Feature Size, Decoder Geometry, and Semantic Specificity in Sparse Autoencoders

Alexander Knox

Review draft | November 6, 2026

Abstract

Sparse autoencoder (SAE) features are usually inspected through their activating examples and decoder directions. We study an additional statistic: feature size, the log standard deviation of encoder pre-activation over text. Across 29 model-layer cells, larger features occupy more crowded and tighter decoder neighbourhoods; the relations also hold in all 27 cells with structural-feature exclusions. Size additionally tracks semantic specificity. The partial correlation with firing-token entropy is negative in all 21 measured JumpReLU-trained cells after structural exclusion, with median -0.29, but becomes small or changes sign in some TopK-trained cells. Blind abstractness ratings support the relationship in several early and middle layers; late-layer ratings and gradient-based context dependence reveal its limits. Matched-sparsity controls at one site show that the geometric signature depends on training recipe: structureless-data nulls reproduce much more of it under TopK than under JumpReLU. Size predicts explanation quality at Gemma 2B layer 12 and identifies token-locked features across dictionaries, although encoder norm alone captures much of this diagnostic signal. We report corpus and dictionary stability, contrary results, and exploratory transfer to held-out model lineages. These findings establish a useful empirical connection between feature scale, local geometry and semantic specificity, while leaving its general causal mechanism unresolved.

1. Introduction

Sparse autoencoders decompose neural activations into a larger collection of sparsely activated latent variables. Bricken et al. demonstrated how this approach can recover interpretable features from a language model [13]. Public dictionaries make it possible to investigate the properties of these latents across many layers and models [1-4]. We use the conventional term feature for an SAE latent; this terminology does not assume a unique, ground-truth decomposition of a model's computation.

A feature has both an encoder, which detects it, and a decoder, which maps its activation back into the model's residual stream. A unit-normalized decoder specifies a direction. The encoder also has a magnitude, and its pre-activation varies across inputs. This variation is inexpensive to measure, yet it is not ordinarily presented as a semantic diagnostic. We ask whether it tells us which features are token-specific, which occupy tight local neighbourhoods, and which are easier to explain.

Our main empirical result is a connection between feature size, local decoder geometry and semantic specificity. Larger features have more similar neighbouring decoder directions in every model-layer cell we measure. A semantic connection also appears, but its strength depends on depth, the dictionary recipe, and the operational definition of abstraction. We therefore distinguish firing-token diversity, rated abstractness and context dependence, and evaluate interpretability through explicit diagnostic tasks.

Three findings organize the study. First, size has a consistent positive association with decoder crowding and tightness across the measured cells. Second, size is associated with lower firing-token entropy in every measured JumpReLU-trained cell, with substantially weaker evidence in some other recipes and at late depth. Third, size is useful as a diagnostic of explanation quality and token

locking, but much of its practical signal is available from encoder norm alone. This third finding is about measured interpretability tasks; it does not establish a general ranking of feature usefulness for steering.

The study combines an initial exploratory investigation with a series of hypotheses recorded before their subsequent tests. The predictions, amendments, failures and inconclusive results remain in the research ledger. Several proposed mechanisms failed. These outcomes constrain the final claims rather than requiring a complete causal account before the empirical findings can be reported.

2. Measurements and study design

2.1 Feature size and geometric quantities

Let x be the model activation at a fixed residual-stream site, e_i the encoder vector for feature i, b_i its bias, and C the covariance of x over an evaluation text pool. We define size before the SAE nonlinearity:

$$s_i = \log \mathrm{std}(e_i^T x + b_i) = 0.5 \log(e_i^T C e_i).$$

Equivalently, size is the sum of log encoder norm and log input standard deviation along the unit encoder direction. The bias does not affect this variance. Size is neither firing frequency nor decoder norm. Decoder vectors are normalized when computing geometry; loader-specific decoder scaling is incorporated into encoder parameters where appropriate. Absolute size is coordinate- and normalization-dependent, so our main comparisons are associations within dictionaries rather than comparisons of absolute units across models.

Decoder density is the mean cosine similarity to a feature's ten nearest decoder neighbours, excluding the feature itself. Tightness is a different local statistic. We greedily partition decoder directions into groups of four, starting with directions having the strongest nearest-neighbour similarity, and add available directions that maximize summed similarity to the group. A feature's tightness is its mean cosine similarity to the other members. Density and tightness consequently share geometric information; their correlation is not independent evidence of semantic structure. Nearest-neighbour decoder cosine has previously been examined by Michaud et al. [9, Appendix E.1]; here we measure both mean-ten-neighbour density and quartet tightness and test their association with encoder pre-activation scale.

2.2 Models, dictionaries and evaluation pools

The main panel contains 27 cells: Gemma 2 2B and 9B, five Gemma 3 sizes (270M, 1B, 4B, 12B and 27B), Llama 3.1 8B, and Qwen3-8B, each measured at three depths. Gemma dictionaries come from Gemma Scope and Gemma Scope 2 [1,2]; Llama and Qwen dictionaries come from their public Scope releases [3,4]. Two additional raw-only cells, GPT-2 small layer 6 and Pythia-70m layer 2, provide smaller-model comparisons. Appendix A lists the 27 main cells and their measured associations.

The 21 Gemma cells use JumpReLU-trained dictionaries. The six Llama/Qwen cells use TopK-trained releases, although their deployed encoders need not use identical gates: the Llama release uses a calibrated threshold after TopK training. The two smaller-model dictionaries use ReLU with an L1 penalty. Architecture labels describe training provenance. Model family and public dictionary recipe are confounded, so cross-family differences alone cannot identify which is responsible.

The standard evaluation pools contain approximately 200,000 residual-stream activations from tokenized public web text drawn from the FineWeb sample. The corrected primary pipelines prepend a beginning-of-sequence token to each 1,024-token chunk and omit the BOS and first content-token positions from saved rows. Model-specific tokenization and saved row-to-token

mappings are respected. The initial Gemma analysis omitted repeated BOS tokens; the affected primary size and firing statistics were rebuilt before the results used here. Historical uncorrected results are not pooled with corrected measurements.

Public dictionaries and their evaluation pools must be distinguished from our own recipe-control dictionaries. The latter were evaluated on text overlapping their training pools in the original R1/R1b program. These controls demonstrate behaviour on that text and support recipe sensitivity; they are not held-out generalization tests. New crossed-model training and a disjoint-text replication remain follow-up work.

2.3 Semantic specificity and interpretability outcomes

Firing-token entropy measures lexical diversity: for each feature with at least 200 firing events, we sample 200 events and compute the entropy of their current-token identities. This fixes the event count while retaining differences in token diversity. It is an operational measure, not a definition of semantic abstraction. A feature can depend on a long context while firing on one token type.

Blind automated rating batteries sample approximately 1,000 features per evaluated cell, stratified by size. Language-model raters judge descriptions or activating snippets on a five-point concrete-to-abstract scale; these are not human annotations. Raters are not given feature size or firing rate. Description and snippet ratings are retained separately when they disagree. The original Gemma rating set was selected using pre-correction size deciles, which limits interpretation of the corrected rated sample as a representative dictionary sample.

Feature nonlocality (FNL) is a published gradient-based measure of context dependence [7]. We join the authors' per-feature values to the identical canonical Gemma Scope dictionaries. Their Protocol B uses 64-token windows, no BOS, and excludes the window-initial influence bin. This is an external measurement on the same features, but with a different context protocol from our main pools.

