REPRESENTATION REDISTRIBUTION IN SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) decompose language model activations into sparse features, but prompt transformations designed to preserve the underlying request can redistribute activity across them. A feature change observed on one transformed prompt need not recur, and even a recurring change need not affect a downstream computation. We introduce the redistribution sensitivity score (RSS), which identifies large feature responses that recur in a consistent direction, and evaluate the resulting supports on variants excluded from selection. Across 27 pretrained SAEs spanning eight language models, the tested prompt transformations produce a reproducible redistribution profile on these excluded variants. We then separate recurring representation change from signed readout contribution and decision failure. For a fixed linear readout, an exact decomposition shows how response magnitude, readout coefficients, directional alignment, and the clean margin determine downstream consequences. In two frozen safety readouts, combining RSS support erosion with the full clean margin yields recurrent failure AUROCs of 0.935 and 0.958 using median supports of 20 and 19 features, compared with 0.965 and 0.964 using full erosion. RSS therefore provides a way to identify recurring sparse representation changes and quantify their contribution to a downstream readout without equating representation change with failure.

1 INTRODUCTION

Sparse autoencoders (SAEs) decompose dense language model activations into sparse features intended to separate signals entangled by superposition (Elhage et al., 2022; Huben et al., 2024). Related inputs can change which features respond, including shifts toward more specific features and changes under prompt or surface form variation (Chanin et al., 2025; Cho & Hockenmaier, 2025; Verma et al., 2026). Such prompt variations can also affect model behavior (Sclar et al., 2024; Mizrahi et al., 2024). Using SAE features to characterize this variation requires distinguishing responses specific to individual inputs from patterns that recur across related inputs, and assessing their relevance to a downstream computation.

Tracking one feature can reveal its response while missing changes elsewhere, whereas aggregate activation distances summarize overall change without identifying the contributing features. We refer to changes in how activity is distributed across the sparse basis as representation redistribution. Observing redistribution on one set of inputs does not establish which feature changes recur beyond those observations, and recurrence alone does not show whether the same changes matter to a downstream computation.

We therefore study recurring redistribution as a distinct object, asking which sparse feature changes repeat across related prompt transformations and persist beyond the variants used to identify them. We introduce the redistribution sensitivity score (RSS), which combines response magnitude with recurrence and directional consistency to select a responsive support. Features and their directions are selected on one set of variants and then held fixed when evaluating disjoint variants. MeanRSS summarizes the response magnitude of the selected support (Figure 1).

Recurring representation change is not the same as downstream effect. A selected feature can receive little readout weight, contributions from different features can cancel, and a change in readout score need not cross the decision threshold. We therefore decompose the contribution of recurrent changes to a fixed linear readout and assess the remaining margin relative to the clean input. We evaluate

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

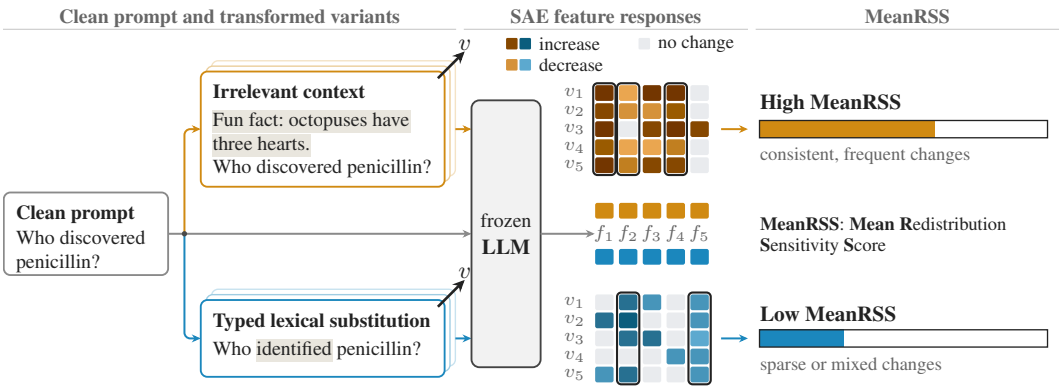


Figure 1: RSS under prompt transformations. In the illustrated five-variant selection, boxed features change in the same direction in at least four variants. MeanRSS averages their absolute response magnitudes.

this connection through offline audits of two frozen safety readouts with known labels. Our main contributions are as follows.

- We introduce RSS to identify sparse feature changes that recur in a consistent direction across related prompt transformations, and evaluate fixed supports on variants excluded from selection. Across 27 pretrained SAEs, the tested transformations yield a reproducible redistribution profile beyond the variants used to select the supports.
- We separate recurring representation change from downstream contribution and decision failure. For a fixed linear readout, an exact decomposition shows how response magnitude, readout weight, directional alignment, and the clean margin connect recurrent changes to readout effects and threshold crossing.
- We show that recurrent changes selected without readout information nevertheless explain substantial held out readout variation. In two frozen safety readouts, RSS supports reconstruct held out logit changes and retain offline audit utility when combined with the clean margin. Median supports of 20 and 19 features yield recurrent failure AUROCs of 0.935 and 0.958, versus 0.965 and 0.964 using full erosion.

2 Related Work

Sparse autoencoder features have been evaluated through reconstruction, sparsity, and downstream use (Bricken et al., 2023; Huben et al., 2024; Gao et al., 2025; Makelov et al., 2025; Karvonen et al., 2025; Kantamneni et al., 2025). Related inputs can also alter which features respond, including absorption into more specific features and changes under prompt or surface form variation (Chanin et al., 2025; Cho & Hockenmaier, 2025; Verma et al., 2026). Feature sensitivity has also been measured as activation reliability across semantically related texts (Tian et al., 2025). RSS instead asks which transformation induced feature changes recur after the support is fixed and evaluates those changes on variants excluded from selection.

Prompt sensitivity studies show that format, ordering, and wording can alter model behavior (Zhao et al., 2021; Sclar et al., 2024; Mizrahi et al., 2024; Chatterjee et al., 2024), while other work studies stable representation directions and SAE responses to optimized perturbations (Dasgupta et al., 2026; Li et al., 2026). We use controlled prompt variation to study recurring redistribution within the sparse basis rather than behavioral sensitivity alone.

Linear probes on SAE activations test whether task relevant information can be recovered from internal representations (Kantamneni et al., 2025), while feature interventions can reveal context dependent downstream effects (Hoang et al., 2026). Our fixed readout analysis serves a different purpose. We hold the fitted readout fixed and measure how much of its prompt induced change is carried by features selected without readout information.

3 METHOD

We select recurrent feature responses without using the readout, then quantify their contributions to a fixed linear readout and the remaining margin.

3.1 REDISTRIBUTION SENSITIVITY SCORE

A large change on one variant need not recur. RSS therefore forms a magnitude based candidate pool, requires recurrence and directional consistency within it, and retains the largest eligible responses.

All quantities are defined for one pretrained SAE, layer, base prompt, and transformation setting. Let x_0 be the clean prompt and x_v, $v \in V$, its transformed variants. At the aligned target position, $a_f(x)$ is the activation of feature f. Define its change and average response magnitude by

$$\delta_{v,f} = a_f(x_v) - a_f(x_0), \; h_f = \frac{1}{|V|} \sum_{v \in V} |\delta_{v,f}|. \tag{1}$$

Here h_f averages absolute changes, so opposing directions do not cancel, and the M largest values form the candidate pool C_M. Supplementary Section S3 specifies operator-level averaging for repeated realizations.

Recurrence and direction are measured from n selector observations. An observation is an individual variant change or an operator's mean change, depending on the protocol. Let n_f^+ count observations greater than τ, and n_f^- count observations less than $-\tau$. Define

$$r_f = \frac{n_f^+ + n_f^-}{n}, \qquad c_f = \frac{\max(n_f^+, n_f^-)}{n_f^+ + n_f^-}, \; s_f = \operatorname{sign}(n_f^+ - n_f^-). \tag{2}$$

Here r_f measures recurrence, c_f directional consistency among responding observations, and s_f their dominant direction; $c_f = 0$ when no observation responds.

Eligible features satisfy $r_f \geq \rho$ and $c_f \geq \kappa$. Let $\mathcal{E}_M = f \in C_M : r_f \geq \rho, \; c_f \geq \kappa$ and $k = \min(K, |\mathcal{E}_M|)$. RSS retains the k eligible features with largest h_f as the responsive support S. Since $\kappa > 1/2$, tied directions are ineligible. For nonempty S,

$$\text{MeanRSS} = \frac{1}{|S|} \sum_{f \in S} h_f. \tag{3}$$

RSS selects the support, whereas MeanRSS summarizes its response magnitude. An empty support makes the summary unavailable rather than zero.

Proposition 1 (Constrained magnitude characterization of RSS)

$$S \in \arg \max_{T \subseteq \mathcal{E}_M \| T| = k} \sum_{f \in T} h_f. \tag{4}$$

If $|\mathcal{E}_M| \geq K$, the same objective is maximized over all dictionary features satisfying the recurrence and directional consistency constraints, up to ties.

The eligibility constraints imply a finite sample directional margin. Let $d_f = \max(n_f^+, n_f^-)$ and $o_f = \min(n_f^+, n_f^-)$ denote the dominant and opposing response counts.

Proposition 2 (Finite sample directional margin) *Every selected feature satisfies*

$$\frac{d_f}{n} = r_f c_f \geq \rho\kappa, \; \frac{d_f - o_f}{n} = r_f(2c_f - 1) \geq \rho(2\kappa - 1). \tag{5}$$

With $n = 5$ and $\rho = \kappa = 0.8$, at least four observations share the dominant direction and at most one opposes it, so changing or removing one observation cannot reverse that direction.

These guarantees apply only to the selection observations, so S and s_f are fixed on V_{sel} and evaluated without recomputation on disjoint V_{ho}.

3.2 Readout Effects of Redistribution

A fixed linear readout lets us isolate the contribution of each feature change exactly. Let $\widetilde{a}_f(x)$ denote activation scaled using clean training data. RSS remains defined on raw feature changes, so probe scaling does not alter S. Supplementary Section S3 gives scaling, probe fitting, and threshold calibration. Define

$$\widetilde{\delta}_{v,f} = \widetilde{a}_f(x_v) - \widetilde{a}_f(x_0),\ z(x) \quad = \mathbf{w}^\top \widetilde{\mathbf{a}}(x) + b,\ m(x) = y[z(x) - t], \qquad m_0 = m(x_0),$$

where $\mathbf{w}, b$ define the fixed readout, $y \in -1, +1$ is the class sign, and t the fixed threshold. Positive margin denotes a correct decision, with m_0 its clean value.

Equal response magnitudes can have different downstream effects because of direction and cancellation. For a variant set U, define average response magnitude A_f^U, alignment η_f^U with the RSS selected direction, and signed margin erosion e_f^U:

$$A_f^U = \frac{1}{|U|}\sum_{v\in U} |\widetilde{\delta}_{v,f}|,\ \eta_f^U \quad = \frac{s_f \sum_{v\in U} \widetilde{\delta}_{v,f}}{\sum_{v\in U} |\widetilde{\delta}_{v,f}|}, \qquad -1 \le \eta_f^U \le 1,\ e_f^U = -y w_f \frac{1}{|U|}\sum_{v\in U} \widetilde{\delta}_{v,f}. \quad (6)$$

A_f^U and e_f^U apply to every readout feature, while η_f^U is defined relative to s_f for $f \in S$. Positive e_f^U erodes the margin; the sign of η_f^U indicates whether the net response over U follows s_f, while $|\eta_f^U|$ decreases with cancellation. We set $\eta_f^U = 0$ when its denominator is zero. These quantities aggregate over U; Section 3.4 treats individual held out variants separately.

Proposition 3 (RSS direction and readout contribution) *For $f \in S$,*

$$e_f^U = -y w_f s_f A_f^U \eta_f^U, \qquad |e_f^U| = |w_f| A_f^U |\eta_f^U| \le |w_f| A_f^U. \quad (7)$$

For nonzero contribution, $\operatorname{sign}(-y w_f s_f)$ *predicts* $\operatorname{sign}(e_f^U)$ *exactly when* $\eta_f^U > 0$.

On the selection variants, write $e_f = e_f^{V_{\text{sel}}}$. The full and selected contributions to average margin erosion are

$$E = \sum_f e_f, \qquad E_S = \sum_{f\in S} e_f, \quad (8)$$

where E is full average margin erosion and E_S the part carried by the RSS support.

3.3 Support Fidelity and Residual Margin

To assess how much margin remains, we combine average erosion with the clean margin:

$$m_{\text{res}} = m_0 - E, \qquad m_{\text{res}}^S = m_0 - E_S,\ m_{\text{res}} \quad = \frac{1}{|V_{\text{sel}}|}\sum_{v\in V_{\text{sel}}} m(x_v). \quad (9)$$

Thus m_{res} is exactly the average transformed margin on V_{sel}, while m_{res}^S replaces full erosion by E_S but retains the full clean margin m_0.

On held out variants, let $e_f^{\text{ho}} = e_f^{V_{\text{ho}}}$. For nonzero total absolute contribution, Margin Contribution Retention (MCR) measures the fraction retained by S:

$$\text{MCR} = \frac{\sum_{f\in S} |e_f^{\text{ho}}|}{\sum_f |e_f^{\text{ho}}|}. \quad (10)$$

Because absolute values are taken after averaging each feature's contribution over V_{ho}, MCR measures retained contribution magnitude rather than signed margin fidelity.

Signed fidelity also depends on transfer of the selected contributions and the net contribution of omitted features. Supplementary Section S1 gives the exact cancellation decomposition and conditions for preserving readout decisions and average margin rankings.

