follow Appendix A, with both blocks and all fourteen cells reported in Tables 13 and 14.

Table 13: **The new blocks have different point preferences.** We average budget-16 individual-latent F1 equally over seven sparsities within each block, with positive Δ favoring uniform-480. Displayed block numbers are one-based and correspond to saved IDs 0 and 1, with individual signs reported descriptively.

Block	Assignment	Top F1	Uniform F1	Δ
1	Behavioral	0.6449	0.6482	$+0.0034$
1	Geometric	0.5646	0.5625	-0.0021
2	Behavioral	0.6324	0.6285	-0.0039
2	Geometric	0.5749	0.5356	-0.0392

The first block reverses in its point average despite having no individual cell with positive behavioral and negative geometric contrasts, whereas the second favors top in both averages despite two cells with that orientation. Equal-block means are -0.0003 for behavioral and -0.0207 for geometric assignment, with descriptive two-sided 95% t_1 intervals of $[-0.0464, 0.0459]$ and $[-0.2564, 0.2150]$, respectively. Recurrence remains unresolved across these two independent blocks, each containing seven sparsities, with secondary budgets and execution records in Supplement S4.

Historical generator pilot remains separate. The earlier fourteen-model standard pilot jointly varied two generator and training draws, sharing initialization and data-order seeds across seven sparsities within each block. We report its mixed point results separately in Table 12, with complete methods, quality and budgets in Supplement S4, because combining it with the new blocks would change the prospectively specified two-block test.

Paired-width study. We train twenty new standard and MP models on the original generator, crossing widths 4,096 and 16,384 with five paired seeds and using 19,531 batches of 2,048 examples (39,999,488 in total), float32 arithmetic with TF32 off and 6,510 warmup steps. Standard targets $L_0 = 30$, while MP permits repeated supports with a maximum of 30 iterations and zero residual threshold. New scoring, assignment and evaluation samples use seeds 922101, 922102 and 922104, and all twenty models meet the eligibility and candidate-count requirements. We average draws

Table 14: All fourteen new cells. VE and achieved L_0 describe model quality and coverage. Contrasts are uniform-minus-top F1 at budget 16. The table preserves adverse cells and is descriptive.

Block	Target L_0	Achieved L_0	VE	Eligible	Δ_A	Δ_D
1	15	15.01	0.731	3426	-0.0143	-0.0180
1	20	20.01	0.767	3775	-0.0055	$+0.0245$
1	25	25.01	0.780	3975	-0.0069	-0.0380
1	30	30.00	0.787	4071	-0.0059	-0.0165
1	35	35.00	0.791	4062	-0.0077	-0.0049
1	40	40.02	0.795	4054	$+0.0441$	$+0.0277$
1	45	45.05	0.798	4046	$+0.0198$	$+0.0101$
2	15	15.01	0.732	3450	-0.0111	-0.1243
2	20	20.02	0.767	3781	$+0.0167$	-0.0564
2	25	25.01	0.780	3971	-0.0019	-0.0195
2	30	30.02	0.787	4066	-0.0149	-0.0358
2	35	35.02	0.791	4061	-0.0037	-0.0019
2	40	40.03	0.794	4050	-0.0139	-0.0137
2	45	45.04	0.797	4040	$+0.0017$	-0.0231

within each model and use the five training seeds as inference units conditional on the shared generator and realized populations, reporting four primary width-16k contrasts with two-sided 98.75% t_4 intervals and width-4k rows as diagnostics.

Table 15: **Paired-width study at selection budget 16.** Uniform-minus-top F1 and 98.75% t_4 intervals across five paired training seeds. Width-16,384 rows are primary. Width-4,096 rows are matched-control diagnostics. None passes the joint reversal criterion.

Family	Width	Δ_A [interval]	Δ_D [interval]
Standard	4,096	$+0.0031$ $[-0.0441, +0.0503]$	-0.0267 $[-0.1387, +0.0852]$
Standard	16,384	-0.0290 $[-0.0941, +0.0361]$	-0.0239 $[-0.1178, +0.0699]$
MP	4,096	$+0.0774$ $[+0.0403, +0.1145]$	-0.0211 $[-0.0976, +0.0553]$
MP	16,384	-0.0282 $[-0.0562, -0.0003]$	-0.1396 $[-0.3528, +0.0737]$

Neither wide family passes both directions in Table 15: wide MP's behavioral contrast is negative, -0.02824 $[-0.05619, -0.00030]$, while its geometric contrast remains unresolved. The 4k MP behavioral gain of $+0.07737$ $[+0.04026, +0.11449]$ likewise accompanies an unresolved geometric direction, and all alternative budgets fail the joint reversal criterion. The paired width change $I_{16k} - I_{4k}$ has descriptive 95% intervals spanning zero for both families, with unequal achieved sparsity further limiting attribution to width or convergence (Table 16).

