A REPORTING STANDARD FOR SPARSE AUTOENCODER ABLATION SCORES

Anonymous authors
Paper under double-blind review

ABSTRACT

Ablation scores for sparse autoencoder latents disagree across dictionaries partly because each dictionary chooses the token at which its own latent is scored. A latent is scored by switching it off at one token and measuring how much the model's output changes, and by convention that token is where the dictionary's latent fires hardest. We isolate the effect with autoencoders trained from a shared initialisation, each differing in a single fitting choice. Across seven base models from four families, the choice of token carries at least 46.6% of the variance in the measured effect, about as much as the choice of latent. We propose a reporting standard: record the measurement position, and when dictionaries are compared, fix it outside all of them. Fixing the tokens makes dictionaries rank latents more consistently, with a bootstrap interval excluding zero on every model at the main corpus size. Both items take a line of evaluation code, and the second needs no more forward passes than letting each dictionary choose. Left to choose, dictionaries rarely agree on where to measure a latent, and agree less as the evaluation corpus grows. Released Qwen-Scope and Gemma Scope dictionaries measure most latents matched across a pair of dictionaries at different tokens, including many that the pair encodes almost identically.

Code: https://anonymous.4open.science/r/sae-artifact-8E6D/

1 INTRODUCTION

Sparse autoencoders (SAEs) turn a language model's activations into a dictionary of named parts, called latents (Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024). To show that a part matters, a practitioner switches a latent off, runs the model again, and reads how much the output changed (Marks et al., 2025; Gao et al., 2025; Cho et al., 2026). The number is reported as a score for the latent and set beside the scores other papers report for other dictionaries.

There is a step between the latent and the score. A latent fires at many tokens across many sequences, and the intervention happens at one of them. The near-universal choice is the token where the latent fires hardest, an instinct inherited from single-neuron neuroscience. That token is computed from the dictionary's own activations, so a second autoencoder fit on the same data places the same latent's maximum somewhere else (Figure 1). The dictionary under comparison thus picks its own measurement site, and a difference in site is read as a difference between dictionaries.

Four terms carry the rest of the paper. An arm is one of six dictionaries trained on one base model from a shared initialisation, each differing from the others in a single fitting choice. A per-arm position is a token an arm selects for a latent from its own activations, as published practice does. A shared position is chosen for each latent outside every arm, and every arm is measured there. The latent-by-arm interaction is the part of the score's variance that depends on the pairing of latent and arm, and a comparison of dictionaries reads it as a difference between them.

That the top-activating token is one of many is well known, and the choice is usually expected to average out across latents. It does not, for three reasons absent from prior work. The disagreement between arms is systematic, so it does not cancel across latents. It belongs to the dictionary, so it loads onto exactly the comparison a reader makes when two dictionaries are scored against each other. And it grows with the evaluation corpus, so more text makes it worse. Only one of a sample of five papers we audit records which token its scores were taken at (§7).

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.

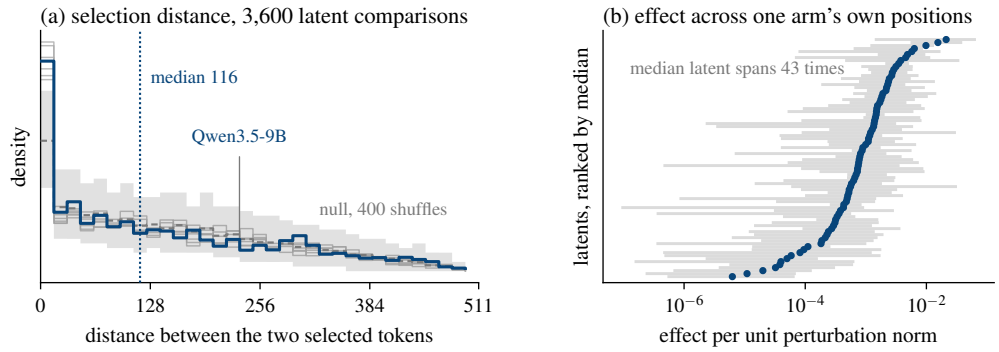


Figure 1: (a) Distance between the tokens two arms select for the same latent, 3,600 latent comparisons over all 15 arm pairs in blue, other models in grey, over a null drawing each arm's choice from the positions that latent fires at, shown as its band from the second to the ninety-eighth percentile over 400 shuffles with its mean dashed. (b) Each latent's smallest and largest effect across the six positions the decoder-free arm selects for itself, dot at the median, latents ranked by that median, so no dictionary difference is in play and the spread is position alone. Qwen3.5-9B, 384 sequences.

Result. We train six arms on each of seven base models, Qwen3.5-9B, Qwen3.5-4B, Qwen3.5-2B, SmolLM3-3B, OLMo-2-1B, Gemma-3-1B and Gemma-2-2B, and measure every arm first at per-arm and then at shared positions. At shared positions nothing but position varies inside a cell, and position carries 46.6% to 62.8% of the variance in the measured effect (Table 1). Fixing the position removes most of the latent-by-arm interaction and raises the reliability of the latent ranking across arms by $+0.130$ to $+0.259$ on every model. In practice, a comparison of dictionaries reads the dictionaries themselves only once all of them are measured at the same tokens. Box 1 is the standard, with what skipping each item costs.

Box 1. Reporting checklist for an ablation score.

- ☐ Record the measurement position, meaning the rule that chose it and the number of positions per latent, at least 8 for $E\rho^2 \geq 0.8$ on every model. Cost of skipping it: position leaves 46.6% to 62.8% of the variance in the score unspecified (§6).
- ☐ When dictionaries are compared, fix the positions outside all of them and measure every dictionary at the same tokens. Cost of skipping it: the reliability $E\rho^2$ (2) of the latent ranking across dictionaries forgoes the gain of $+0.130$ to $+0.259$ that fixing the positions brings, with an interval excluding zero on every model (§6). A latent with no token at which every dictionary fires keeps each dictionary's score at the item 1 positions, flagged as not comparable across dictionaries, beside the share of latents that have such a token.

Both items take one line of evaluation code and no retraining. A paper that kept its activation caches can meet item 1 retrospectively. Item 2 needs ablation passes at the shared positions, as many as a per-arm evaluation.

Contributions.

- The two-item reporting standard of Box 1 and a recommendation on normalisation stated with the single comparison behind it (§4). Prior work on evaluation reliability samples seeds and data splits (Bouthillier et al., 2021; Dodge et al., 2019), and for autointerpretability scores the evaluator's pipeline choices (van der Ben et al., 2026). The facet here is the measurement position, which the system under evaluation selects.
- The measurements behind both items. Released Qwen-Scope and Gemma Scope dictionaries select different tokens for a large share of the latents they encode almost identically (§5). A controlled six-arm design on seven base models shows that holding the position fixed removes most of the variance a comparison would attribute to the dictionaries (§6). Bolukbasi et al. (2021) show that reading a unit off its maximally activating inputs can

manufacture an interpretation. We show the same selection moves a score, and by how much.