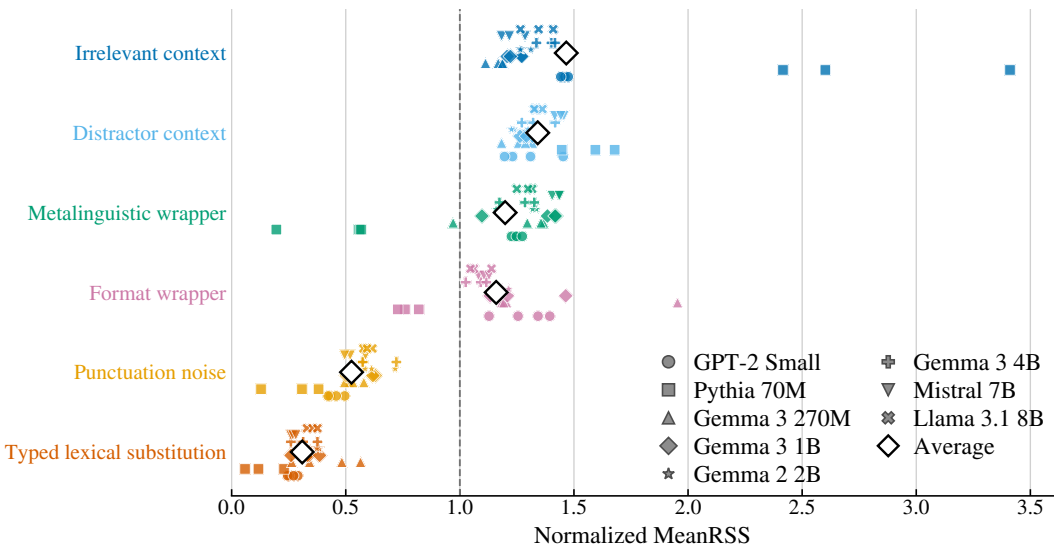


Figure 2: Normalized MeanRSS across six prompt transformations and 27 pretrained SAEs. Each SAE's six setting means are divided by their average, so the plot compares transformation profiles rather than absolute activation scales. Large outlined diamonds show averages across SAEs.

3.4 Localization, Readout Counterfactuals, and Selective Review

Average contributions do not resolve individual transformed decisions, so we evaluate the same support on each held out variant. Linearity gives the exact logit decomposition and the margin obtained by restoring selected readout inputs to their clean values:

$$z(x_v) - z(x_0) = \sum_f w_f \, \widetilde{\delta}_{v,f}, \tag{11}$$

$$m_{\mathrm{cf}}(x_v) = m_0 + y \sum_{f \notin S} w_f \widetilde{\delta}_{v,f}. \tag{12}$$

Restricting the first sum to S reconstructs the support's logit shift. For an erroneous transformed decision, $m_{\mathrm{cf}}(x_v) > 0$ means that restoring the selected readout inputs reverses the error. This counterfactual changes only the readout computation, not the language model's activations or generated behavior. For offline selective review, cases are ranked by increasing residual margin computed on selection variants and evaluated against errors on held out variants.

4 Experiments and Results

4.1 Systematic and Persistent Redistribution Across Pretrained SAEs

We first test whether the tested transformations produce a reproducible redistribution profile across pretrained SAEs. The study uses 1,000 English prompts across 12 content categories, generated once and fixed before SAE inference. The six settings contrast context and wrapper transformations, which alter surrounding material or presentation, with punctuation and lexical edits that modify the preserved request locally. We evaluate 27 pretrained SAEs spanning GPT-2, Pythia, Gemma 2 and 3, Mistral, and Llama 3.1 across multiple layers. We normalize the six MeanRSS setting means within each SAE to compare transformation profiles rather than absolute response scales. Each setting uses five operators with three realizations each; mean changes within each operator supply the selector observations, and the evaluated token remains aligned to the preserved target span. Supplementary Section S3 and Table S3 give the full protocol, selector parameters, and SAE specifications.

The normalized context and wrapper mean exceeds the local edit mean in all 27 SAEs, with an average difference of 0.874 and a 95% interval of $[0.867, 0.881]$ from bootstrap resampling of base

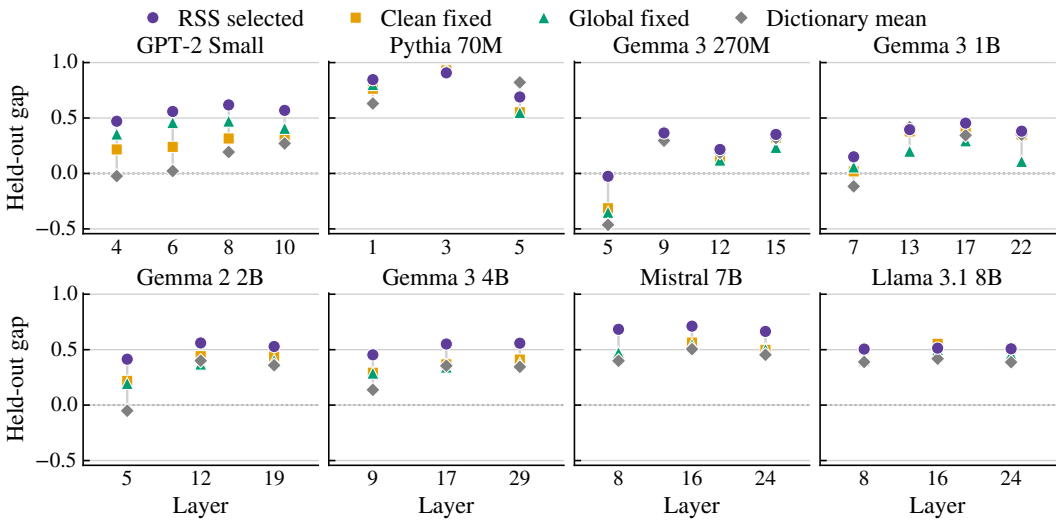


Figure 3: Response contrast on variants excluded from feature selection. Each point is the normalized context and wrapper mean minus the local edit mean for the indicated feature set and SAE layer. Clean fixed: features selected from clean prompts. Global fixed: features selected by activity across the clean corpus. Dictionary mean: equal average over all SAE features.

prompts with the SAE panel fixed. The pattern also holds across prompts: after aggregation across SAEs, the context and wrapper mean exceeds the local edit mean for all 1,000 prompts, and every context or wrapper setting exceeds both local edits for 997. Supplementary Section S4 reports support organization and robustness to feature scaling and unavailable cases. Across four intensity levels, the median prompt level Spearman correlations between intensity and MeanRSS are 0.95 for punctuation edits, 0.74 for irrelevant context, and 0.50 for the format wrapper. Supports are selected separately at each intensity, so this tests magnitude calibration rather than persistence of the same features. Supplementary Table S12 gives the intensity definitions.

We next test persistence beyond selection at two levels: whether a fixed support preserves the transformation profile on held out variants, and whether its features preserve their selected directions. Figure 3 evaluates the first question. RSS gives stronger context and wrapper versus local edit separation than each comparison method in 24–26 of 27 SAEs and gives the largest contrast in 22. The profile also remains stable under disjoint realization and operator holdouts (Supplementary Section S4). Directional consistency on the selection observations, including the finite sample guarantee in Proposition 2, does not guarantee the same direction on new variants. On held out variants, the response rate in the selected direction is 0.952 for RSS and 0.758 for random supports matched in size and candidate pool (Supplementary Tables S16–S17).

4.2 Feature Persistence and Functional Relevance in Frozen Readouts

To test whether recurring responses also affect a downstream computation, we audit two frozen linear safety readouts. The functional study uses 1,800 balanced WildGuardMix prompts (Han et al., 2024), split 1,200/300/300 into training, validation, and test sets. Probe fitting, regularization, threshold calibration, and layer selection use only clean training and validation data; the resulting readouts are frozen before transformed test evaluation. We use Gemma 3 4B layer 17 and Llama 3.1 8B layer 24, with selection and clean performance reported in Supplementary Table S18.

For each test prompt and transformation setting, one realization per operator forms V_{sel} and the other two form disjoint V_{ho}. The five selection variants determine the RSS support and ranking signals, while persistence, readout contributions, and failure are evaluated on the ten held out variants. This tests transfer to new realizations of the same operators without selecting on their outcomes. We define recurrent readout failure as at least two of the ten held out margins being negative. Requiring a correct clean prediction, available support, and complete outcomes leaves 1,524 Gemma and 1,445 Llama prompt and setting cases. Confidence intervals use bootstrap resampling clustered by

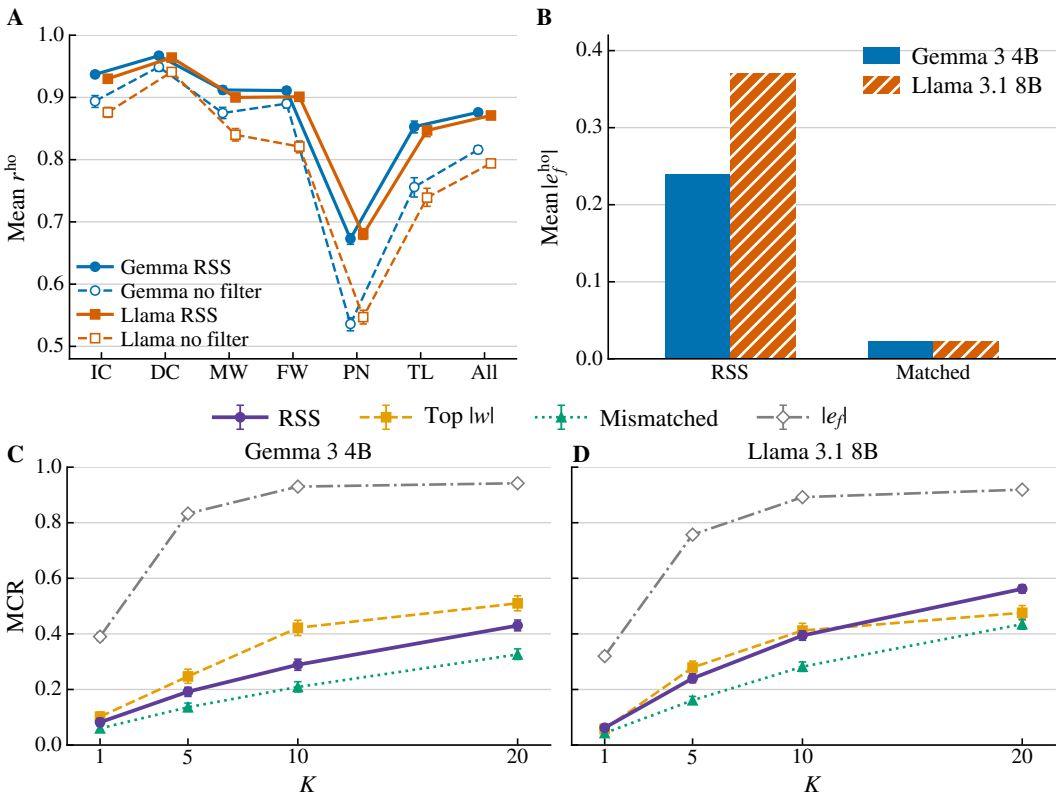


Figure 4: Feature validation on held out variants. A, response rate in the direction chosen on selection variants, comparing RSS with magnitude alone. B, mean $|e_f^{\mathrm{ho}}|$ for RSS features and unselected features matched by readout coefficient magnitude. C–D, MCR as the support cap K varies. The $|e_f|$ ranking uses contributions from the selection variants.

base prompt. Large changes might already be stable without the RSS recurrence and consistency constraints. We therefore compare RSS with magnitude-only selection matched to the RSS support size, removing the recurrence and consistency constraints. For $f \in S$, held out response in the selected direction is

$$r_f^{\mathrm{ho}} = \frac{1}{|V_{\mathrm{ho}}|} \sum_{v \in V_{\mathrm{ho}}} \mathbf{1}[s_f \, \delta_{v,f} > \tau] \, . \tag{13}$$

Mean held out response rates are 0.876 versus 0.816 for Gemma and 0.871 versus 0.794 for Llama (Figure 4A). Thus the RSS constraints improve persistence beyond magnitude selection.

Large contribution could instead reflect a large readout coefficient. After matching RSS features to unselected features by $|w_f|$, mean absolute held out contribution is 0.240 versus 0.022 for Gemma and 0.370 versus 0.022 for Llama, with RSS larger in 96.9% and 98.0% of matched pairs (Figure 4B). The sign rule in Proposition 3 agrees with the sign of the contribution averaged over held out variants in 97.4% and 97.1% of cases.

At $K = 20$, RSS retains 43.0% and 56.2% of absolute held out contribution, versus 32.6% and 43.5% for RSS supports selected for another transformation. RSS does not maximize retention: the readout aware $|e_f|$ ranking, which uses selection variant contributions, retains more than 90% in both models (Figure 4C–D). RSS therefore retains useful contributions without selecting for the readout, although selection by readout contribution retains more.

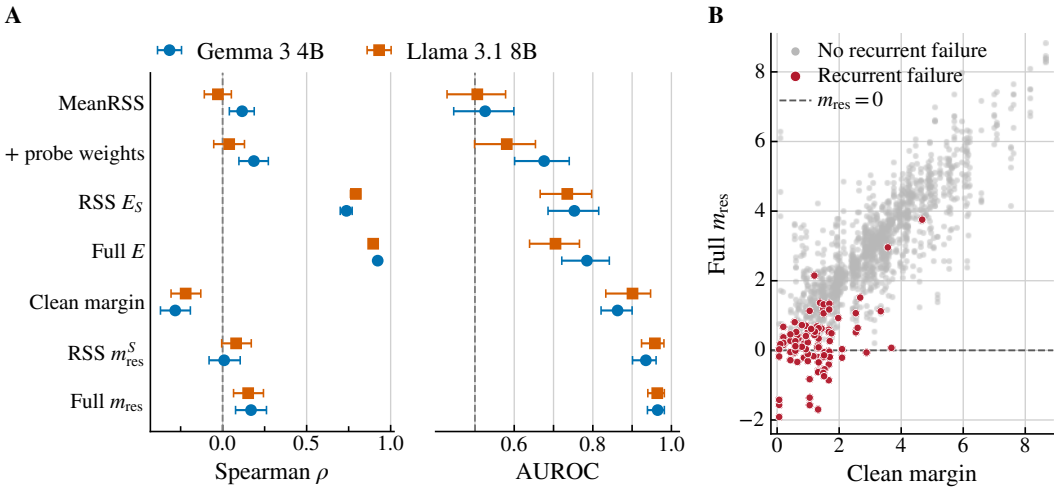


Figure 5: Readout change and recurrent readout failure on held out variants. A, Spearman correlation with held out margin erosion and AUROC for recurrent failure. Correlation measures erosion transfer, while AUROC measures failure ranking; residual margin correlations are oriented toward increasing risk. B, clean margin versus full residual margin for Gemma 3 4B, colored by held out recurrent failure. The dashed line marks zero average selection margin.