The token-locking diagnostic labels a feature when one token identity accounts for at least 80% of its top 50 activations, among features with at least 50 firings. We use the term token-locked rather than treating every negative label as a validated concept feature. Explanation quality is assessed separately with a describe-then-predict task: descriptions are generated from five top snippets and evaluated on ten held-out snippets, five firing and five non-firing. Balanced accuracy is computed per feature. This task follows the broader automated explanation-and-scoring approach of Bills et al. [10] and is closely related to the detection evaluation of Paulo et al. [11]. Our five-example description set, balanced ten-example evaluation and size-stratified sampling define the protocol used here; the contribution is the measured association with size, rather than a new explanation-scoring method.

2.4 Exclusions, uncertainty and scope

Headline analyses exclude features flagged as structural. The frozen rule flags a feature when at least half of its top 50 input tokens are whitespace, digits, punctuation or special tokens, or when at least half of its ten highest direct-unembedding tokens are special tokens or whitespace. The output-side component is a screening rule, not a validated account of causal output behaviour. Raw results are reported alongside exclusions in the archived analyses.

Most dictionary-wide associations are partial Spearman correlations controlling for log firing rate, implemented by residualizing ranked variables on ranked covariates. The main AUD1 panel uses approximate Fisher-z intervals with a Spearman adjustment; targeted tests use feature bootstrap intervals. These intervals describe uncertainty under feature-level sampling assumptions. Features share neighbourhoods and training history, so they are not independent model or training replications. We report effect sizes, cell-level patterns and exceptions, and do not interpret 21 related Gemma cells as 21 independent model families.

3. Size, geometry and semantic specificity

3.1 Larger features occupy tighter decoder neighbourhoods

After structural exclusion, the size-density partial correlation is positive in all 27 main cells, with median 0.651 and range 0.239 to 0.870. The size-tightness partial is also positive in all 27, with median 0.515 and range 0.188 to 0.840. The corresponding raw associations remain positive in all 29 cells, including GPT-2 and Pythia. Figure 1 shows the full main panel, rather than selecting representative positive cases.

The association is local. In Gemma 2B layer 12, randomly chosen decoder pairs remain approximately orthogonal, while the largest-size decile has closer nearest neighbours than the smallest-size decile. The finding concerns local crowding, not a claim that all large features share a common global direction.

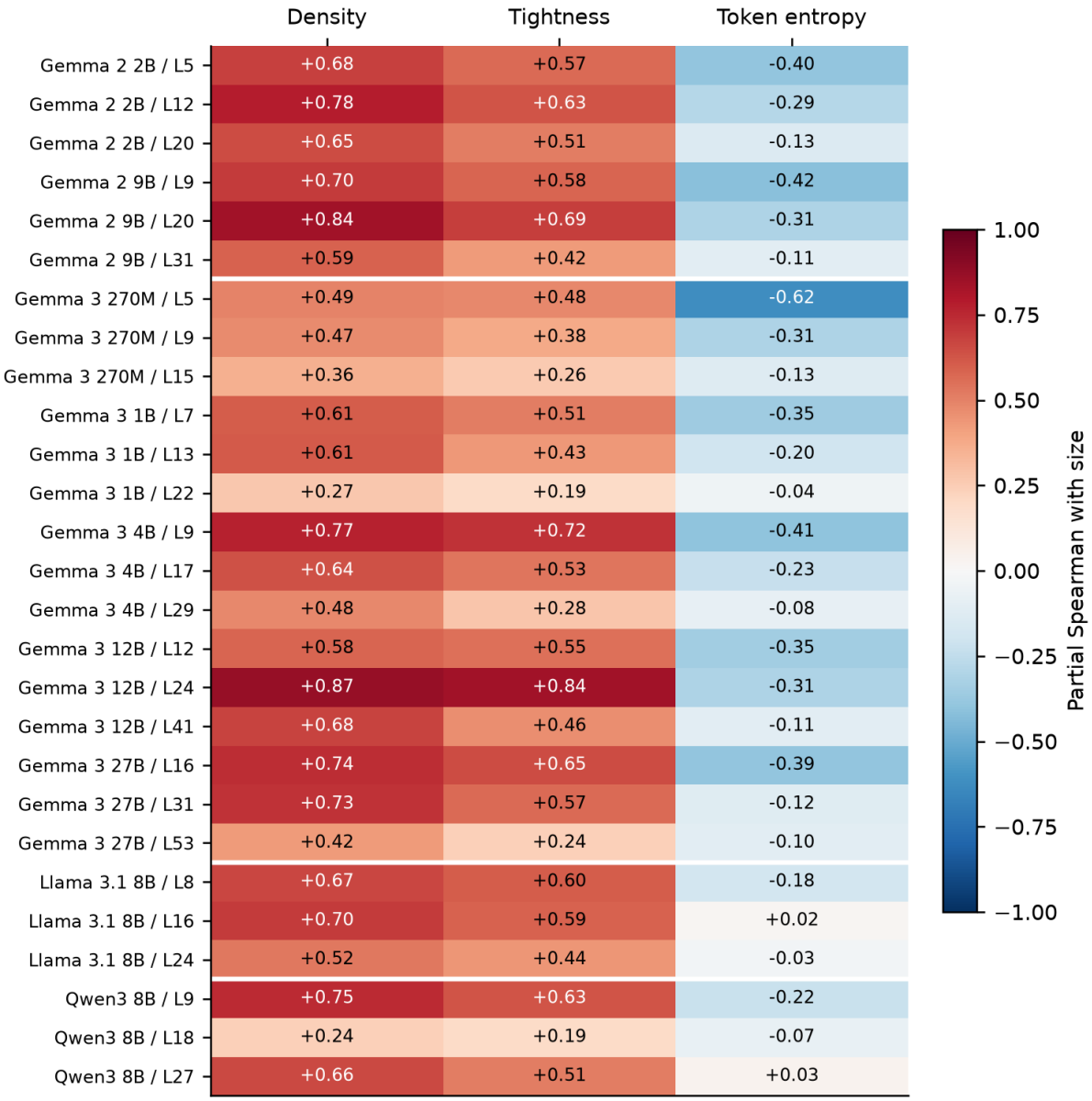


Figure 1. Partial correlations with feature size across the 27 main cells, controlling for log firing rate and excluding structural features. Each row is one model-layer cell; the dictionary is the unit of descriptive comparison. Decoder density and tightness have positive associations throughout. The token-entropy association is weaker and sometimes changes sign in the TopK-trained cells. Values come directly from the archived AUD1 structural-excluded results.

3.2 The semantic connection is strong but scoped

The size-token-entropy partial correlation is negative in 25 of 27 main cells after structural exclusion. Across all 27, its median is -0.199 and its range is -0.621 to +0.034. Restricting to the 21 JumpReLU-trained Gemma cells, all 21 correlations are negative with reported intervals excluding zero, and their median is approximately -0.29. This restriction was examined after recipe controls; it is a descriptive subgroup result rather than a new independent confirmatory test.

The relationship is strongest in many early and middle Gemma cells and weaker late. In Gemma 3, excluding structural features leaves a median of about 40% of the raw late-layer effect. Some residual late effects are small, including approximately -0.04 at Gemma 3 1B's late layer. Qwen layer 27 and Llama layer 16 have near-zero positive values after exclusion; Llama layer 24 is also close to zero. These exceptions are part of the result.

