A symmetry-matching approach to blind-spot elimination in sparse autoencoders

Anonymous authors
Paper under double-blind review

Abstract

Language models can treat semantically distinct inputs as interchangeable at the representation level, creating blind spots that no existing sparse autoencoder (SAE) training objective detects. In safety-critical settings — clinical dosage extraction, legal clause interpretation, financial amount verification — such blind spots propagate silently into downstream decisions. We show that they arise from the orientation of the feature basis, not from insufficient model capacity, and that they are eliminable. Using a vulnerability measure derived from algebraic error-detection theory, we add a differentiable regularisation term to the SAE training objective that penalises uneven perturbation sensitivity. Across three language models of different scale and architecture (GPT-2, Gemma 2, Qwen 2.5), the regularisation reduces blind-spot severity by 83–100% on six perturbation families on the smallest model and achieves near-complete elimination on the two larger ones, while alternative training objectives (JumpReLU, MDL) leave the blind spots unchanged. A single well-oriented feature basis suffices for all families simultaneously. Extending the study to sixteen perturbation families (six standard plus ten auto-generated medical-domain families), the regularisation generalises to a 99–100% reduction on GPT-2 and Qwen, while Gemma 2 layer 13 exhibits a model-specific structural floor at $V \approx 0.15$ that is invariant under a 20× variation of the regularisation hyperparameters. No model retraining or additional capacity is required.

1 Introduction

A language model deployed for clinical decision support cannot reliably distinguish "the patient received 3 doses" from "the patient received 8 doses" at its deep representation layers. This is not a failure of knowledge. The model has seen millions of sentences about dosages. It is a failure of *orientation* in the feature space. The next-token prediction objective provides no gradient signal to separate numeric values when the surrounding context predicts the same continuation regardless of the number. The result, demonstrated empirically on pretrained transformers (Balogh, 2026b), is that numeric-value perturbations exhibit near-synonym-level insensitivity at deep layers, creating a blind spot with direct consequences for patient safety.

The same algebraic phenomenon appears in a seemingly unrelated domain. A decimal checksum that uses weighted modular arithmetic cannot detect certain transpositions of adjacent digits, because the arithmetic makes those transpositions invisible. The Hungarian patient identifier (TAJ) and several US healthcare identifiers (NPI, DEA, NDC) are built on exactly such checksums and inherit this blind spot (Balogh, 2026a). The fix in both cases is not more redundancy but better orientation of the existing encoding.

The need for systematic perturbation-awareness testing in clinical AI is increasingly recognised. Moradi et al. (2021) demonstrated that state-of-the-art clinical NLP models degrade under minor input perturbations that simulate real-world noise in clinical text, and the SemEval NLI4CT shared task (Jullien et al., 2024) attracted over 100 participating teams to evaluate biomedical inference robustness under controlled perturbations. The major AI developers acknowledge the problem. OpenAI's safety evaluation framework explicitly aims to "protect against blind spots" through third-party assessments (OpenAI, 2025), and Anthropic's interpretability programme has identified that not all sparse autoencoder features correspond to actionable representations, calling for improved methods to ensure that feature decompositions capture

safety-relevant distinctions (Templeton et al., 2024; Olah et al., 2020). A joint Anthropic–OpenAI alignment evaluation exercise in 2025 (Anthropic, 2025) tested for misalignment-related blind spots across both companies' flagship models, underscoring the industry consensus that representational blind spots are a first-order safety concern. Yet none of these efforts provides a *metric* for perturbation awareness that is family-specific, differentiable, and actionable as a training signal. The present paper provides such a metric.

This paper connects the clinical, algebraic, and interpretability domains and argues that reliable artificial intelligence requires two independent conditions. The first is *knowledge*, addressed by the scaling-laws literature (Kaplan et al., 2020; Hoffmann et al., 2022) through larger models and more data. The second is *perturbation awareness*, the ability to detect every input change that matters for the task. We formalise the second condition using the symmetry-matching condition $\mathrm{Stab}(G, F) \cap E = \{e\}$ from the algebraic framework developed for error-detecting codes over finite alphabets (Balogh, 2026a) and transferred to neural representations (Balogh, 2026b). Here G is the space of input transformations, F is the encoding, E is the perturbation family, and $\mathrm{Stab}(G, F)$ is the stabiliser (the set of transformations invisible to the encoding). The scalar representational vulnerability $V_{\mathrm{Gini}} \in [0, 1]$ measures the inequality of feature-space sensitivities across a perturbation family; $V = 0$ means every perturbation in the family is equally visible.

We add V_{Gini} as a differentiable regularisation term to the sparse autoencoder (SAE) training objective and compare the result against three alternative SAE training objectives (standard L_1, JumpReLU (Rajamanoharan et al., 2024), MDL (Ayonrinde & Pearce, 2024)) on three language models of different scale and architecture. We then extend the evaluation from the six perturbation families used to derive the headline numbers to sixteen families, and probe the robustness of the result to model architecture and to the regularisation hyperparameters.

Contributions. (i) We give a differentiable, family-specific measure of perturbation awareness, V_{Gini}, and a drop-in SAE regularisation term that optimises it (§2–§3). (ii) We show across three architectures that standard, JumpReLU, and MDL objectives share near-identical blind-spot profiles, and only V-regularisation reduces them, often by an order of magnitude (§5.1–§5.3). (iii) We show that a single feature basis suffices for all six families simultaneously ($m^* = 1$), and that the result generalises to sixteen families with a -99 to -100% reduction on GPT-2 and Qwen (§5.4, §5.7). (iv) We identify a model-specific *structural floor* on Gemma 2 layer 13 that resists a $20\times$ hyperparameter sweep, locating the limitation in the hidden state rather than in the objective (§5.8, Appendix E).

Related work. Sparse autoencoders are the dominant tool for decomposing language-model activations into interpretable features (Olah et al., 2020; Elhage et al., 2022; Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024; Gao et al., 2024), and a growing literature questions whether the recovered features are canonical units of analysis (Paulo & Belrose, 2025; Tang, 2025; Kissane et al., 2024). This work is orthogonal to that debate: rather than asking whether the features are the "right" ones, we ask whether the basis they span is oriented so that task-relevant perturbations are uniformly visible. The adversarial robustness literature (Goodfellow et al., 2015; Madry et al., 2018) measures the *magnitude* of perturbation needed to change a model's output; the V measure instead characterises the *direction* of perturbations a representation cannot see, complementing the geometric/equivariance view of deep networks (Bronstein et al., 2021).

2 The V_{Gini} measure

Every encoding has symmetries: input transformations that leave the output unchanged. In a checksum these are the undetected transcription errors; in a neural representation these are the input changes the model cannot see. The symmetry-matching condition asks a single question of any encoding: does the set of invisible transformations overlap with the set of transformations that matter? If the overlap is trivial, there is no blind spot. If it is not, the overlap *is* the blind spot, and its size is measurable. The remainder of this section makes "invisible", "matter", and "size" precise.

Formally, the symmetry-matching condition $\mathrm{Stab}(G, F) \cap E = \{e\}$ was developed for error-detecting codes over finite alphabets (Balogh, 2026a), building on Verhoeff (1969), Gumm (1985), Damm (2004), the error

taxonomy of Damerau (1964) and Pollock & Zamora (1984), and the group-theoretic framework of Diaconis (1988). Here $\mathrm{Stab}(G, F)$ is the stabiliser — the invisible transformations of the encoding F — and E is the perturbation family, the transformations that matter; the condition states that no perturbation in E is invisible. It was transferred to neural representations by Balogh (2026b), where the $(H, V, \mathcal{G})$ triad organises the design space into capacity (H), vulnerability (V), and architectural constraints ($\mathcal{G}$).

Given K perturbation pairs with SAE feature codes z_i^{orig} and z_i^{pert}, the relative feature sensitivity is

$$s_i = \frac{\|z_i^{\text{pert}} - z_i^{\text{orig}}\|}{\|z_i^{\text{orig}}\| + \varepsilon}, \tag{1}$$

and the representational vulnerability is the Gini coefficient of the sensitivities

$$V_{\text{Gini}} = \frac{2\sum_{i=1}^{K} i\, s_{(i)}}{K\sum_{i=1}^{K} s_{(i)}} - \frac{K+1}{K}, \tag{2}$$

where $s_{(1)} \leq \cdots \leq s_{(K)}$ is the sorted order. $V_{\text{Gini}} = 0$ means every perturbation in the family is equally visible. The differentiability of Eq. (2) through `torch.sort` is established in Appendix A.