Table 16: Paired-width-study quality and descriptive absolute F1, budget 16. Upper panel: range of achieved sparsity and arithmetic means across five seeds on S. Alive means active at least once, eligible uses the selection rule. Lower panel: mean selected F1, averaging 32 uniform draws within each seed. These descriptive baselines compare the observed conditions. Fixed budgets do not imply equal convergence, and choosing a policy requires separate validation. All twenty quality rows remain in Supplement S4.

Family / width	L_0 range	Mean alive	Mean eligible	Mean VE
Standard / 4,096	29.80–30.09	4096.0	3947.2	0.7733
Standard / 16,384	42.90–43.95	15818.6	4916.4	0.8563
MP / 4,096	29.81–29.83	4096.0	4095.0	0.8382
MP / 16,384	26.25–26.28	16377.2	2514.0	0.9869

Family	Width	Assignment	Top	Uniform	M	Random
Standard	4,096	Behavioral	0.6225	0.6256	0.6217	0.5860
Standard	4,096	Geometric	0.5864	0.5597	0.5578	0.5317
Standard	16,384	Behavioral	0.7623	0.7333	0.7653	0.7124
Standard	16,384	Geometric	0.6760	0.6521	0.7146	0.6408
MP	4,096	Behavioral	0.7440	0.8214	0.7522	0.5299
MP	4,096	Geometric	0.6275	0.6064	0.5780	0.4617
MP	16,384	Behavioral	0.9758	0.9475	0.9395	0.7643
MP	16,384	Geometric	0.8258	0.6862	0.5321	0.5236

D DEPENDENCE INTERVENTION AND LABEL TRANSFER

We use label transfer to mean evaluating labels fitted on one assignment domain in another evaluation domain.

Paired design and its failed selector endpoint. We train twelve standard SAEs across two generator blocks, two training replicates and three conditions: the original hierarchy and correlation, a sampling control reproducing the original correlated concept sampler, and independent Bernoulli concepts. Within each block and replicate we pair the dictionary, bias and initialization, using the 200M-example recipe with target $L_0 = 30$ and estimating independent probabilities on 400,000 original rows. The resulting marginal match is approximate, and we retain every final model and its sealed 256-candidate selection in these exploratory studies motivated by earlier results. In-house freezes do not make these studies confirmatory.

The initial study changes the training and measurement populations together. For AUROC selections, intervention-minus-original changes in $I = \Delta_D - \Delta_A$ (Eq. 2) are -0.007069, -0.023849, $+0.006210$ and $+0.030661$. Their median absolute change of 0.015459 is below the largest control change of 0.059561, and their signs disagree, so the criterion requiring concordant signs and a median absolute change larger than every control change fails. Point reversals occur in 0/4 original conditions, 2/4 controls and 0/4 independent conditions, leaving no original-condition reversal for this intervention to remove.

Common populations and full-feature census. We then evaluated every model on shared original and independent assignment and evaluation populations, each containing 240,000 rows per role and block, and sealed 24 fitted maps before drawing the new evaluation samples. The census covers all 4,096 features, with no zero, invalid or tied maxima among 98,304 measurements across features and domains. Original and control training yield 6,183 and 6,213 disagreements out of 16,384 features on original assignment data, falling to 3,815 and 3,816 on independent assignment data. Independently trained models yield 23 and 22 disagreements on the two populations, with none hidden in their large ineligible feature group. Models remain the comparison units, and feature counts describe the outcomes within them.

In one post-hoc example, dependence block 1 and replicate 1 (saved ID b1r1), changing assignment data alone moves the behavioral contrast from -0.004210 to $+0.020241$ on fixed original E, along-

side an unchanged geometric contrast of -0.035396. We report this change in point preference descriptively.

Original and control models achieve $L_0 \simeq 30$ and VE 0.786 on original evaluation data, compared with $L_0 \simeq 48$ and VE 0.290 on independent data. Independently trained models achieve $L_0 = 31.28$ with VE 0.543 on original data and $L_0 = 30.00$ with VE 0.541 on independent data. On the actual shared populations, 96.20–96.52% of concept-rate ratios fall within 25%, compared with 98.50–98.65% in the pre-training checks, so greater assignment agreement does not by itself establish better recovery: reconstruction quality and residual marginal differences also vary.

Stable labels and held-out consequences. For original and control models, disjoint half-sample refits give 91.0–92.5% same-label agreement on original data and 97.3–98.3% on independent data, compared with 70–73% agreement between domains. Two new paired assignment and evaluation panels retain 93.77–94.78% and 98.07–98.71% of labels in the respective domains. To measure the held-out consequence for each feature, we define

$$g_j(A, E) = F1_E(j, a_j^{\text{beh}}(A)) - F1_E(j, a_j^{\text{geo}}).$$

Here $a_j^{\text{beh}}(A)$ and a_j^{geo} are concept IDs, and the per-feature difference g contributes to selector interaction through $I = -\sum_j (w_{Uj} - w_{Tj}) g_j$ for fixed selection weights. We define the disagreeing subset on A before computing its g values and never filter it by the sign of g (Table 17).