Blind automated ratings support a semantic interpretation beyond the entropy proxy in several cells. At Gemma 2B layer 12, the coefficient of rated abstractness in a regression of size, controlling for log rate, is negative (-0.168 per standard deviation of rated abstractness). Related rated tests replicate at Gemma 2B layer 5, Gemma 2 9B layers 9/20, and the sampled Llama feature batteries. Controls for firing-token norm and normalized inputs preserve the relationship. These regressions use different feature sets and statistics from the dictionary-wide rank correlations and should not be compared numerically as if they were the same estimand.

Late-layer ratings are less reliable. Some Gemma description-based ratings reverse sign while snippet ratings follow the entropy-based pattern. We do not select the rater that agrees with the hypothesis. The divergence limits the claim that one scalar captures abstraction at every depth.

3.3 Context dependence supplies a different check

At Gemma 2B layer 12, size is negatively associated with published FNL after controlling for rate and excluding structural features: partial correlation -0.40, with a reported interval of [-0.46, -0.35]. The corresponding relation is weak at layer 5 and absent at layer 20. Our prediction of a strong negative relation at every tested layer failed. FNL and firing-token entropy are themselves negatively associated in these cells, demonstrating that they cannot be used interchangeably as abstraction scores.

4. Training recipe constrains the interpretation

4.1 Structureless-data controls

We trained dictionaries on Gaussian activations matched to the real pool's mean and covariance and on activations shuffled independently across coordinates. The latter preserves coordinate marginals while removing cross-coordinate dependence; the Gaussian null retains second-order geometry but changes higher-order structure. Neither control is a universal model of non-semantic data.

At Gemma 2B layer 12, matched-sparsity JumpReLU nulls reproduce 38% and 4% of the real size-density correlation and 21% and 4% of the real size-tightness correlation, respectively. In the tested TopK recipe, the same two nulls reproduce 76%/91% of size-density and 65%/85% of size-tightness. Table 1 gives the underlying correlations. Thus the observed geometric relation can be largely generated by a dictionary recipe on these nulls, and its presence alone does not demonstrate semantic organization.

Recipe	Data	Size / density	Size / tightness	Tightness / density
JumpReLU	Real	0.581	0.497	0.867
JumpReLU	Gaussian	0.219	0.103	0.530

Recipe	Data	Size / density	Size / tightness	Tightness / density
JumpReLU	Shuffled	0.022	0.020	0.671
TopK	Real	0.660	0.519	0.863
TopK	Gaussian	0.499	0.335	0.677
TopK	Shuffled	0.601	0.440	0.728

Table 1. Partial correlations controlling for log firing rate at Gemma 2B layer 12. JR denotes JumpReLU. JR null sparsity was retuned to approximately 80 active features; TopK uses k = 64. Real JR correlations are averages across the two reference dictionaries. Each null condition has one seed. Reconstruction quality differs across real and null inputs.

The JumpReLU result supports an excess over these particular nulls at this site. It does not establish that real semantic structure is necessary at every model or layer. Conversely, the TopK result is obtained with our training implementation, not the exact Llama Scope or Qwen-Scope recipes. The public TopK columns are therefore susceptible to a method explanation, but we have not directly established its size there. Tightness-density is substantially reproduced by nulls under both recipes, consistent with their shared geometric definition.

4.2 Recipe and developmental checks

Our own JumpReLU dictionaries reproduce the negative size-entropy relationship at Gemma 2B layer 12. At approximately matched public-dictionary sparsity, two seeds give partial correlations -0.311 and -0.308. Across the tested JumpReLU sparsity sweep the sign remains negative. Our TopK runs produce different relationships, including near-zero effects at the larger training budgets. Because recipes also differ in objectives and optimization conventions, these experiments identify recipe sensitivity rather than isolating the nonlinearity as the sole cause.

A Pythia-160m checkpoint series trained with the same JumpReLU recipe has negative size-entropy correlations from step 1,000 onward. The completed random-initialization control has a positive correlation, +0.123 [0.106, 0.141]. Its formal verdict is indeterminate because the registered decision rule covered a near-zero band and a negative band, but not this positive outcome. The observed sign contrast argues against the negative association being inevitable under this recipe on every residual stream. It does not identify an emergence time between initialization and step 1,000, and the checkpoint comparisons retain a sparsity confound.

4.3 Stability without a complete mechanism

Feature size ranks are stable across new chat, German Wikipedia and code pools at Gemma 2B layers 5, 12 and 20: raw correlations with the original pool range from 0.827 to 0.986, with greater drift late. The size-density, size-tightness and size-entropy signs survive in all nine corpus-layer combinations. Across matched Gemma Scope features at widths 16k and 65k, size rank correlation is approximately 0.95. This matching result concerns matched features and does not characterize features without a reliable match.

The causal mechanism remains unresolved. Covariance along decoder directions accounts for part of the associations, and real-data crowding interventions change the size-entropy relationship only partially. Training dynamics give small temporal asymmetries. A direction-matched surgery control shows that an apparent response to reduced crowding can instead be a generic refit after changing decoder direction. Statistical mediation terms share measurements with size and cannot identify causal pathways. The empirical signature is therefore more secure than any proposed explanation of its origin.

5. Practical interpretability diagnostics

5.1 Explanation quality and token locking

At Gemma 2B layer 12, the describe-then-predict experiment returns evaluable predictions for 494 of 500 sampled features. Size correlates with per-feature balanced accuracy at 0.259 [0.177, 0.338], and the partial correlation given rate is 0.243. The effect is modest and concentrated partly in the smallest deciles, whose explanations are less predictive. Figure 2 shows this result together with the broader token-locking comparison. It establishes a measured relationship with explanation quality at this site, not a cross-model guarantee of interpretability.

For the original Gemma token-locking diagnostic, size gives AUROC 0.841. Across the 27-cell baseline comparison, median raw AUROC is 0.770 for size, 0.765 for encoder norm and 0.708 for density. The difference between two medians is not the median paired improvement: the median within-cell advantage of size over norm is approximately 0.03. With structural features excluded, the corresponding medians are approximately 0.762, 0.754 and 0.710. Size is cheap but requires a text pool; encoder norm alone requires only weights.

On the separately rated concrete-versus-abstract subset at Gemma 2B layer 12, size reaches AUROC 0.885 while log rate alone reaches 0.854. Thus a strong classification score does not always imply a large incremental benefit over a simple baseline. Token-locking and rated abstractness are separate tasks with different sampling and label definitions.