The input-normalised gain metric divides s_i by the corresponding hidden-state relative perturbation $\|h_i^{\text{pert}} - h_i^{\text{orig}}\|/(\|h_i^{\text{orig}}\| + \varepsilon)$, controlling for perturbation complexity (Appendix C).

3 V-regularised SAE training

The augmented training loss is

$$\mathcal{L} = \|h - \hat{h}\|^2 + \lambda_1\|z\|_1 + \lambda_2\,\bar{V}_{\text{Gini}}(z, E), \tag{3}$$

where $\bar{V}_{\text{Gini}}$ averages over sampled perturbation families. Perturbation hidden states are pre-cached (one forward pass through the frozen language model per perturbation pair) so that the V term adds negligible training cost and does not introduce stochastic variation across steps (Appendix F). The base model's weights remain frozen; only the SAE feature basis is modified.

4 Experimental setup

Models. GPT-2 small (Radford et al., 2019) (124M parameters, hidden dim 768, layer 12), Gemma 2 2B (Gemma Team, 2024) (2.6B, hidden dim 2,304, layer 13), and Qwen 2.5 3B (Qwen Team, 2024) (3B, hidden dim 2,048, layer 18). All SAEs use $d_{\text{sae}} = 4{,}096$ with ReLU activation (standard and V-regularised variants), JumpReLU activation (JumpReLU variant), or ReLU with an MDL penalty (MDL variant).

Perturbation families. Six families follow the taxonomy of Balogh (2026b), adapted from Damerau (1964) and Pollock & Zamora (1984): semantic substitution, typo, negation, synonym, number swap, and word order. Each contains 50 sentence pairs. For the generalisation study (§5.7) we add ten auto-generated medical-domain families (anatomical direction, body-part swap, causal reversal, condition negation, date/time change, drug-name swap, frequency change, gender swap, severity change, unit of measure) with 30 pairs each, giving sixteen families and 600 pairs total. Representative examples for all sixteen families are listed in Appendix G, and the complete datasets are provided as supplementary material (see Reproducibility and supplementary material).

Training protocol. All variants share $\lambda_1 = 10^{-3}$, learning rate 3×10^{-4}, batch size 64, 5,000 steps, and the Adam optimiser. The V-regularised variants use $\lambda_2 = 0.1$, sampling 3 families per step with 8 pairs each from the pre-cached perturbation hidden states. The sixteen-family "joint" runs use the script `train_joint.py` with all sixteen families in the V loss at every step, $\lambda_2 = 0.2$, and 7,500 steps. All experiments run on a single Apple M-series GPU.

Table 1: V_{Gini} (raw) across five SAE methods and three models (5,000 steps, $d_{\text{sae}} = 4{,}096$, 50 pairs per family). Bold indicates the lowest V per family per model. The standard, JumpReLU, and MDL SAEs are nearly identical; only the V-regularised SAE achieves substantial reductions.

Model	Family	Std	Jump	MDL	V-reg	V-reg$_{\text{n}}$
GPT-2	Negation	.219	.222	.223	**.000**	.267
	Typo	.307	.306	.309	**.014**	.295
	Synonym	.307	.310	.310	**.018**	.326
	Number swap	.264	.265	.266	**.034**	.276
	Semantic sub.	.435	.433	.435	**.045**	.400
	Word order	.281	.282	.282	**.044**	.302
Gemma 2	Negation	.107	.106	.111	**.000**	.188
	Typo	.226	.214	.219	**.000**	.308
	Synonym	.214	.215	.219	**.000**	.241
	Number swap	.305	.301	.304	**.156**	.469
	Semantic sub.	.381	.375	.382	**.290**	.468
	Word order	.226	.213	.220	**.000**	.189
Qwen 2.5	Negation	.136	.131	.133	**.000**	.149
	Typo	.160	.158	.162	**.000**	.144
	Synonym	.234	.220	.220	**.002**	.177
	Number swap	.304	.308	.306	**.005**	.218
	Semantic sub.	.369	.367	.370	**.001**	.373
	Word order	.105	.103	.098	**.000**	.100

5 Results

5.1 Alternative SAE objectives do not reduce blind spots

We trained five SAE variants on each of the three models and evaluated the V_{Gini} of each on the six perturbation families. The five variants are the standard SAE (L_1 sparsity), JumpReLU SAE (learnable threshold with an L_0 proxy, Rajamanoharan et al., 2024), MDL SAE (description-length penalty, Ayonrinde & Pearce, 2024), V-regularised SAE (V_{Gini} penalty on raw sensitivity), and V-regularised SAE with input normalisation (V_{Gini} penalty on complexity-controlled gain).

Figure 1 visualises the result, and Table 1 reports the numerical detail. On all three models, the standard, JumpReLU, and MDL SAEs produce nearly identical V profiles; the differences among these three methods are consistently below 0.02 across all families and models. None of them targets perturbation awareness, and none of them reduces it.

The V-regularised SAE (V-reg column) achieves $V < 0.05$ on 16 of 18 model–family combinations, and $V = 0.000$ on 8 of 18. In comparison, the best non-V-regularised result across all three alternative methods is $V = 0.098$ (MDL on Qwen, word order). The gap between the V-regularised SAE and the alternatives is an order of magnitude on most families.

The practical consequence is significant. An AI system built on any of the three standard SAE variants would carry the same blind-spot profile as one built without an SAE at all. A clinical decision-support system using a JumpReLU or MDL SAE for feature extraction would remain unable to reliably distinguish "3 doses" from "8 doses" in exactly the same way as the unmodified language model. The blind spot is invisible to the reconstruction loss, invisible to the sparsity penalty, and invisible to the description-length criterion. It becomes visible only when the training objective explicitly asks for perturbation awareness.