Table 17: **Held-out consequences depend on assignment and evaluation domains.** Mean behavioral-minus-geometric F1 over eight original and control models and two fresh panels. Columns average all 4,096 features or each model's subset disagreeing on its corresponding A. These are descriptive model means. Independently trained models have only 3–9 disagreements and all-feature changes of at most approximately 0.0001.

Assignment A	Evaluation E	All features	Disagreeing subset
Original	Original	+0.0428	+0.1125
Original	Independent	+0.0004	+0.0012
Independent	Original	+0.0259	+0.1113
Independent	Independent	+0.0345	+0.1480

Keeping the original feature IDs and weights, we observe 0/32 AUROC and 6/32 native point reversals in original-trained models, compared with 10/32 and 26/32 in controls and 0/64 across both scores in independent models. Each set of 32 comparisons combines four models, two sample panels and four pairs of assignment and evaluation domains, yielding dependent descriptive outcomes that do not change the failed original dependence-study criterion.

Hierarchy groups explain cancellation in this panel. Our post-hoc relation classification identifies ancestor–descendant pairs in 12,364 of 12,396 original-domain disagreements and 7,604 of 7,631 independent-domain disagreements among original and control models (99.74% and 99.65%, respectively). The independently trained condition contains too few disagreements for comparable percentages. Original assignment labels evaluated on independent data have mean $g = -0.113$ for parent labels, with 92.9% negative, and positive means for child and deeper descendant labels. Averaging the census and both new panels equally, 53.9% return to the geometric label after an independent-domain refit, of which 97.8% have parent labels. Another 41.7% retain their behavioral labels and have positive means in all four domains. These groups account for the descriptive cancellation in Figure 4, but do not establish label equivalence or identify a mediator of selector reversal.

Factorial training separates two configuration components. We train sixteen models across two blocks and two replicates with hierarchy and correlation each present or absent, pairing initialization and using the same standard recipe. We decided to add the second eight-model replicate after seeing the first results, making that replication outcome-informed. The correlation-only condition rebuilds thresholds from empirical marginals, while hierarchy combines parent gating (a child requires its

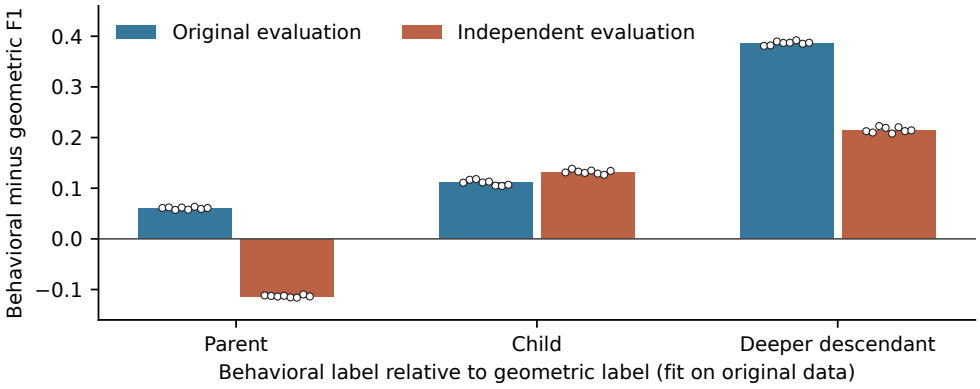


Figure 4: **Opposing label groups in the original in-house standard-SAE panel at target $L_0 = 30$.** Behavioral labels are fitted on original A. The geometric map and model remain fixed. Bars show mean held-out g within each relation category for the eight original and control models, with points marking each model's mean across two new panels. The negative parent-label contrast offsets positive child and descendant contrasts on independent E in these categories, which we selected for analysis after observing the outcomes. The points are model summaries, not independent confidence intervals. Rare remaining relations remain in the reproduction archive.

Table 18: **Factorial disagreement on common assignment data.** Four models per row, 4,096 determined features per model. Descriptive fractions. This is the sixteen-model in-house panel, separate from the earlier twelve models.

Training configuration	Original A	Independent A
Hierarchy + low-rank correlation	38.45%	23.40%
Hierarchy only	37.82%	23.78%
Low-rank correlation only	0.14%	0.13%
Neither	0.13%	0.12%

parent), sibling exclusion and probability compensation, producing configurations with differing dependence strengths.