- What the standard adds to five published measurements, with the size of each unreported latitude taken from the decomposition above (§7).

2 Related work

Dictionaries. Sparse autoencoders decompose a residual stream into an overcomplete set of sparsely active directions (Bricken et al., 2023; Cunningham et al., 2024), motivated by superposition (Elhage et al., 2022), with the linear representation view refined since (Park et al., 2024; Engels et al., 2025). The line has scaled to production models (Templeton et al., 2024), to TopK and JumpReLU objectives (Gao et al., 2025; Rajamanoharan et al., 2024b), to gated and batched variants (Rajamanoharan et al., 2024a; Bussmann et al., 2024), to objectives that select for downstream effect (Braun et al., 2024), to transcoders (Kissane et al., 2024; Dunefsky et al., 2024), and to public dictionary suites (Lieberum et al., 2024).

Evaluating dictionaries. SAEBench (Karvonen et al., 2025) consolidated the metrics the field reports, and a run of results has since asked what they discriminate. Paulo & Belrose (2025) show that seed alone substantially changes which features are learned, TopK more than ReLU, and Song et al. (2026) propose a pairwise dictionary correlation to measure that consistency across runs. Heap et al. (2025) find automated interpretability scores do not separate trained transformers from random ones, and descriptions built from a feature's activating inputs miss its causal effect on outputs (Gur-Arieh et al., 2025). Korznikov et al. (2026) show that random baselines match fully trained SAEs on interpretability, sparse probing and causal editing, and Chanin (2026) finds that SAEBench's TPP and SCR metrics fail more than one of three reliability checks at their canonical settings. Evaluations supplying external ground truth are the partial exception (Karvonen et al., 2024; Chaudhary & Geiger, 2024; Kantamneni et al., 2025), and on AxBench SAEs are not competitive with simple baselines for steering or concept detection (Wu et al., 2025). A second line reports the units themselves unstable under width and sparsity (Leask et al., 2025; Chanin et al., 2025). The position confound is orthogonal to both lines, since it survives holding the latent identity fixed by construction (§6). For autointerpretability scores, van der Ben et al. (2026) find that variance from the evaluation corpus, the draw of activating examples, the explainer model and the wording of the explanation collectively exceeds variance from the SAE architecture, and propose a variance decomposition and a minimum reporting checklist. We decompose the variance of ablation scores, along a facet chosen by the dictionary under evaluation rather than by the evaluator.

Ablation as a causal readout. Zeroing a latent and reading a magnitude is the standard single-latent causal probe (Marks et al., 2025; Gao et al., 2025; Cho et al., 2026), distinct from interchange interventions scored by match rate (Huang et al., 2024; Makelov et al., 2024), a separate readout family outside this standard (§9). It inherits its logic from activation patching (Meng et al., 2022; Wang et al., 2023), later automated (Conmy et al., 2023) and approximated by gradients for scale (Syed et al., 2024; Kramár et al., 2024). That literature established that the patching metric and the corrupted baseline change the conclusion (Zhang & Nanda, 2024). What it does not supply is the facet one level down, which position the intervention is applied at, and that facet differs from the others in being chosen by the dictionary. Downstream uses inherit whatever the measurement selects (Chalnev et al., 2024; Farrell et al., 2024). Cho et al. (2026) is the closest published methodology, a single-latent ablation with a KL readout related to depth.

Measurement reliability. How much of an observed difference is the object and how much is the measurement procedure is the question of generalizability theory (Cronbach et al., 1972; Brennan, 2001; Shavelson & Webb, 1991). The failure at issue is the one Clark (1973) named, treating a sampled facet as fixed, which psycholinguistics answered with crossed random-effects designs (Baayen et al., 2008; Judd et al., 2012). Work on benchmark variance in machine learning samples seeds and splits (Bouthillier et al., 2021; Dodge et al., 2019; Henderson et al., 2018; Card et al., 2020). The measurement position is a facet it does not sample.

3 SETUP

We measure the causal effect of a single SAE latent by ablating its contribution to the residual stream and reading the change in the model's next-token distribution. For latent f with decoder direction d_f and activation a_f at position t, we replace $h_t \leftarrow h_t - a_f d_f$ at the SAE's hook point and record

$$\mathrm{KL}_t = D_{\mathrm{KL}}(p(\cdot \mid h) \,\|\, p(\cdot \mid h')) \,, \tag{1}$$

evaluated at position t. Group comparisons use the Hodges–Lehmann estimator, the median of pairwise ratios, with a bootstrap interval.

Base models. The design runs on seven base models from four families, newest first: Qwen3.5-9B-Base, Qwen3.5-4B-Base and Qwen3.5-2B-Base (2026) (Qwen Team, 2026), SmolLM3-3B-Base (July 2025) (Bakouch et al., 2025), OLMo-2-0425-1B (April 2025) (Team OLMo, 2025), Gemma-3-1B (March 2025) (Gemma Team, 2025) and Gemma-2-2B (June 2024) (Gemma Team, 2024). Each dictionary sits at the block nearest 46% relative depth. That is layer 11 of 24 on Qwen3.5-2B, 17 of 36 on SmolLM3, 7 of 16 on OLMo-2 and 12 of 26 on both Gemma models. The larger Qwen3.5 rungs use layer 15 of 32 and extend one family upward, so that a trend in scale is separable from a family effect. The body quotes Qwen3.5-9B as its worked example and summarises the other models by the range across all seven. The two Gemma models are kept because released dictionaries exist for both, Gemma Scope (Lieberum et al., 2024) for Gemma-2-2B and Gemma Scope 2 (McDougall et al., 2025) for Gemma-3-1B. The released-dictionary analysis uses Qwen-Scope (Deng et al., 2026) on Qwen3.5 and Gemma Scope on Gemma-2-2B. Qwen3.5 is a hybrid stack in which a quarter of the blocks carry full attention, which does not change an autoencoder fit on the residual stream.

Token policy. Position 0 is dropped on every model before any statistic. The Qwen3.5, SmolLM3 and OLMo-2 tokenizers prepend no BOS token, so there the dropped position is the first real token of each sequence, removed to keep one token policy with Gemma. On Gemma position 0 holds the BOS token, whose residual the dictionary reconstructs poorly.

Choosing shared positions. The per-arm rule takes each arm's own top-activating tokens. The shared rule takes one arm-symmetric set per latent, ranking candidates by the minimum gated activation across arms. The latent then fires in every dictionary at every selected position, and no dictionary serves as the reference. Appendix C gives a second rule built not to select on inter-arm agreement.

Shared initialisation of the arms. A latent-wise comparison requires that latent i denote the same thing in every dictionary. SAE training does not provide that, since Paulo & Belrose (2025) find that autoencoders differing only in random seed share roughly 30% of their features, with TopK architectures, which we use, more seed dependent than ReLU ones.