4.3 Frozen Readout Failure Depends on Alignment and Remaining Margin

A feature response can be persistent and functionally relevant without pushing the readout toward an error. We therefore compare response magnitude, weighted magnitude, signed erosion, and residual margin against held out erosion and recurrent readout failure (Figure 5A).

Although probe AUROC remains above 0.90 under every transformation, 4.2–9.5% of decisions change. MeanRSS correlates weakly with held out margin erosion, and absolute probe weighting adds only modest information. Full erosion estimated on selection variants correlates with held out erosion at 0.923 for Gemma and 0.896 for Llama. The corresponding RSS support correlations are 0.737 and 0.793 (Supplementary Table S25). Thus the selected responses retain substantial information about erosion, though less than the full readout.

Clean margin alone gives recurrent failure AUROCs of 0.863 for Gemma and 0.901 for Llama. Subtracting RSS support erosion from the full clean margin gives AUROCs of 0.935 and 0.958, using median supports of 20 and 19 features. The full residual margins give 0.965 and 0.964 (Figure 5; Supplementary Table S26). These are offline rankings using known class labels, with erosion computed on selection variants and recurrent failure measured on held out variants.

Failures concentrate at low full residual margin (Figure 5B), and every case with negative full residual margin has recurrent readout failure in this cohort. Because residual margin is computed on selection variants whereas recurrent failure is defined on held out decisions, this association reflects transfer rather than an identity.

4.4 Compact Supports Preserve Readout Change and Audit Utility

Because failure ranking also uses the full clean margin, we separately test how much readout change the selected support reconstructs on individual held out variants.

RSS reconstructs held out logit changes with R^2 of approximately 0.58 and 0.63. Restoring selected readout inputs to their clean values reverses 72.2% and 71.8% of transformed probe errors (Figure 6A–B). Mismatched supports perform worse, whereas globally unstable and random features yield R^2 of only about 0.01–0.02. These comparisons localize a substantial part of the readout change to the selected responses (Supplementary Table S29). The restoration test acts only inside the readout.

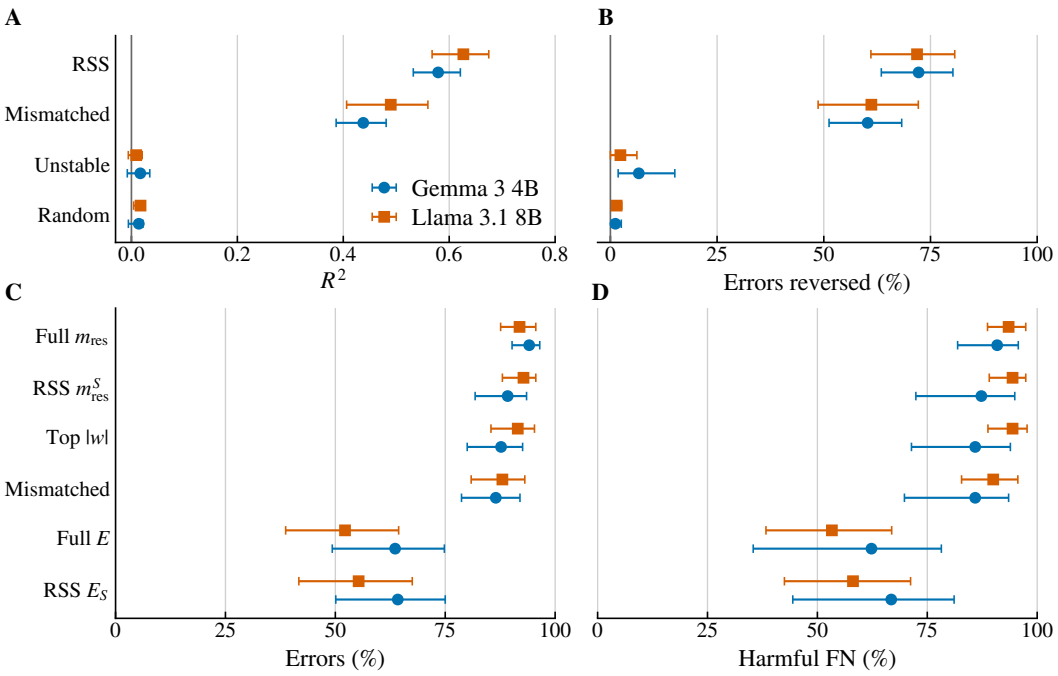


Figure 6: Utility of RSS supports selected for each prompt and transformation within the frozen readout. A, R^2 for reconstructed held out logit changes. B, readout errors reversed by restoring selected feature values. C–D, errors and harmful false negatives captured at a 20% offline review budget using known labels. Error bars show 95% bootstrap confidence intervals.

At a 20% offline review budget, the RSS residual margin captures 89.2% of Gemma errors and 92.8% of Llama errors, together with 87.3% and 94.4% of harmful false negatives (Figure 6C–D). The median supports contain 20 and 19 features, compared with 415 and 862 nonzero coefficients in the full readouts. Supplementary Tables S27 and S31 give the full comparison and uncertainty intervals. The compact object here is the set of changes used for each case, not a single reduced probe that replaces the full clean computation.

5 Discussion and Conclusion

The results establish recurring redistribution as a distinct property of sparse representations under prompt variation. Rather than asking only whether an individual feature responds or how far the representation moves in aggregate, RSS identifies sparse changes that recur across related transformations and remain observable on variants excluded from selection. The shared profile across pretrained SAEs further shows that these recurring changes can form structured patterns within a fixed sparse basis. A second implication is that recurring representation change and downstream effect are related but distinct quantities. A recurrent feature response affects a fixed readout through its response magnitude, readout weight, directional alignment, and the clean margin. This decomposition turns redistribution from a representational observation into a quantity that can be related directly to downstream computation, while explaining why large changes need not cause failure and smaller aligned changes can matter near a decision boundary.

Most importantly, the recurring changes are selected without readout information, yet they explain a substantial part of held out readout variation. Their value therefore extends beyond describing representation change: the same supports retain meaningful contribution, reconstruct logit changes, and localize changes that can alter readout decisions. This provides a direct bridge from recurrent sparse representation change to functional consequence within a fixed downstream computation. The results concern the tested transformations and frozen linear readouts, with the main functional holdout using new realizations of the same operators. Within this setting, RSS identifies recurring changes first and evaluates their downstream relevance separately.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability of the manuscript, to check the correctness and internal consistency of mathematical derivations and proofs, and to assist with experimental plotting.

REPRODUCIBILITY STATEMENT

Sections 3.1–3.4 define RSS, MeanRSS, readout contributions, residual margins, MCR, and the readout counterfactual. Sections 4.1–4.2 specify the experimental splits, comparisons, and bootstrap evaluation. Supplementary Section S1 gives the derivations, while Supplementary Section S3 specifies the transformation and alignment protocols, selector parameters, model and SAE settings, probe fitting, and threshold calibration. Supplementary Table S2 summarizes the main selection and uncertainty parameters, and Supplementary Table S3 lists the pretrained SAE specifications.

REFERENCES

Joseph Bloom. GPT2-Small-SAEs-Reformatted. Hugging Face model repository, 2024. URL https://huggingface.co/jbloom/GPT2-Small-SAEs-Reformatted.

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 Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems*, volume 38, pp. 82318–82355, 2025.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14550–14565. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.852.

Ikhyun Cho and Julia Hockenmaier. Toward efficient sparse autoencoder-guided steering for improved in-context learning in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28961–28973, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1474. URL https://aclanthology.org/2025.emnlp-main.1474/.

Agnibh Dasgupta, Abdullah Tanvir, and Xin Zhong. Invariant features in language models: Geometric characterization and model attribution. *arXiv preprint arXiv:2605.06458*, 2026.

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. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

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

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*, 2025.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track*, volume 37, pp. 8093–8131, 2024. URL https://arxiv.org/abs/2406.18495.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024. URL https://arxiv.org/abs/2410.20526.

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

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

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? a case study in sparse probing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29018–29049. PMLR, 2025.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29223–29264, 2025.

Aaron J. Li, Suraj Srinivas, Usha Bhalla, and Himabindu Lakkaraju. Evaluating adversarial robustness of concept representations in sparse autoencoders. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5940–5957. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.eacl-long.279.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. In *International Conference on Learning Representations*, pp. 33588–33636, 2025.

Callum McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, and Neel Nanda. Gemma scope 2 - technical paper. Google DeepMind Technical Report, 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.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. doi: 10.1162/tacl_a_00681.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024.

Claire Tian, Katherine Tian, and Nathan Hu. Measuring sparse autoencoder feature sensitivity. *arXiv preprint arXiv:2509.23717*, 2025. URL https://arxiv.org/abs/2509.23717.

Curt Tigges. pythia-70m-deduped__res-sm_processed. Hugging Face model repository, 2024. URL https://huggingface.co/ctigges/pythia-70m-deduped__res-sm_processed.

Aastha A K Verma, Anwoy Chatterjee, Mehak Gupta, and Tanmoy Chakraborty. Multilingual language models encode script over linguistic structure. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33685–33719, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1554. URL https://aclanthology.org/2026.acl-long.1554/.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021.

SUPPLEMENTARY MATERIAL

The supplementary material provides the proofs, definitions of plotted quantities, complete protocols, and extended representation and readout results. Table S1 summarizes the notation used throughout these analyses.

Table S1: Core notation. Quantities are defined for one pretrained SAE, layer, base prompt, and transformation setting unless stated otherwise.

Symbol	Meaning
x_0, x_v	clean prompt and its vth transformed variant
$\delta_{v,f}$	change in SAE feature f from x_0 to x_v
h_f	mean absolute feature change used for RSS magnitude ranking
S	responsive support selected by RSS
MeanRSS	average h_f over the responsive support
$\widetilde{\delta}_{v,f}$	standardized feature change used by the frozen probe
A_f^U	mean absolute standardized response of feature f on variant set U
η_f^U	directional factor on U relative to the direction selected by RSS
$m(x)$, m_0	signed margin and clean signed margin
e_f^U	feature contribution to margin erosion on variant set U
E, E_S	full and RSS support margin erosion on V_{sel}
m_{res}, m_{res}^S	full and RSS support residual margin
MCR	fraction of absolute held out feature contribution retained by S

S1 MATHEMATICAL PROPERTIES AND PROOFS

We first prove the selector properties and the readout contribution factorization. We then separate transfer error from omitted contributions and derive conditions for preserving readout decisions and average margin rankings. Selector thresholds and other protocol parameters are listed in Supplementary Table S2.

PROOF OF PROPOSITION 1

By construction, $\mathcal{E}_M$ is the set of candidate features that satisfy the recurrence and directional consistency constraints, and RSS retains the k members with largest h_f. Suppose another k-feature subset $T \subseteq \mathcal{E}_M$ had a larger value of $\sum_{f\in T} h_f$. Then at least one feature in $T \setminus S$ would have larger h_f than a feature in $S \setminus T$, contradicting the ranking rule. This proves Equation 4.

If $|\mathcal{E}_M| \geq K$, every feature outside C_M has magnitude no larger than the candidate pool cutoff, while C_M already contains at least K eligible features. No eligible feature outside C_M can therefore increase the constrained top-K objective. Ties at the candidate pool cutoff can produce additional maximizers with the same objective value.

PROOF OF PROPOSITION 2

For a selected feature,

$$d_f + o_f = n_f^+ + n_f^- = n r_f$$

and

$$\frac{d_f}{d_f + o_f} = c_f.$$

Therefore

$$\frac{d_f}{n} = r_f c_f$$

and

$$\frac{d_f - o_f}{n} = \frac{(d_f + o_f)(2c_f - 1)}{n} = r_f(2c_f - 1).$$

Using $r_f \geq \rho$ and $c_f \geq \kappa > 1/2$ gives Equation 5.

For $n = 5$ and $\rho = \kappa = 0.8$, at least four observations respond and at least 80% of the responding observations must share the dominant direction. Hence $d_f \geq 4$ and $o_f \leq 1$, so $d_f - o_f \geq 3$. Arbitrarily changing one observation can reduce this difference by at most two, leaving it positive; deleting one observation can reduce it by at most one. Neither operation can reverse the dominant direction.

Population Response Statistics

The finite sample directional margin is a property of the selected observations. Population response statistics provide a separate view of why magnitude alone does not determine directional recurrence because variation around a mean response also matters. Under a distribution of changes, let $\mu_f = \mathbb{E}[\delta_f]$ and $v_f = \mathrm{Var}(\delta_f)$. Jensen's inequality, the triangle inequality, and Cauchy–Schwarz give

$$|\mu_f| \leq \mathbb{E}|\delta_f| \leq |\mu_f| + \sqrt{v_f}. \tag{14}$$

If $|\mu_f| > \tau$, Cantelli's inequality gives

$$\Pr[\mathrm{sign}(\mu_f)\delta_f > \tau] \geq \frac{(|\mu_f| - \tau)^2}{v_f + (|\mu_f| - \tau)^2}. \tag{15}$$

These bounds relate absolute response magnitude, variation, and the probability of a response in the direction of the population mean. They do not supply a population guarantee for a direction selected from finite data and do not enter Proposition 2.

For completeness, let $a = |\mu_f| - \tau > 0$ and $X = \mathrm{sign}(\mu_f)(\delta_f - \mu_f)$. Then $\mathbb{E}X = 0$, $\mathrm{Var}(X) = v_f$, and $\mathrm{sign}(\mu_f)\delta_f \leq \tau$ implies $X \leq -a$. Cantelli's inequality gives

$$\Pr[X \leq -a] \leq \frac{v_f}{v_f + a^2},$$

which yields Equation 15.