Across four block and replicate units, the hierarchy effect on original assignment data ranges from +37.16 to +38.77 percentage points, compared with -0.49 to $+0.81$ for correlation, and remains +23.14 to +23.89 points on independent assignment data. VE on original assignment data is 0.786–0.787 with hierarchy versus 0.520–0.544 without it, accompanying these differences in labels. The block-1 marginal check occurred after training, and the new panels had already been examined in the earlier study, as recorded with all sixteen quality and count rows in Supplement S5. Table 18 measures total effects of the training configuration and measurement population in this experiment. It does not identify a unique mixing mechanism or rule out correlation effects in other generators.

E Persistence across sparsity and hierarchy shape

We test whether the difference between hierarchical and independent training persists when sparsity or hierarchy shape changes. These extensions are exploratory. Eight standard models cross target sparsities 15 and 45, two blocks and hierarchy-only or independent training, while eight more cross deeper and wider hierarchies with the same two conditions at target 30. We retain the 200M-example recipe and pair one new replicate per setting to factorial replicate 0, counting the reused target-30 reference only once. New hierarchy-shape assignment and evaluation data are shared within each shape and block, whereas sparsity studies reuse the earlier held-out panels.

We call a feature determined when it is active at least once and has positive best F1, unique maxima under both assignment rules and a valid decoder direction, and report d/n for d disagreements among n determined features. The all-feature bounds $[d/4096, (d + 4096 - n)/4096]$ allow the remaining features either label relationship, while relation shares condition on determined disagreements. For the lone undetermined sparsity feature, this gives $1960/4095$ and bounds $[1960/4096, 1961/4096]$, preserving its uncertainty in the all-feature result.

On original assignment data, hierarchy-minus-independent differences are 47.8–48.0 percentage points at target 15 and 67.0–67.5 at target 45. The deeper shape, with 256 roots, branching 2 and depth 4, yields 57.3–58.4 points, while the wider shape, with 128 roots, branching 8 and depth 2, yields 3.5–3.9 points. Activity, hierarchy membership, marginals and reconstruction also differ across conditions, so this is not an isolated depth or matched-quality effect. Supplement S5 reports the quality and marginal checks.

Table 19: Persistence across sparsity and hierarchy shape, with model quality on original-domain assignment data. We average the two block-specific model summaries within each setting and condition after computing disagreement per model, retaining bounds for the one undetermined target-15 hierarchy feature. Targets 15 and 45 change sparsity, while Deep and Wide change hierarchy shape at target 30, with the reported quality measures varying across conditions. No feature-level uncertainty is inferred.

Setting	Arm	Disagreement (%)	Achieved L_0	VE
15	Hierarchy	48.02	15.18	0.730
15	Independent	0.15	15.06	0.212
45	Hierarchy	67.40	45.45	0.797
45	Independent	0.15	47.45	0.667
Deep	Hierarchy	58.01	29.97	0.773
Deep	Independent	0.11	30.64	0.329
Wide	Hierarchy	3.85	29.92	0.711
Wide	Independent	0.15	30.54	0.585

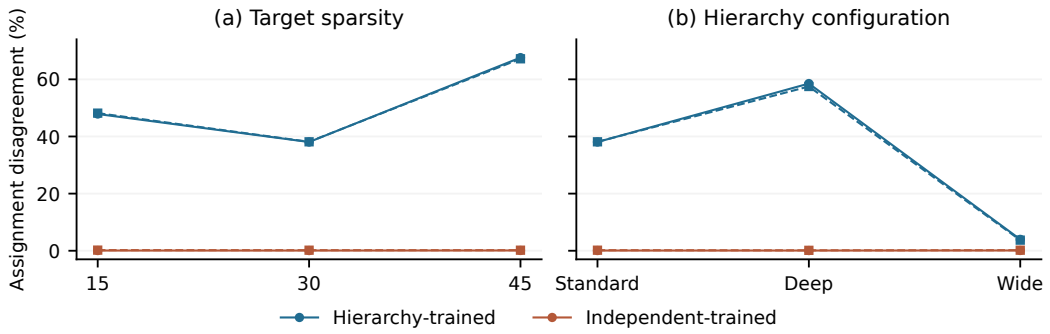


Figure 5: **The training contrast survives new sparsities and hierarchy configurations.** Higher means the behavioral and geometric labels disagree on more features. Disagreement on common original-domain assignment data, among determined features. Lines show the two paired blocks descriptively, reusing factorial replicate 0 for the original shape at target 30 and adding eight models per extension. In panel (b), "Standard" denotes the original hierarchy shape, not an SAE family. Low-rank correlation is off, while populations and concept rates differ across shapes (Appendix E).