Reliability. We treat latents and arms as crossed random factors. Writing v_a for the variance between latents, v_{ab} for the latent-by-arm interaction, and v_e for the variance within a cell, the coefficient for a latent measured at n positions (Cronbach et al., 1972; Brennan, 2001) is

$$E\rho^2 \;=\; \frac{v_a}{v_a + v_{ab} + v_e/n}, \tag{2}$$

It is the expected squared correlation between a ranking of latents taken on one set of arms and one taken on another. The interaction v_{ab} sits in its denominator, so a convention that inflates v_{ab} depresses $E\rho^2$ arithmetically rather than as a separate finding. Components are estimated by method of moments and clipped at zero. A clipped point value marks the boundary rather than measuring the component, so there we report the 97.5th percentile of its latent-level bootstrap as its bound. Put simply, the more a score depends on which dictionary measured a latent, the less a ranking of latents carries over from one dictionary to another.

Scale and parameters. Six TopK SAEs per base model share initialisation seed 0, width $16{,}384$ and $k = 82$. Each is then trained with one different, individually defensible choice: decoder free, soft frozen at $\tau = 0.80$, soft frozen at $\tau = 0.90$, learning rate $10\times$ lower, sparsity $k = 41$, or a reshuffled corpus order. Every arm sees the same 12M tokens of WikiText-103 (Merity et al., 2017), giving 42

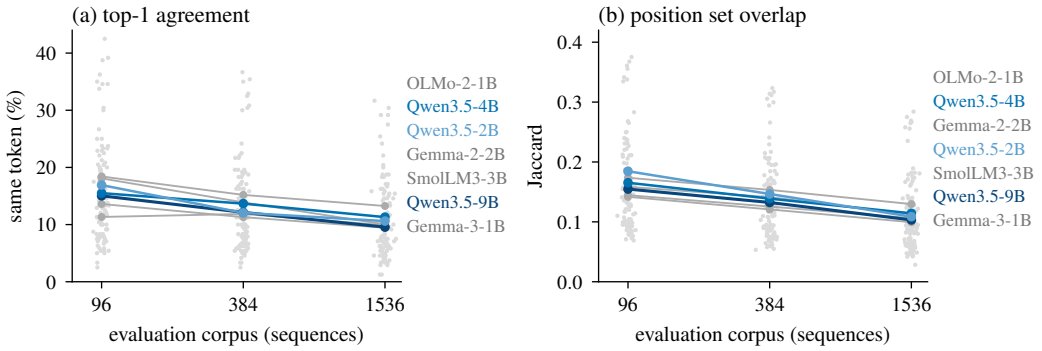


Figure 2: (a) Share of latents for which two arms select the same single top-activating token, against evaluation corpus size, mean over the 15 arm pairs for each base model, Qwen3.5 in blue, others in grey, every pair of every model as a faint dot. (b) Jaccard overlap of the six position sets, same axes.

trained dictionaries. Evaluation runs on WikiText-2 raw at 96, 384 and 1536 sequences of 512 tokens. The rungs are nested, each smaller one the first sequences of the larger, and no sequence repeats within a rung. Live latents are intersected across the six arms, and a uniform sample of 240 is drawn from the intersection through a fixed permutation of the full identifier space (Appendix H). Every arm is measured on exactly that sample at up to six positions per latent, a crossed latent-by-arm design with replication on both factors. Released-dictionary comparisons use the Qwen-Scope sparsity pair at layer 11 of Qwen3.5-2B and 15 of Qwen3.5-9B, and six Gemma Scope dictionaries at layer 12 of Gemma-2-2B, three widths and two sparsity targets, all 15 pairs.

4 The reporting standard

An ablation score is comparable across papers when a reader can tell that two scores were taken the same way. Box 1 states the two items that make this possible, and the crossed design of §3 measures what each is worth. Item 1 rests on the within-cell share of variance at shared positions, which is estimated away from any boundary on every model (§6). Item 2 rests on arms left to choose agreeing on a latent's top token for only 11.3% to 15.2% of latents (§5), and on the gain in $E\rho^2$ once the token is fixed (§6). Item 2 covers the latents that fire in every dictionary at a common token, 22% to 50% of those sampled here. The others are reported per dictionary at the positions item 1 records, flagged as not comparable across dictionaries, together with the share of latents that item 2 covers.

A recommendation on normalisation. Ablation perturbs the residual stream by $a_f d_f$, and activation magnitude covaries with firing frequency. The median perturbation norm runs 3.60 in the rarest decile against 14.76 in the most frequent on Gemma-2-2B, so raw effect sizes compare interventions of unequal size. From one comparison on one model, we recommend reporting the effect per unit perturbation norm beside the raw effect. The comparison behind it asks whether a Gemma Scope dictionary's rarest activation-frequency decile carries more causal mass than a frequency-matched decile of a tied random dictionary. Raw KL says less (HL $= 0.683$, 95% CI $[0.459, 0.961]$) and KL per unit norm says more (HL $= 1.687$, $[1.171, 2.498]$), on identical latents and an identical control (Appendix A). The choice can therefore decide a sign, and normalisation sits beside the standard as a recommendation. For a practitioner, meeting the standard is a small change to evaluation code that lets another lab take the same score the same way.

5 Measurement position selected by the dictionary

Item 2 rests on a premise that needs no training of ours: two dictionaries fitted on the same data select different tokens for the same latent.

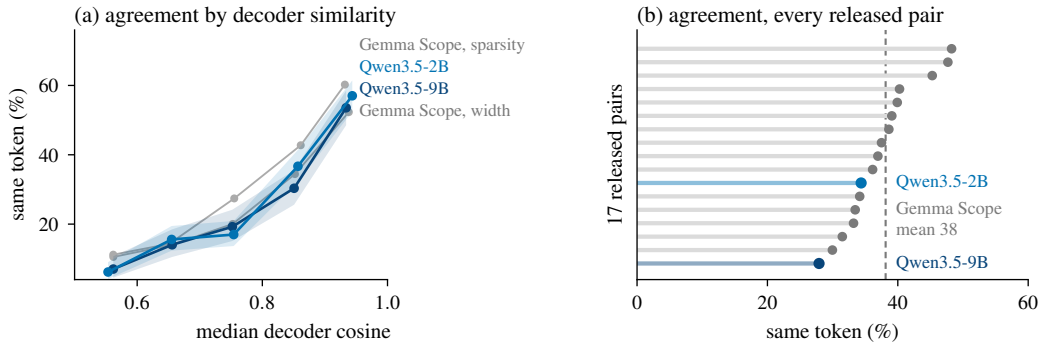


Figure 3: Released dictionaries matched by mutual-nearest-neighbour decoder cosine, Qwen-Scope on Qwen3.5-2B and 9B in blue, Gemma Scope on Gemma-2-2B (layer 12) in grey. (a) Same-token share of matched latents in five disjoint cosine bands, with Wilson intervals. (b) The same share for the 15 Gemma Scope pairs, ranked, mean dashed, and each Qwen-Scope pair band-weighted.