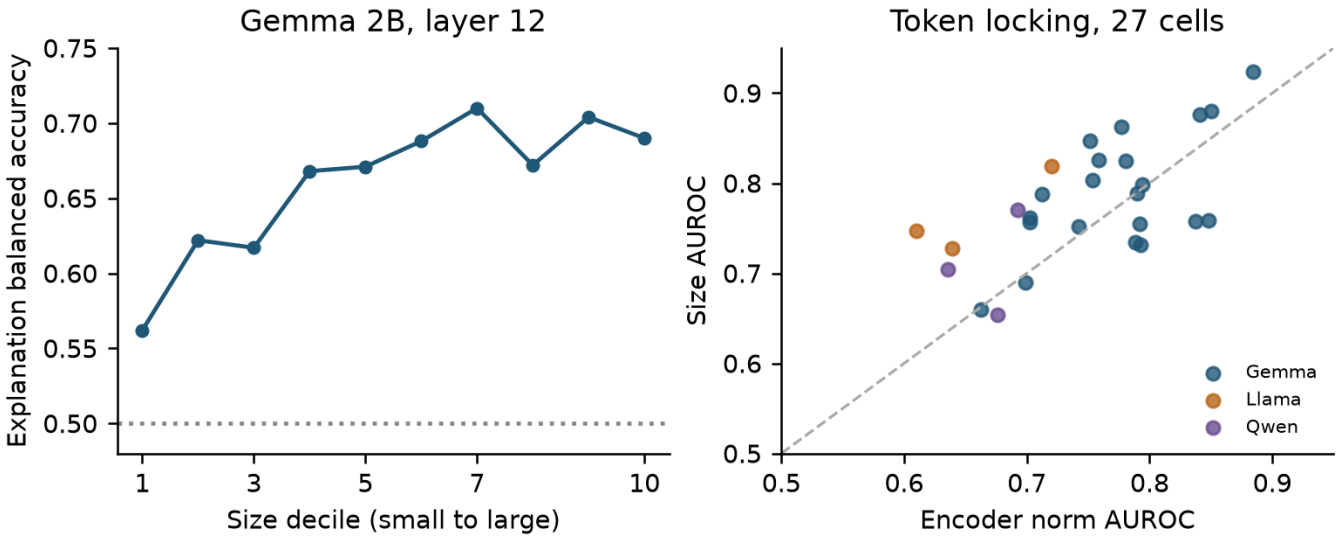


Figure 2. Left: explanation balanced accuracy by size decile at Gemma 2B layer 12 (494 features overall; descriptive decile means). Right: token-locking AUROC for size and encoder norm across the 27-cell structural-excluded baseline comparison. Each point is one cell, and the diagonal marks equal performance. A high AUROC on this label should not be interpreted as a direct test of general conceptual understanding.

5.2 Exploratory transfer to held-out model lineages

Within-dictionary prediction can benefit from dictionary-specific regularities. We therefore added an exploratory transfer test using five weights-only inputs: log encoder norm, density, maximum nearest-neighbour cosine, tightness and encoder-decoder cosine. We held out all Gemma generations, sizes and layers together, then held out Llama and Qwen in turn. Targets are never used to train or tune a model for their held-out lineage. This is a reuse of existing measurements on new prediction splits, not a fresh-data replication.

With original-statistic inputs, a fixed boosted-tree predictor attains family-median token-locking AUROCs of 0.703 for held-out Gemma, 0.679 for Llama and 0.733 for Qwen. The corresponding

norm-only predictors give 0.700, 0.634 and 0.613. A linear five-input baseline reaches 0.734, 0.665 and 0.733. Performance depends on the target: for token entropy, the five-input boosted model does not consistently improve on norm alone; for firing rate, its Llama transfer is near zero with original-statistic inputs. These failures are reported, not removed by choosing a different model separately for each test lineage.

Outcome	Held-out lineage	Cells	Norm only	Five inputs	Linear
Token locking (AUC)	Gemma	21	0.700	0.703	0.734
Token locking (AUC)	Llama	3	0.634	0.679	0.665
Token locking (AUC)	Qwen	3	0.613	0.733	0.733
Token entropy	Gemma	21	0.259	0.224	0.256
Token entropy	Llama	3	0.167	0.130	0.252
Token entropy	Qwen	3	0.081	0.221	0.214
Log firing rate	Gemma	21	-0.030	0.387	0.293
Log firing rate	Llama	3	-0.478	-0.033	0.544
Log firing rate	Qwen	3	-0.186	0.298	0.340
Current-token determinacy	Gemma	21	0.455	0.348	0.291
Current-token determinacy	Llama	3	0.290	0.447	0.509
Current-token determinacy	Qwen	3	0.471	0.454	0.531
Next-token determinacy	Gemma	21	0.325	0.326	0.351
Next-token determinacy	Llama	3	0.445	0.467	0.493
Next-token determinacy	Qwen	3	0.376	0.365	0.456

Table 2. Exploratory transfer with an entire lineage held out. Values are medians over cells within the held-out group. Token locking uses AUROC; other outcomes use within-cell Spearman correlation. Five-input boosted trees are compared with a fitted encoder-norm-only model and a linear five-input baseline. These are prediction results, not steering experiments.

A pre-specified secondary representation ranks the five inputs within each dictionary, using the held-out dictionary's unlabeled weights. It improves some transfer results and worsens others. Training targets for regression are ranked within training cells; training samples are capped at 2,000 features per cell and balanced by lineage and cell. Appendix B specifies the models and includes the secondary results. There are only three lineages, with family and recipe confounded, so these results do not support precise population-level claims about unseen families.

5.3 Boundaries of the diagnostic claim

Input and output feature scores need not identify the same useful features, as studied by Arad et al. [8]. Our size statistic is not their output score. Improved selection for a particular steering task requires a separate intervention evaluation, including collateral effects. Neither token-locking AUROC nor explanation accuracy establishes that application.

An exploratory cached-activation intervention in Appendix D tests a narrower functional question. At matched target activation increases, protecting nearby decoder features reduces collateral SAE activation energy relative to an encoder-direction baseline, with small additional input norm. This is evidence for geometry-informed activation control in the tested dictionaries; it does not evaluate downstream language-model behaviour or actual weight edits.

A preparation-stage audit found a defect in our secondary ablation-concentration measurement: zero-change windows were assigned maximum concentration. We therefore withhold all output-concentration inferences, including the apparent size-concentration reversal and its relationship with context dependence. Appendix B documents the defect. This measurement is not used in the size, decoder geometry, token entropy, abstractness-rating or explanation-quality calculations.

6. Related work

Public SAE releases supply the dictionaries needed for our comparisons [1-4]. Their training choices are part of the measurement system, rather than neutral instruments. Work on dense SAE latents shows that frequent activation can have meaningful functional roles [5]; this motivates controlling for firing rate rather than equating frequency with feature quality. Our size measure is a pre-activation standard deviation, distinct from the frequency of nonzero activations.

Cosine-scored SAEs address the influence of input magnitude on inner-product encoders [6]. Our results are compatible with an input-norm confound: the size statistic includes both encoder norm and input spread, and is not norm-free. The rated robustness tests reduce this concern for their particular associations, while the structureless-data controls show that training can also generate geometry. We do not infer that magnitude is always semantic or that suppressing it would necessarily remove useful information.

Feature nonlocality provides an independent, gradient-based account of context dependence [7]. Its partial agreement with size and disagreement with firing-token entropy are informative boundaries on our semantic interpretation. Li et al. [12] describe spatial modularity, associations between co-occurrence and geometric locality, and anisotropy in SAE feature dictionaries. Michaud et al. [9] study feature manifolds and scaling; their Appendix E.1 directly examines nearest-neighbour decoder cosine in Gemma 2B layer 12. These studies establish relevant geometric structure. We ask how local geometry covaries with encoder pre-activation scale and measured semantic specificity across dictionaries.