In Figure 6, $g_{\text{orig,ind}}$ is behavioral-minus-geometric F1 on independent-domain E, averaged over features whose labels disagree when fitted on original-domain A; orig and ind denote these two domains (Appendix D). We also evaluate all 175 released checkpoints, including twelve calibration models and five checkpoints per family and sparsity cell, retaining bounds for Matryoshka's undetermined features. Figure 6 shows that JumpReLU disagreements are predominantly between different trees (87–94%), with only 4–12% joining ancestors and descendants. Released standard models at target 30 have $g_{\text{orig,ind}} = +0.052$, compared with the in-house near-zero result, while MP remains positive throughout the grid. Thus the parent-label cancellation observed in-house varies across released families, which lack hierarchy-off training counterparts for isolating architecture effects. Achieved sparsity and reconstruction quality also co-vary with family, so these observational differences do not isolate architecture. Supplement S5 retains all 35 rows of achieved sparsity, VE, undetermined counts and relation shares, and Table 19 summarizes extension quality.

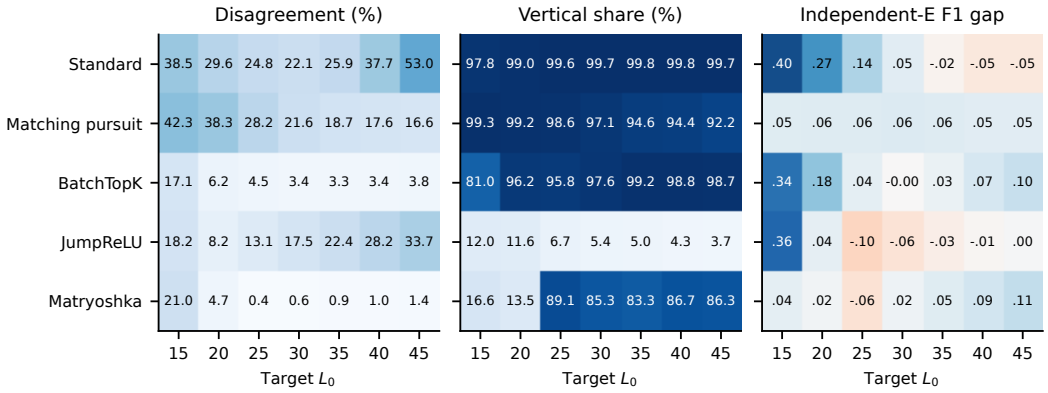


Figure 6: **Assignment disagreement and its consequences vary by family and sparsity.** Each cell averages five checkpoint summaries. Left: determined disagreement fraction on original A. Middle: "Vertical share" is the ancestor–descendant share among those disagreements. Right: behavioral-minus-geometric F1 on independent E, restricted to original-A disagreements. Positive gaps favor the behavioral label in these descriptive cell averages, whose horizontal axes show target sparsity and whose achieved sparsities and reconstruction quality are reported separately.

On independent assignment data, sparsity-study hierarchy contrasts are 47.8–48.1 percentage points at target 15 and 33.0–35.8 at target 45. For the fresh-panel disagreeing subset, $g_{\text{orig,ind}}$ is $+0.441$

and $+0.439$ in blocks 0 and 1 at target 15, compared with -0.032 and -0.029 in the same order at target 45, with the aggregate combining relation groups that can differ in sign. The deeper and wider shapes yield independent-domain contrasts of 51.1–52.3 and 6.6–6.9 points, respectively, while mean active-concept counts are 51.6 for the deeper shape, 30.7 for the wider shape and 34.5 for the original. Approximate marginal matching and these quality differences leave the effects of hierarchy depth entangled with other changes in the generator.

F NATURAL AUDIT AND SECONDARY DETECTORS

The direct natural audit favors top-50 under both assignments on one SAE and concept bank. The separate country-detector comparison remains unresolved, and the first-letter detection and intervention criteria are unmet.

Model, bank and selected outcomes. We extract Gemma-2-2B activations with the Gemma Scope SAE at these revisions:

Model: `c5ebcd40d208330abc697524c919956e692655cf`.
SAE: `fd571b47c1c64851e9b1989792367b9babb4af63`.
Path: `layer_12/width_16k/average_l0_82/params.npz`.

The checkpoint name labels the release variant; it is not a measurement of achieved sparsity on these city data. We encode $z = (hW_{\text{enc}} + b_{\text{enc}})\,1[hW_{\text{enc}} + b_{\text{enc}} > \theta]$ without subtracting decoder bias, where θ is the per-latent JumpReLU threshold and 1 is an indicator. Model inference uses BF16 and SDPA, and SAE arithmetic uses float32 with TF32 off, reading the final city-span token before answer generation.

We assign disjoint city sets to development (964), scoring (778), assignment (515), refitting (524) and evaluation (771), with eight scoring contexts giving 6,224 rows. We draw 256 of the 796 features that meet the eligibility requirements of at least 100 positive rows and active fraction at most 0.25, then compare native top-50 scores with uniform-50 scores at budget 16 using exact cutoff ties. Top-50 averages the largest 50 positive activations, while we average uniform's selected outcomes over 32 separately ranked draws.