Proof of Proposition 3

For $A_f^U > 0$, Equation 6 gives

$$s_f A_f^U \eta_f^U = s_f \left(\frac{1}{|U|} \sum_{v \in U} |\widetilde{\delta}_{v,f}| \right) \frac{s_f \sum_{v \in U} \widetilde{\delta}_{v,f}}{\sum_{v \in U} |\widetilde{\delta}_{v,f}|} = \frac{1}{|U|} \sum_{v \in U} \widetilde{\delta}_{v,f},$$

because $s_f^2 = 1$. Substitution into Equation 6 gives

$$e_f^U = -y w_f s_f A_f^U \eta_f^U.$$

The triangle inequality gives

$$|\eta_f^U| = \frac{\left| \sum_{v \in U} \widetilde{\delta}_{v,f} \right|}{\sum_{v \in U} |\widetilde{\delta}_{v,f}|} \leq 1,$$

so

$$|e_f^U| = |w_f| A_f^U |\eta_f^U| \leq |w_f| A_f^U.$$

When the contribution is nonzero, its sign differs from $\mathrm{sign}(-y w_f s_f)$ only through the sign of η_f^U, proving the final statement. The zero-response case follows from the definitions $A_f^U = \eta_f^U = e_f^U = 0$.

Held Out Support Fidelity and Preservation

The main text separates held out residual-margin error into transfer of the selected support and signed contribution omitted from it. We first make the role of cancellation explicit. Let

$$E^{\mathrm{ho}} = \sum_f e_f^{\mathrm{ho}}, \qquad E_S^{\mathrm{ho}} = \sum_{f \in S} e_f^{\mathrm{ho}},$$

and define the omitted absolute mass and cancellation factor

$$L_S^{\mathrm{ho}} = \sum_{f \notin S} |e_f^{\mathrm{ho}}|, \qquad \chi_S^{\mathrm{ho}} = \frac{\left|\sum_{f \notin S} e_f^{\mathrm{ho}}\right|}{L_S^{\mathrm{ho}}}. \tag{16}$$

We set $\chi_S^{\mathrm{ho}} = 0$ when $L_S^{\mathrm{ho}} = 0$. By construction, $0 \le \chi_S^{\mathrm{ho}} \le 1$. Values near one indicate little cancellation among omitted contributions, while values near zero indicate strong cancellation. The following identities involving MCR apply when the total absolute held out contribution is nonzero, as required by its definition.

Proposition 4 (Held out support fidelity) *The signed contribution omitted from S satisfies the exact identity*

$$\left|E^{\mathrm{ho}} - E_S^{\mathrm{ho}}\right| = \chi_S^{\mathrm{ho}}(1 - \mathrm{MCR}) \sum_f |e_f^{\mathrm{ho}}|. \tag{17}$$

Moreover,

$$\left|\bar{m}_{\mathrm{ho}} - m_{\mathrm{res}}^S\right| \le \left|E_S^{\mathrm{ho}} - E_S\right| + \chi_S^{\mathrm{ho}}(1 - \mathrm{MCR}) \sum_f |e_f^{\mathrm{ho}}|. \tag{18}$$

The support transfer term can further be bounded as

$$\left|E_S^{\mathrm{ho}} - E_S\right| \le \sum_{f \in S} |w_f| \left(\left|A_f^{\mathrm{ho}} - A_f^{\mathrm{sel}}\right| + A_f^{\mathrm{sel}} \left|\eta_f^{\mathrm{ho}} - \eta_f^{\mathrm{sel}}\right|\right), \tag{19}$$

where $A_f^{\mathrm{sel}} = A_f^{V_{\mathrm{sel}}}$, $A_f^{\mathrm{ho}} = A_f^{V_{\mathrm{ho}}}$ and likewise for η_f.

Proof. By definition,

$$E^{\mathrm{ho}} - E_S^{\mathrm{ho}} = \sum_{f \notin S} e_f^{\mathrm{ho}},$$

so Equation 16 gives

$$\left|E^{\mathrm{ho}} - E_S^{\mathrm{ho}}\right| = \chi_S^{\mathrm{ho}} L_S^{\mathrm{ho}}.$$

Equation 10 implies

$$L_S^{\mathrm{ho}} = (1 - \mathrm{MCR}) \sum_f |e_f^{\mathrm{ho}}|,$$

which proves Equation 17. Since $\bar{m}_{\mathrm{ho}} = m_0 - E^{\mathrm{ho}}$ and $m_{\mathrm{res}}^S = m_0 - E_S$,

$$\bar{m}_{\mathrm{ho}} - m_{\mathrm{res}}^S = (E_S - E_S^{\mathrm{ho}}) + (E_S^{\mathrm{ho}} - E^{\mathrm{ho}}).$$

The triangle inequality together with Equation 17 proves Equation 18.

For the transfer term, Proposition 3 gives

$$E_S^{\mathrm{ho}} - E_S = -y \sum_{f \in S} w_f s_f \left(A_f^{\mathrm{ho}} \eta_f^{\mathrm{ho}} - A_f^{\mathrm{sel}} \eta_f^{\mathrm{sel}}\right).$$

For each selected feature,

$$\left|A_f^{\mathrm{ho}} \eta_f^{\mathrm{ho}} - A_f^{\mathrm{sel}} \eta_f^{\mathrm{sel}}\right| \le \left|A_f^{\mathrm{ho}} - A_f^{\mathrm{sel}}\right| + A_f^{\mathrm{sel}} \left|\eta_f^{\mathrm{ho}} - \eta_f^{\mathrm{sel}}\right|,$$

because $|\eta_f^{\mathrm{ho}}| \le 1$. Summing after multiplication by $|w_f|$ proves Equation 19.

Corollary 1 (Worst case accounting bounds) *Since $\chi_S^{\mathrm{ho}} \le 1$,*

$$\left|E^{\mathrm{ho}} - E_S^{\mathrm{ho}}\right| \le (1 - \mathrm{MCR}) \sum_f |e_f^{\mathrm{ho}}|. \tag{20}$$

On the selection variants,

$$\left|m_{\mathrm{res}} - m_{\mathrm{res}}^S\right| = \left|\sum_{f \notin S} e_f\right| \le \sum_{f \notin S} |e_f|. \tag{21}$$

MCR controls omitted absolute contribution, whereas signed fidelity also depends on cancellation. Setting the cancellation factor to one gives the worst case bound, which does not use any cancellation among omitted features.

Preserving an average margin does not by itself preserve every individual decision. For a held out variant v, compare the omitted signed contribution with the margin reconstructed from selected changes. Define

$$\widehat{m}_v^S = m_0 + y \sum_{f \in S} w_f \widetilde{\delta}_{v,f}. \tag{22}$$

Corollary 2 (Readout decision preservation) *If*

$$\left| \sum_{f \notin S} y w_f \widetilde{\delta}_{v,f} \right| < |\widehat{m}_v^S|, \tag{23}$$

then

$$\operatorname{sign} m(x_v) = \operatorname{sign} \widehat{m}_v^S.$$

Proof. Linearity gives

$$m(x_v) = \widehat{m}_v^S + y \sum_{f \notin S} w_f \widetilde{\delta}_{v,f}.$$

Under Equation 23, the omitted signed contribution is too small to cross zero, so the two margins have the same sign.

A sufficient condition for preserving an ordering between cases is that their estimated margins differ by more than their possible transfer and omission errors. For each case i, define the held out fidelity radius

$$B_i = \left| E_{S_i}^{\mathrm{ho}} - E_{S_i} \right| + \chi_{S_i}^{\mathrm{ho}} (1 - \mathrm{MCR}_i) \sum_f |e_{i,f}^{\mathrm{ho}}|. \tag{24}$$

Proposition 4 gives $|\bar{m}_{\mathrm{ho},i} - m_{\mathrm{res},i}^{S_i}| \le B_i$.

Corollary 3 (Held out ranking preservation) *For cases i and j,*

$$m_{\mathrm{res},j}^{S_j} - m_{\mathrm{res},i}^{S_i} > B_i + B_j \quad \Longrightarrow \quad \bar{m}_{\mathrm{ho},j} > \bar{m}_{\mathrm{ho},i}. \tag{25}$$

Proof. The fidelity radii imply

$$\bar{m}_{\mathrm{ho},i} \le m_{\mathrm{res},i}^{S_i} + B_i, \qquad \bar{m}_{\mathrm{ho},j} \ge m_{\mathrm{res},j}^{S_j} - B_j.$$

Under Equation 25, the second lower bound exceeds the first upper bound, proving the result. This condition preserves the ordering of average held out margins, not the number of individual variants with negative margin.

Dilution Across the Full Dictionary

Suppose k of F dictionary features respond with average magnitude a, while the remaining features have negligible change. The mean over the full dictionary is approximately $(k/F)a$. When $k \ll F$, averaging over the full dictionary reduces the response by the factor k/F, while averaging over the responsive support preserves the scale a.

Exact Readout Decomposition

For standardized activations,

$$z(x_v) - z(x_0) = \mathbf{w}^\top \left[\widetilde{\mathbf{a}}(x_v) - \widetilde{\mathbf{a}}(x_0) \right] = \sum_f w_f \, \widetilde{\delta}_{v,f}. \tag{26}$$

Restricting the final sum to $f \in S$ gives the reconstruction used in the localization analysis. Replacing the selected transformed activations by their clean values subtracts $\sum_{f \in S} w_f \widetilde{\delta}_{v,f}$ from the transformed logit. Multiplying by the label sign and adding the clean margin yields Equation 12. Thus the readout counterfactual reverses a transformed probe error exactly when the contribution remaining outside S leaves positive signed margin. The language model itself is unchanged.

S2 Definitions of Plotted Quantities

This section specifies the quantities plotted in the main figures, separating representation profiles, feature contributions, and readout outcomes. All averages use available cases only. For Figures 2 and 3, quantities are defined within one SAE before aggregation across SAEs.

To compare response profiles across SAEs, normalize each setting mean by the average across the six settings. For transformation setting g, let

$$R_g = \mathrm{mean}_i \ \mathrm{MeanRSS}_{i,g}, \qquad \widetilde{R}_g = \frac{R_g}{\frac{1}{6}\sum_{g'} R_{g'}}, \tag{27}$$

where i indexes prompts. Figure 2 plots $\widetilde{R}_g$ for each SAE. Thus the six values for each SAE have mean one.

For the transfer comparison, evaluate each selected or comparison feature set on held out variants before computing the transformation contrast. For a feature set A, define its held out response under setting g as

$$H_g(A) = \mathrm{mean}_i \ \mathrm{mean}_{f \in A_{i,g}} \ \mathrm{mean}_{v \in V_{\mathrm{ho}}} |\delta_{v,f}|. \tag{28}$$

We normalize $H_g(A)$ across the six settings in the same way as Equation 27, giving $\widetilde{H}_g(A)$. Let $\mathcal{C}$ denote the four context and wrapper settings and $\mathcal{L}$ the two local edits. The ordinate in Figure 3 is

$$\Delta(A) = \frac{1}{4}\sum_{g \in \mathcal{C}} \widetilde{H}_g(A) - \frac{1}{2}\sum_{g \in \mathcal{L}} \widetilde{H}_g(A). \tag{29}$$

Figure 4A reports the cohort mean of the held out response rate in the selected direction, r_f^{ho}, from Equation 13. Panel B reports the cohort mean of $|e_f^{\mathrm{ho}}|$ for RSS features and features matched by coefficient magnitude. Panels C–D use MCR from Equation 10.

To evaluate transfer of the selection variant erosion estimates, define full held out erosion as

$$E^{\mathrm{ho}} = \sum_f e_f^{\mathrm{ho}} = m_0 - \frac{1}{|V_{\mathrm{ho}}|}\sum_{v \in V_{\mathrm{ho}}} m(x_v). \tag{30}$$

Figure 5A reports Spearman correlation with E^{ho} and AUROC for recurrent failure. For residual-margin signals, the correlation sign is reported in the direction of increasing risk.

Localization compares the full logit change with the sum over the evaluated support. For each held out variant, let

$$d = z(x_v) - z(x_0), \qquad \widehat{d}_A = \sum_{f \in A} w_f \widetilde{\delta}_{v,f}.$$

Figure 6A reports the standard coefficient of determination R^2 of $\widehat{d}_A$ against d over the evaluated held out rows. Panel B reports the fraction of transformed probe errors ($m(x_v) < 0$) for which the corresponding readout-counterfactual margin becomes positive. Panels C–D report the fractions of all held out probe errors and harmful false negatives, respectively, contained in the highest-risk 20% of cases under each ranking signal.

S3 Frozen Protocols and Reproducibility

The protocols specify the input transformations and target alignment before SAE analysis, then separate clean probe fitting from transformed evaluation. Table S2 lists the selection and uncertainty parameters.

Representation Corpus Protocol

In the separate five-variant analysis, all 90,000 expected prompt, SAE, and setting cases across the smaller four models are evaluated. Eighteen cases (0.02%), all under the two local edit settings, have no eligible feature and remain unavailable rather than assigned zero. This case count is distinct from the number of transformed input texts in the main protocol with 15 variants.

Table S2: Selection and uncertainty parameters. The main representation protocol has five operator observations derived from fifteen variants per setting. The readout audit selects from five variants and evaluates on ten excluded variants.

Parameter	Value
Selector observations, n	5
Response rate threshold, ρ	0.8
Directional consistency threshold, κ	0.8
Response threshold, τ	10^{-5}
Candidate features, M	400
Support cap, K	20
Target token policy	final non special in target span
Bootstrap resamples	5,000

Table S3 lists the checkpoint and architecture details for all 27 pretrained SAEs. The GPT-2 and Pythia SAEs use released checkpoints from Bloom (2024); Tigges (2024); the Gemma 2 and Gemma 3 checkpoints are from Gemma Scope (Lieberum et al., 2024) and Gemma Scope 2 (McDougall et al., 2025), respectively; the Mistral checkpoints use the residual SAEs from Engels et al. (2025), and the Llama checkpoints are from Llama Scope (He et al., 2024).

Table S3: Pretrained SAE specifications. "n.r." means not reported.

