
Your Intervention Was Too Narrow: Mechanism Width as a Precondition for Interpretability-Driven Discovery

Sumit Yadav Basanta Joshi

Pulchowk Campus, Institute of Engineering, Tribhuvan University, Nepal
076bct088.sumit@pcampus.edu.np, basanta@ioe.edu.np

Abstract

Interpretability-driven discovery rests on interventions: a part is credited with a function when perturbing it changes behavior, and dismissed when nothing happens. We show that the field’s standard intervention unit (one direction, one layer, one attention head) recovers only $\approx 10\%$ of the achievable causal effect in three unrelated testbeds, one task each: rank-1 steering in vision encoders (0.111), single-layer ablation in Llama-3.1-8B (0.096), and single-head sufficiency in GPT-2’s IOI circuit (0.086). The mechanism’s width k^* can be measured *before* intervening, and widening the intervention to k^* recovers 95–100%. The effect is about width itself: a 12-layer band does $7.5\times$ the additive damage prediction, random subsets track contiguous bands, and pushing past k^* destroys object identity. Across 12 language models, whether a single-unit test works is predicted by a cliff in the writer-score ladder ($\rho = 0.81$). Width is necessary, not sufficient: at a documented *readout boundary*, 81% of a representation gap closes while the consumer’s output moves ≈ 0 , though it reads the attribute when prompted. A negative result at standard width is an underpowered test, not evidence of absence; a cheap protocol (sweep width, compare k_{repr}^* to k_{eff}^* , run matched controls) tells the difference before a discovery claim is made.

1 Introduction

Interpretability-for-discovery holds that models internalize regularities humans have not yet articulated, and that we can read them back out [Kim et al., 2018, Olah et al., 2020]. Because probes detect information a model may never use [Belinkov, 2022, Alain and Bengio, 2017], the evidentiary standard is the *intervention*: believe the interpretation only if perturbing the part moves behavior [Meng et al., 2022, Wang et al., 2023]. Interventions license positive claims, and just as often negative ones. We steered the vector, we ablated the layer, nothing changed; the feature must not be causally involved.

This paper is about a systematic flaw in that inference. The standard intervention unit is one thing: one CAV direction [Kim et al., 2018, Turner et al., 2023], one residual layer [Gromov et al., 2024, Men et al., 2024], one attention head [Wang et al., 2023]. If the mechanism occupies k^* units, and we find k^* between 2 and 65, a width-1 intervention tests a $1/k^*$ sliver of it, and its null is an *underpowered test*, not evidence of absence. Width 1 recovers 8.6% (GPT-2 heads), 9.6% (Llama layers), and 11.1% (DINOv2 orientation) of what matching width achieves (Fig. 1a).

Our contributions:

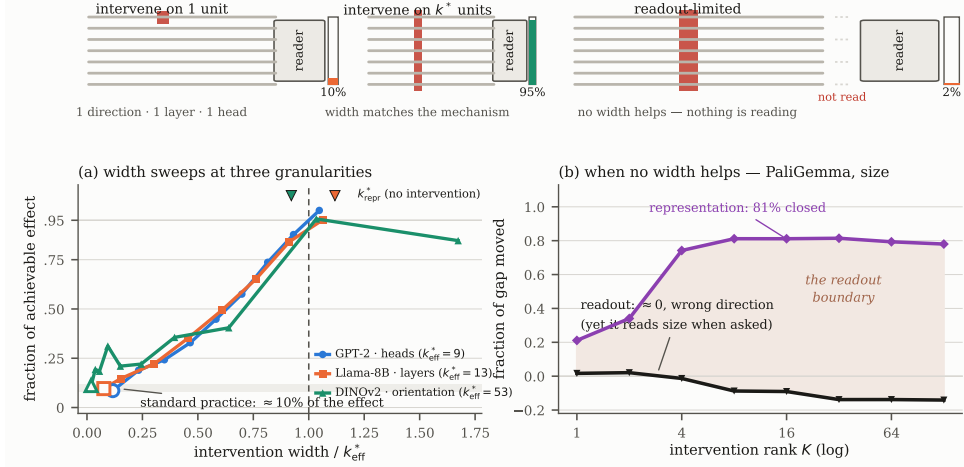


Figure 1: **The unit of intervention must match the width of the mechanism.** *Top:* a mechanism occupies k^* units (directions, layers, or heads); intervening on one unit leaves the rest carrying the signal, intervening on k^* units captures it, and when nothing downstream reads the signal, no width helps. *(a)* Fraction of the achievable effect versus normalized intervention width for three granularities in three unrelated systems: best head-coalition of each size in GPT-2’s IOI circuit (exhaustive 512-coalition lattice), contiguous layer-band removal in Llama-3.1-8B (100 queries), rank- K orientation steering in DINOv2. The x -axis is normalized by each curve’s own k_{eff}^* , so passing through $(1, 0.95)$ holds *by construction*. The evidence is the rest: width 1 (open markers) recovers 0.086–0.111, and the representation-only width k_{repr}^* (triangles on top; no intervention used) lands at $0.92\times$ (orientation) and $1.15\times$ (layers) of k_{eff}^* . The GPT-2 head curve has no representation-side width. *(b)* The readout boundary: at PaliGemma’s vision tower (L24), rank- K steering closes up to 81% of the size representation gap while the frozen language model’s answer moves ≈ 0 at every rank. When simply *asked*, the same model reports the attribute with a +0.32 probability gap (beige band).

1. **Width sweeps at three granularities** (§3): monotone everywhere, width 1 far below saturation. Across 12 language models spanning four families, single-unit success is predicted by a cliff in the writer ladder ($\rho = 0.81$, $p = 0.001$; Fig. 2).
2. **Width, not magnitude or contiguity** (§4): a 12-layer band does $7.5\times$ the additive prediction; random subsets track contiguous bands; widening past k^* collapses identity ($0.97 \rightarrow 0.15$ at $7k^*$).
3. **The readout boundary** (§5): an edit moving a within-tower probe to 0.82 of the gap leaves captions invariant at every rank and site, though prompting proves the LM reads the attribute; at width 1 a random direction beats the concept vector (0.65 vs. 0.22).
4. **A diagnostic** (§6): k_{repr}^* vs. k_{eff}^* ; agreement means width-limited, k_{repr}^* without k_{eff}^* means readout- or basis-limited. Table 1 classifies 17 pairs.

For discovery the stakes are concrete: “the model does not use X ” backed by a width-1 null is exactly the finding that sends a domain scientist looking elsewhere, wrongly.

2 Two widths

Setup. Three testbeds. *(i) Vision feature space:* DINOv2, CLIP, and SigLIP (base) on Kubric renders that vary object size and orientation; steering swaps the top- K attribute directions toward a target prototype, $h' = h + P_K(t^* - h)$; we report the fraction of the source $\rightarrow$ target readout gap closed, and nearest-centroid object identity. *(ii) Depth:* Llama-3.1-8B-Instruct with k contiguous decoder layers made identity (residual skip); effect is $1 - \text{top-1 agreement with the intact model}$, normalized by the destroyed-model reference. *(iii) Circuit:* GPT-2 small’s IOI circuit; the best coalition of each size from an exhaustive lattice over all 2^9 subsets of the nine circuit heads, effect $(v(S) - v_\emptyset) / (v_{\text{full}} - v_\emptyset)$. Interventions have a direction and the lattice gives both: single heads built

up from empty (sufficiency, Fig. 1a) recover at most 0.086; single heads torn from the intact circuit (necessity, the usual knockout) cost 0.12–0.51, median 0.27. Either direction, a width-1 null misses most of the mechanism.

Two definitions of width. For an attribute with targets t (e.g. $\{\text{scale}\}$, or $\{\sin \theta, \cos \theta\}$ for orientation):

- k_{repr}^* : on the centered representation, rank SVD components by $\sum_t \text{corr}^2$ with the targets; k_{repr}^* is the smallest K reaching 95% of total relevance. *No intervention required.*
- k_{eff}^* : the smallest intervention width reaching 95% of the maximum achievable interventional effect, read off the width sweep. k_{eff}^* exists only if the sweep saturates (≥ 0.9 of the gap); Table 1 labels width-limited when $k_{\text{eff}}^* \leq 1.4 k_{\text{repr}}^*$, basis-limited when no saturating k_{eff}^* appears despite a small k_{repr}^* . These constants are analytic choices of the very kind this paper is about, so we vary them: over a 3^4 grid (0.90/0.95/0.99 for both 95% thresholds, saturation 0.85/0.90/0.95, ratio 1.2/1.4/2.0), no Table 1 row changes between width-limited and basis- or readout-limited, and width-1 fractions use no threshold at all (App. B, Table 2).