The frozen bank contains Africa, Asia, Europe, North America, South America, China, Russia and the US, with overlapping labels permitted. We fit both assignment rules on each assignment sample, maximizing detection F1 for behavioral assignment and signed cosine to normalized positive-minus-negative mean residual directions for geometric assignment. We use signed cosine because each proxy is a positive-minus-negative direction. These residual directions serve as concept proxies rather than known synthetic factors, with ties resolved by concept order and the resulting maps denoted A_{fit} and A_{refit} for the first and second samples.

Table 20: Direct natural study: absolute selection-weighted F1. All-positive mean ranks candidates by mean encoder activation over all positive scoring examples, then evaluates the selected features at the same budget and with the same tie rule.

Map	Assignment	Top	Uniform	Random	All-positive mean
A_{fit}	behavioral	0.2969	0.2695	0.1257	0.2805
A_{fit}	geometric	0.2775	0.2366	0.1130	0.2509
A_{refit}	behavioral	0.2975	0.2649	0.1201	0.2782
A_{refit}	geometric	0.2844	0.2459	0.1161	0.2591

Supported same-direction result and assignment uncertainty. We estimate uncertainty with twenty thousand paired country-stratified entity resamples that preserve observed country counts, obtaining four Bonferroni 98.75% percentile bands conditional on the maps and fixed setting. All four favor top in both the primary and predefined second bootstrap-seed run (Table 5), with absolute F1 in Table 20 and AUROC contrasts in Table 21. Cross-sample selection-weighted label agreement measures the weight on features retaining their labels across fits: top and uniform yield 0.7500 and 0.7031 for behavioral assignment and 0.8125 and 0.7148 for geometric assignment.

Bootstrap stability instead measures the selection weight retaining its point-map label when we resample A and refit assignments. For uniform selection, behavioral and geometric stability are 0.799 and 0.849 on the first sample and 0.788 and 0.846 on the refit, so the earlier requirement that all values exceed 0.80 remains unmet. We report assignment uncertainty under the revised protocol, with bands from resampling assignment data alone or jointly with evaluation data remaining negative conditional on the concept bank. Native paired interactions are -0.0134 $[-0.0293, 0.0023]$ and -0.0059 $[-0.0162, 0.0045]$ under post-hoc 95% bands, alongside candidate-map disagreement of 62.9% and 59.4% in the two samples. Thus top preference does not imply identical assignments.

Post-hoc AUROC rescoring. After inspecting E, we rescore the fixed candidates while retaining the bank, maps, budget and entities, using raw readout $x^\top d_j/\|d_j\|$ without absolute value, residual normalization or bias subtraction. The top and uniform scorers retain their respective sets of 50 positive rows and share 4,000 inactive references per candidate, with exact integer AUROC midranks.

Table 21: Post-hoc natural AUROC rescoring. Conditional bands use the same 20,000 paired E resamples, correcting four contrasts within this scorer. The analysis is exploratory, with native-score rows reported in Table 5.

Score	Map	Assignment	Top	Uniform	Δ	98.75% band
AUROC	A_{fit}	behavioral	0.2146	0.1532	-0.0613	$[-0.0778, -0.0455]$
AUROC	A_{fit}	geometric	0.1893	0.1405	-0.0488	$[-0.0642, -0.0342]$
AUROC	A_{refit}	behavioral	0.2074	0.1495	-0.0578	$[-0.0739, -0.0424]$
AUROC	A_{refit}	geometric	0.1987	0.1495	-0.0492	$[-0.0656, -0.0341]$

All four AUROC bands favor top despite positive paired interactions of $+0.0125$ $[0.0018, 0.0231]$ and $+0.0086$ $[0.0004, 0.0164]$, showing that a changed interaction can accompany an unchanged selector preference. Selection-weight overlap $\sum_j \min(w_{Tj}, w_{Uj})$ falls from 0.6191 under native scores to 0.1209 under AUROC, while overlap between scorers is 0.1948 for top and 0.1230 for uniform. Candidate random F1 remains 0.1257 and 0.1130 under behavioral and geometric assignment on the first sample, and 0.1201 and 0.1161 on the refit, with both AUROC selectors descriptively exceeding these baselines. This rescoring uses the existing inference outputs and concept bank.

Chronology and scope. The study covers one SAE, site, prompt family and bank, with supported top preference under both assignments. Supplement S8 records the code-freeze chronology, the analysis-code hash check added after outcomes, and reporting corrections.

Separate country detector comparison. For twelve-country detection on 329 RAVEL test cities, we use eight positive and 32 negative development labels to select 32 candidates per country by label-based screening, then rank those candidates using 288 activations without labels. An OR-of-two detector predicts a concept when either selected feature is active, allowing us to compare top-eight and uniform-eight scores based on the eight largest or eight randomly sampled active examples. Their matched-fraction counterparts use $q = \min(1, 8/37)$ of each feature's positive examples, where 37 is the median number of active examples.