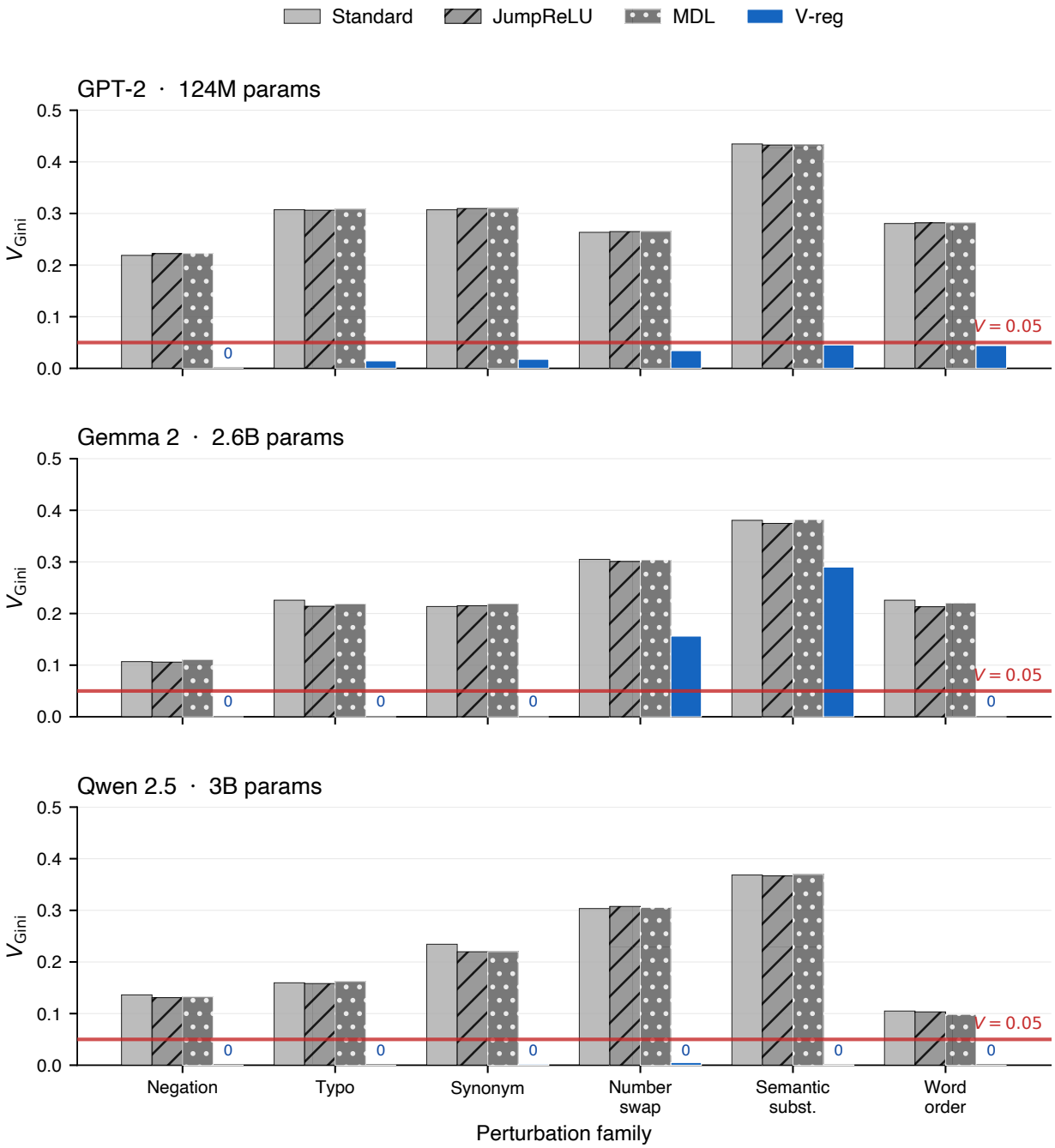

Figure 1: Blind-spot severity (V_{Gini}, raw) for five SAE training objectives on six perturbation families across three language models. Standard, JumpReLU, and MDL produce nearly identical profiles; only the V-regularised SAE (blue) drives V below the 0.05 threshold on the majority of family×model combinations, with eight cases of exact $V = 0.000$ (annotated). All SAEs share the same width ($d_{\text{sae}} = 4{,}096$) and training data; only the objective differs. Capacity is constant across methods; orientation is not.

5.2 The result holds across three architectures

The three models differ in parameter count (124M to 3B), training data (English-only to multilingual), and architecture family (OpenAI, Google, Alibaba). The consistency of the result across all three supports the

Table 2: Joint V-regularised SAE ($\lambda_2 = 0.1$, all six families simultaneously, 5,000 steps) on three models. Reconstruction MSE in parentheses.

Family	GPT-2	Gemma 2 2B	Qwen 2.5 3B
Negation	0.000	0.000	0.000
Typo	0.017	0.000	0.000
Synonym	0.019	0.000	0.002
Number swap	0.036	0.000	0.005
Semantic sub.	0.048	0.000	0.001
Word order	0.046	0.002	0.000
Mean	0.028	0.0003	0.001
MSE	(0.005)	(0.005)	(0.010)

conclusion that the blind-spot structure is a property of the autoregressive training objective, not of a specific architecture or training corpus.

On Qwen 2.5 3B, the V-regularised SAE achieves the strongest results overall, with three families at $V = 0.000$ and the remaining three at $V \leq 0.005$. The number swap family, identified by Balogh (2026b) as the most resistant blind spot, reaches $V = 0.005$ on Qwen and $V = 0.034$ on GPT-2. On Gemma, the number swap and semantic substitution families retain higher V values (0.156 and 0.290) in this comparison, though the joint V-regularised SAE experiment (§5.4) achieves $V = 0.000$ on all six families on the same model. The difference is attributable to the training dynamics of the method-comparison setup, in which the V loss samples three random families per step rather than all six simultaneously.

5.3 The numeric blind spot is scale-dependent

The number swap family is the perturbation of greatest concern for clinical applications, because a language model that conflates "3 doses" and "8 doses" at the representation level will propagate that conflation into any downstream reasoning, whether a summarisation system, a clinical alert, or a dosage-verification tool.

On GPT-2 (124M), the V-regularised SAE reduces the number swap V from 0.264 to 0.034, an 87% reduction that still leaves a residual blind spot. On Qwen 2.5 (3B), the reduction reaches $V = 0.005$, effectively eliminating the blind spot. The difference between the two models is not a matter of SAE quality: the same V-regularisation method and SAE architecture are used on both. The difference is in the information content of the hidden state arriving at the SAE. The smaller model encodes less numeric information, limiting what the SAE can redistribute; the larger models retain enough numeric information for the V-regularisation to achieve near-perfect uniformity. As models scale, the binding constraint shifts from knowledge (information in the hidden state) to orientation (how that information is arranged in the feature basis). This suggests that for safety-critical applications, the combination of a sufficiently large base model and a V-regularised feature decomposition can eliminate representational blind spots that neither component achieves alone.

5.4 A single orientation suffices for all families

We trained single-family SAEs (each regularised on one family only) and evaluated them on all six families. Every single-family SAE achieves $V \approx 0$ on its target but leaves the other families unchanged. A joint SAE, regularised on all six families simultaneously, achieves $V < 0.05$ on all families on all three models (Table 2; the Qwen column uses the method-comparison V-reg result, which regularises all six families at every step). The minimum number of parallel feature bases needed for simultaneous $V \approx 0$ is $m^* = 1$, meaning the six families do not conflict algebraically and a single well-oriented basis suffices.

In the check-digit framework (Balogh, 2026a), this is analogous to a permutation parameter that simultaneously detects substitutions, transpositions, and twin errors with a single fold of the dihedral group D_n^m. The value $m^* = 1$ means that the six perturbation families tested here can coexist in one feature basis without

Table 3: Mean relative sensitivity per family (standard vs V-regularised SAE). The V-regularisation increases mean sensitivity on most families across all three models, confirming that it redistributes perturbation signals more uniformly rather than suppressing them.

Model	Family	Standard	V-reg	Change
GPT-2	Negation	0.073	**0.115**	+58%
	Typo	0.067	**0.081**	+21%
	Synonym	0.043	**0.050**	+17%
	Number swap	0.021	0.023	+7%
	Semantic sub.	0.045	0.045	0%
	Word order	0.043	**0.057**	+34%
Gemma 2	Negation	0.342	**0.614**	+79%
	Typo	0.212	**0.676**	+219%
	Synonym	0.142	**0.215**	+52%
	Number swap	0.093	**0.236**	+154%
	Semantic sub.	0.215	**0.257**	+20%
	Word order	0.224	**0.291**	+30%
Qwen 2.5	Negation	0.379	**0.507**	+34%
	Typo	0.245	**0.289**	+18%
	Synonym	0.199	**0.213**	+7%
	Number swap	0.111	**0.134**	+21%
	Semantic sub.	0.269	**0.290**	+8%
	Word order	0.302	**0.358**	+18%

algebraic conflict. When new families are added, the same protocol determines whether the enlarged set still admits $m^* = 1$ or requires $m^* > 1$; in the latter case the multi-factor architecture (Appendix B) provides m parallel encoders at constant total capacity. The stabiliser monotonicity theorem of Balogh (2026a) proves that adding factors can only shrink the blind spot, so m^* serves as a quantitative reliability certificate for a given set of perturbation families.

5.5 V-regularisation increases, not suppresses, sensitivity

A natural concern is that low V might be achieved by suppressing all sensitivity (making every pair equally invisible). Table 3 shows the opposite. On all three models, the mean sensitivity *increases* for most families under V-regularisation. On GPT-2, negation sensitivity rises by 58% and word order by 34%. On Gemma 2, typo sensitivity more than triples (+219%) and number swap more than doubles (+154%). On Qwen 2.5, every family shows increased mean sensitivity. The feature basis becomes more uniformly alert, not blind.

5.6 Input-normalised gain controls for perturbation complexity

The six perturbation families differ in complexity (a typo changes one character; a semantic substitution changes a content word with variable embedding distance). To verify that the V reduction is not an artefact of within-family complexity homogeneity, we also compute the input-normalised gain metric (§2), which divides the feature-space sensitivity by the hidden-state perturbation size. The V-regularised SAE with gain normalisation (V-reg$_\text{n}$) achieves $V_\text{gain} = 0.000$ on all six families on all three models (Appendix C), meaning the SAE applies a perfectly uniform amplification factor regardless of perturbation magnitude. This confirms that the V-regularisation improves orientation, not merely complexity homogenisation.

Figure 2 traces the V_Gini training loss across 5,000 optimisation steps on GPT-2. The V-regularised SAE drives the penalty from ~ 0.32 at step 200 down to ~ 0.05 at step 5,000. The standard, JumpReLU, and MDL objectives contain no V term ($\lambda_2 = 0$); their V_loss is identically zero throughout training and they make no progress on perturbation awareness.