Model	Layers	SAE checkpoint	Width	Hook	Architecture / sparsity
GPT-2 Small	4, 6, 8, 10	`GPT2-Small-SAEs-Reformatted`	24,576	`resid_pre`	ReLU, 32×; L1 coef. 8×10^{-5}
Pythia-70M	1, 3, 5	`pythia-70m-deduped__res-sm`	32,768	`resid_post`	ReLU, standard; sparsity n.r.
Gemma 3 270M	5, 9, 12, 15	`gemma-scope-2-270m-pt`	16,384	`resid_post`	JumpReLU; $L_0 = 18$ at layer 5 and 20 elsewhere
Gemma 3 1B	7, 13, 17, 22	`gemma-scope-2-1b-pt`	16,384	`resid_post`	JumpReLU; $L_0 = 18$ at layer 7 and 20 elsewhere
Gemma 3 4B	9, 17, 29	`gemma-scope-2-4b-pt`	65,536	`resid_post`	JumpReLU; nominal $L_0 = 53, 60, 60$
Mistral 7B	8, 16, 24	`mistral_7b_layer_L`	65,536	`resid_pre`	ReLU, 16×; measured $L_0 = 92, 74, 75$
Llama 3.1 8B	8, 16, 24	`Llama3.1-L{8,16,24}R-8x`	32,768	`resid_post`	JumpReLU, 8×; measured $L_0 = 50$
Gemma 2 2B	5, 12, 19	`gemma-scope-2b-pt-res width_65k`	65,536	`resid_post`	JumpReLU; measured $L_0 = 53, 72, 63$

SAE Specifications

The four larger base models, which contribute twelve of these SAEs, use bfloat16 precision without quantization, while their SAEs use float32. Checkpoint preprocessing is unchanged. Gemma 3 4B and Gemma 2 2B use none, Mistral uses per token constant norm rescaling, and Llama Scope uses checkpoint specific rescaling using dataset statistics. These settings follow the released checkpoints and are not used as a common normalization across models.

Prompt Provenance and Transformations

The 1,000 clean prompts are generated once, before any SAE inference, using GPT-5.4 snapshot `gpt-5.4-2026-03-05` at temperature 0.6 with fixed acceptance criteria for length (5–30 words), naturalness, plausibility, category match, and duplication against previously generated prompts.

Irrelevant context prepends topically unrelated benign material. Distractor context prepends a benign competing task while retaining the original request as operative. Metalinguistic wrappers alter response mode or presentation. Format wrappers place the request in structured envelopes. Punctuation noise inserts local punctuation without changing the lexical sequence. Typed lexical substitution

replaces noncritical words from a fixed type compatible pool. The preserved target phrase appears exactly once in every valid variant.

TARGET TOKEN ALIGNMENT

Every prompt records the character span of the preserved target. For each clean prompt and variant, the evaluated position is the final non special token overlapping that span. Prefixes and wrappers can shift the token index, but the evaluated surface span remains fixed. No mapping failures are observed across all 90,000 cases in the five-variant analysis.

HANDLING REPEATED REALIZATIONS

Section 3.1 gives the generic formulation. In the analysis using 15 variants, each case contains five transformation operators with three realizations per operator. Exact duplicate texts are collapsed within each operator before selection. Let U_o be the set of unique texts for operator o and $\delta_{t,f}$ the feature change for text t. The mean change for operator o is

$$\bar{\delta}_{o,f} = \frac{1}{|U_o|} \sum_{t \in U_o} \delta_{t,f}, \tag{31}$$

and the candidate magnitude is

$$h_f = \frac{1}{5} \sum_{o} \left[\frac{1}{|U_o|} \sum_{t \in U_o} |\delta_{t,f}| \right]. \tag{32}$$

An operator responds when $|\bar{\delta}_{o,f}| > 10^{-5}$. A feature is eligible when it responds in at least four of five operators, the dominant direction share is at least 0.8, and it lies in the top $M = 400$ by h_f. MeanRSS averages the absolute changes over operators and the selected support,

$$\mathrm{MeanRSS} = \frac{1}{5|S|} \sum_{o} \sum_{f \in S} \left[\frac{1}{|U_o|} \sum_{t \in U_o} |\delta_{t,f}| \right], \tag{33}$$

so that repeated realizations of the same text are collapsed before averaging. This expression reduces to Equation 3 when every operator has the same number of unique realizations. If duplicate removal leaves different numbers of unique realizations, the expression keeps the five operators equally weighted rather than weighting them by the number of remaining texts. The table of selected supports is available for 161,996 of 162,000 cases; unavailable cases remain missing.

SAFETY CORPUS AND PROBE TRAINING

The safety corpus contains 900 harmful and 900 benign WildGuardMix prompts. Exact and near duplicates are grouped before the balanced split, and every prompt's transformed variants remain in the same split. For each feature, the scale estimated from clean training data is

$$\sigma_f = \sqrt{\frac{1}{N_{\mathrm{tr}}} \sum_{i=1}^{N_{\mathrm{tr}}} a_f(x_i)^2 + 10^{-12}}, \tag{34}$$

and features with $\sigma_f < 10^{-8}$ are removed. The retained activations are standardized as $\widetilde{a}_f(x) = a_f(x)/\sigma_f$. These standardized clean training activations feed an elastic-net logistic probe with objective

$$\frac{1}{N} \sum_{i=1}^{N} \log\bigl(1 + \exp[-y_i(\mathbf{w}^\top \widetilde{\mathbf{a}}_i + b)]\bigr) + \lambda \left[0.9\|\mathbf{w}\|_1 + \frac{0.1}{2}\|\mathbf{w}\|_2^2 \right]. \tag{35}$$

The intercept is unpenalized. Clean validation AUROC selects λ, with a sparsity tie break inside a tolerance of 0.002. A threshold calibrated on clean benign validation logits targets at most 5% empirical false-positive rate. No transformed or test outcome enters unit, regularization, or threshold selection.

S4 Extended RSS Atlas and Calibration

Profiles Across Models and Layers

Figure S1 resolves the aggregate transformation profile by model and layer. It tests whether the context and wrapper versus local edit separation is shared across the sampled SAEs rather than driven by a subset of layers.

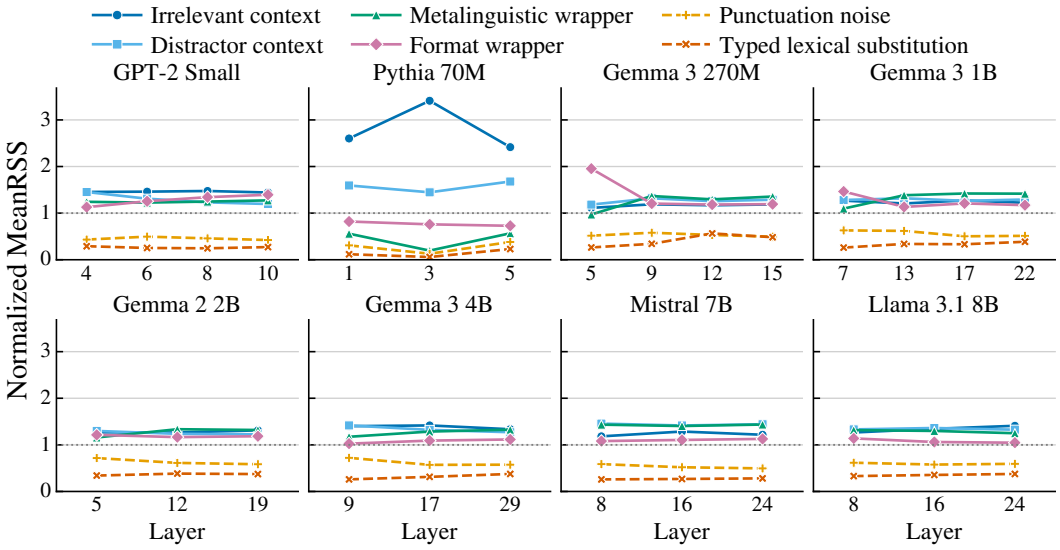


Figure S1: Normalized MeanRSS profiles for the six transformations across all sampled layers of the eight models. Normalization follows Figure 2.

Support Organization

Similar response magnitudes need not involve the same features. Figure S2 therefore compares selected feature identities across transformation settings using Jaccard overlap.

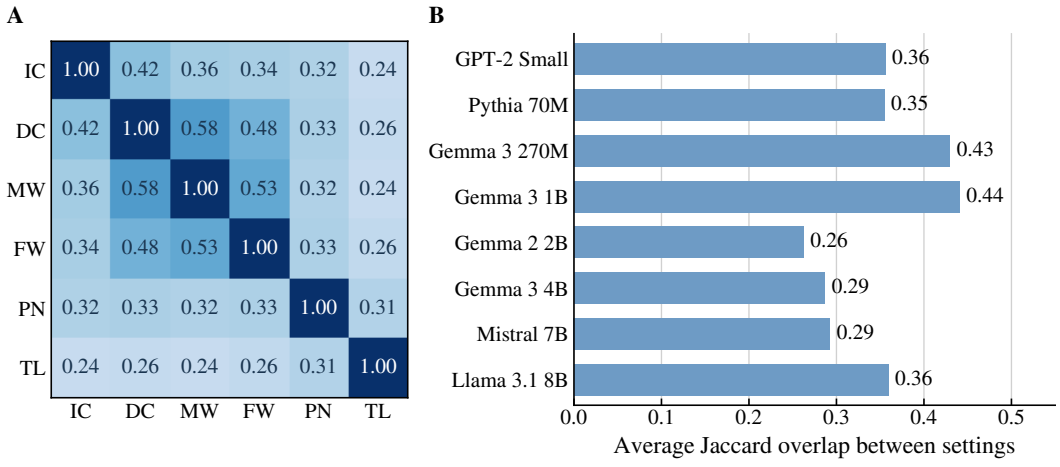


Figure S2: Overlap between the feature supports selected under different transformations, averaged over all units and 1,000 prompts. A, pairwise Jaccard overlap between transformations; B, averages within each model. IC denotes irrelevant context, DC distractor context, MW metalinguistic wrapper, FW format wrapper, PN punctuation noise, and TL typed lexical substitution.

SUPPORT AVAILABILITY AND ACTIVE FEATURE CHANGES

Tables S4 and S6 report support availability and the changes in active SAE coordinates between clean and transformed prompts. Availability is nearly complete across all settings. Context and wrapper transformations also change the active feature set more strongly than the two local edits.

Table S4: Support availability and size across transformation settings. Each setting contains 27,000 expected groups.

Setting	Available	Mean $\|S\|$	Median $\|S\|$	$\|S\| = 20$
Irrelevant context	27,000	19.36	20	90.3%
Distractor context	27,000	18.77	20	82.1%
Metalinguistic wrapper	27,000	19.06	20	86.6%
Format wrapper	27,000	18.56	20	82.1%
Punctuation noise	27,000	18.01	20	71.1%
Typed lexical substitution	26,996	16.18	20	54.6%

Table S5 shows that the treatment of the four unavailable groups does not change the aggregate comparison.

Table S5: Sensitivity to the treatment of unavailable groups. Gap is the context and wrapper mean minus the local-edit mean; brackets give 95% confidence intervals.

Treatment	Gap	95% CI	Order
Omitted	0.8740	[0.8671, 0.8810]	Reference
Assigned zero	0.8741	[0.8673, 0.8810]	Unchanged

For this analysis an active coordinate is a nonzero entry in the frozen sparse SAE encoding. For a clean prompt and one transformed prompt, let A_0 and A_v denote their sets of active coordinates. We report $|A_0 \cap A_v|/|A_0 \cup A_v|$ for overlap, $|A_v \setminus A_0|/|A_v|$ for the newly active fraction, and $|A_0 \setminus A_v|/|A_0|$ for the inactive fraction. When both sets are empty, overlap is undefined. When the transformed set is empty, the newly active fraction is undefined, and when the clean set is empty, the inactive fraction is undefined. Within each operator, exact repeated texts are collapsed before averaging; unique realizations are equally weighted and the five operators are then equally weighted. Table S6 reports these quantities together with the positive and negative shares of absolute activation change restricted to the RSS support. The two shares partition the support-level absolute change whenever that change is nonzero.

Table S6: Changes in active SAE coordinates between clean and transformed prompts. Overlap is Jaccard similarity, newly active is normalized by the transformed active set, and inactive is normalized by the clean active set. Positive and negative shares partition the absolute activation change within the RSS support.

Setting	Overlap	Newly active	Inactive	Positive	Negative
Irrelevant context	0.383	0.462	0.469	0.368	0.632
Distractor context	0.361	0.490	0.487	0.304	0.696
Metalinguistic wrapper	0.394	0.453	0.451	0.375	0.625
Format wrapper	0.381	0.496	0.478	0.230	0.770
Punctuation noise	0.692	0.198	0.187	0.406	0.594
Typed lexical substitution	0.776	0.133	0.133	0.381	0.619

An alternative selector applies the rule in Section 3.1 directly to all 15 variants without grouping them by operator. It produces mean off-diagonal Jaccard 0.316, compared with 0.354 for the main selector. Across the same SAE and transformation combinations, the two selectors have a Spearman correlation of 0.986 for normalized MeanRSS and a Jaccard overlap of 0.698 for supports selected under the same transformation. The transformation profiles are therefore nearly identical even though the exact selected features differ.

Content Categories and Model Composition

To check whether prompt composition or a particular model drives the profile, Figure S3 separates the results by content category, and Figure S4 removes each model in turn. Table S7 gives the aggregate normalized profile, and Table S8 reports consistency across prompts after aggregation across SAEs.