Firing positions and agreement between arms. On the 384-sequence corpus a released Gemma Scope dictionary's latents are live at a median of 103 positions each, so the convention selects one position among dozens to hundreds. The right comparison is therefore a shuffle null, which draws each arm's choice at random from the positions that latent actually fires at (Figure 1a). On Qwen3.5-9B that gives 3.9% against the observed 12.1%, so the arms agree about three times as often as chance and are not picking arbitrary tokens. At 384 sequences the arms have the same single top-activating position for 11.3% to 15.2% of latents across the seven base models, and their six position sets overlap at Jaccard 0.121 to 0.153.

The premise at production scale. Qwen's released Qwen3.5 dictionaries (Deng et al., 2026) supply only a sparsity pair, $k = 100$ against 50 at equal width. That pair agrees for 34% of randomly drawn matched pairs on Qwen3.5-2B and 28% on Qwen3.5-9B (Appendix B). Across the 15 pairs of released Gemma Scope dictionaries, agreement over each pair's whole matched population runs 30% to 48% (Figure 3b). Even the best-agreeing pair of released dictionaries measures a latent at a different token about half the time. Agreement rises with decoder similarity, from about one in ten in the lowest cosine band to 60.2% for the pair differing in sparsity and 52.3% for the pair differing in width above 0.90. Even where two released dictionaries encode a latent almost identically, they measure it at different tokens a large fraction of the time. Weighting each band by its share of matched pairs gives 42% and 35% respectively for a randomly drawn pair (Appendix B). The Qwen3.5 pair therefore agrees no more often than the Gemma Scope pair that also differs in sparsity, so the premise holds on the production dictionaries of both labs.

Agreement against evaluation corpus size. Across three corpora spanning a sixteenfold range, top-position agreement is lower at 1536 sequences than at 96 on every model (Figure 2), falling monotonically on all but SmolLM3. Jaccard falls in step on every model. More text means more places for two dictionaries to disagree about, so the usual expectation that more evaluation data stabilises a measurement runs the wrong way.

Consequence for the design. Because each arm selects its own positions, position is nested within latent and arm rather than crossed with latent. The latent-by-arm interaction then mixes the latent behaving differently under a different fit with the arms having chosen different tokens. Only 49 of 5,436 (latent, position) cells are common to all six arms in a per-arm measurement on Gemma-2-2B. The confound therefore cannot be resolved by subsetting and requires re-measuring under the shared mode of §3. Appendix E gives the per-arm decomposition itself, on an earlier Gemma-2-2B sample at 96 sequences.

			latent×arm (%)		$E\rho^2$	paired gain	position
model	year	latents	per-arm	shared (bound)	per-arm→shared	[95% CI]	(%)
Qwen3.5-9B	2026	115	9.5	0.0 ($\leq$0.2)	0.628→0.802	+0.174 [+0.080, +0.288]	57.9
Qwen3.5-4B	2026	120	8.6	0.0 ($\leq$0.0)	0.610→0.767	+0.156 [+0.075, +0.235]	62.8
Qwen3.5-2B	2026	106	11.9	0.0 ($\leq$1.6)	0.604→0.808	+0.201 [+0.128, +0.283]	57.4
48M tokens	2026	72	14.4	0.7 ($\leq$3.6)	0.578→0.831	+0.242 [+0.141, +0.333]	51.1
SmolLM3-3B	2025	88	12.7	0.0 ($\leq$0.5)	0.555→0.796	+0.234 [+0.161, +0.318]	59.3
OLMo-2-1B	2025	93	8.9	0.0 ($\leq$0.0)	0.542→0.807	+0.259 [+0.144, +0.390]	58.6
Gemma-3-1B	2025	53	11.9	2.4 ($\leq$6.5)	0.634→0.767	+0.138 [+0.024, +0.290]	56.4
Gemma-2-2B	2024	93	7.6	0.0 ($\leq$0.4)	0.738→0.869	+0.130 [+0.074, +0.208]	46.6

Table 1: The crossed design at 384 evaluation sequences. In every row, fixing the position shrinks latent×arm to a fraction of its per-arm value and raises $E\rho^2$. Newest base model first, with release years (months and checkpoints in §3). Within a row the per-arm and shared estimates use identical latents at identical replication and differ only in where the measurement was taken. Latent×arm is v_{ab} of (2) as a share of total variance, with the shared estimate followed by its bootstrap bound (§3). The gain is the paired bootstrap gain in $E\rho^2$, resampling latents for both designs together, so it matches its own two coefficients only to within bootstrap error. Position is the within-cell share of variance under shared positions.

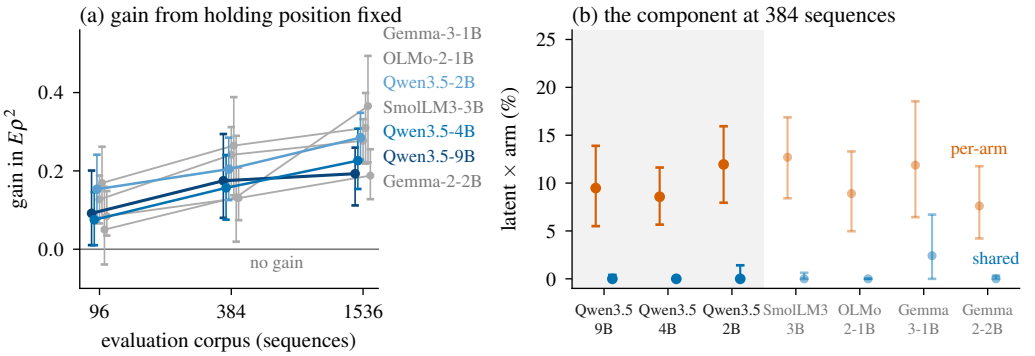


Figure 4: (a) Paired gain in $E\rho^2$ (2) from holding the position fixed, every base model at three evaluation corpora, Qwen3.5 in blue, with the 95% interval of a latent-level bootstrap resampling both designs together. (b) The latent-by-arm component per position mode at 384 sequences, same bootstrap, Qwen3.5 columns on the shaded band and the other models paler.

6 Comparison at a fixed measurement position

The interaction under shared positions. Under per-arm positions on the full sample, the latent-by-arm component is 15.2% on Qwen3.5-9B, and 13.6% to 24.8% across the seven base models. The shared design retains only the latents firing in every arm at a common position, so both designs are compared on those latents (Table 1). On that population the per-arm interaction is 7.6% to 12.7% of the variance. Under shared positions its method-of-moments estimate is clipped to zero on every model except Gemma-3-1B, and the bootstrap bounds it at or below 1.6% on those models. On Gemma-3-1B it is estimated away from the boundary, at 2.4% (bound 6.5%). Once the convention is controlled it is a small fraction of what it was, and the latents rank nearly the same way whichever dictionary produced them.