One concept, three instantiations. Width is always the number of intervention units (directions, layers, or heads) in the smallest intervention reaching 95% of the achievable effect: k_{eff}^* has this single operational definition in every testbed. What changes with the unit is only the effect metric, which must be native to the system (gap closed, top-1 agreement, logit difference), and how k_{repr}^* is read without intervening: SVD relevance for directions, probe-transfer band width for layers (App. B). Because every effect is normalized to its own maximum, curves are compared in shape and in their width-1 fraction, never in raw units. The two widths are logically independent: one is about where information sits, the other about what a causal handle can grasp. Their relation is the diagnostic. For DINOv2 orientation they agree without being told to: $k_{\text{repr}}^* = 49$ from the representation alone, $k_{\text{eff}}^* = 53.2$ from the sweep, and this is the clean non-circular case. For heads, the natural representation-side count is the number of heads passing a role criterion ($\geq 0.5 \times \text{max previous-token score}$, upstream of the reader), but a sweep over criterion heads alone cannot test that count. We therefore also select heads with no role score at all, greedily adding the upstream head whose *joint* ablation most reduces the reader (App. B, Table 3). The criterion is not a reliable width: in the four GPT-2 models the causal mechanism spans 10–22 heads and the criterion set delivers only 0.20–0.51 of its effect (in GPT-2 XL it shares 0/9 heads with the greedy set), while in Pythia-410m the criterion overcounts (5 heads, $k_{\text{eff}}^* = 2$). A role criterion is itself a width choice, so for heads k_{eff}^* must be measured. For IOI, the nine heads were found by greedy elimination over all 144, not by role; with no size cap, that search needs 7–8 heads for 90–95% of the logit difference, matching $k_{\text{eff}}^* = 8.6$.

3 Width sweeps at three granularities, then twelve models

Figure 1a is the central measurement: intervention width, normalized by each mechanism’s own k_{eff}^* , against effect. Passing through (1, 0.95) is guaranteed by the normalization, so the finding is everything left of it: every curve is monotone, width 1 recovers 0.086/0.096/0.111, and saturation is far away everywhere we measured. Two quantities in the panel are *not* set by the normalization: the width-1 fraction, and where the intervention-free k_{repr}^* lands ($0.92 \times$ and $1.15 \times k_{\text{eff}}^*$, triangles). We do not claim one functional form (the shapes differ; see the seven sigmoids of Fig. 2a), and each testbed is one task with its own metric. What is common is the inference error: the same pattern in three settings, not yet a law.

Does this generalize? We ranked 12 language models’ previous-token *writers* by score and ablated the top- k jointly against size- and layer-matched controls (Fig. 2a). The result is bimodal: 6 models realize more than half the joint effect at $k=1$ and the rest realize less than 0.18, and which group a model falls in is predicted by its ladder cliff. GPT-2 small, the model on which canonical head-level circuit results were established, is the outlier that made single-head tests look adequate. Its writer ladder falls off a cliff ($0.992 \rightarrow 0.547$) that no other model reproduces at its scale (Fig. 2b; Spearman $\rho = 0.81$, $p = 0.001$, $n=12$), and induction-head counts grow with depth ($r = +0.95$). Two cautions: for the smallest models a high $k=1$ fraction is partly a width effect (n_w is small), though Qwen2-0.5B realizes 0.88 at $k=1$ of *eight* writers, so the cliff is not reducible to $1/n_w$; and this induction testbed is a different circuit, task, and direction from the IOI sufficiency curve, which