We adapt the Tversky Monosemanticity Score (TMS; Filus & Pokuciński, 2026) by averaging Jaccard overlap of thresholded active-feature sets across example pairs, while input coherence averages activation-weighted cosine similarity between distinct normalized input vectors. Task mean difference ranks mean positive-minus-negative activation, and task screening F1 ranks detection F1 on the same 40 labels, alongside M from Eq. 4 and T_{all} using all active and inactive examples. Because OR is nonlinear, we average whole tied sets when evaluating detectors (Table 22). The primary uniform-eight-minus-top-eight contrast of $+0.062$ $[-0.008, 0.155]$ fails its criterion, and the matched-fraction contrast is $+0.121$ $[0.030, 0.235]$, but both matched-fraction selectors score below M and input coherence.

We fit logistic coefficients on the original selected IDs using the same 40 screening labels, training-only standardization, balanced weights, $C = 1$ and threshold 0.5. The dense probe uses all residual dimensions, while original top-eight uses ID tie breaking, explaining its different OR score. We treat small differences among probe scores descriptively, retain cities across alternate prompts and report the complete protocols in Supplement S6.

Table 22: **Natural selection depends on the comparator and detector.** Country macro F1, two-feature budget. Left: tie-averaged sampling, with 32 uniform draws. Right: probes fitted on the same original frozen IDs as their OR comparator. Original top-eight used an ID tie break, explaining its different OR value. Every probe uses the same 40 screening labels. The dense probe uses all residual dimensions, giving it greater input capacity than the two-feature methods.

Sampling / OR	F1	Frozen selection	OR F1	Probe F1
Top-eight	0.227	Top-eight	0.218	0.223
Uniform-eight	0.288	T_{all}	0.393	0.455
Top-q	0.235	M	0.416	0.488
Uniform-q	0.357	Adapted TMS	0.430	0.494
M	0.416	Input coherence	0.573	0.699
Input coherence	0.573	Task mean difference	0.343	0.704
Random	0.230	Task screening F1	0.561	0.697
Dense residual probe, same 40 labels				0.726

The separate first-letter detection study asks whether selected features' activity identifies a word's initial letter across 26 letters and 624 held-out words (24 per letter). It fails the frozen criterion requiring a 0.05 F1 margin over the strongest cheap comparator, at most 0.01 precision loss and a positive paired-letter interval. In a twelve-letter causal pilot, we ablate selected features to test first-letter prediction, choosing intervention strength on reserved data by a required ten-point accuracy reduction. The primary all-active direct-AUC score, based on target-active versus control-active decoder readouts, achieves the required reduction for none of the target letters, while secondary T_{all} reduces target accuracy by 1.5 points versus 7.3 for task mean difference. These intervention results concern the first-letter task, separately from the country-detection comparisons above. Detection and fitted probes do not establish steering success.

G Worked example and descriptive scale

G.1 Reproducible selection and verification of latent 341

We fixed the standard family, central target sparsity $L_0 = 30$ and seed 0, then took the smallest latent ID with nonzero selection weight for either top-480 or uniform-480 at budget 16 whose geometrically assigned concept is an ancestor of its behavioral assignment. This post-hoc rule selects latent 341 from 15 qualifying latents to illustrate the relation, using ID order rather than effect size and carrying no claim of random or representative sampling.

The generator path is $97 \to 519 \to 2207$, and latent 341 is active on 1,092 examples in the 240,000-example original evaluation population. Concept 2207 occurs 618 times, with the latent active on 617 of those examples, whereas concept 97 occurs 9,193 times, with the latent active on 977 of those examples, yielding behavioral precision, recall and F1 of 0.565018, 0.998382 and 0.721637 and geometric values of 0.894689, 0.106277 and 0.189985. With the activation threshold unchanged, uniform selection gives the latent weight $1/16$ in all 32 draws and top selection gives it zero weight, giving this latent a contribution of $(1/16)(0.189985 - 0.721637) = -0.033228$ to its checkpoint interaction.

We checked the pinned decoder and generator vectors directly, obtaining signed cosine $+0.579680$ for concept 97 and $+0.558422$ for concept 2207, with concept 97 the global absolute-cosine maximizer. Both cosines are positive and the geometric preference is modest, while independent assignment refits with seeds 919202 and 919203 both retain concept 2207 with assignment-sample F1 of 0.724858 and 0.718710. These refits assess sample stability of the synthetic-factor labels on the fixed trained model.