Gain in reliability. The generalizability coefficient rises from 0.542 to 0.738 under per-arm positions to 0.767 to 0.869 under shared ones. The paired gain of +0.130 to +0.259 at 384 sequences has an interval excluding zero on every model. Across the three corpora the gain grows with corpus size on every model (Figure 4a), and its interval excludes zero in 20 of the 21 model and corpus cells. The exception is Gemma-3-1B at 96 sequences. Part of the growth is the balanced design admitting

paper	readout	positions per latent	position stated	normalised
Bricken et al. 2023	per-feature ablation, logit change	one	no	no
Marks et al. 2025	single-latent ablation, KL	one	no	no
Templeton et al. 2024	per-feature zero clamp	one curated token	yes	no
Gao et al. 2025	per-latent ablation, sparsity of effect vector	aggregate	no	scale free
Cho et al. 2026	single-latent ablation, KL and logit score	one	no	no

Table 2: What each design states, by the text test of §7. Appendix G gives the full audit and the four papers excluded as a different readout family.

more, and less positionally consistent, latents as the corpus grows. Appendix D separates that part from the rest on a latent set held fixed across all three corpora of the two Gemma models.

The retained subpopulation. The shared design keeps only latents firing in every arm at a common position. From 35% to 64% of the 240 sampled have at least one such position, and the balanced set behind every cell of Table 1 is 22% to 50% of the 240, so every controlled number is a claim about that subpopulation. The filter is applied to both designs, so it cannot by itself create a difference between them. On the two Gemma models the excluded latents are the less positionally consistent ones. On Gemma-3-1B those first qualifying at the largest corpus carry far more latent-by-arm variance than those balanced from the start (Appendix D).

A second shared-position rule. The shared rule ranks candidates by the minimum activation across arms, which makes it arm symmetric and also makes it select away from the tokens the arms most disagree about. A second rule takes the union of every arm's own top positions and keeps only tokens where all six arms have nonzero activation (Appendix C, Table 3). Under it the shared interaction stays at or within a point of the boundary and the gain's interval still excludes zero on every model except Gemma-3-1B. On Gemma-3-1B the interaction rises but stays below its per-arm value, and the gain's interval includes zero. On that model part of the controlled estimate depends on the selectivity of the rule.

Variance carried by position within a latent. In the shared rows every arm sits at the same tokens, so the within-cell term is position alone rather than position confounded with each arm's own selection. It is 46.6% to 62.8% of variance at 384 sequences, against 34.4% to 51.5% for the between-latent term (Figure 1b shows it on single latents). Over all 21 model and corpus cells it is 46.6% to 65.7%. Unlike the interaction it sits nowhere near the variance boundary, so it is estimated directly. Which of a latent's own firing positions the measurement lands on moves the answer about as much as which latent was measured. The reliability therefore depends on how many positions each latent is measured at. With $n = k$ in (2), $E\rho^2$ is 0.354 to 0.525 at the single position of published practice and reaches 0.8 at 4 to 8 positions (Appendix F). Relative position, activation magnitude and within-cell activation rank together explain almost none of the per-arm residual at 96 sequences. Their R^2 is 0.003 on Qwen3.5-9B, 0.004 on Qwen3.5-4B, 0.005 on Qwen3.5-2B, 0.001 on SmolLM3, 0.007 on OLMo-2, 0.008 on Gemma-3-1B and 0.005 on Gemma-2-2B. The position effect is not a proxy for these covariates, and holding position fixed removes it without having to predict it.

7 The standard applied to published measurements

Table 2 records, for a sample of five papers confirmed to use single-latent zero ablation with a magnitude readout, what their methods sections state. The last two columns follow a test on the text. A paper records the position when the sections we reviewed state which token the intervention was applied at. It normalises when they divide the readout by the size of the perturbation or use a readout that is scale free by construction.

Value of each item. Four of the five designs do not record the position, so each would add item 1. The score each reports leaves unspecified the share of variance that position carries in our design (§6). Templeton et al. record a curated token per latent and so already meet item 1. Item 2 binds only where dictionaries are compared, and there it moves the whole latent-by-arm component. Four of the five do not normalise their readout, which the recommendation of §4 would add, and the readout of Gao et al. is scale free by construction. For a reader, most published ablation scores become comparable across papers once their methods state the position.

8 Implications

For authors reporting an ablation number. Choose the evaluation positions once, outside the dictionary, record them, and pass the same positions to every dictionary in the comparison. If a comparison across dictionaries is the point, report the latent-by-arm component under both position modes, with its bootstrap bound. That pair of numbers tells a reader how much of the difference is the dictionaries. If the effect is compared across latents of different firing frequency, report it per unit perturbation norm beside the raw value, since the two can disagree in sign (§4).

For readers comparing numbers across papers. Two ablation numbers from different papers on different dictionaries are not comparable unless both state the position. Where they do not, the usable content is the within-paper ranking rather than the magnitude, and even that ranking carries the lower $E\rho^2$ of an uncontrolled design (Table 1).

For benchmark builders. A metric that ablates one latent and reads a magnitude should fix the evaluation positions in the benchmark rather than leave them to the dictionary under test.

For dictionary trainers. A training change justified by a causal metric is informative only once the positions are held fixed across the before and after dictionaries. Our arms differ by one fitting choice, so a real training change likely shows more instability than measured here.

9 Scope of the evidence

Reach of the audit. The audit reads methods sections, so it establishes what a paper states rather than what its authors did.

Dictionary scale. The controlled arms are 12M-token TopK dictionaries of width 16,384, far smaller than production suites trained on billions of tokens. They share an initialisation, which no public suite provides. The gain is measured on these arms, and on released dictionaries only top-token agreement (§5). Six further Qwen3.5-2B arms see 48M tokens each, four times the budget. Their position share and gain (Table 1, 48M row) fall inside the 95% intervals of the 12M arms, and their top agreement falls outside. At four times the budget top agreement is lower, so on this base model longer training sharpens the disagreement the standard corrects rather than removing it.

Models and readouts covered. The base models run from 1B to 3B parameters and on Qwen3.5 up to 9B, in four families, at one relative depth. The results cover metrics that zero a latent's contribution and read a magnitude, and leave interchange-based scores such as RAVEL outside them (§2). The normalisation recommendation rests on one comparison on Gemma-2-2B.

Bias in the controlled numbers. Shared initialisation, clipping at the variance boundary and the retained subpopulation of §6 each shrink the measured contribution of the convention. A comparison across seeds, widths and architectures would show a larger one. On the Gemma models the retained subpopulation excludes the latents where position matters most. Only the selectivity of the shared rule runs the other way, on Gemma-3-1B alone (Appendix C).

10 Conclusion

An ablation score is comparable across papers when it reports the position it was taken at (Box 1, item 1), which carries 46.6% to 62.8% of its variance. A comparison across dictionaries is informative when those positions are fixed outside them (item 2), which raises the reliability of the ranking by $+0.130$ to $+0.259$, with an interval excluding zero on every model. We also recommend reporting the effect per unit perturbation norm beside the raw effect.

ETHICS STATEMENT

This paper makes claims about what five published papers report. Every pipeline variant we compare was written by us from a published description. The groups whose results we re-examine released artifacts that made re-examination possible. The subject matter is a measurement convention in interpretability tooling. No models were trained beyond small sparse autoencoders on public text, no capability was elicited, and no human subjects or personal data are involved.