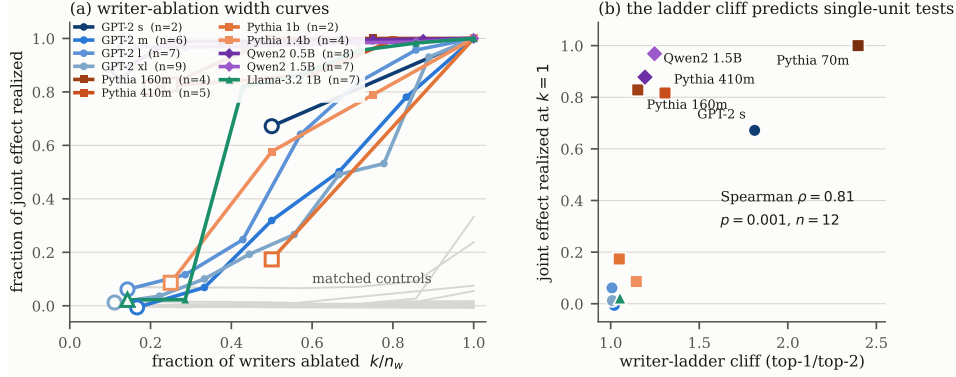


Figure 2: **Universality across 12 models** (GPT-2, Pythia, Qwen2, Llama families). (a) Ablating the top- k previous-token writers versus size- and layer-matched control heads (grey): the effect grows with width in every model. Llama-3.2-1B hides a redundant writer triple: nothing at $k=1, 2$, a cliff at $k=3$. (b) Whether the single-unit test works is predicted by the writer-ladder cliff (top-1/top-2 score): every model above 1.15 realizes most of the joint effect at $k=1$, every model below fails ($\rho = 0.81$, $p = 0.001$).

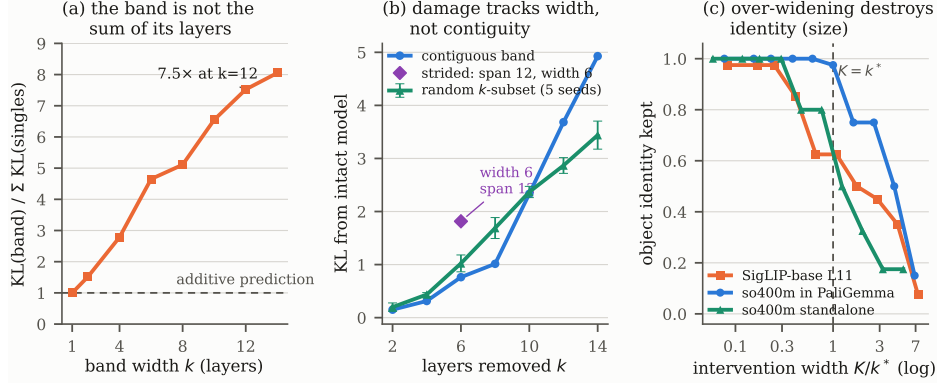


Figure 3: **Controls.** (a) KL of removing a k -layer band versus the additive prediction from the same layers’ single-layer ablations (Llama-3.1-8B, 100 queries): superadditivity reaches $7.5\times$ at $k=12$. (b) Base Llama-3.1-8B on 32 prompts: random k -subsets drawn from the survivable region (5 seeds, mean $\pm$ sd) track contiguous bands; a strided band spanning 12 layers with width 6 behaves like width 6, not span 12. (c) Steering size wider than k^* collapses object identity in three independent measurements; matching width is a sweet spot, not a monotone knob.

is why one model can fail there and succeed here. The cliff is computable from attention patterns alone: a representation-side quantity that predicts whether a single-unit test can work *before* anyone runs it. Fig. 2 normalizes by the criterion writers’ joint effect, which can be a fraction of the whole mechanism; against the criterion-free greedy mechanism (Table 3) the cliff predicts the width-1 fraction more tightly ($\rho = 0.89$, $p = 0.007$, $n=7$).

4 The effect tracks width, not magnitude or position

Three objections, three controls (Fig. 3). “*Wider interventions just perturb more.*” The additive extrapolation from single-layer sweeps predicts 1.85 KL at $k=12$; the band measures 13.9 ($7.5\times$). More diagnostic is the crossover: top-1 agreement sits above the independence prediction until $k \approx 10$ and below it after: individually redundant, collectively indispensable. “*It’s the specific contiguous region.*” Random k -subsets sampled from the interior (layers 2–29) track contiguous damage, within about 25% and with the same growth (KL 2.87 ± 0.15 vs. 3.68 at $k=12$). Sliding the 12-layer band across interior starts 4–16 moves KL only within 3.3–4.1; the band is catastrophic only at the ends (6.0 swallowing layers 0–1, 7.5 swallowing 30–31), so position matters exactly where

single-layer sweeps already say it does. A strided band with the *span* of 12 but the *width* of 6 lands at 1.82 KL, between contiguous 6 (0.76) and contiguous 12 (3.68) and closer to its width than to its span. “Then why not always intervene maximally wide?” Because over-widening destroys what the intervention was supposed to spare: pushing size past k^* drops object identity from 0.97 (at k^*) to 0.15 (at $7k^*$), reproduced in three separate systems (Fig. 3c). And the lift is not a generic high-rank shove: a *random* K -dimensional subspace at $K=55$ closes 0.24 ± 0.20 of the orientation gap where the attribute subspace closes 0.96 (five seeds), just as matched-magnitude random directions are inert at rank 1 [cf. Maskey et al., 2026]. Width must be *matched*, not maximized.