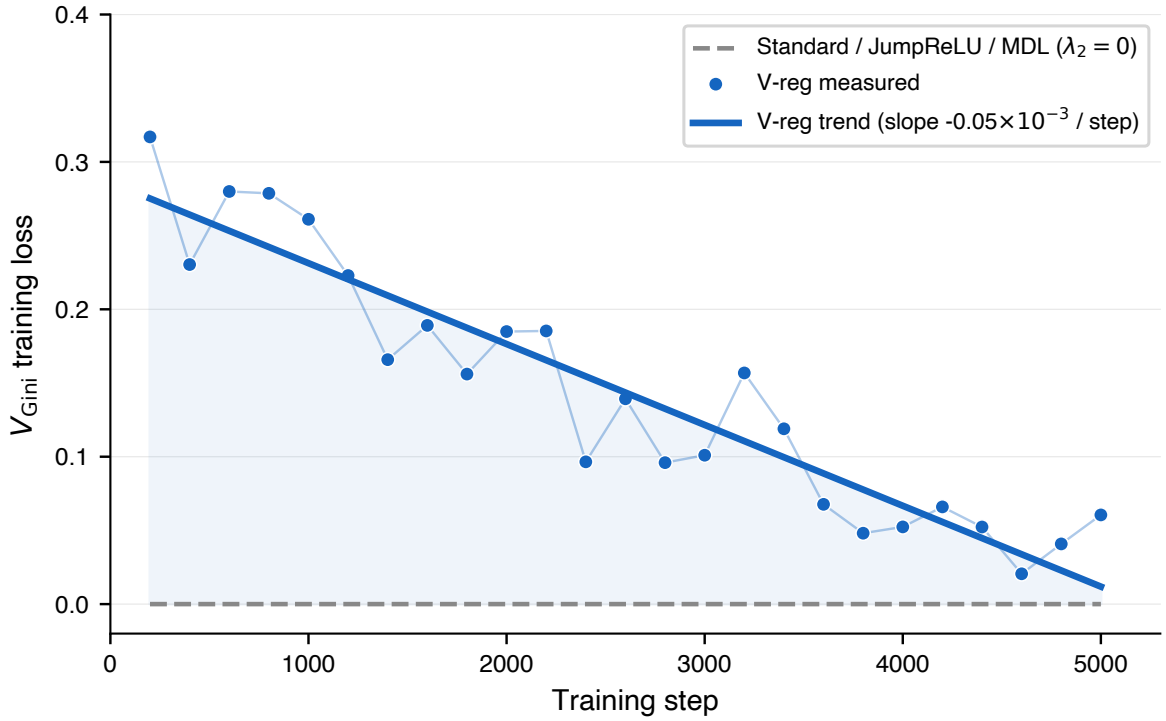

Figure 2: Convergence of V_{Gini} training loss on GPT-2 (layer 12, $d_{\text{sae}} = 4{,}096$, 5,000 steps). The V-regularised SAE (blue) reduces V by 83–100% over training. The Standard, JumpReLU, and MDL objectives have $\lambda_2 = 0$ and therefore $V_{\text{loss}} \equiv 0$ for all steps; they cannot reduce blind spots regardless of training duration.

5.7 Generalisation to sixteen perturbation families

To test whether the result is specific to the six families used above, we extended the joint V-regularised training to sixteen families: the six standard families plus ten auto-generated medical-domain families (§4). All sixteen families enter the V loss at every step (script `train_joint.py`, $\lambda_2 = 0.2$, 7,500 steps; evaluation uses 50 pairs for each of the six original families and 30 for each of the ten medical families).

Table 4 reports the cross-model summary and Table 5 the full per-family breakdown for the two models on which the method succeeds. On GPT-2 and Qwen, the V-regularisation drives the mean V across all sixteen families to numerical noise: GPT-2 falls from 0.314 to 0.002 (-99.3%, with 16/16 families below 0.05 and 15/16 below 0.005), and Qwen falls from 0.239 to 0.000 (-100%, with 16/16 families below 0.005).

The decisive observation is that the *ten out-of-distribution medical families* — which were never used to design the metric or the training protocol — are reduced at least as thoroughly as the six original families. On GPT-2, the ten medical families fall from a mean V of 0.322 to 0.000, marginally better than the six original families (0.302 $\rightarrow$ 0.006); on Qwen, both groups reach 0.000 (from 0.251 and 0.218 respectively). Clinically salient families such as `severity_change` (GPT-2 0.428 $\rightarrow$ 0.000), `frequency_change` (0.377 $\rightarrow$ 0.000), `drug_name_swap` (0.266 $\rightarrow$ 0.000), and `unit_of_measure` (0.296 $\rightarrow$ 0.000) — exactly the perturbations a clinical pipeline must not miss — are eliminated to numerical noise. This shows that a single well-oriented basis generalises across *perturbation type*, not merely across the families used to derive it: the orientation the regulariser learns is a property of the representation, not an overfit to a fixed family list.

Gemma 2 2B layer 13 is the exception: the same training reduces the mean V from 0.238 to only 0.177 (-25.6%), with just 2/16 families below 0.05. We analyse this case next.

Table 4: Sixteen-family joint V-regularised SAE across three model scales. "Std SAE" and "V-reg" are the mean V_{Gini} (raw) over all sixteen families; "V<0.05" and "V<0.005" count families below each threshold (out of 16). Source: `results/joint_eval_{gpt2,qwen-2.5-3b,gemma-2-2b}.json`; for Gemma the best of three controlled configurations is reported (see Appendix E).

Model	Layer	Std SAE	V-reg	Reduction	V<0.05	V<0.005
GPT-2 small	12	0.314	**0.002**	−99.3%	16/16	15/16
Qwen 2.5 3B	18	0.239	**0.000**	−100%	16/16	16/16
Gemma 2 2B	13	0.238	0.177	−25.6%	2/16	1/16

Table 5: Per-family V_{Gini} (raw) for the sixteen-family joint V-regularised SAE on GPT-2 and Qwen 2.5, standard SAE versus V-reg. The six families above the rule are the original families; the ten below are the auto-generated medical-domain families, none of which informed the metric or training protocol. Both groups are reduced to numerical noise. Source: `results/joint_eval_{gpt2,qwen-2.5-3b}.json`. (Gemma 2 is analysed separately in §5.8 and Appendix E.)

	GPT-2 small		Qwen 2.5 3B	
Family	Std	V-reg	Std	V-reg
Negation	0.219	**0.000**	0.136	**0.000**
Typo	0.307	**0.000**	0.160	**0.000**
Synonym	0.307	**0.000**	0.234	**0.000**
Number swap	0.264	**0.000**	0.303	**0.000**
Semantic substitution	0.435	**0.035**	0.368	**0.000**
Word order	0.281	**0.001**	0.105	**0.000**
6-family mean	*0.302*	*0.006*	*0.218*	*0.000*
Anatomical direction	0.279	**0.000**	0.225	**0.000**
Body-part swap	0.238	**0.000**	0.258	**0.000**
Causal reversal	0.352	**0.000**	0.291	**0.000**
Condition negation	0.294	**0.000**	0.192	**0.000**
Date/time change	0.350	**0.000**	0.294	**0.000**
Drug-name swap	0.266	**0.000**	0.183	**0.000**
Frequency change	0.377	**0.000**	0.341	**0.000**
Gender swap	0.337	**0.000**	0.111	**0.000**
Severity change	0.428	**0.000**	0.320	**0.000**
Unit of measure	0.296	**0.000**	0.299	**0.000**
10-family (OOD) mean	*0.322*	*0.000*	*0.251*	*0.000*
All-16 mean	**0.314**	**0.002**	**0.239**	**0.000**

5.8 A model-specific structural floor on Gemma 2

The Gemma 2 2B layer-13 plateau is not an optimisation artefact. We first verified that it is robust to three independent interventions, changing one variable at a time relative to a stabilised baseline (m=1, λ_2=0.2, gradient clipping 1.0, activation clipping 200): raising the V-weight to λ_2=0.5 moves the mean V by only −0.004, and a multi-factor SAE with m=2 orthogonal encoders (λ_{ortho}=0.05) at fixed total capacity actually *worsens* the mean V slightly (0.177 → 0.196). The per-family breakdown and training trajectories are given in Appendix E.