Automated interpretation also predates this study. Bills et al. [10] generate and score natural-language explanations of neurons, and Paulo et al. [11] extend automated explanation and evaluation to SAE features, including discrimination between activating and non-activating contexts. Our explanation-quality experiment uses that general evaluation framework to test a size association. As these distinctions and the steering work [8] emphasize, predicting input activation is a different outcome from predicting the causal effects of an intervention.

Takehi et al. [14] report early weight-space programming experiments in which a general-purpose LLM reads or writes neural-network weights. Their larger demonstrations include tools and forward-pass feedback, with mixed outcomes across tasks. We cite this work as a research blog, not an established general capability. It motivates asking whether measured weight geometry can help specify targeted functional changes. Appendix D tests that narrower question using SAE dictionary weights and a fixed quadratic optimization, rather than an LLM editing base-model weights. Its connection to weight-space programming is therefore prospective.

7. Limitations and conclusion

The strongest claim is descriptive: feature size has a consistent relationship with local decoder geometry across the measured cells and a useful, scoped relationship with semantic specificity. Its abstraction connection depends on the metric, depth and dictionary recipe. Public releases confound model identity with training method; the null training controls cover one model-layer site with one seed per null. Own-dictionary evaluation overlaps training text. Feature-level intervals do not capture training-seed uncertainty or all dependence among neighbouring features.

Explanation quality was tested directly at one site using language-model descriptions and predictions. The rated samples are stratified and are not interchangeable with full-dictionary entropy samples. Early rating records do not consistently retain exact model identifiers, and the original abstractness prompt was reconstructed for a later rater check; the available prompts and this provenance limitation are retained in the archive. Structural filtering removes substantially different shares across releases and includes a direct-unembedding screening component. FNL uses a different context protocol. The output-concentration analysis was withdrawn after a measurement audit; no downstream language-model steering benefit is established. Appendix D measures collateral SAE activation changes in one model family; its tightness association does not replicate across depths, and size does not consistently predict benefit. Shared source text across model evaluations and the limited number of lineages constrain the transfer interpretation.

Within those limits, size and encoder norm offer practical starting points for inspecting SAE features. The empirical geometry and semantic associations warrant reporting without a universal causal account. The next decisive questions are whether matched SAE recipes reconcile model differences and whether these cheap diagnostics improve feature selection on independently evaluated interpretability tasks. They are extensions of the present findings rather than assumptions needed to state them.

AI assistance disclosure

Generative AI tools assisted hypothesis refinement, experimental design, literature discovery, code development, analysis and manuscript preparation. Language models also generated experimental abstractness ratings, feature descriptions and activation predictions, as described in Section 2.3. The available prompts and scoring procedures are retained in the research archive; missing primary-rater model identifiers are disclosed as a limitation in Section 7. These automated measurements are distinguished from human judgments. The author is responsible for reviewing the scientific claims and takes responsibility for the final released work.

References

[1] Tom Lieberum et al. 2024. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. https://arxiv.org/abs/2408.05147

[2] Google DeepMind. Gemma Scope 2: public interpretability tools and release documentation for Gemma 3. https://deepmind.google/models/gemma/gemma-scope/ (accessed November 4, 2026).

[3] Zhengfu He et al. 2024. Llama Scope: Extracting Millions of Features from Llama-3.1-8B with Sparse Autoencoders. https://arxiv.org/abs/2410.20526

[4] Boyi Deng et al. 2026. Qwen-Scope: Turning Sparse Features into Development Tools for Large Language Models. https://arxiv.org/abs/2605.11887

[5] Xiaoqing Sun et al. 2025. Dense SAE Latents Are Features, Not Bugs. https://arxiv.org/abs/2506.15679

[6] Silen Naihin and Lev Stambler. 2026. Size Doesn't Matter: Cosine-Scored Sparse Autoencoders. https://arxiv.org/abs/2606.15054

[7] Chuqiao Lin, Shivaji Sondhi and Xiao-Liang Qi. 2026. Measuring Semantic Abstractness of SAE Features via Nonlocality. https://arxiv.org/abs/2608.10537

[8] Dana Arad, Aaron Mueller and Yonatan Belinkov. 2025. SAEs Are Good for Steering -- If You Select the Right Features. https://arxiv.org/abs/2505.20063

[9] Eric J. Michaud, Liv Gorton and Tom McGrath. 2025. Understanding sparse autoencoder scaling in the presence of feature manifolds. https://arxiv.org/abs/2509.02565

[10] Steven Bills et al. 2023. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html

[11] Gonçalo Paulo, Alex Mallen, Caden Juang and Nora Belrose. 2024; revised 2025. Automatically Interpreting Millions of Features in Large Language Models. https://arxiv.org/abs/2410.13928

[12] Yuxiao Li, Eric J. Michaud, David D. Baek, Joshua Engels, Xiaoqing Sun and Max Tegmark. 2025. The Geometry of Concepts: Sparse Autoencoder Feature Structure. Entropy 27(4), 344. https://doi.org/10.3390/e27040344

[13] Trenton Bricken et al. 2023. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. https://transformer-circuits.pub/2023/monosemantic-features/index.html

[14] Rikiya Takehi, Jayden Teoh, Alex Zhang and Omar Khattab. 2026. Sparks of Weight-Space Programming: LLMs can natively read & write neural network weights. Research blog, November 2. https://rikiyat.github.io/blog/weight-space-programming/ (accessed November 6, 2026).

Appendix A. Cell-level results and reproducibility

Model / layer	Recipe	N entropy	Size / density	Size / tightness	Size / entropy
Gemma 2 2B / L5	JumpReLU	7,626	+0.685	+0.566	-0.395
Gemma 2 2B / L12	JumpReLU	9,206	+0.782	+0.628	-0.288
Gemma 2 2B / L20	JumpReLU	8,844	+0.651	+0.508	-0.135
Gemma 2 9B / L9	JumpReLU	11,225	+0.702	+0.583	-0.421
Gemma 2 9B / L20	JumpReLU	9,263	+0.836	+0.689	-0.308
Gemma 2 9B / L31	JumpReLU	10,510	+0.589	+0.422	-0.105
Gemma 3 270M / L5	JumpReLU	4,179	+0.495	+0.481	-0.621
Gemma 3 270M / L9	JumpReLU	5,546	+0.473	+0.380	-0.306
Gemma 3 270M / L15	JumpReLU	5,025	+0.358	+0.261	-0.127
Gemma 3 1B / L7	JumpReLU	4,341	+0.608	+0.515	-0.352
Gemma 3 1B / L13	JumpReLU	6,635	+0.614	+0.431	-0.199
Gemma 3 1B / L22	JumpReLU	7,103	+0.269	+0.188	-0.045
Gemma 3 4B / L9	JumpReLU	5,803	+0.768	+0.719	-0.406
Gemma 3 4B / L17	JumpReLU	6,707	+0.637	+0.532	-0.229
Gemma 3 4B / L29	JumpReLU	8,243	+0.478	+0.279	-0.082
Gemma 3 12B / L12	JumpReLU	5,472	+0.585	+0.549	-0.349
Gemma 3 12B / L24	JumpReLU	3,128	+0.870	+0.840	-0.308
Gemma 3 12B / L41	JumpReLU	8,401	+0.678	+0.462	-0.107
Gemma 3 27B / L16	JumpReLU	6,306	+0.742	+0.654	-0.394
Gemma 3 27B / L31	JumpReLU	6,680	+0.725	+0.566	-0.123
Gemma 3 27B / L53	JumpReLU	8,604	+0.423	+0.240	-0.105
Llama 3.1 8B / L8	TopK-trained	4,880	+0.668	+0.602	-0.181
Llama 3.1 8B / L16	TopK-trained	9,821	+0.700	+0.593	+0.021
Llama 3.1 8B / L24	TopK-trained	9,631	+0.524	+0.444	-0.027
Qwen3 8B / L9	TopK-trained	9,492	+0.749	+0.630	-0.222
Qwen3 8B / L18	TopK-trained	4,284	+0.239	+0.188	-0.075
Qwen3 8B / L27	TopK-trained	4,775	+0.658	+0.512	+0.034

Table A1. Primary structural-excluded associations. N is the number of eligible features in the size-token-entropy comparison; geometry may use more features. All correlations control for log firing rate. Layer labels identify residual-stream sites in the saved analysis configuration. The table is generated from the stored AUD1 results rather than transcribed from the narrative draft.