5 The readout boundary: when no width helps

Widening fixes underpowered interventions, not a consumer that is not listening. Steering PaliGemma’s tower (L24, ranks 1–128) dissociates (Fig. 1b; detail in Fig. 4, appendix): the size representation gap closes to 81% while p (“large”) moves -0.045, in the wrong direction, and captions do not change. This holds at three injection sites and under question-conditioned prompting. Yet the LM *can* read size. Asked directly, its answer probability differs by +0.32 between small and large objects. The attribute is present, and it is consumed through a pathway our subspace does not intersect. One mundane alternative is that the LM reads size from spatial extent (how many patch tokens the object covers), which a pooled push cannot change. We tested it: replacing the entire L24 token field with the large-class mean field flips the answer (p from 0.03 to 0.52, past the 0.38 large-object baseline), and a position-shuffled version of the same field, which destroys spatial layout while preserving content, flips it just as well (0.59). The consumer reads L24 token content and layout is not the variable; the wall is specific to the pooled rank- K handle. *Readout-limited* therefore means the handle misses the variable the consumer uses: k_{repr}^* exists (13), no k_{eff}^* at any width of this handle, and switching handles, not widening, is the escape.

Width-1 practice here is worse than underpowered: the rank-1 size CAV moves p (“large”) less than a norm-matched random direction (0.22 vs. 0.65), so a study reporting only the CAV arm would have published a causal effect its own control refutes.

A second refusal is instructive: DINOv2’s *size* never exceeds 0.39 of the gap at $K \approx k^*$ at *any* of the eleven injection depths, while orientation closes at k^* in all three encoders (worst 0.80). Size has $k_{\text{repr}}^* \approx 12$ –16 but stable rank near 1: a curved, nearly one-dimensional manifold a linear handle grips poorly (*basis-limited*). Width is necessary; basis and readout alignment complete the checklist.

6 Protocol, related work, limitations

The protocol. Before reporting an interventional null (or a positive): (1) measure k_{repr}^* ; (2) sweep width past it and locate k_{eff}^* ; (3) run matched controls (random units of equal width and magnitude, plus a prompted capability check); (4) classify: saturates near k_{repr}^* means width-limited; representation moves but the consumer does not means readout-limited; linear handles failing on a low-stable-rank manifold means basis-limited. Only then interpret the null.

Related work. That decodable information need not be causally potent is established [Belinkov, 2022, Hase et al., 2023, Makelov et al., 2023, Elhage et al., 2022]; width-1 steering unreliability is well documented [Tan et al., 2024, Liu, 2026, Wu et al., 2025], and Billa [2026] and Braun et al. [2025] predict where rank-1 steering works from representation geometry. Our delta is the variable all of these leave free: we measure *width* before intervening, show matching it converts failures to successes across granularities and families, and classify the residue when it does not. Multi-dimensional units are established practice: circular features [Engels et al., 2025], learned alignment subspaces [Geiger et al., 2024, Wu et al., 2023], rank- k edits [Wu et al., 2024], full-rank erasure [Belrose et al., 2023], multi-direction steering [Oozer et al., 2025]. Those methods choose or learn a width; we measure the width the mechanism demands and test whether it was the binding constraint (Sharma et al. [2026] uses rank to predict side effects, not required width). On circuits, self-repair confounds single-head nulls [McGrath et al., 2023, Rushing and Nanda, 2024], coalition level tests of IOI exist [Shi et al., 2024], and concurrent work recovers backup structure by conditional co-ablation [Gong et al., 2026]; we show the regularity underneath, at every granularity. Layer pruning showed single middle layers removable [Gromov et al., 2024, Men et al., 2024]; we measure the collective indispensability this conceals. The mirror dissociation, steerable but not decodable, also occurs

[Nadaf, 2026]. Calls for identification rigor [Lin and Liu, 2026, Asiaee, 2026] get a runnable instance: the width sweep is the missing power analysis.