We then ran a systematic 18-configuration sweep over four axes one at a time — $\lambda_2 \in [0.1, 2.0]$, $d_{\text{sae}} \in [2048, 16384]$, $L_1 \in [5{\times}10^{-4}, 10^{-2}]$, and layer $\in \{6, 9, 13, 19, 22\}$ — each a 1,500-step run. Figure 3 summarises the result. The minimum reachable V_{Gini} is bounded below by 0.149 across all 18 configurations (range 0.149–0.194). Three of the four axes are nearly flat: a 20× variation of L_1 moves $V_{\min}$ by only 0.0004,

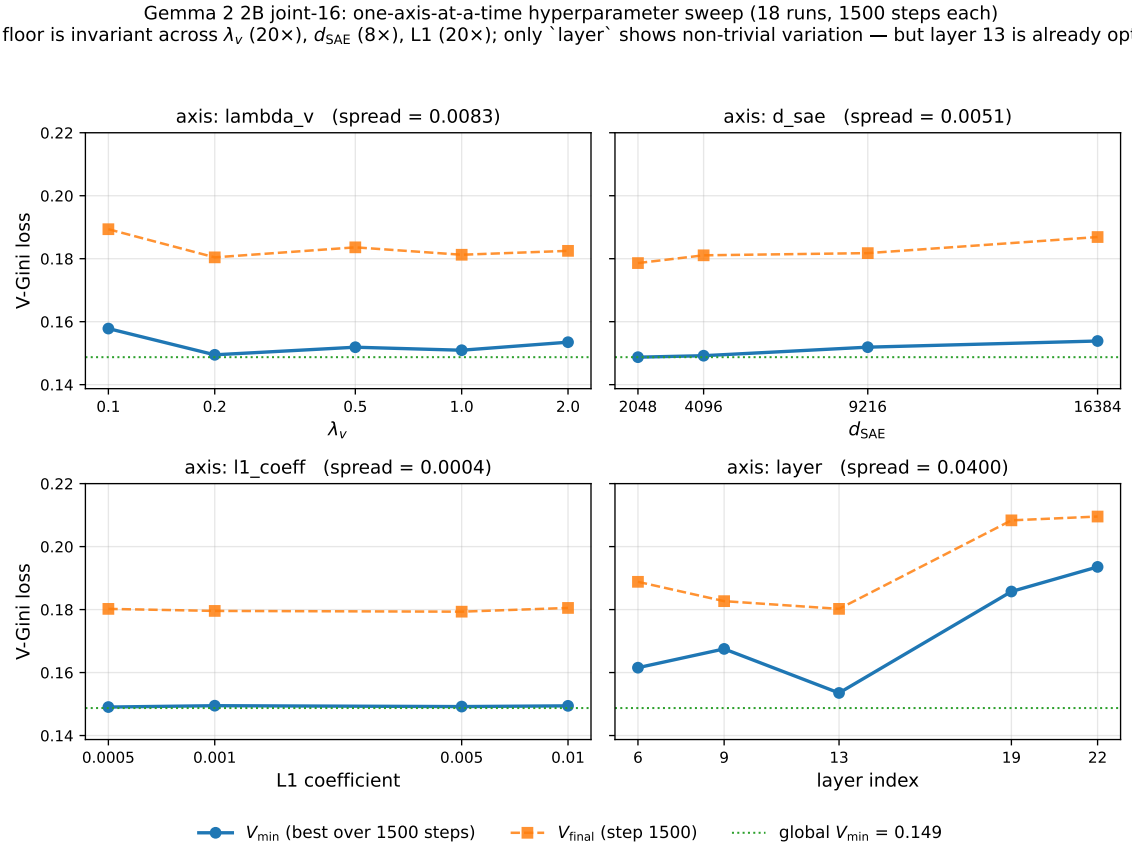

Figure 3: Gemma 2 2B hyperparameter smoke sweep (18 runs, 1,500 steps each). Each panel varies one axis (λ_2, d_{sae}, L_1, layer) and plots the minimum V_{Gini} reached. The floor at $V_{\min} \approx 0.149$ is invariant to a 20× variation of λ_2, d_{sae}, and L_1; only the layer axis moves $V_{\min}$ appreciably, and layer 13 is already optimal. Per-axis values are tabulated in Appendix E.

a 8× variation of d_{sae} by 0.0051, and a 20× variation of λ_2 by 0.0083. The only axis with non-trivial variation is the layer (spread 0.0400), and it moves in the "wrong" direction — layer 13 is already the best of the five tested layers, with layers 19 and 22 plateauing higher.

The interpretation is that the layer-13 residual representation of Gemma 2 2B contains a blind-spot subspace that is invariant under generic SAE reparametrisation with V-regularisation. The limitation is not all-or-nothing — the `word_order`, `gender_swap`, and `synonym` families do reach $V \approx 0$ — but is concentrated in semantically rich families (`semantic_substitution`, `number_swap`, `frequency_change`, `date_time_change`) whose layer-13 representations appear to be encoded in a non-redistributable superposition. Resolving this floor would require structurally different mechanisms (non-Frobenius orthogonality couplings, layer-specific factor decompositions, or upstream representation interventions), not hyperparameter tuning. This refines, rather than undermines, the central claim: V-regularisation removes blind spots wherever the hidden state carries the requisite information, and where it does not, the residual V becomes a diagnostic of a model-specific representational limit.

6 Discussion

The central finding is that blind spots in language model representations are an orientation problem, not a capacity problem. The five SAE variants tested here share the same capacity ($d_{\text{sae}} = 4{,}096$) and training data. They differ only in the training objective. The standard, JumpReLU, and MDL objectives produce

nearly identical V profiles because none of them targets perturbation awareness. Only the V-regularisation targets it, and only the V-regularisation reduces it. The implication is that perturbation awareness is an independent axis of the SAE design space, orthogonal to reconstruction fidelity, sparsity, and description length.

This finding has immediate consequences for AI development practice. Model developers currently evaluate SAEs on reconstruction error and downstream task performance, neither of which captures perturbation awareness. A model that scores well on both metrics can still carry severe blind spots. The V measure provides a third evaluation axis that developers can integrate into their testing pipelines alongside existing metrics. Concretely, a team building a clinical-facing language model could specify a set of safety-relevant perturbation families (dosage changes, negation of contraindications, unit-of-measure substitutions), measure the V profile of their feature decomposition, and apply V-regularisation to any family that exceeds a safety threshold. The regularisation adds a single term to the SAE training loss and requires no changes to the base model, the training data, or the deployment infrastructure.

The sixteen-family generalisation strengthens this picture decisively. On GPT-2 and Qwen the regularisation eliminates blind spots not only on the six families used to derive the metric but on ten held-out medical families — including the dosage, frequency, severity, and unit-of-measure changes that a clinical pipeline must not miss — which reach a mean V of 0.000 (Table 5). That the out-of-distribution families are reduced as thoroughly as the in-sample ones indicates the learned orientation is a property of the representation rather than an overfit to a fixed family list, and it is the result we would lead with in deployment: a single basis, regularised once, closes blind spots across perturbation types it was never shown. The Gemma 2 layer-13 floor adds an important qualification. Where a hidden state lacks the information to distinguish a perturbation family, no reorientation of the feature basis can manufacture it; the residual V then measures a representational limit of the upstream model rather than a deficiency of the SAE objective. This is consistent with the scale-dependence observed in §5.3 and gives practitioners a concrete diagnostic: a floor that survives a hyperparameter sweep points to the choice of model or layer, not to the regulariser.

The urgency of this problem extends beyond individual systems. The WHO projects a global shortage of 11 million health workers by 2030 (World Health Organization, 2025), a gap increasingly expected to be bridged by AI-assisted clinical workflows and machine-to-machine data pipelines that operate with minimal human oversight. In such pipelines a corrupted or adversarially poisoned input ("80 mg" instead of "8 mg") propagates undetected if the feature decomposition treats the two values as equivalent, and traditional syntactic data-quality checks cannot catch semantic-level poisoning that exploits representational blind spots.

The V measure also makes blind spots quantifiable for governance purposes. If a deployed clinical system has $V > 0$ on a safety-relevant perturbation family, that blind spot is a measurable quantity, not an anecdotal risk. Regulators can specify V thresholds, and developers can demonstrate compliance by reporting the V profile of their system at each release. The adversarial robustness literature measures perturbation magnitude (how much perturbation is needed to fool the system) but not perturbation direction (which specific input changes the system cannot see); the V measure fills this directional gap.

The V-regularisation is a post-hoc intervention that improves perturbation awareness without retraining the base model and without risking degradation of its knowledge. This separation of knowledge and orientation is particularly valuable in regulated domains such as healthcare, where model retraining requires extensive re-validation and regulatory approval.