The accompanying review archive contains the manuscript source, PDF builder, figure inputs, aggregate results, analysis scripts for the exploratory follow-ups, source checksums and the historical prediction ledger. It supports reconstruction of the displayed figures and tables. Full regeneration of the underlying activation pools additionally requires the original model/dictionary weights, text preprocessing and experiment code. A public release location has not yet been deposited for this review version; local file paths are not public availability claims.

Size, entropy and rate use aligned feature indices within each dictionary. The exploratory transfer outputs were checked against source arrays for exact feature IDs, masks and labels; every saved score was independently recomputed, and held-out lineages were checked for disjointness from fitting lineages. These checks establish consistency of that analysis, not an independent audit of the original model extraction code. The reproducibility package distinguishes those levels of validation.

Appendix B. Exploratory follow-ups

The transfer predictor uses HistGradientBoosting with 100 iterations, 15 maximum leaf nodes, learning rate 0.1, no early stopping and seed 20261104. The norm-only baseline uses the same estimator with one input. Linear baselines use training-only standardization followed by Ridge with alpha 1 for regression or logistic regression with C = 1 for classification. Fitting uses at most 2,000 deterministically sampled features per cell, with equal total weights for training lineages and then their cells. Regression labels are percentile ranks within each training cell. Evaluation uses every eligible feature in the held-out cell.

The primary representation uses the original five weight statistics. The secondary representation replaces each input by its within-dictionary percentile, including at the target dictionary; this is unlabeled target-weight normalization. Firing rate, size and covariance-derived statistics are never predictor inputs. Token-entropy labels require at least 200 firing events, and token-locking labels at least 50. Current- and next-token determinacy are adjusted explained-variance statistics (epsilon squared) for pre-activation grouped by current or next token identity; token types with fewer than five pool occurrences are pooled. These are token-association measures, not downstream intervention effects. All reported transfer cells require at least 100 eligible features. FNL is available only for Gemma and cannot support the same lineage-transfer test. Output-concentration targets are withheld following the measurement audit below.

Outcome	Held-out lineage	Cells	Norm only	Five inputs	Linear
Token locking (AUC)	Gemma	21	0.689	0.656	0.756
Token locking (AUC)	Llama	3	0.614	0.644	0.665
Token locking (AUC)	Qwen	3	0.664	0.697	0.725
Token entropy	Gemma	21	0.183	0.164	0.369
Token entropy	Llama	3	0.144	0.211	0.254
Token entropy	Qwen	3	0.105	0.198	0.196
Log firing rate	Gemma	21	-0.031	0.271	0.240
Log firing rate	Llama	3	0.396	0.541	0.518
Log firing rate	Qwen	3	0.142	0.427	0.471
Current-token determinacy	Gemma	21	0.459	0.375	0.376
Current-token determinacy	Llama	3	0.465	0.498	0.511
Current-token determinacy	Qwen	3	0.526	0.569	0.579
Next-token determinacy	Gemma	21	0.332	0.317	0.304
Next-token determinacy	Llama	3	0.440	0.496	0.509
Next-token determinacy	Qwen	3	0.426	0.439	0.460

Table B1. Secondary transfer with within-dictionary percentile inputs. Metric definitions match Table 2. This representation uses unlabeled target weights; it is not a separate fresh-data replication.

Measurement audit: output concentration withheld

The historical ablation pipeline normalized the absolute change in log probabilities, q = |delta log p| / sum |delta log p|, and computed concentration as 1 + sum(q log q) / log V. It clamped a zero denominator to a small positive value. For a zero-change window this gives q = 0 and concentration = 1, although a normalized effect distribution is undefined. Features selected from full-context firings often stopped firing when re-evaluated in the shorter ablation windows; these windows were included in feature averages.

Across the five audited Gemma/Llama cells, the Spearman correlation between saved concentration and the fraction of windows no longer firing ranges from 0.99967 to 0.99997. In each cell, every feature that never fires in the ablation windows has concentration exactly 1 (8 to 24 features per cell). Thus the reported concentration is dominated by the no-effect fraction. Reproducibility across ablation runs does not validate its interpretation as output selectivity.

We withdraw concentration-based size and geometry associations, the exploratory FNL-concentration comparison, and concentration prediction scores from the paper's evidentiary claims. Original results remain in the historical archive, explicitly marked as affected. A future evaluation must define the outcome for zero-effect events, report their frequency separately, and validate event-level effects before interpreting conditional concentration. An algebraic adjustment of saved averages is not presented as a validated repair. The core measurements and the five transfer outcomes reported here do not use the affected concentration values.

Appendix C. Prediction ledger and contrary evidence

The archived ledger contains 48 entries: 13 with a qualified or unqualified PASS label, 25 FAIL, five MIXED, two PENDING, one INFEASIBLE, one UNRESOLVED and one INDETERMINATE. PASS includes rows with explicitly failed secondary predictions. The ledger is a historical account of heterogeneous tests, not a success rate or a single family of independent statistical tests. Predictions and amendments were recorded at different stages of an adaptive research program; not every current analysis was prospectively specified at the start of the project.

The table below preserves the recorded outcome labels. Pending studies are not evidence for the current claims. Main-text interpretation follows the completed data and stated limitations, even when an older ledger description uses stronger causal wording. In particular, the random-init control remains formally indeterminate, the crowding mechanism remains unresolved, and the new transfer analyses are exploratory. Rows marked AUDIT contain concentration-dependent results whose original interpretations are withdrawn; the historical verdict is not a current validation of those claims. Other outcomes in a mixed row are not automatically invalidated.