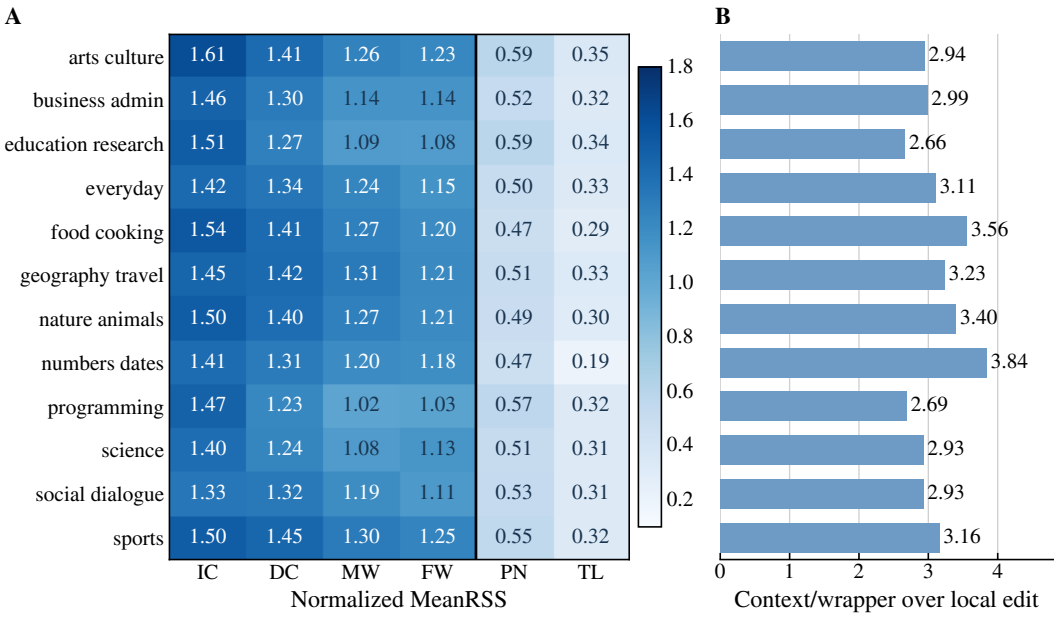


Figure S3: Normalized MeanRSS by prompt topic. A, topic level means for each transformation; B, the ratio of the context and wrapper mean to the local edit mean for each topic. Transformation abbreviations as in Figure S2.

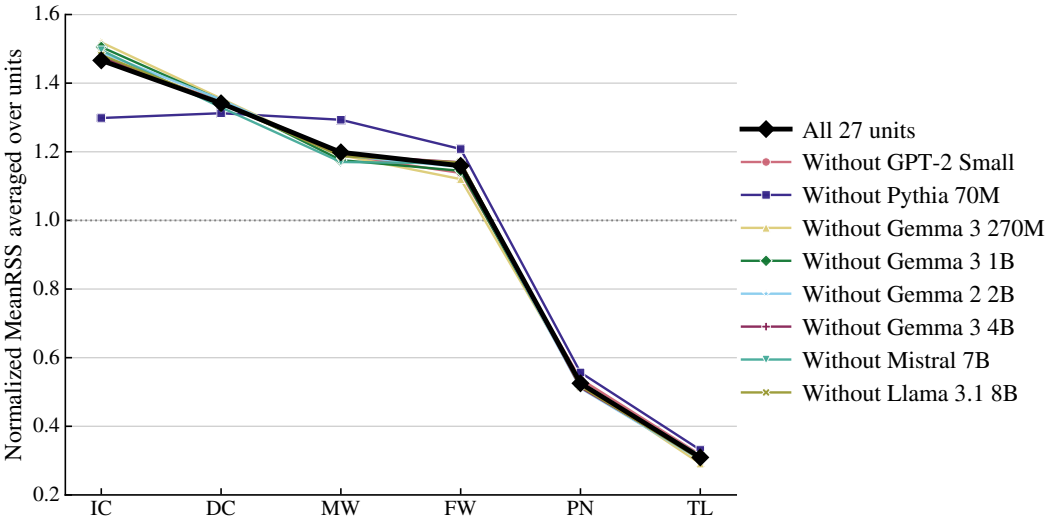


Figure S4: Profile of the six transformations after removing one model at a time, compared with the profile over all units. Transformation abbreviations as in Figure S2.

The prompt counts in Table S8 distinguish a positive mean group difference from the stronger condition that every context or wrapper setting exceeds both local edits.

Table S7: Aggregate normalized MeanRSS profile with equal weight for each SAE. Within each SAE, the six setting means are divided by their average. Brackets give 95% confidence intervals.

Setting	Normalized MeanRSS
Irrelevant context	1.466 [1.459, 1.474]
Distractor context	1.341 [1.337, 1.346]
Metalinguistic wrapper	1.199 [1.192, 1.205]
Format wrapper	1.159 [1.154, 1.164]
Punctuation noise	0.525 [0.519, 0.532]
Typed lexical substitution	0.309 [0.305, 0.314]

Table S8: Prompt-level consistency over the 1,000 representation-corpus prompts.

Criterion	Prompts	Share
Context/wrapper mean > local-edit mean	1,000 / 1,000	100.0%
Every context/wrapper setting > both local edits	997 / 1,000	99.7%

Selector Diagnostics

Figure S5 compares the main selector, which uses operator means for recurrence and direction, with selection from all fifteen variants individually. Support sizes and overlaps show how the selected features differ, while the normalized MeanRSS comparison tests whether those differences change the transformation profile.

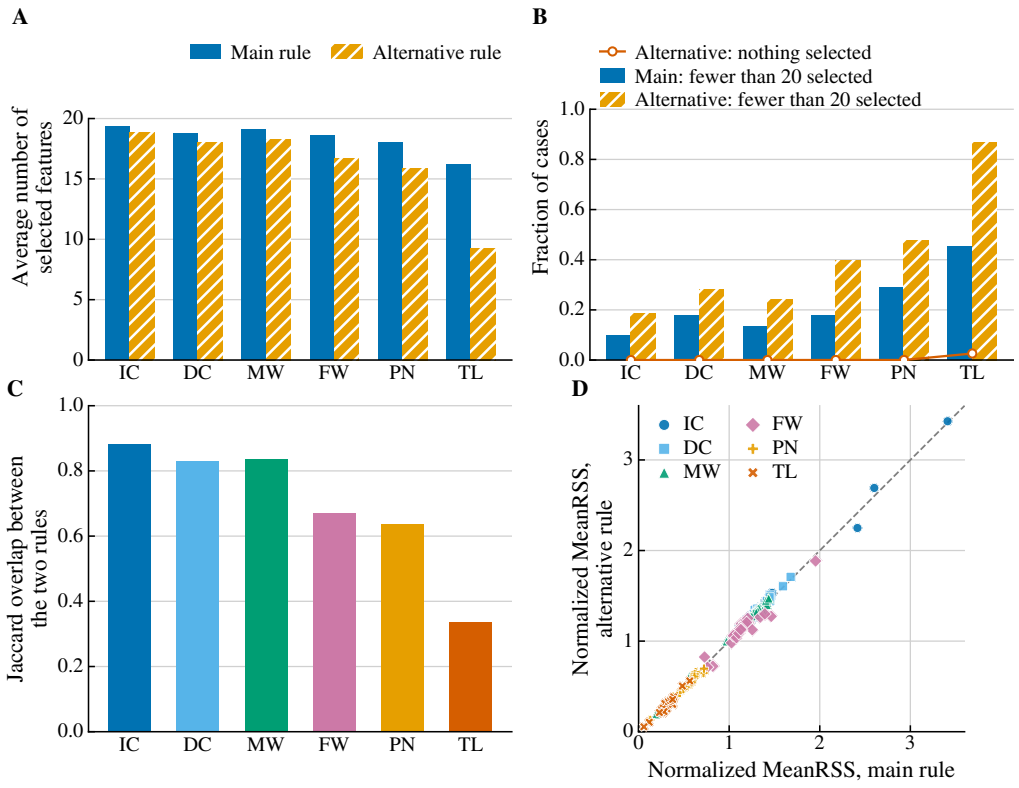


Figure S5: Comparison of the main feature selection rule with an alternative rule that treats all fifteen variants individually. A, mean support size; B, rates of undersized and empty selections; C, Jaccard overlap between the supports chosen by the two rules; D, unit level normalized MeanRSS under the two rules. Transformation abbreviations as in Figure S2.

The main selector is available for 161,996 of 162,000 cases, with mean selected count 18.32 and directional consistency 0.959 weighted by the selected features. The alternative selector has mean selected count 16.18 and directional consistency 0.944.

Selector and Summary Robustness

Table S9 varies selector parameters to test whether the profile depends on the default recurrence, consistency, or support cap. Table S10 tests a different dependency by comparing 25 combinations of magnitude summaries and normalization schemes.

Feature scales can also affect magnitude ranking. We therefore repeat RSS after scaling each feature change by its activation scale on the 1,000 clean representation prompts. The scale is the root mean square activation on these prompts; features below 10^{-8} are excluded, and retained changes are divided by the same scale with 10^{-12} added inside the square root. Table S11 reports the resulting profile for $\tau \in \{0.001, 0.01, 0.05, 0.1\}$.

Table S9: Selector parameter ablation on a stratified prompt pool. The reference uses $\rho = \kappa = 0.8$ and $K = 20$; other rows change one parameter. Gap is the context and wrapper mean minus the local-edit mean, with 95% confidence intervals in brackets.

Setting	Spearman	Minimum	Gap	Mean $\lvert S\rvert$	Jaccard
Reference	1.00	1.00	0.626 [0.604, 0.646]	13.2	1.00
$\rho = 0.6$	1.00	0.94	0.632 [0.610, 0.652]	13.8	0.90
$\kappa = 0.6$	1.00	0.77	0.595 [0.571, 0.617]	17.4	0.61
$K = 10$	1.00	0.89	0.702 [0.680, 0.723]	8.6	0.75
$K = 50$	1.00	0.94	0.537 [0.514, 0.558]	17.5	0.89

The robustness counts distinguish preservation of the complete six-setting order from preservation of only its top or bottom setting. They should not be read as identical evidence for all three comparisons.

Table S10: Transformation-ranking robustness across 25 combinations of magnitude summaries and normalization schemes.

Criterion	Specifications
Full six-setting order preserved	17 / 25
Top setting preserved	20 / 25
Bottom setting preserved	25 / 25

Table S11: RSS after scaling feature changes using clean-prompt activations. Profile Spearman is relative to the main result; Gap is the context and wrapper mean minus the local-edit mean.

τ	Available	Profile Spearman	Gap	Positive units
0.001	99.99%	0.943	0.835	27/27
0.01	99.94%	0.943	0.817	27/27
0.05	99.65%	0.943	0.756	27/27
0.1	99.40%	0.943	0.706	27/27

Intensity Protocol

Table S12 defines the four levels for each intensity series and reports the realized token and edit counts. The levels increase transformation strength within a setting, rather than place different transformation settings on a common numerical scale.

Support selection is performed independently at every level, so the support is selected separately for each prompt and intensity level. Rank correlations are computed within each prompt, SAE, and transformation before averaging equally across SAEs.

Table S12: Intensity-level definitions and realized values. Tokens and edits are mean counts.

Setting and level	Definition	Tokens	Edits
Irrelevant context 1	4–8 added tokens	4.00	0
Irrelevant context 2	12–16 added tokens	12.00	0
Irrelevant context 3	24–32 added tokens	24.00	0
Irrelevant context 4	48–64 added tokens	48.00	0
Format wrapper 1	single label or quote	3.24	10.4
Format wrapper 2	envelope with two fields	8.92	28.2
Format wrapper 3	nested wrapper	15.96	44.2
Format wrapper 4	nested multi-field wrapper	28.96	85.4
Punctuation edits 1	exactly 1 edit	1.00	1
Punctuation edits 2	exactly 2 edits	1.99	2
Punctuation edits 3	exactly 4 edits	3.99	4
Punctuation edits 4	exactly 8 edits	7.96	8

Selection–Evaluation Split Sensitivity

We vary how the fixed total of 15 variants is divided between selection and evaluation. This tests sensitivity to the selection and evaluation allocation, rather than a monotone increase in total sample size. Table S13 reports reliability, profile stability, group sign stability, and support overlap. A larger selection set necessarily leaves fewer variants for evaluation.

Table S13: Split-half reliability by selection size. Brackets give 95% confidence intervals. Profile is stability against the 15-variant reference; Sign is transformation-group sign stability; Jac. is split-to-split support Jaccard.

Sel.	Eval.	Reliability	Profile	Sign	CV	Jac.
3	12	0.773 [0.763, 0.782]	0.805	0.960	0.221	0.393
5	10	0.802 [0.792, 0.810]	0.827	0.964	0.189	0.539
8	7	0.714 [0.703, 0.724]	0.773	0.946	0.268	0.527
10	5	0.681 [0.670, 0.691]	0.765	0.946	0.279	0.654

Table S14 compares selection methods at the same selection and evaluation sizes. RSS need not maximize every measure. For example, the dictionary mean has higher reported reliability with three selection variants, whereas RSS has higher group sign stability.

Table S14: Reliability by selection method and selection size. Column definitions follow Table S13.

Method	Sel.	Reliability	Profile	Sign
RSS support	3	0.773	0.805	0.959
RSS support	5	0.802	0.827	0.964
RSS support	8	0.714	0.773	0.946
RSS support	10	0.681	0.765	0.946
Clean support	3	0.772	0.809	0.921
Clean support	5	0.752	0.792	0.918
Clean support	8	0.619	0.709	0.889
Clean support	10	0.568	0.679	0.886
Global support	3	0.744	0.776	0.922
Global support	5	0.761	0.781	0.928
Global support	8	0.621	0.703	0.887
Global support	10	0.591	0.694	0.892
Dictionary mean	3	0.861	0.908	0.952
Dictionary mean	5	0.798	0.865	0.937
Dictionary mean	8	0.695	0.791	0.918
Dictionary mean	10	0.617	0.726	0.891

HELD OUT SUPPORT VALIDATION

The following tests separate persistence on new realizations from persistence on an operator omitted during selection. Table S15 uses different realizations for selection and evaluation while keeping the selected support fixed. Table S16 tests whether this support is more informative than random features from the same candidate pool and of the same size. Table S17 then excludes an entire operator from selection.

Table S15: Fixed-support evaluation with one realization per operator used for selection. Profile is Spearman correlation with the 15-variant reference; Gap is the context and wrapper mean minus the local-edit mean; Pos. units is the number of SAEs with positive Gap; Group sign and Pairwise are prompt-level ordering rates.

Fold	Avail.	Profile	Gap	Pos. units	Same dir.	Group sign	Pairwise
0	99.985%	1.000	0.909	27/27	0.940	99.92%	93.49%
1	99.985%	1.000	0.896	27/27	0.957	99.86%	92.00%
2	99.989%	1.000	0.897	27/27	0.959	99.87%	92.15%

Table S16 reports both absolute response and agreement with the direction chosen during selection. The random comparison is matched in support size and candidate pool.