The framework connects naturally to cognitive science. Human perceptual blind spots, including change blindness and inattentional blindness (Simons & Chabris, 1999), can be interpreted through the same algebraic lens: input changes that fall in the stabiliser of the current attentional state remain undetected. Looking forward, the SAE training objective is revealed as a multi-dimensional Pareto problem (reconstruction, sparsity, perturbation awareness, compressibility) whose full surface remains unmapped; the function $V(\lambda_2, d_{\text{sae}}, \text{scale})$ defines a second scaling law complementing the knowledge-focused laws (Kaplan et al., 2020); and determining the orientation cost m^* for broader perturbation taxonomies would characterise the algebraic limits of single-basis feature decomposition.

7 Conclusion

Representational blind spots in language models are eliminable by reorienting the feature basis, without retraining the model or adding capacity. A single differentiable regularisation term, optimising the Gini coefficient of perturbation sensitivities, reduces blind-spot severity by an order of magnitude where standard, JumpReLU, and MDL objectives leave it untouched. The headline result is the generalisation: on GPT-2 and Qwen 2.5 a single jointly regularised basis drives the mean V across *sixteen* perturbation families to numerical noise (-99.3% and -100%), and the ten held-out medical families — dosage, frequency, severity, drug-name, and unit-of-measure changes — are eliminated as completely as the six families used to derive the metric. Where the method does not reach zero — the Gemma 2 layer-13 structural floor, robust to a $20\times$ hyperparameter sweep — the residual vulnerability is a property of the hidden state, turning V into a diagnostic of model-level representational limits rather than a failure of the objective. Perturbation awareness is thus both an actionable training signal and a measurable safety certificate.

Broader Impact Statement

This work targets a safety-relevant failure mode of language models used in clinical and other high-stakes settings: the inability to distinguish input changes (numeric dosages, negations, unit-of-measure substitutions) that matter for the task. The V measure and its regularisation are intended to make such blind spots measurable and reducible, and we expect the primary impact to be positive — enabling developers and regulators to quantify and close representational vulnerabilities before deployment. We caution that a low V on a tested family is not a guarantee of safety on untested families, and that the Gemma 2 floor shows the method cannot manufacture information a model does not encode. The V profile should therefore complement, not replace, downstream clinical validation and human oversight. The perturbation families used here are synthetic and English-language; deployment in other languages or clinical sub-domains requires re-deriving family-specific perturbation sets.

Reproducibility and supplementary material

This paper is self-contained: the main text states all settings needed to reproduce the headline results, and Appendices A–G give the full derivations, per-family and per-axis tables, and example sentence pairs for all sixteen perturbation families.

As anonymised supplementary material (a single archive, well within the TMLR 100 MB limit) we provide: (i) the perturbation-pair datasets — the six original families (50 pairs each) and the ten medical families (30 pairs each), 600 pairs in total; (ii) the per-family and per-axis result files behind every table and figure (the `joint_eval_*` and `smoke_sweep_summary` records); and (iii) all code for training, evaluation, and the hyperparameter sweep. The pipeline uses PyTorch and the HuggingFace `transformers` library and reproduces on a single consumer GPU. Activation caches and SAE checkpoints are regenerated by the scripts on first run and are therefore not bundled. A non-anonymous archival deposit (with a DOI) will be linked in the camera-ready version.

On the use of AI assistance

During the preparation of this work the author used a large language model assistant to help formalise the V_{Gini} loss, draft the training and evaluation scripts, identify and correct errors in intermediate results, and refine the prose. The conceptual framework, experimental design, and all scientific conclusions are the author's own.

References

Anthropic. Findings from a pilot anthropic–openai alignment evaluation exercise, 2025. `https://alignment.anthropic.com/2025/openai-findings/`.

Kola Ayonrinde and Michael T. Pearce. Interpretability as compression: Reconsidering SAE explanations of neural activations with MDL-SAEs. Preprint, 2024. https://arxiv.org/abs/2410.11179.

Csaba Balogh. A general theory of error detection via symmetry breaking. Preprint, 2026a. https://doi.org/10.5281/zenodo.20129750.

Csaba Balogh. Orientation over capacity: A stabiliser-theoretic analysis of representational vulnerability in neural networks. Preprint, 2026b. https://doi.org/10.5281/zenodo.20247188.

Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Trenton Bricken, Adly Templeton, Joshua Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. Technical report, Anthropic, Transformer Circuits Thread, 2023. https://transformer-circuits.pub/2023/monosemantic-features.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. Preprint, 2021. https://arxiv.org/abs/2104.13478.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Fred J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7:171–176, 1964. doi: 10.1145/363958.363994.

H. Michael Damm. *Total anti-symmetrische Quasigruppen*. PhD thesis, Philipps-Universität Marburg, 2004.

Persi Diaconis. *Group Representations in Probability and Statistics*. IMS Lecture Notes–Monograph Series Vol. 11. Institute of Mathematical Statistics, Hayward, CA, 1988.

Nelson Elhage, Tristan Hume, Catherine Olsson, et al. Toy models of superposition. Technical report, Anthropic, Transformer Circuits Thread, 2022. https://transformer-circuits.pub/2022/toy_model.

Leo Gao, Tom Dupré la Tour, Henk Tillman, et al. Scaling and evaluating sparse autoencoders. Preprint, 2024. https://arxiv.org/abs/2406.04093.

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

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

H. Peter Gumm. A new class of check-digit methods for arbitrary number systems. *IEEE Transactions on Information Theory*, 31(1):102–105, 1985. doi: 10.1109/TIT.1985.1056991.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. *Advances in Neural Information Processing Systems*, 35:30016–30030, 2022.

Maël Jullien, Marco Valentino, Hannah Frost, et al. Semeval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of SemEval*, pp. 1258–1279, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. Preprint, 2020. https://arxiv.org/abs/2001.08361.

Connor Kissane, Robert Krzyzanowski, Arthur Conmy, and Neel Nanda. Enhancing neural network interpretability with feature-aligned sparse autoencoders. Preprint, 2024. https://arxiv.org/abs/2411.01220.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Milad Moradi, Kathrin Blagec, and Matthias Samwald. Deep learning models are not robust against noise in clinical text. Preprint, 2021. https://arxiv.org/abs/2108.12242.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024.001, 2020. doi: 10.23915/distill.00024.001.

OpenAI. Strengthening our safety ecosystem with external testing, 2025. https://openai.com/index/strengthening-safety-with-external-testing/.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders do not find canonical units of analysis. Preprint, 2025. https://arxiv.org/abs/2502.04878.

Joseph J. Pollock and Antonio Zamora. Automatic spelling correction in scientific and scholarly text. *Communications of the ACM*, 27:358–368, 1984. doi: 10.1145/358027.358048.

Qwen Team. Qwen2.5: A party of foundation models. Preprint, 2024. https://arxiv.org/abs/2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, et al. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. Preprint, 2024. https://arxiv.org/abs/2407.14435.

Daniel J. Simons and Christopher F. Chabris. Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception*, 28:1059–1074, 1999. doi: 10.1068/p281059.

Yibo Tang. A unified theory of sparse dictionary learning in mechanistic interpretability. Preprint, 2025. https://arxiv.org/abs/2512.05534.

Adly Templeton, Tom Conerly, Jonathan Marcus, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Technical report, Anthropic, Transformer Circuits Thread, 2024. https://transformer-circuits.pub/2024/scaling-monosemanticity.

Jacobus Verhoeff. *Error Detecting Decimal Codes*. Mathematical Centre Tracts No. 29. Mathematisch Centrum, Amsterdam, 1969.

World Health Organization. Health workforce: global strategy and governance, 2025. https://www.who.int/teams/health-workforce/global-strategy.

A Differentiability of V_{Gini}

The Gini coefficient (Eq. 2) involves a sort, which is not differentiable in the classical sense because infinitesimal perturbations do not change the ordering. However, V_{Gini} is a piecewise-linear function of the unsorted values $s_1, \ldots, s_K$, and its subgradient is well-defined almost everywhere (undefined only when two values are exactly equal, a measure-zero event). In our implementation we use `torch.sort`, which returns both the sorted values and the permutation; its backward pass propagates gradients to the unsorted inputs via the inverse permutation, equivalent to the straight-through estimator for discrete permutations and standard in differentiable sorting (Blondel et al., 2020).

The gradient of V_{Gini} with respect to the unsorted sensitivity s_j is

$$\frac{\partial V_{\mathrm{Gini}}}{\partial s_j} = \frac{2\,\pi(j)}{K \cdot S} - \frac{2\sum_{i=1}^{K} i\, s_{(i)}}{K \cdot S^2}, \tag{4}$$

where $\pi(j)$ is the rank of s_j and $S = \sum_i s_i$. It is well-defined whenever $S > 0$ and no two values are exactly equal, both of which hold in practice due to the ε in Eq. (1) and the continuous nature of the hidden-state activations. The gradient propagates through the encoder by the chain rule, with the ReLU non-differentiability at zero handled by the standard subgradient convention.