REPRODUCIBILITY STATEMENT

Code and result files are available at https://anonymous.4open.science/r/sae-artifact-8E6D/. The repository holds the result files behind each table, figure and number in the text, the scripts that produce them, and a check that compares the manuscript with those files. Latent sampling and bootstrap seeding are specified in Appendix H. The measurement uses 42 SAEs, six per base model, each trained on 12M tokens, on Qwen3.5-9B, Qwen3.5-4B, Qwen3.5-2B, SmolLM3-3B, OLMo-2-1B, Gemma-3-1B and Gemma-2-2B, and six further Qwen3.5-2B SAEs on 48M tokens each. Training and evaluation used NVIDIA A100 and H100 GPUs, with a single 16 GB GPU for the smallest models. Base models and Qwen-Scope and Gemma Scope dictionaries are public. Training data is WikiText-103, large enough for the 12M-token budget, and evaluation data is WikiText-2 raw, whose 11,843 paragraphs passing our length filter yield 3,432 distinct 512-token sequences, about 1.8M tokens, so the largest rung of the nested corpus ladder, 1536 sequences, uses 45% of what is available and repeats no sequence. The seed is 0 throughout, the six arms of a base model start from one initialisation, and the two position modes receive different positions and nothing else. A second training of the Gemma-2-2B arms under another software environment, again with seed 0, gave variance components within 1.8 percentage points of the first, so the components are stable although the pipeline is not bitwise deterministic.

AI USE STATEMENT

Large language model assistants served three purposes in this work, and no other. Writing: they edited the text for clarity, concision and consistency. Research ideation and execution: they helped brainstorm and refine hypotheses and experimental designs, wrote and debugged the code that trains the autoencoders, runs the ablations and computes the variance decomposition, and helped interpret results. Drafting: they wrote first drafts of sections, which we then revised. Every hypothesis, design, analysis and claim was checked by the authors, and every reported number is regenerated by the released code from the released result files. The authors are responsible for the final content of this paper, including any text, claims and artifacts produced with generative AI assistance.

REFERENCES

R. Harald Baayen, Douglas J. Davidson, and Douglas M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412, 2008. doi: 10.1016/j.jml.2007.12.005.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: Smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for BERT. *arXiv preprint arXiv:2104.07143*, 2021. URL https://arxiv.org/abs/2104.07143.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting

for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2021. URL https://arxiv.org/abs/2103.03098.

Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. Identifying functionally important features with end-to-end sparse dictionary learning. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. URL https://arxiv.org/abs/2405.12241.

Robert L. Brennan. *Generalizability Theory*. Statistics for Social and Behavioral Sciences. Springer, New York, 2001. doi: 10.1007/978-1-4757-3456-0.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, Anthropic, October 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024. URL https://arxiv.org/abs/2412.06410.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9263–9274, 2020. doi: 10.18653/v1/2020.emnlp-main.745. URL https://arxiv.org/abs/2010.06595.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. Improving steering vectors by targeting sparse autoencoder features. *arXiv preprint arXiv:2411.02193*, 2024. URL https://arxiv.org/abs/2411.02193.

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

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. URL https://arxiv.org/abs/2409.14507.

Maheep Chaudhary and Atticus Geiger. Evaluating open-source sparse autoencoders on disentangling factual knowledge in GPT-2 small. *arXiv preprint arXiv:2409.04478*, 2024. URL https://arxiv.org/abs/2409.04478.

Seonglae Cho, Zekun Wu, Kleyton Da Costa, Rishi Kalra, Ilham Wicaksono, and Adriano Koshiyama. Are single-token sparse autoencoder features causally necessary? Layer-depth and SAE-family effects. *arXiv preprint arXiv:2607.20596*, 2026. URL https://arxiv.org/abs/2607.20596.

Herbert H. Clark. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12(4):335–359, 1973. doi: 10.1016/s0022-5371(73)80014-3.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. URL https://arxiv.org/abs/2304.14997.

Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. John Wiley and Sons, New York, 1972.

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

Boyi Deng, Xu Wang, Yaoning Wang, Yu Wan, Yubo Ma, Baosong Yang, Haoran Wei, Jialong Tang, Huan Lin, Ruize Gao, Tianhao Li, Qian Cao, Xuancheng Ren, Xiaodong Deng, An Yang, Fei Huang, Dayiheng Liu, and Jingren Zhou. Qwen-Scope: Turning sparse features into development tools for large language models. *arXiv preprint arXiv:2605.11887*, 2026. URL https://arxiv.org/abs/2605.11887.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194, 2019. doi: 10.18653/v1/d19-1224. URL https://arxiv.org/abs/1909.03004.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.11944.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022. URL https://arxiv.org/abs/2209.10652.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2405.14860.

Eoin Farrell, Yeu-Tong Lau, and Arthur Conmy. Applying sparse autoencoders to unlearn knowledge in language models. *arXiv preprint arXiv:2410.19278*, 2024. URL https://arxiv.org/abs/2410.19278.

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

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Yoav Gur-Arieh, Roy Mayan, Chen Agassy, Atticus Geiger, and Mor Geva. Enhancing automated interpretability with output-centric feature descriptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5757–5778, 2025. doi: 10.18653/v1/2025.acl-long.288. URL https://aclanthology.org/2025.acl-long.288/.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Automated interpretability metrics do not distinguish trained and random transformers. *arXiv preprint arXiv:2501.17727*, 2025. URL https://arxiv.org/abs/2501.17727.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. doi: 10.1609/aaai.v32i1.11694. URL https://arxiv.org/abs/1709.06560.

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://arxiv.org/abs/2402.17700.

Charles M. Judd, Jacob Westfall, and David A. Kenny. Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103(1):54–69, 2012. doi: 10.1037/a0028347.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? A case study in sparse probing. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2502.16681.

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 37 (NeurIPS)*, 2024. URL https://arxiv.org/abs/2408.00113.

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

Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, and Neel Nanda. Interpreting attention layer outputs with sparse autoencoders. *arXiv preprint arXiv:2406.17759*, 2024. URL https://arxiv.org/abs/2406.17759.

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. URL https://arxiv.org/abs/2602.14111.

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

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

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, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://arxiv.org/abs/2408.05147.

Aleksandar Makelov, George Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. *arXiv preprint arXiv:2405.08366*, 2024. URL https://arxiv.org/abs/2405.08366.

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

Callum McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, and Neel Nanda. Gemma Scope 2 - Technical Paper. Technical report, Google, 2025. URL https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/gemma-scope-2-helping-the-ai-safety-community-deepen-understanding-of-complex-language-model-behavior/Gemma_Scope_2_Technical_Paper.pdf.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022. URL https://arxiv.org/abs/2202.05262.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1609.07843.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2311.03658.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. *arXiv preprint arXiv:2501.16615*, 2025. URL https://arxiv.org/abs/2501.16615.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

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

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

Richard J. Shavelson and Noreen M. Webb. *Generalizability Theory: A Primer*. Sage Publications, Newbury Park, CA, 1991.