Test	Recorded outcome	Verdict
§1 R-ST	No early/mid sign change; late Gemma 3 keeps 22-49% (median 40%)	PASS
§2 R1b	L0 ≈ 80: −0.311 / −0.308 (CIs exclude 0), two seeds; negative at L0 17-360	PASS
§3 M4b	9/12 (JumpReLU 0/3 at equal density); crowding adds under both (TopK +2.3→+4.2, JumpReLU −8.7→+1.5)	MIXED
§3 R2	Next month	PENDING
§4 O1c	1 of 3: depth ρ +0.29 [−0.26, +0.64]; Gemma 3 4/4; gaps +0.16/+0.10	FAIL
§4 O1 (f)	f1 late 0/3, all 37%; f2 ρ +0.17 [−0.23, +0.60]	FAIL
§4 CH1 Gemma 2B [AUDIT]	Rises L5→L12, falls at L20 (−0.02 [−0.03, −0.01]); conc ~ o +0.45 to +0.58	FAIL

Test	Recorded outcome	Verdict
§4 CH2 Gemma 2B [AUDIT]	−0.48 / −0.58 / −0.22 (opposite sign); survives KL control	FAIL
§4 CH1 Llama [AUDIT]	L16: conc ~ o +0.54 (holds); size ~ conc −0.41 (opposite to "same pattern"); L8/L24 not run	FAIL
§4 CH4b redo	6% coherent vs 50% for small features; 47% of top decile structural	FAIL
§4 G4d'	Not run (Kaggle quota reset)	PENDING
§5 Llama L16 mismatch	Big +0.318 [0.208, 0.395]; small −0.003; dial also fails (1.22×)	PASS
§5 Llama L8/L24 dial	L8 ×2 0.98 [0.85, 1.13] p 0.57; L24 ×2 1.05 [0.87, 1.19] p 0.35: both not-a-unit nulls	FAIL
§6 Full-set reruns	P1 0.083 (27 cells); P2 fails; P3 (ii) ρ +0.74; (iii) 40%	PASS
§7 U1-x	0.71-0.87 in 22/24 cells; no late drop; exclusion does not help	FAIL
CH1-repro [AUDIT]	0.963 / 0.972; next_share 0.52	PASS
AN1 (PREREG)	Axis relations universal (27/27, 26/27, 26/27); mediation in 2/27 cells (median 0.24); depth part +0.78 PASS	FAIL
G4a' (PREREG)	Max density drop 2.7% across μ 0.1-3 (penalty ~0.1% of loss); fingerprint unchanged	INFEASIBLE
M4a-Llama-L8/L24 (PREREG)	L8 0.90 / 0.88; L24 1.15 / 0.97 (L16 1.52); all mediator	PASS
CH2-Llama (PREREG) [AUDIT]	−0.41 / −0.46 (CIs excl. 0); conc ~ o +0.54 / +0.58	PASS
CH2-Llama-L24 (PREREG) [AUDIT]	−0.33 / −0.35 (CIs excl. 0); conc ~ o +0.53 / +0.55	FAIL
SPLIT-ε^2 (PREREG)	Mediator at Gemma 2B L12 and Llama L16, both directions; proportion 1.10 / 1.56 (not < 1)	FAIL
NL1 (PREREG)	R-ST: FNL~size −0.16 / −0.40 / −0.01, FNL~log‖e‖ +0.08 / −0.32 / −0.15 (L5/L12/L20); |FNL~size| > |tokH~size| at L12	FAIL
NULL-GEO JumpReLU (PREREG)	Hold: size~density 29% / opposite, size~tightness 11% / opposite. Fail: tightness~density 56% / 65%, size~mean-on 62% (Gaussian) / 8%	MIXED
VIS-SAE (PREREG)	log‖e‖~density −0.59 (flip), tight~density +0.75, size~density +0.02, size~tight +0.04 (EV 0.779)	FAIL
VIS-SAE-2 (PREREG)	D1: log‖e‖~density +0.63, tight~density +0.80, size~density +0.42, size~tight +0.35 (EV 0.74); D2, C failed EV gate (unscored)	PASS (D1)
BIM1 (PREREG) [AUDIT]	L12: ST1 ratio 9.6×, ε^2_cur Δ +0.047, word class 94% vs 50%; tokH and density reversed, ablation n.s.; L5 same; L20 / Llama no POINT group	PASS (L12, L5)
CORP1 (PREREG)	size ρ 0.83-0.99 raw, 0.79-0.98 R-ST; no sign flipped or lost	PASS
GEO-DEC (PREREG) [AUDIT]	tokH 27/27 ≥ size; ε^2_cur ≥ size 15/27; rate readable 27/27 (0.67); FNL 3/3; conc 3/5; ST1 late 7/7	MIXED
DICT1 (PREREG)	size +0.95, density +0.90, tightness +0.70; size~density +0.78 vs +0.75; L0 variants size 0.90-0.96; Llama 8x/32x size 0.78	PASS
GEO-DEC-X (PREREG)	rate 0.30+0.27 = 0.57; FNL log‖e‖ 0.63; token entropy cos(enc,dec) 0.42 (all families); ST1 log‖e‖ 0.34 vs cos(enc,dec) 0.25	FAIL
BIM1-X (PREREG)	POINT in g2b L5/L12 (68%/48%) and g9 L9/L20 (50%/34%) only; 0% in all late, Llama, Qwen and all 15 Gemma 3 cells; ST1 enrichment 3.7-9.6×	PASS
MECH-A (PREREG)	size~θ median −0.18, log‖e‖~θ −0.29, θ/‖e‖~density −0.50 (opposite sign in 21/21), size~density cut 0.8%	FAIL

Test	Recorded outcome	Verdict
MECH-C1 (PREREG)	size~norm_mean median −0.16 (negative in 24/27); cut −0.1% (with norm_std 2.9%)	FAIL
MECH-C2 (PREREG)	size~share late +0.42, early |0.55|; cut ≥ 20% in 18/29; median cut 37% (Gemma 3 56%, Gemma 2 18%, Llama 3%, Qwen 12%)	FAIL
MECH-C3 (PREREG)	size~β median +0.10 (−0.60 to +0.76; sign varies by family); cut 0.7% (Qwen3 23%)	FAIL
MECH-C4 (PREREG)	size~cos(e,μ_on) median +0.28 (Gemma 2B L5/L12 ≈ +0.5, L20 and Llama ≈ 0); density~NN template cos −0.06; cut 1.7%	FAIL
MECH-C2b (PREREG, post hoc)	Gemma 3 cut peaks at k=20 (64%); projected size~density +0.61 → +0.35, < 0.3 in 7/15; non-Gemma-3 max |Δ| 0.04; whitening lowers all families	MIXED
MECH-D (PREREG)	density~E_nn ≥ 0.3 in 4/29; size~E_nn median +0.16; cut median 0.2% (max 14%, pythia)	FAIL
MECH-E (PREREG)	size~logE ≥ 0.5 in 6/27 (median +0.37); density~logE +0.27; cut median 2.9% (max 23%); Gaussian null cut 25%	FAIL
MECH-W (PREREG)	share_enc median 0.51 (≥ 0.6 in 10/29; spread ≥ 0.6 in 4/29); whitened 0.57 (rise 19/29); Gaussian null −0.42; own real 0.24	MIXED
MECH-S (PREREG)	(1) +0.33 / +0.44; (2) +0.23 / +0.40; (3a) +0.94 / +0.76; (3b) +0.36 / +0.45; cut 1.7% / 10.6%	FAIL
MECH-B (PREREG)	(1) 9/29, median +0.28; (2) 6/29, median −0.06 (≤ +0.05 in 20/29); co-firing +0.12 better; Gaussian null +0.06 vs real −0.23	FAIL
§4 MECH-1	P2 lead +0.067 / +0.069, P3 cross +0.062 / +0.064 (two seeds, CIs exclude 0). Secondary FAILED: P4 gradient flow 0.30-0.38, reversed vs early density; P5/P6 step-0 −0.04. share_enc 0.24 at L0 80, 0.58 at L0 30, 0.71 with Gemma Scope conventions at L0 38, 0.64 with them at L0 65 (arm 3)	PASS (P2 and P3; P4-P6 failed)
MECH-3 (PREREG)	A: DiD Δlog‖e‖ −0.0087 (CI −0.0090, −0.0085; sham +0.0004); B: Δdensity −0.0001 (CI ±0.0007)	UNRESOLVED
NULL-GEO TopK arm (PREREG)	size~density 76% / 91%, size~tight 65% / 85%, tight~density 78% / 84%, size~meanon 69% / 76%	FAIL
NULL-GEO matched-L0 (PREREG amendment)	size~density 38% / 4% (+9 / +15 pp); size~meanon Gaussian 62% → 41% (−21 pp); tight~density 61% / 77%	FAIL
CH3-INIT (PREREG)	+0.123 [+0.106, +0.141] (L0 29.9); trained checkpoints −0.29 to −0.42	INDETERMINATE