G.2 MAGNITUDE ESTIMATOR AND ITS CONDITIONING

Let $d_{cs,\ell}$ be the uniform-minus-top F1 contrast at checkpoint seed s in family and sparsity cell c, under assignment $\ell \in \{A, D\}$. With n_c available checkpoints in that cell, define

$$a_c = \frac{1}{n_c} \sum_s |d_{cs,D} - d_{cs,A}|, \tag{7}$$

$$b_{c,\ell} = \binom{n_c}{2}^{-1} \sum_{s<t} |d_{cs,\ell} - d_{ct,\ell}|, \qquad b_c = \tfrac{1}{2}(b_{c,A} + b_{c,D}). \tag{8}$$

We compute $\overline{a} = \mathrm{mean}_c\, a_c$ and $\overline{b} = \mathrm{mean}_c\, b_c$, then take the panel ratio $\overline{a}/\overline{b}$ so that cells receive equal weight before division. Seed changes compare distinct checkpoints, with the fresh panel containing 35 cells and 163 checkpoints: six cells retain three seeds after calibration exclusions and the others retain five. Checkpoint pairs contribute to the descriptive average without being treated as independent inference units.

Across all cells, $100\overline{a} = 4.664$ and $100\overline{b} = 2.909$, giving $1.603\times$. The assignment change a_c exceeds b_c in 19/35 cells and both $b_{c,A}, b_{c,D}$ in 17/35. For the fresh evaluation of the original primary 105-model subset, the values are 6.530 and 3.808 points ($1.715\times$). This descriptive reanalysis does not replace the failed original primary sampling test. For the fresh all-panel comparison, the separate behavioral and geometric denominator values are 2.323 and 3.495 points, and the ratios using the pooled denominator are 1.59, 1.60 and 1.69 at budgets 8, 16 and 32. The earlier held-out evaluation of the same checkpoints gives 1.69 at budget 16.

Sensitivity and uncertainty scope. The root-mean-square ratio is 1.64, and pairwise squared checkpoint differences equal twice the sample variance within each fixed assignment. Leaving out each seed label globally yields a descriptive range of 1.53–1.73 for the fresh all-panel ratio of 1.603. Inference over possible targets would require a sampling model beyond the two fixed assignment rules. Our seed comparison measures variation in the complete checkpoint pipeline, including candidates, references and scoring draws, with shared training-stream coupling unverified. Regenerating that pipeline for the same checkpoint changes the contrast by 2.938 points on average across targets and cells, giving two observed realizations from which to describe sensitivity, but too few to decompose its variance.

H REPRODUCTION AND RECORD INDEX

Download and extract `supplementary_code.zip`. The saved Gemma audit arrays are included, so no separate cache download is needed for these checks. With Python 3.10+, run from the extracted `supplementary_code` folder:

```
python -m pip install numpy scipy
python -B check_results/check_reported_numbers.py
python -B check_results/fixed_selection_audit/run_all.py \
    --output check_results/output/audit
```

The first command reconstructs scale estimates, worked counts, common-seed bounds, natural contrasts and generator-block summaries from saved outputs, while the second exercises the fixed-selection audit on the complete fresh standard and MP panels and both natural scorers. These commands replay saved metrics and weights; they do not retrain models or rerun neural inference. The archive README and Supplement S8 specify their scope. Supplements S1–S3 hold selection, diagnostic and extension records, and S4–S6 hold records of retraining, dependence interventions, persistence across sparsity and hierarchy shape, and natural studies. The reproduction archive retains input hashes and the evidence inventory.

I CLOSEST PRIOR-WORK DISTINCTIONS

Top-example inspection uses native activation means (Sharma, 2026), while sampling affects interpretability metrics (Oikarinen et al., 2025; Paulo & Belrose, 2026) and feature rankings (van der

Ben et al., 2026). We evaluate decoder-readout AUROC and native activation scores separately to distinguish these scoring definitions.

Evaluation choices also matter outside SAE research: disentanglement metrics assess different properties (Locatello et al., 2020), aggregation changes which NLP system ranks first (Peyrard et al., 2021), and a combined change to matching, thresholding and normalization changes instance-segmentation winners (Stillwagon et al., 2026). We instead hold the metric, aggregation, selections and binary activity fixed while changing assigned targets. Our random baseline selects within a trained SAE, differing from comparisons against random directions or activations (Korznikov et al., 2026), and our fixed-selection comparison differs from supervised steering selection (Jørgensen & Hansen, 2026). Supplement S2 defines the adapted ranking baselines used here.

Individually unstable features can reproduce at the subspace level (Gerasimov et al., 2026). Our refits hold the basis fixed and find that stable labels carry more than 99% of standard's interaction, locating this result within a trained basis while leaving ambiguity across training runs open. Prior work on hierarchy and correlated-feature mixing (Chanin et al., 2025b;a) already motivates this setting. We use the factorial to separate training-configuration effects from those of the population used to fit labels, with extended source-specific discussion in Supplement S7 and readout diagnostics in Supplement S2.3.