Xiangchen Song, Aashiq Muhamed, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona T. Diab, Virginia Smith, and Kun Zhang. Mechanistic interpretability should prioritize feature consistency in sparse autoencoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2172–2210, 2026. doi: 10.18653/v1/2026.acl-long.99. URL https://aclanthology.org/2026.acl-long.99/.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://arxiv.org/abs/2310.10348.

Team OLMo. 2 OLMo 2 furious. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2501.00656.

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

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. URL https://arxiv.org/abs/2607.19386.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2211.00593.

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, 2025. URL https://arxiv.org/abs/2501.17148.

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

APPENDIX CONTENTS

A The sign reversal under magnitude normalisation . . . 15
B Position agreement between released production dictionaries . . . 15
C A second shared-position rule . . . 16
D Composition of the balanced sample across corpus sizes . . . 17
E Variance decomposition under per-arm positions . . . 17
F Reliability against the number of positions per latent . . . 17
G The full literature audit . . . 18
H Latent sampling and bootstrap seeding . . . 18

A THE SIGN REVERSAL UNDER MAGNITUDE NORMALISATION

The recommendation of §4 rests on this comparison, on Gemma-2-2B. The contrast is whether a dictionary's rarest activation-frequency decile carries more or less causal mass than a frequency-matched decile of an untrained, tied random dictionary. Under raw KL the rare decile is worse than random (HL $= 0.683$, 95% CI $[0.459, 0.961]$, $n = 40$ against 35, Mann–Whitney $p = 0.039$). Under KL per unit perturbation norm the identical latents and the identical control are better than random (HL $= 1.687$, 95% CI $[1.171, 2.498]$, $p = 0.011$). Both intervals exclude one. The two readings share the sample, the model and the dictionary, and differ only in whether the intervention's own size is accounted for.

Frequency and causal mass. Activation frequency predicts causal mass in neither the trained nor the control arm once magnitude is accounted for. The Spearman correlation between frequency bin and KL per unit norm is -0.065 ($p = 0.313$, $n = 240$ latents) on the trained dictionary and $+0.036$ ($p = 0.612$, $n = 198$) on the control, against $+0.258$ ($p < 0.001$) for raw KL on the trained dictionary. The apparent frequency dependence here is magnitude rather than frequency, and under raw KL it would read as a rarity finding.

B POSITION AGREEMENT BETWEEN RELEASED PRODUCTION DICTIONARIES

§5 states the result. This appendix gives the matching procedure behind it, the threshold sensitivity, and the band weighting.

Method. We take two Gemma Scope dictionaries on the same base model and layer differing in width or sparsity target, and match latents across them by decoder cosine, since released dictionaries share no index correspondence. Matching is mutual nearest neighbour above a stated cosine threshold, and we report how many pairs survive at each of several thresholds rather than choosing one: a permissive threshold admits pairs that are not the same feature, and a strict one selects the most duplicated directions. On the surviving pairs we compute each dictionary's own top-activating positions under its own gate and report mean Jaccard overlap and top-position agreement, the same two quantities as §5. The Qwen3.5 release is read the same way, with two differences its format imposes. It is TopK rather than JumpReLU, so each latent's activation is taken from the full top-k encode rather than from its own column, and it ships one width per model, so the only pair is the sparsity pair $k = 100$ against 50 at width 32768, a strictly narrower test than the Gemma Scope grid.

Result. On pairs matched above cosine 0.70, median cosine 0.90 for the pair differing in sparsity and 0.87 for the pair differing in width, so near-duplicate directions, the two dictionaries pick the same single top-activating position for 42.7% and 38.8% of latents, at Jaccard 0.404 and 0.361. Both are far below full agreement, on dictionaries trained independently at production scale. On the Qwen3.5-2B sparsity pair at layer 11 the same threshold gives 40.3% at median cosine 0.88, and pairs above cosine 0.90 agree for 57.0%, and on Qwen3.5-9B at layer 15 the figures are 38.4% at 0.85 and 53.6%.

Agreement within cosine bands. Matched pairs are the latents whose decoder directions already agree most, so they should agree about position more than a typical pair would. Loosening the threshold does not show this, since the eligible pool grows but the sample stays dominated by

near duplicates, so we instead sample within disjoint cosine bands (Figure 3). Agreement rises monotonically with decoder similarity, from 11.2% and 10.5% in the 0.50 to 0.60 band to 60.2% and 52.3% above 0.90.

Weighting of the bands. The bands are not equally populated, and reporting the low one alone would describe a tail as though it were the centre. Across the 15 pairs, mutual nearest neighbours have median cosine 0.854, and only 3.9% of them fall in the 0.50 to 0.60 band while 34.7% sit above 0.90. Weighting each band's agreement by its share gives 42% and 35% for a randomly drawn matched pair, consistent with the pair-level figures in Figure 3 computed a different way. The same weighting over the Qwen-Scope sparsity pair's own band shares gives 34% on Qwen3.5-2B and 28% on Qwen3.5-9B. The modal matched pair is a near duplicate, and even where two released dictionaries encode a latent almost identically they measure it at different tokens a large fraction of the time.

C A SECOND SHARED-POSITION RULE

The rule used throughout ranks each latent's candidate positions by the minimum gated activation across the six arms. That is what makes it arm symmetric, since no arm serves as the reference, and it is also selective in a direction that matters: a token where five arms fire hard and one barely fires ranks below a token where all six fire moderately, so the rule prefers positions of high inter-arm agreement and avoids the contested ones. Since the contested positions are what this paper is about, some of the reported reduction in the interaction could be that selection rather than the position control it is attributed to.

The comparator. For each latent, take the union of every arm's own n top-activating positions, so each arm's argmax is a candidate whatever the other arms' activations there. From that union keep only the tokens at which every arm has nonzero activation. Without this the design stops being crossed, and an arm contributing a null effect at a token where it does not fire would sit at the log floor and dominate the decomposition. Rank the survivors by the mean activation across arms: the maximum would privilege whichever arm fires hardest and reintroduce an arm as reference, and the minimum would rebuild the rule under test. The result is arm symmetric and non-selective with respect to inter-arm agreement, while still guaranteeing that every arm fires where it is measured.

	latents retained		shared latent×arm (%)		paired gain [95% CI]	
model	min-gated	union	min-gated	union	min-gated	union
Qwen3.5-9B	151	129	0.0	0.0	$+0.174$ $[+0.080, +0.288]$	$+0.140$ $[+0.081, +0.224]$
Qwen3.5-4B	154	141	0.0	0.0	$+0.156$ $[+0.075, +0.235]$	$+0.161$ $[+0.100, +0.224]$
Qwen3.5-2B	149	132	0.0	0.0	$+0.201$ $[+0.128, +0.283]$	$+0.177$ $[+0.101, +0.266]$
SmolLM3-3B	144	124	0.0	0.9	$+0.234$ $[+0.161, +0.318]$	$+0.232$ $[+0.138, +0.341]$
OLMo-2-1B	139	126	0.0	0.0	$+0.259$ $[+0.144, +0.390]$	$+0.279$ $[+0.174, +0.413]$
Gemma-3-1B	85	66	2.4	6.3	$+0.138$ $[+0.024, +0.290]$	$+0.037$ $[-0.046, +0.171]$
Gemma-2-2B	131	118	0.0	0.0	$+0.130$ $[+0.074, +0.208]$	$+0.085$ $[+0.037, +0.151]$

Table 3: The two shared-position rules at 384 sequences. Retained counts are latents with at least one qualifying position under each rule. The union columns are computed on the balanced cube of latents retained under the union rule, which is smaller than the minimum-gated one.