B Multi-factor SAE architecture

The multi-factor SAE implements the D_n^m analogy from check-digit theory. It consists of m parallel encoders, each mapping the hidden state $h \in \mathbb{R}^{d_{\text{in}}}$ to a factor code $z^{(k)} \in \mathbb{R}^{d_{\text{factor}}}$, and a shared decoder mapping the concatenation $z = [z^{(1)}; \ldots; z^{(m)}]$ back to $\hat{h}$. Total capacity is held constant at $d_{\text{factor}} \times m = d_{\text{sae}}$, so the comparison between m=1 and m>1 isolates the effect of factorisation at fixed capacity. Families are assigned to factors round-robin ($i \bmod m$); the V loss for factor k is computed on $z^{(k)}$ only, while evaluation computes V_{Gini} on the concatenated code. On both GPT-2 and Gemma 2 2B, $m^* = 1$: a single encoder regularised on all six families achieves $V < 0.05$ on all of them. The architecture was tested at $m = 1, 2, 3, 6$ to confirm the m=1 result is not an artefact of the specific factorisation. (An orthogonality penalty $\lambda_{\text{ortho}} \| {W^{(j)}}^\top W^{(k)} \|_F^2$ encourages the factors to span complementary subspaces.)

C Input-normalised gain and perturbation complexity

The raw sensitivity s_i (Eq. 1) does not account for the magnitude of the input perturbation. The input-normalised gain divides the feature-space sensitivity by the hidden-state sensitivity,

$$g_i = \frac{\|z_i^{\text{pert}} - z_i^{\text{orig}}\| / \|z_i^{\text{orig}}\|}{\|h_i^{\text{pert}} - h_i^{\text{orig}}\| / \|h_i^{\text{orig}}\|}, \tag{5}$$

and V_{Gini} of the g_i measures whether the SAE applies a uniform amplification factor across all pairs. The V-reg$_{\text{n}}$ variant (trained with $\lambda_2 \cdot V_{\text{gain}}$) achieves $V_{\text{gain}} = 0.000$ on all six families on all three models (Table 6). The trade-off is that V-reg$_{\text{n}}$ does not minimise the raw V (its raw values are comparable to or higher than the standard SAE on some families, Table 1); the two variants occupy different points on the Pareto surface of perturbation awareness.

Table 6: V_{Gini} (gain) for the V-reg$_{\text{n}}$ variant. All values are 0.000, indicating perfect gain uniformity.

Family	GPT-2	Gemma 2	Qwen 2.5
Negation	0.000	0.000	0.000
Typo	0.000	0.000	0.000
Synonym	0.000	0.000	0.000
Number swap	0.000	0.000	0.000
Semantic sub.	0.000	0.000	0.000
Word order	0.000	0.000	0.000

The six families differ in input-level complexity, measured by the coefficient of variation (CV) of the hidden-state perturbation magnitudes (Table 7). Semantic substitution has the highest CV on all three models (0.84–0.91); negation has the lowest on Gemma (0.23). The fact that V-reg achieves $V \approx 0$ even on high-CV families (semantic substitution on Qwen: $V = 0.001$, CV= 0.89) shows that it overcomes input-level complexity heterogeneity rather than merely benefiting from it.

D Reconstruction quality across methods

On GPT-2, the V-reg MSE (0.0069) is about twice the standard (0.0034), a modest cost for a 10× reduction in V. On Gemma 2, V-reg achieves the lowest MSE of all five methods (0.0119 vs standard 0.0130), indicating that perturbation-aware orientation can sometimes improve reconstruction. On Qwen 2.5, V-reg

Table 7: Input perturbation complexity per family (CV of $\|h^{\text{pert}} - h^{\text{orig}}\|/\|h^{\text{orig}}\|$ across 50 pairs).

Family	GPT-2	Gemma 2	Qwen 2.5
Negation	0.530	0.229	0.285
Typo	0.583	0.309	0.298
Synonym	0.640	0.263	0.340
Number swap	0.581	0.577	0.435
Semantic sub.	0.914	0.841	0.891
Word order	0.622	0.373	0.209

Table 8: Reconstruction MSE and sparsity (fraction of active features) across five SAE methods and three models. All share $d_{\text{sae}} = 4{,}096$ and 5,000 steps.

Model	Method	MSE	Frac active
GPT-2	Standard	0.0034	0.352
	JumpReLU	0.0035	0.333
	MDL	0.0065	0.337
	V-reg	0.0069	0.344
	V-reg$_{\text{n}}$	0.0033	0.329
Gemma 2	Standard	0.0130	0.072
	JumpReLU	0.0222	0.097
	MDL	0.0250	0.076
	V-reg	0.0119	0.094
	V-reg$_{\text{n}}$	0.0142	0.071
Qwen 2.5	Standard	0.0052	0.172
	JumpReLU	0.0020	0.204
	MDL	0.0038	0.248
	V-reg	0.0104	0.162
	V-reg$_{\text{n}}$	0.0040	0.160

has the highest MSE (0.0104), but the absolute level remains low. The V-reg$_{\text{n}}$ variant consistently achieves reconstruction comparable to or better than the standard, because the gain-normalised V loss places less pressure on the encoder weights.

E Detailed analysis of the Gemma 2 layer-13 floor

This appendix provides the per-family and per-axis detail behind §5.8. All runs use `train_joint.py` (single factor) or `train_joint_mfactor.py` (multi-factor), the sixteen families of `joint16_pairs.json`, and 7,500 steps unless stated.

E.1 Three controlled configurations

Starting from a stabilised baseline (a) m=1, λ_2=0.2, gradient clip 1.0, activation clip 200, we change exactly one variable: (b) raises λ_2 to 0.5; (c) uses m=2 orthogonal factors (λ_{ortho}=0.05) at the same total capacity. Table 9 gives the per-family V_{raw} over 50 evaluation pairs. Raising λ_2 barely moves the floor (mean V 0.177 $\rightarrow$ 0.173), and the multi-factor variant is slightly *worse* (0.196), although it does halve the gain-normalised vulnerability ($V_{\text{gain}} \approx 0.14 \rightarrow 0.08$).

All three runs plateau in the 0.12–0.20 band by step 500 and never escape it (Table 10).

Table 9: Per-family V_{raw} for the three Gemma 2 2B controls (50 eval pairs/family). (a) λ_2=0.2, m=1; (b) λ_2=0.5, m=1; (c) λ_2=0.2, m=2. Source: `results/gemma-2-2b/joint16/lambda_sweep_eval.json`.

Family	(a) lv02	(b) lv05	(c) m=2
word_order	0.000	0.000	0.135
gender_swap	0.045	0.044	0.071
synonym	0.057	0.057	0.148
negation	0.087	0.100	0.101
condition_negation	0.148	0.136	0.152
severity_change	0.151	0.146	0.194
drug_name_swap	0.190	0.191	0.148
causal_reversal	0.201	0.188	0.210
unit_of_measure	0.204	0.191	0.217
anatomical_direction	0.207	0.203	0.220
date_time_change	0.217	0.225	0.229
body_part_swap	0.228	0.221	0.219
typo	0.228	0.220	0.194
frequency_change	0.267	0.263	0.278
number_swap	0.295	0.281	0.278
semantic_substitution	0.307	0.299	0.339
mean V_{raw}	**0.177**	**0.173**	**0.196**
V<0.05	2/16	2/16	0/16
V<0.005	1/16	1/16	0/16

Table 10: V-loss training trajectories for the three Gemma controls.

Step	(a) λ_2=0.2	(b) λ_2=0.5	(c) m=2
1	0.358	0.332	0.358
500	0.190	0.178	0.166
2000	0.151	0.140	0.143
4000	0.176	0.124	0.127
6000	0.180	0.164	0.178
7500	**0.166**	**0.157**	**0.169**

E.2 Systematic 4-axis hyperparameter sweep

To rule out hyperparameter tuning as a route past the floor, we ran an 18-configuration one-axis-at-a-time sweep (orchestrator `smoke_sweep_gemma.py`; each run a 1,500-step `train_joint.py` call otherwise identical to baseline (a)). $V_{\min}$ is the lowest V-loss during the trajectory (logged every 250 steps); V_{final} is the value at step 1,500. Table 11 reports all four axes.