Table S16: Held out response for RSS and random supports matched in candidate pool and support size. Brackets give 95% confidence intervals.

Support	Absolute response	95% CI	Same direction	95% CI
RSS	107.670	[106.749, 108.597]	0.952	[0.951, 0.952]
Random, 100 draws	5.797	[5.748, 5.846]	0.758	[0.758, 0.759]

Across the 81 unit-fold comparisons, RSS has the larger context and wrapper versus local edit difference in 76.

In Table S17, the withheld operator contributes no selection observations. Comparing requirements of four versus three responding operators among the remaining four makes the availability and direction persistence tradeoff explicit.

Table S17: Evaluation with one transformation operator excluded from support selection. Gap is the context and wrapper mean minus the local-edit mean.

Selection requirement	Available	Same direction	Gap
4 of 4 operators	99.42%	0.894	0.925
3 of 4 operators	100.00%	0.848	0.896

S5 EXTENDED AUDIT RESULTS

Mismatched RSS uses a support selected for another transformation setting. For each case, comparison settings are traversed in the frozen cyclic order, repeated feature indices are removed by first occurrence, and the resulting support is truncated to the size of the current RSS support; the comparator is unavailable if there are too few unique features.

PROBE PERFORMANCE

Table S18 reports the selected probes and their clean test performance. Figure S6 compares clean and transformed AUROC under each setting while keeping the fitted coefficients and calibrated threshold fixed.

Table S18: Safety probes and clean test performance. Brackets give 95% confidence intervals for AUROC; FNR and FPR are false-negative and false-positive rates.

Model	Unit	λ	Weights	Clean AUROC	Harmful FNR	Benign FPR
Gemma 3 4B	L17	0.01	415	0.941 [0.914, 0.963]	23.3%	7.3%
Llama 3.1 8B	L24	0.001	862	0.919 [0.884, 0.949]	33.3%	6.0%

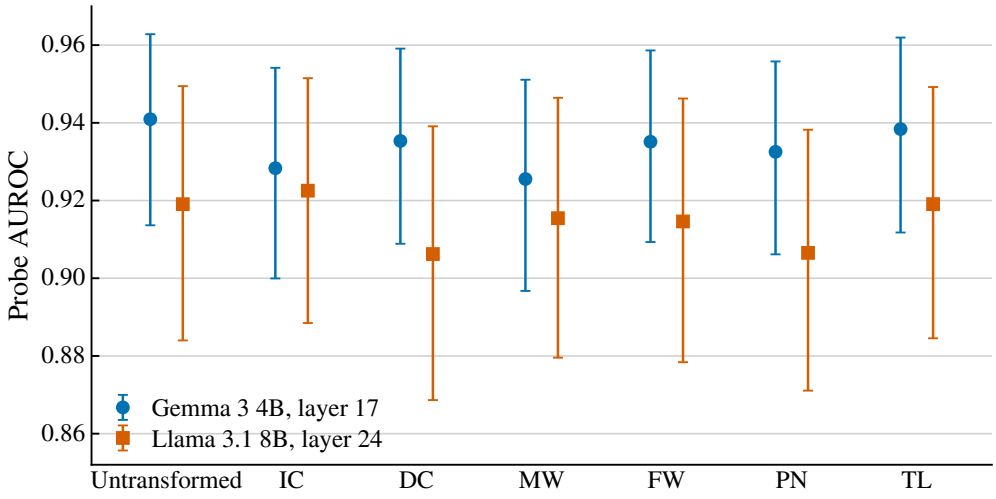


Figure S6: Probe AUROC for the two safety readouts on untransformed test prompts and under each transformation. Transformation abbreviations as in Figure S2. Error bars show 95% bootstrap confidence intervals.

Results by Transformation

The transformation profile describes response magnitude, but it need not order readout effects in the same way. Figure S7 and Table S19 compare normalized MeanRSS with held out margin erosion, decision changes, and probe AUROC for each setting.

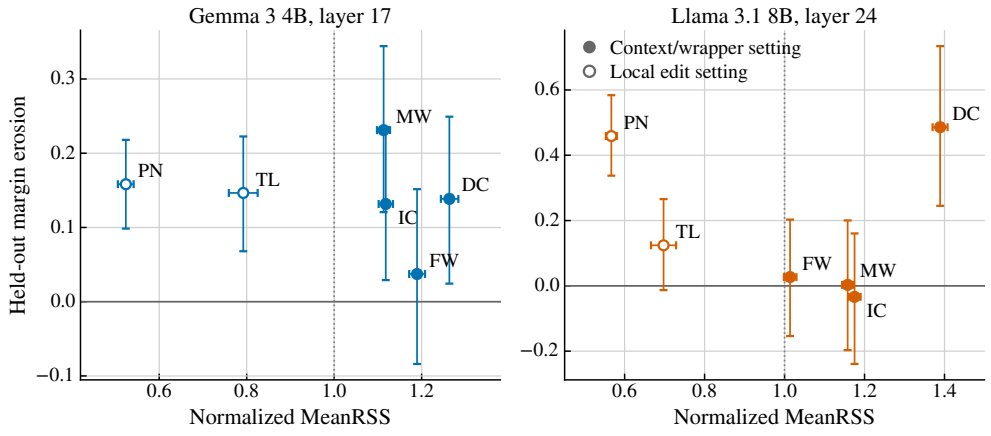


Figure S7: Normalized MeanRSS and margin erosion on held out variants for each transformation, evaluated on the audited safety units. Transformation abbreviations as in Figure S2. Error bars show 95% bootstrap confidence intervals.

For Gemma, the difference in margin erosion between the context and wrapper transformations and the local edits is $+0.338$ $[0.236, 0.443]$ for benign prompts and -0.374 $[-0.498, -0.259]$ for harmful prompts. Thus the same transformation group is associated with greater erosion for one

Table S19: MeanRSS and probe outcomes by transformation setting. Brackets give 95% confidence intervals for normalized MeanRSS and erosion. Erosion is mean held out margin loss; Flip is the fraction of predictions that differ from the clean prediction.

Model	Setting	Norm. MeanRSS	Erosion	Flip	AUROC
Gemma 3 4B	Irrelevant context	1.118 [1.102,1.134]	0.132 [0.029,0.235]	6.1%	0.928
	Distractor context	1.263 [1.244,1.284]	0.138 [0.024,0.249]	5.6%	0.935
	Metalinguistic wrapper	1.113 [1.098,1.128]	0.231 [0.121,0.344]	6.6%	0.926
	Format wrapper	1.190 [1.172,1.208]	0.037 [-0.084,0.152]	8.0%	0.935
	Punctuation noise	0.524 [0.506,0.541]	0.158 [0.098,0.218]	4.2%	0.933
	Typed lexical substitution	0.792 [0.759,0.825]	0.146 [0.068,0.223]	4.6%	0.938
Llama 3.1 8B	Irrelevant context	1.175 [1.160,1.190]	-0.034 [-0.239,0.161]	7.5%	0.923
	Distractor context	1.389 [1.370,1.408]	0.486 [0.245,0.734]	9.5%	0.906
	Metalinguistic wrapper	1.158 [1.144,1.172]	0.003 [-0.197,0.200]	7.8%	0.915
	Format wrapper	1.014 [0.998,1.030]	0.027 [-0.154,0.203]	6.9%	0.915
	Punctuation noise	0.567 [0.553,0.581]	0.459 [0.338,0.584]	7.1%	0.907
	Typed lexical substitution	0.697 [0.666,0.729]	0.124 [-0.013,0.266]	5.5%	0.919

class and less erosion for the other after projection through the probe. A shared representation profile does not imply a shared direction of readout effect.

Intensity Analysis on the Safety Corpus

The safety intensity analysis tests whether increases in representation response are accompanied by increases in margin erosion. Figure S8 reports both quantities, and Table S20 separates rank correlation from monotonicity across all four levels.

The monotonicity denominators use only clean-correct prompts among those eligible for the intensity analysis. For Gemma, these subsets contain 254 prompts for irrelevant context and format wrappers and 222 for punctuation noise. For Llama, they contain 241 and 212 prompts, respectively. The numerators count prompts for which the evaluated mean response is nondecreasing across all four levels; they do not use the full eligible counts as denominators.

Table S20: Prompt-level intensity calibration in the safety corpus. Monotonic counts use prompts whose mean response is nondecreasing across the four levels.

Model	Setting	Eligible	MeanRSS ρ	MeanRSS mono.	Erosion ρ	Erosion mono.
Gemma 3 4B	Irrelevant context	300	0.8	78/254	0.4	51/254
	Format wrapper	300	1.0	203/254	0.4	63/254
	Punctuation noise	265	1.0	137/222	0.4	37/222
Llama 3.1 8B	Irrelevant context	300	0.8	67/241	0.2	39/241
	Format wrapper	300	1.0	192/241	0.4	51/241
	Punctuation noise	265	1.0	138/212	0.6	62/212

Full Risk Results

Table S21 compares risk signals on the common evaluation cohort. Correlation with held out erosion evaluates transfer of margin change, whereas AUROC and average precision (AP) evaluate ranking of recurrent failure. These are distinct targets, so a signal can track erosion well without giving the strongest failure ranking.

Relevance of the Selected Features on New Variants

Tables S22–S24 separately assess direction persistence, contribution after matching coefficient magnitude, and the fraction of absolute contribution retained. These comparisons distinguish stable response selection from selection optimized for a fixed readout.

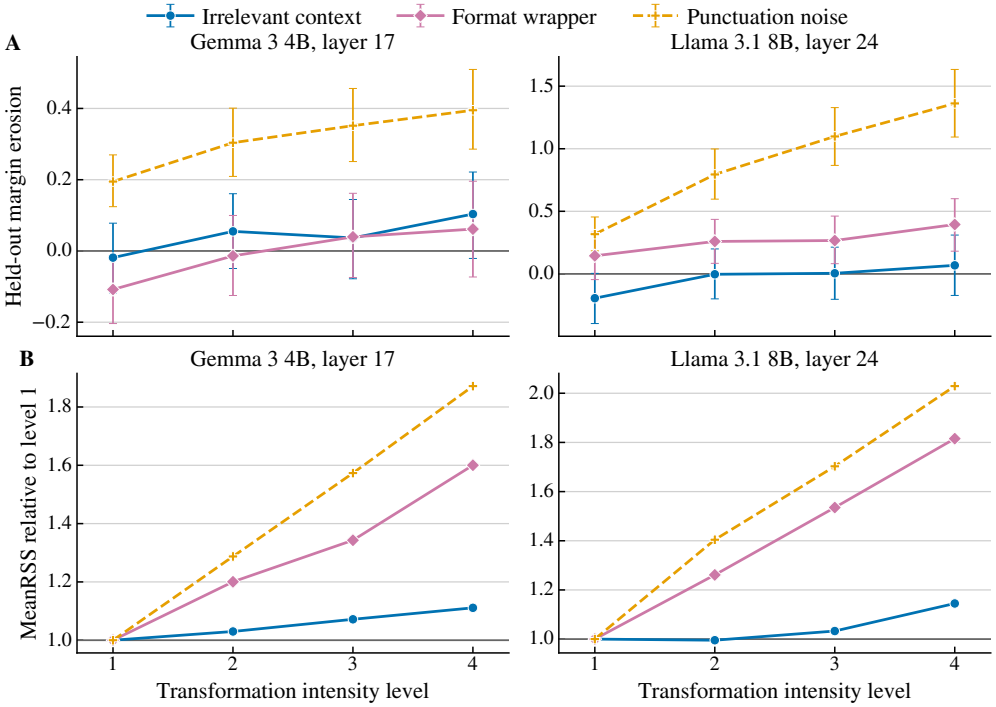


Figure S8: Effect of transformation strength on the audited safety units. A, margin erosion under three transformations rebuilt at four increasing intensity levels, with 95% bootstrap confidence intervals. B, MeanRSS for the same series, relative to the weakest level.

Table S21: Risk signals on the common held out cohort. Brackets give 95% confidence intervals. Erosion ρ is Spearman correlation with held out margin erosion; AUROC and AP use recurrent failure. Residual-margin correlation is reported in the direction of increasing risk.

Model	Risk signal	Erosion ρ	Failure AUROC	AP
Gemma 3 4B	Clean margin	-0.282 [-0.370,-0.192]	0.863 [0.821,0.900]	0.294
	Full ℓ_1	0.101 [0.025,0.173]	0.520 [0.444,0.595]	0.072
	Full cosine	0.083 [0.006,0.157]	0.486 [0.398,0.572]	0.063
	MeanRSS	0.116 [0.040,0.188]	0.526 [0.446,0.599]	0.069
	Probe exposure	0.104 [0.029,0.178]	0.597 [0.510,0.683]	0.107
	MeanRSS $\times$ exposure	0.130 [0.049,0.210]	0.606 [0.524,0.689]	0.091
	+ probe weights	0.186 [0.097,0.272]	0.676 [0.601,0.740]	0.125
	Margin erosion E	0.923 [0.909,0.935]	0.785 [0.720,0.841]	0.198
	Residual margin m_{res}	0.169 [0.077,0.261]	0.965 [0.940,0.982]	0.805
Llama 3.1 8B	Clean margin	-0.220 [-0.307,-0.130]	0.901 [0.833,0.947]	0.505
	Full ℓ_1	0.039 [-0.046,0.123]	0.466 [0.372,0.554]	0.079
	Full cosine	-0.024 [-0.104,0.060]	0.528 [0.427,0.627]	0.136
	MeanRSS	-0.028 [-0.109,0.052]	0.506 [0.429,0.578]	0.086
	Probe exposure	0.041 [-0.029,0.110]	0.569 [0.482,0.652]	0.105
	MeanRSS $\times$ exposure	0.011 [-0.067,0.091]	0.570 [0.483,0.648]	0.109
	+ probe weights	0.039 [-0.053,0.130]	0.581 [0.499,0.654]	0.127
	Margin erosion E	0.896 [0.877,0.912]	0.705 [0.639,0.767]	0.191
	Residual margin m_{res}	0.152 [0.065,0.243]	0.964 [0.940,0.982]	0.850

Table S22: Held out response in the direction selected by RSS compared with magnitude-only selection. Brackets give 95% confidence intervals.