Table C1. Historical prediction/outcome ledger, generated from the stored 48-row record. Detailed operational criteria, amendments and numeric results remain in the accompanying research archive. AUDIT marks concentration-dependent historical claims; see Appendix B. This table is not a substitute for those criteria.

Appendix D. Exploratory geometry-informed activation control

D.1 Motivation and preceding pilot

Motivated by weight-space programming [14], we asked whether dictionary geometry can help design a selective activation intervention. This follow-up was developed after the main study and is outside the 48-entry historical ledger. Each pilot's design was recorded before computing its outcomes, but the second was motivated by the first; neither is an independent confirmation of the main paper's hypotheses.

The preceding pilot injected equal-norm positive and negative decoder directions into 256 cached rows per layer, using 128 randomly sampled live, non-structural features in Gemma 2B layers 5, 12 and 20. At an input norm of 1% of the median residual norm, tightness correlated negatively with

other SAE gate changes after controlling encoder norm and firing rate: -0.201, -0.245 and -0.136, respectively. Exploratory intervals excluded zero at layers 5 and 12, opposite to the proposed positive-interference relationship. Equal input norm did not guarantee equal target effect, motivating the matched-effect design below. The original pilot and contrary result are retained in the archive.

D.2 Matched target design

The second pilot retained the same 128 source features per layer without selecting on earlier intervention outcomes. We sampled 8,192 new cached rows per layer, excluding the preceding pilot's rows, using seed 20261107 plus layer number. Each source was evaluated on its first at most 32 baseline-active sampled rows; at least eight were required. This yielded 76, 99 and 89 included sources at layers 5, 12 and 20. These rows come from the existing evaluation corpus, not a new independent dataset. Layer 12 was primary; other depths were descriptive checks.

Let e_i and d_i denote the source encoder and decoder. Four intervention directions satisfy e_i^T v = 1: the decoder direction d_i/(e_i^T d_i), the minimum-input-norm encoder direction e_i/||e_i||^2, geometry protection, and random protection. Geometry protection selects the 32 nearest unit decoder neighbours among live non-structural features with finite size and geometry; random protection samples 32 other eligible features uniformly. Both exclude the source. With their unit encoder vectors as columns of U, the protection direction solves:

$$\text{minimize } ||v||^2 + 10\,||U^T v||^2, \text{ subject to } e_i^T v = 1.$$

We compute v proportional to (I + 10 U U^T)^(-1)e_i and normalize to the constraint, using a 32-by-32 Woodbury solve. Neighbour count and penalty were fixed without tuning. Perturbing x to x + delta_i v increases an already-active JumpReLU source by exactly delta_i, where delta_i = 0.10 exp(s_i) at the primary dose; 0.05 and 0.20 are fixed sensitivities. The direction is constructed from weights, while the dose uses the previously measured activation scale. All arms use identical contexts and target increments.

The primary outcome is mean squared off-target activation change, summed over all other eligible SAE features and divided by delta_i squared. The target is always excluded; unchanged coordinates contribute zero. Secondary outcomes are other-gate changes and required input norm. This outcome does not use the withdrawn concentration metric. Ratios are exponentiated median paired log ratios, with 500 feature-bootstrap resamples. Intervals ignore neighbour dependence and do not represent model-family or training-seed replication.

D.3 Results and boundaries

Layer	Sources	Geometry / encoder energy	Geometry / random energy	Extra input norm
5	76	0.894 [0.807, 0.994]	0.997 [0.889, 0.999]	2.2%
12 (primary)	99	0.751 [0.657, 0.818]	0.800 [0.708, 0.872]	3.3%
20	89	0.585 [0.501, 0.619]	0.757 [0.695, 0.837]	4.0%

Table D1. Primary-dose matched-target pilot. Energy ratios below one favour geometry protection. Brackets give exploratory 95% feature-bootstrap intervals. Input cost is the median percentage increase in perturbation norm relative to the encoder baseline. Comparisons use identical source features and active contexts within each layer.

Geometry protection reduced median paired collateral energy relative to the encoder by 11%, 25% and 42% at layers 5, 12 and 20, with 2-4% additional input norm. It improved 76%, 79% and 93% of included sources. The advantage over random protection was substantial at layers 12 and 20 but only 0.3% at layer 5. Geometry/encoder energy ratios remained below one at all three doses; layer 5 was close to parity at the smallest dose (0.991). The simple decoder also performed competitively,

with decoder/encoder ratios 0.833, 0.994 and 0.623 and median input-norm overheads of 29%, 22% and 30%. A post-outcome descriptive comparison gives geometry/decoder ratios 0.999, 0.817 and 0.970; the clearest selectivity advantage over this baseline is at layer 12.

At layer 12, tighter neighbourhoods predicted greater protection benefit: the partial Spearman correlation with log(geometry/encoder energy), controlling encoder norm and firing rate, was -0.276 [-0.457, -0.065], or -0.279 after additionally controlling size. This pattern did not replicate across depths: layer 5 gave +0.112 [-0.107, +0.324] and layer 20 +0.100 [-0.144, +0.321]. Size did not consistently predict benefit; all three adjusted intervals included zero. Its association with input cost is partly built into the dose definition through exp(s_i). Abstractness and explanation quality were not tested in this intervention.

The positive result is that decoder neighbourhoods can guide a fixed intervention algorithm toward lower collateral SAE activation energy at matched target effect in these dictionaries. It does not establish a general chain from size or abstraction through tightness to programmability. The energy metric is dictionary-unit-dependent, other features need not represent unrelated concepts, and rare sources were excluded by the active-context requirement. Geometry-arm energy rank agreement between disjoint context halves was only 0.181, 0.367 and 0.033 at layers 5, 12 and 20; individual-feature forecasts are consequently uncertain. No base-model forward passes or weight edits were performed, so downstream behaviour and weight-space programming capability remain untested.

The complete CPU calculation took 29.2 seconds with three threads. Independent direct float64 re-encoding of two sources per layer across all four arms reproduced gate counts exactly and energy within 0.1%, with matched target increments to relative error below 1e-10. Every included direction satisfied its target constraint and input-norm lower bound; both protection directions satisfied the constrained optimum conditions. Protocols, per-source outcomes, source identifiers, sampled-row identifiers, direction arrays and validation records are retained locally. The compact review package includes protocols, outcomes, summaries and scripts, while activation/direction arrays and external weights are outside that package.