The L_1 axis is the most consistent (20× variation → $V_{\min}$ spread 0.0004): the sparsity prior cannot redistribute the layer-13 vulnerability. The d_{sae} axis is next tightest (8× capacity → 0.0051), and the *smallest* SAE (d_{sae}=2048) attains the lowest $V_{\min}$; expansion does not unstick the floor. The λ_2 axis (20× penalty → 0.0083) shows the optimiser finds the same low-V region with a larger objective value. Only the layer axis varies appreciably, and layer 13 is already the best of the five tested. Extending training to 7,500 steps does not improve on the 1,500-step floor either. We conclude the floor is a structural property of the layer-13 hidden states, invariant under generic SAE reparametrisation with V-regularisation.

Table 11: 18-run Gemma 2 2B hyperparameter sweep. The minimum reachable V_{Gini} is 0.149 globally; three axes are nearly flat and only the layer axis varies appreciably. Source: `results/gemma-2-2b/joint16/smoke_sweep_summary.json`.

Axis	Value	V_{final}	$V_{\min}$	recon	L_1
λ_2	0.1	0.1894	0.1578	0.0222	0.6295
	0.2	0.1804	**0.1495**	0.0324	0.6136
	0.5	0.1836	0.1519	0.0201	0.6050
	1.0	0.1813	0.1509	0.0169	0.6293
	2.0	0.1825	0.1535	0.0224	0.6820
	spread		**0.0083**		
d_{sae}	2048	0.1786	**0.1487**	0.0248	0.9996
	4096	0.1811	0.1492	0.0279	0.6324
	9216	0.1818	0.1519	0.0346	0.3293
	16384	0.1869	0.1539	0.0540	0.1822
	spread		**0.0051**		
L_1	5e-4	0.1802	0.1490	0.0221	0.6335
	1e-3	0.1796	0.1495	0.0188	0.6151
	5e-3	0.1793	0.1492	0.0293	0.5693
	1e-2	0.1805	0.1494	0.0204	0.4998
	spread		**0.0004**		
layer	6	0.1888	0.1615	0.0218	0.5132
	9	0.1827	0.1675	0.0175	0.6514
	13	0.1802	**0.1535**	0.0140	0.6135
	19	0.2083	0.1857	0.0472	0.9756
	22	0.2095	0.1935	0.0583	1.0091
	spread		**0.0400**		
global	(18 runs)		**0.1487–0.1935**		

E.3 Robustness to numerical precision and activation clipping

Two implementation choices specific to Gemma's heavy-tailed activations might in principle inflate the floor, and we rule both out with a 2×2 control (λ_2=0.2, 1,500 steps). First, the model is trained natively in bfloat16, whereas the activation pipeline defaults to float16; we therefore re-ran the cache build and training in bfloat16. This does not lower the floor — $V_{\min}$ moves from 0.149 (float16) to 0.165 (bfloat16), i.e. marginally *higher* — consistent with the largest activation component ($\approx 1.3 \times 10^3$) being far below the float16 overflow threshold (6.6×10^4). Second, the per-sample L_2 activation clip (used to stabilise the outlier dimensions) rescales each side of a perturbation pair independently and is therefore not scale-invariant; removing it leaves $V_{\min}$ unchanged at 0.149, consistent with only 4/2000 samples exceeding the clip threshold. Neither numerical precision nor activation clipping is the source of the plateau.

F Pre-caching of perturbation hidden states

An early V-reg run on Gemma 2 2B that computed the V term with live forward passes through the frozen LM at every step produced an anomalously high reconstruction MSE (0.879 vs the standard's 0.020), caused by stochastic variation in the perturbation hidden states (batching and padding). Switching to pre-cached perturbation hidden states — one forward pass per pair, stored and reused at every step — dropped the MSE to 0.012, below the standard SAE, and four of six families reached $V = 0.000$. The lesson is that the V term should operate on deterministic, pre-cached hidden states, analogous to standard activation caching for the reconstruction loss. For production use, λ_2 should be selected on a validation set via the Pareto frontier of reconstruction MSE versus mean V_{Gini}.

G Perturbation family descriptions

The six original families contain 50 sentence pairs each; the ten medical-domain families (used in the sixteen-family study of §5.7) contain 30 pairs each, for 600 pairs in total. Representative examples for every family follow; the complete datasets are provided as supplementary material (see Reproducibility and supplementary material).

G.1 Original families

Semantic substitution. "The cat sat on the mat…" → "The dog sat on the mat…"; "…very hot and sunny…" → "…very cold and sunny…"; "…drove her car…" → "…drove her truck…"; "The doctor examined…" → "The nurse examined…"; "…from the small shop" → "…from the small store".

Typo (adjacent-character transposition). "…admitted to the hospital…" → "…hopsital…"; "…published the results…" → "…publsihed…"; "…an important message…" → "…mesasge…"; "…their assignment…" → "…thier…"; "…explained the concept…" → "…explaind…".

Negation. "The test result is positive…" → "…is not positive…"; "The drug is safe…" → "…is not safe…"; "The system is working…" → "…is not working…"; "The patient is responding…" → "…is not responding…"; "The data is consistent…" → "…is not consistent…".

Synonym. "…very big…" → "…very large…"; "…happy to hear…" → "…glad to hear…"; "…very hard to complete…" → "…very difficult…"; "She started working…" → "She began working…"; "…fix the broken machine…" → "…repair…".

Number swap. "…received 3 doses…" → "…8 doses…"; "…5 hospitals…" → "…9 hospitals…"; "…2 pills…" → "…7 pills…"; "…scored 90 points…" → "…40 points…"; "…reached 30 degrees…" → "…80 degrees…".

Word order (adjacent-word swap). "The black cat…" → "The cat black…"; "She quickly finished…" → "She finished quickly…"; "The old man…" → "The man old…"; "The bright sun…" → "The sun bright…"; "The young student…" → "The student young…".

G.2 Medical-domain families (out-of-distribution)

These ten auto-generated families were held out from the design of the metric and training protocol; the sixteen-family study (§5.7, Table 5) evaluates generalisation to them. Three representative pairs per family follow.

Anatomical direction (anterior↔posterior, proximal↔distal, dorsal↔ventral). "The mass is located in the *anterior* mediastinum" → "…*posterior* mediastinum"; "Pain is felt in the *proximal* part of forearm" → "…*distal* part…"; "The lesion is on the *dorsal* surface of hand" → "…*ventral* surface…".

Body-part swap (left↔right). "…pain in the *left* knee" → "…*right* knee"; "Fracture detected in the *right* femur" → "…*left* femur"; "Swelling observed in the *left* ankle" → "…*right* ankle".

Causal reversal. "The fever was caused by the infection" → "The infection was caused by the fever"; "Pain *increased* after taking the medication" → "Pain *decreased*…"; "…resulted in significant *improvement*" → "…significant *deterioration*".

Condition negation. "The patient *has* a history of diabetes" → "…*has no* history of diabetes"; "Allergies to penicillin *are* documented" → "*No* allergies…documented"; "There *is* evidence of metastatic disease" → "There *is no* evidence…".

Date/time change. "... scheduled for *Monday* morning" → "... *Friday* morning"; "... admitted on *January* 15" → "... *March* 15"; "Follow-up in *2* weeks" → "... *6* weeks".

Drug-name swap. "... prescribed *metformin* daily" → "... *lisinopril* daily"; "Start *ibuprofen* 400 mg..." → "Start *acetaminophen* 400 mg..."; "... taking *warfarin* for clots" → "... *aspirin* for clots".

Frequency change. "Take the medication *once* daily" → "... *three times* daily"; "Apply the cream *twice daily*" → "... *once weekly*"; "... given every *two weeks*" → "... every *two months*".

Gender swap (he↔she, his↔her). "... reported *he* felt dizzy" → "... *she* felt dizzy"; "*She* was admitted to the ward" → "*He* was admitted..."; "*His* blood pressure was elevated" → "*Her* blood pressure...".

Severity change (mild→severe, minor→major). "... *mild* chest pain" → "... *severe* chest pain"; "... *moderate* difficulty breathing" → "... *extreme* difficulty..."; "... *slight* swelling" → "... *significant* swelling".

Unit of measure (mg→g, mL→L, cm→mm). "The dosage is 500 *mg*..." → "... 500 *g*..."; "Inject 0.5 *mL*..." → "Inject 0.5 *L*..."; "... measured 2.3 *cm*..." → "... 2.3 *mm*...".