Model	RSS	Magnitude only
Gemma 3 4B	0.876 [0.872, 0.879]	0.816 [0.810, 0.822]
Llama 3.1 8B	0.871 [0.867, 0.874]	0.794 [0.788, 0.800]

Table S23 compares absolute feature contributions after matching readout coefficient magnitude. Direction agreement compares the observed contribution sign with $\mathrm{sign}(-yw_f s_f)$, including the class sign and readout coefficient as well as the RSS selected direction.

Table S23: Held out feature contribution after matching selected and unselected features by readout coefficient magnitude. Contribution columns report mean absolute contribution. Direction agreement compares observed contribution sign with $\mathrm{sign}(-yw_f s_f)$. Brackets give 95% confidence intervals for the paired difference.

Model	RSS contribution	Matched contribution	Difference	Direction agreement
Gemma 3 4B	0.2397	0.0221	0.2177 [0.2005, 0.2356]	97.4%
Llama 3.1 8B	0.3702	0.0220	0.3483 [0.3235, 0.3739]	97.1%

Table S24 compares MCR at $K = 20$. The $|e_f|$ ranking uses contributions computed from selection variants, not the held out contributions used in the reported MCR.

Table S24: Margin Contribution Retention at $K = 20$. Mismatched RSS uses a support selected for another transformation; $|e_f|$ ranks features by contributions computed from V_{sel}. Brackets give 95% confidence intervals.

Model	RSS	Mismatched RSS	$\lvert e_f \rvert$
Gemma 3 4B	0.430 [0.411, 0.450]	0.326 [0.308, 0.346]	0.942 [0.936, 0.948]
Llama 3.1 8B	0.562 [0.547, 0.577]	0.435 [0.417, 0.453]	0.919 [0.912, 0.925]

Table S24 reports MCR on V_{ho}. For a fixed case, all supports share the same denominator, so higher MCR means less omitted absolute contribution. Equation 17 shows that signed error additionally depends on cancellation through χ_S^{ho}.

Table S25 compares full and support erosion on the common evaluation cohort. Correlation measures agreement with held out erosion; AUROC uses erosion as a score for recurrent failure.

Table S25: Margin erosion results for the full readout and RSS support. Erosion ρ is correlation with held out erosion, and Failure AUROC ranks recurrent failure using the corresponding erosion score. Brackets give 95% confidence intervals.

Model	Feature set	Erosion ρ	Failure AUROC
Gemma 3 4B	Full readout	0.923 [0.909, 0.935]	0.785 [0.721, 0.842]
	RSS support	0.737 [0.700, 0.771]	0.753 [0.686, 0.815]
Llama 3.1 8B	Full readout	0.896 [0.878, 0.911]	0.705 [0.639, 0.766]
	RSS support	0.793 [0.761, 0.820]	0.735 [0.666, 0.797]

Table S26 adds the full clean margin to each erosion estimate. This evaluates the information retained when selected changes are used to estimate the remaining margin, rather than treating the support as a replacement for the clean readout.

Table S26: Residual margin results for the same cohort. Both feature sets use the full clean margin. Correlation with held out erosion is oriented toward increasing risk; AUROC evaluates recurrent failure. Brackets give 95% confidence intervals.

Model	Feature set	Residual ρ	Residual AUROC
Gemma 3 4B	Full readout	0.169 [0.077, 0.261]	0.965 [0.939, 0.982]
	RSS support	0.009 [-0.081, 0.105]	0.935 [0.901, 0.961]
Llama 3.1 8B	Full readout	0.152 [0.065, 0.243]	0.964 [0.940, 0.982]
	RSS support	0.080 [-0.006, 0.171]	0.958 [0.924, 0.981]

Table S27: Summary of the full frozen readout and RSS supports selected for each case on the common evaluation cohort. Size is the full nonzero coefficient count or median support size. Both residual margin signals use the full clean margin. Errors and Harmful FN are capture rates at 20% offline review with known labels.

Model	Feature set	Size	Erosion ρ	Residual AUROC	Errors	Harmful FN
Gemma 3 4B	Full readout	415	0.923	0.965	94.1%	90.9%
	RSS support	20	0.737	0.935	89.2%	87.3%
Llama 3.1 8B	Full readout	862	0.896	0.964	91.9%	93.5%
	RSS support	19	0.793	0.958	92.8%	94.4%

Additional Llama Layers

We apply the same functional protocol to Llama 3.1 8B layers 8 and 16 to assess whether the compact support results extend beyond the selected layer. Table S28 reports direction persistence, contribution retention, residual margin AUROC, localization, and restoration. The metrics vary across layers, so persistence and retained audit utility should not be treated as interchangeable measures.

Table S28: Functional results for two additional Llama 3.1 8B layers under the main audit protocol.

Layer	$\lvert S\rvert$	Same direction	MCR	Residual AUROC	R^2	Restoration
8	20	0.960	0.628	0.905	0.814	0.566
16	20	0.970	0.578	0.806	0.692	0.807

S6 Localization, Selective Review, and Training Details

Localization and Readout Counterfactuals

The localization comparison tests whether RSS selects changes relevant to the current transformation rather than merely unstable or arbitrary features. Table S29 evaluates all listed supports on the same rows from V_{ho}. In both models, RSS gives higher reconstruction R^2 and reverses more probe errors than a support selected for another transformation. Features that are frequently unstable across the corpus and random features of the same size contribute little to either result.

Table S29: Reconstruction and readout-counterfactual results. Baseline supports are matched to RSS support size; brackets give 95% confidence intervals. Q_{res} is the share of probe logit change recovered from the support.

Model	Support	Size	R^2	Errors reversed	Q_{res}
Gemma 3 4B	RSS support	20	0.579 [0.532, 0.621]	0.722 [0.635, 0.803]	0.618
	Mismatched setting	20	0.438 [0.386, 0.481]	0.603 [0.512, 0.683]	0.368
	Global instability	20	0.017 [-0.008, 0.035]	0.067 [0.018, 0.151]	0.000
	Random matched	20	0.014 [-0.006, 0.022]	0.012 [0.002, 0.026]	0.002
Llama 3.1 8B	RSS support	19	0.627 [0.568, 0.675]	0.718 [0.610, 0.807]	0.800
	Mismatched setting	19	0.490 [0.406, 0.560]	0.611 [0.487, 0.721]	0.590
	Global instability	19	0.009 [-0.006, 0.020]	0.023 [0.000, 0.062]	0.000
	Random matched	19	0.017 [0.004, 0.022]	0.015 [0.007, 0.027]	0.002

Selective Review

Table S30 summarizes offline selective review with the Gemma probe trained only on clean prompts, using the same cohort as the risk prediction analysis. Random is the average of 100 random rankings, with the replicate average computed inside each shared prompt cluster bootstrap draw. The residual margin ranks cases in ascending order. For a review budget c, residual error is the error rate among cases left unreviewed after the highest-risk $\lfloor cN \rfloor$ cases are removed. The normalized partial area under the residual error curve (pAURC) is the trapezoidal area at budgets

$c \in \{0, 0.05, 0.10, 0.20, 0.30, 0.40\}$ divided by 0.40. We report $\Delta\mathrm{pAURC}$ as the mean random ranking pAURC minus the pAURC of the listed ranking, so positive values indicate lower residual error than random.

Table S30: Selective-review summary for the Gemma probe. Capture columns give the fraction of errors reviewed at the stated budget. Residual error is reported at a 20% review budget. pAURC is the normalized partial area under the residual error curve from 0 to 40% review, and Δ pAURC is the random ranking pAURC minus the listed value; brackets give 95% confidence intervals.

Ranking signal	Capture@10	Capture@20	Harmful FN@20	Residual error	Δ pAURC
Residual margin m_{res}	85.1%	94.1%	90.9%	0.25%	+0.0276 [0.0175,0.0385]
Low clean margin	41.3%	68.9%	60.9%	1.30%	+0.0186 [0.0117,0.0267]
+ probe weights	32.1%	47.6%	49.5%	2.20%	+0.0109 [0.0046,0.0180]
MeanRSS	11.0%	18.8%	18.2%	3.40%	+0.0007 [-0.0026,0.0046]

Table S31 gives the corresponding capture rates at the 20% review budget.

Table S31: Selective review at a 20% budget for the rankings in Figure 6C–D. Values are capture percentages; brackets give 95% confidence intervals.

Model	Ranking signal	Errors	Harmful FN
Gemma 3 4B	Residual margin m_{res}	94.1 [90.2, 96.5]	90.9 [81.9, 95.7]
	RSS residual margin m^S_{res}	89.2 [81.8, 93.5]	87.3 [72.4, 94.9]
	Mismatched RSS residual	86.5 [78.7, 92.0]	85.9 [69.8, 93.5]
	Margin erosion E	63.6 [49.3, 74.8]	62.3 [35.4, 78.2]
	RSS margin erosion E_S	64.2 [50.1, 75.0]	66.8 [44.4, 81.1]
Llama 3.1 8B	Residual margin m_{res}	91.9 [87.6, 95.6]	93.5 [88.7, 97.4]
	RSS residual margin m^S_{res}	92.8 [88.0, 95.6]	94.4 [89.1, 97.4]
	Mismatched RSS residual	88.0 [80.9, 93.1]	90.0 [82.8, 95.6]
	Margin erosion E	52.2 [38.7, 64.4]	53.3 [38.3, 66.9]
	RSS margin erosion E_S	55.3 [41.7, 67.5]	58.1 [42.5, 71.2]

Probe Training with MeanRSS Sampling

The frozen readout analysis shows that large redistribution alone is not a failure score. A separate question is whether exposing a probe to examples with large redistribution can help during training. We compare MeanRSS sampling with uniform sampling under the same augmentation budget, rather than infer a training benefit from the frozen audit.

Every augmented probe uses 1,200 clean examples and 6,000 transformed examples, corresponding to five transformed examples per training prompt. The MeanRSS method samples prompt and transformation pairs according to MeanRSS computed from the training prompts, with exponent $\gamma = 2$. The uniform method samples from the same pool with equal probability. Scaling statistics and feature masks remain those of the clean probe. Figure S9 and Table S32 report the results.

For Gemma, the paired gains in the run shown in Figure S9A are $+0.0124$ $[0.0024, 0.0233]$ in clean AUROC, $+0.0094$ $[0.0007, 0.0195]$ in macro AUROC, $+0.0127$ $[0.0007, 0.0243]$ in worst setting AUROC, and a reduction of 4.7 percentage points $[1.6, 7.8]$ in harmful false-negative rate under the worst setting. Across five runs, however, Gemma macro AUROC differences average $+0.0011$ with standard deviation 0.0051. The macro harmful false-negative rate averages 17.65% for MeanRSS sampling and 20.45% for uniform sampling. The corresponding Llama rates are 23.55% and 24.36%.

These averages show fewer harmful prompts missed in the reported runs, but the AUROC evidence is mixed. Table S32 reports lower clean, mean, and worst setting AUROC for Llama under MeanRSS sampling, and Gemma's average macro AUROC gain across runs is small relative to its variation. The training extension therefore does not establish a consistent AUROC advantage over uniform sampling.

Table S32: Retrained-probe results for Figure S9A. Clean, Mean, and Worst are AUROC; Gap is clean minus worst AUROC.

Model	Method	Clean	Mean	Worst	Worst FNR	Gap
Gemma 3 4B	MeanRSS sampling	0.945	0.943	0.938	0.191	0.0067
Gemma 3 4B	Uniform	0.933	0.934	0.926	0.238	0.0071
Llama 3.1 8B	MeanRSS sampling	0.911	0.910	0.901	0.256	0.0100
Llama 3.1 8B	Uniform	0.915	0.913	0.907	0.263	0.0082

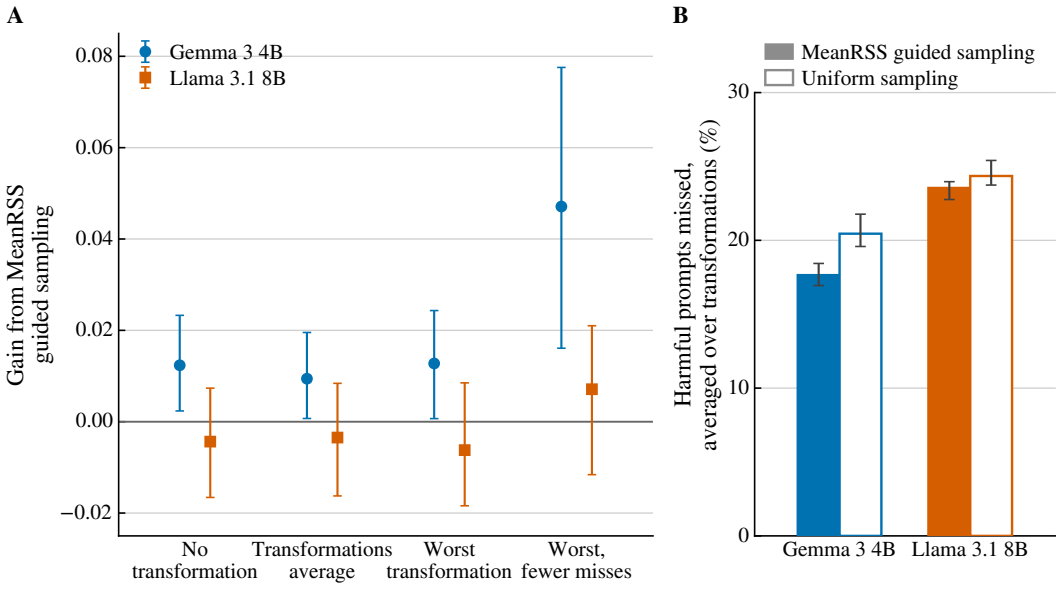


Figure S9: Augmentation guided by MeanRSS compared with uniform sampling. A, differences between the two sampling methods in clean, mean, and worst transformation AUROC, together with the reduction in harmful false negative rate under the worst transformation. Error bars show 95% bootstrap confidence intervals. B, harmful false negative rate averaged over the six transformations across five runs. Bars show the run mean and whiskers the observed range.