Result. Retention falls under the union rule on every model, as expected, since contested tokens are where some arm is more likely silent (Table 3). The interaction under the union rule is at the variance boundary on Gemma-2-2B, Qwen3.5-2B, OLMo-2, Qwen3.5-4B and Qwen3.5-9B, and 0.9% on SmolLM3, and the paired gain keeps an interval excluding zero on every model but Gemma-3-1B. On Gemma-3-1B the interaction rises from 2.4% to 6.3% and the gain's interval includes zero.

Interpretation. On every model but Gemma-3-1B the two rules agree, so candidate selection is not what removes the interaction there. On Gemma-3-1B roughly half the reduction survives the change of rule and the remainder was selection, and its gain under the union rule, at 36 latents, cannot be separated from the loss of power the stricter retention causes. The agreement is shown for these two rules, and neither is derived from a definition of what the estimand should be.

D COMPOSITION OF THE BALANCED SAMPLE ACROSS CORPUS SIZES

The paired gain grows with evaluation corpus on every model (§6). Two things could produce that: the effect strengthening with data, or the balanced design admitting different latents at different corpus sizes. On the two Gemma models the second is real and measurable. On Gemma-3-1B the 38 latents that first qualify at 1536 sequences have median positional consistency 0.084 against 0.155 for those already balanced at 96, and carry 32.3% latent×arm against 10.6%, and on Gemma-2-2B the corresponding figures are 0.105 against 0.114 and 18.8% against 8.7%. The newcomers are the sparser, less positionally consistent latents, exactly where position selection has most room to disagree.

Restricting all three corpora to the latents balanced in every one of them holds the sample fixed, so any remaining trend is about data. The gain then reads +0.046, +0.082 and +0.176 on 32 Gemma-3-1B latents, and +0.082, +0.110 and +0.110 on 62 Gemma-2-2B latents. The growth is therefore part composition and part not, in proportions that differ by model. The gain grows across the retained samples, and the effect on a fixed sample grows less.

The Gemma-2 tie at 384 and 1536. The fixed-latent gain reads +0.110 at 384 sequences and +0.110 at 1536 on the same 62 latents, which could suggest a ceiling, but the components underneath the two rows differ. The per-arm interaction share rises from 6.7% to 8.7%, and both coefficients move, from 0.776 to 0.763 per-arm and from 0.888 to 0.874 shared. Their differences land 0.0005 apart, and the narrower of the two bootstrap intervals is 301 times that gap. The tie is therefore arithmetic rather than a plateau, and the fixed-set gain is reported as rising from the smallest corpus and unchanged thereafter.

E VARIANCE DECOMPOSITION UNDER PER-ARM POSITIONS

The decomposition below is measured on Gemma-2-2B with every arm at its own top-activating positions. Position is nested within arm there, so the latent×arm term mixes the latent behaving differently under a different fit with the arms having chosen different tokens, and the two are not separable after the fact. The crossed design of Table 1 separates them.

component	2 arms (extrapolated)	6 arms (direct)
latent	18.1%	16.9%
arm	0.4%	3.2%
latent × arm	9.7%	12.5%
within-cell (position)	71.8%	67.4%
$E\rho^2$ (one-arm design)	0.455	0.415

Gemma-2-2B, 96-sequence evaluation corpus, 200 latents × 6 arms × 6 positions per cell. This is an earlier sample than the 240 latents of §3, at a smaller corpus than the 384 sequences of §6, so its latent×arm share is not a cell of the range quoted there. The two-arm column compares a trained arm against a soft-frozen arm and carries one degree of freedom on the arm factor.

Within-cell term under per-arm positions. At 67.4% it is larger than the 46.6% of §6, and the two are not interchangeable. Every arm sits at its own top-activating positions here, so the within-cell term mixes position with each arm's selection, which is the confound this design cannot resolve. The body quotes the shared-position figure, where position is the only thing varying inside a cell.

F RELIABILITY AGAINST THE NUMBER OF POSITIONS PER LATENT

The components of (2) are estimated once on each model's shared cube at 384 sequences, and $E\rho^2$ is evaluated with $n = k$ positions per latent. Re-estimating the coefficient on random k-position subsets of the same cube, for k from 2 to 6, reproduces the projection. Table 4 gives the coefficient at one position, the smallest k reaching 0.8 and 0.9, and the limit as k grows, which is below one only where the shared interaction is estimated away from the boundary.

model	$E\rho^2$ at $k = 1$	k for 0.8	k for 0.9	limit
Qwen3.5-9B	0.404	6	14	1.000
Qwen3.5-4B	0.354	8	17	1.000
Qwen3.5-2B	0.413	6	13	1.000
SmolLM3-3B	0.394	7	14	1.000
OLMo-2-1B	0.410	6	13	1.000
Gemma-3-1B	0.398	8	30	0.941
Gemma-2-2B	0.525	4	9	1.000

Table 4: Reliability of the latent ranking under shared positions against the number k of positions per latent, 384 sequences.

G THE FULL LITERATURE AUDIT

Table 2 in the body gives the columns corresponding to the item and the recommendation the paper measures. The full audit also recorded special-token handling, near-duplicate structure and unembedding alignment. No paper states its special-token handling, and none reports duplication or alignment in a way tied to its causal effect measurement, though Bricken et al. and Templeton et al. discuss feature splitting broadly across dictionary sizes, and Cho et al. report a per-layer Δ-logit structure whose relation to unembedding alignment as defined here is not established. For Cho et al. we also implemented the single-token detector at the stated specification, which returned 0.41 to 0.44 of the reported prevalence on our corpus. No result in this paper uses that detector. Four further papers were considered and excluded as a different readout family rather than as negative findings: Makelov et al. 2024 (edits 2, 4 or 6 features simultaneously), SAEBench TPP and SCR (set ablation of several latents per concept), Korznikov et al. 2026 (RAVEL interchange match rate), and Rajamanoharan et al. 2024b (JumpReLU, whole-dictionary substitution). Exclusions are recorded so that the denominator of the audit is visible.

H LATENT SAMPLING AND BOOTSTRAP SEEDING

Bootstrap seeding. Each bootstrap interval is seeded from its own inputs, so adding or removing an unrelated quantity leaves every other interval unchanged.

Latent sampling. Latents are drawn by one fixed permutation of the full identifier space, keeping the first 240 that are live in every arm. The result is a uniform sample of the live-set intersection, and a rerun whose intersection shifts by one latent changes the sample by at most that latent. Two evaluation files are compared only when they share at least 80% of their latents.