Limitations. One task per testbed, each with its own intervention and metric, so the three curves agree in shape, not in units; cross-task replication is the main open item. k^* is basis-relative (correlation-PC k^* is an upper bound) and orientation’s two-target relevance inflates its k^* ; the layer-side probe-transfer width is coarser than its feature-space counterpart and carries its own threshold (0.9 of self-accuracy); and teacher-forced divergence overstates harm to output quality: on Llama-3.1-405B, cutting 1–3 of 126 layers makes 94–98% of 1024-token greedy roll-outs diverge (median onset token 101) and never reconverge, yet no quality metric moves (distinct- n , repetition, type–token ratio, code parse rate; paired permutation $p > 0.05$, $n=100$). Distribution shift is not incoherence, so our width effects are statements about the computation, not about fluency. Code, artifacts, and per-number provenance are available at <https://github.com/rockerritesh/width-artifacts>.

Responsible use. Calibrated evidence reduces false discovery claims; width-matched steering is also a stronger manipulation primitive than rank-1 (we study only innocuous attributes). The mitigation is symmetric: audit robustness at matched width, not width 1.

References

- Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR Workshop*, 2017. arXiv:1610.01644.
- Amir Asiaee. Certified interventional fidelity: Anytime-valid, adaptive evaluation of causal claims in mechanistic interpretability. In *Uncertainty in Artificial Intelligence*, 2026.
- Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.
- Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, 2023.
- Jayadev Billa. Predicting where steering vectors succeed. *arXiv preprint arXiv:2604.15557*, 2026.
- Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un)reliability of steering vectors in language models. *arXiv preprint arXiv:2505.22637*, 2025.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. arXiv:2209.10652.
- Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, 2025.
- Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, 2024.
- Zhiren Gong, Zihao Zeng, Chau Yuen, and Wei Yang Bryan Lim. Conditional co-ablation: Recovering self-repair backups in transformer circuits. *arXiv preprint arXiv:2607.01940*, 2026.
- Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A. Roberts. The unreasonable ineffectiveness of the deeper layers. *arXiv preprint arXiv:2403.17887*, 2024.
- Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, 2023.
- Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, 2018.

- Ze Zheng Lin and Fengming Liu. Position: Mechanistic interpretability must disclose identification assumptions for causal claims. *arXiv preprint arXiv:2605.08012*, 2026.
- Ming Liu. Decodable but not corrected by fixed residual-stream linear steering: Evidence from medical LLM failure regimes. *arXiv preprint arXiv:2605.05715*, 2026.
- Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. *arXiv preprint arXiv:2311.17030*, 2023.
- Utsav Maskey, Sumit Yadav, Mark Dras, and Usman Naseem. SafeConstellations: Mitigating over-refusals in LLMs through task-aware representation steering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.
- Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.
- Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. ShortGPT: Layers in large language models are more redundant than you expect. *arXiv preprint arXiv:2403.03853*, 2024.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.
- Mohammed Suhail B Nadaf. Steerable but not decodable: Function vectors operate beyond the logit lens. *arXiv preprint arXiv:2604.02608*, 2026.
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001.
- Narmeen Oozeer, Luke Marks, Shreyans Jain, Fazl Barez, and Amirali Abdullah. Beyond linear steering: Unified multi-attribute control for language models. *arXiv preprint arXiv:2505.24535*, 2025.
- Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In *International Conference on Machine Learning*, 2024.
- Angira Sharma, Christian Schroeder de Witt, Philip Torr, Anisoara Calinescu, and Jialin Yu. A low-rank subspace analysis of LLM interventions. *arXiv preprint arXiv:2606.14388*, 2026.
- Claudia Shi, Nicolas Beltran-Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David Blei. Hypothesis testing the circuit hypothesis in LLMs. In *Advances in Neural Information Processing Systems*, 2024.
- Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, 2024.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.
- Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.
- Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in Neural Information Processing Systems*, 2023.
- Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems*, 2024.
- Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *International Conference on Machine Learning*, 2025.

A The diagnostic matrix and supplementary evidence

Table 1 classifies every mechanism–system pair we measured. Figure 4 gives the per-rank readout-boundary detail summarized in Fig. 1b.

Table 1: **The diagnostic matrix.** k_{repr}^* : from the representation alone. k_{eff}^* : width reaching 95% of max effect, where the sweep saturates (≥ 0.9 ; thresholds in §2). $\dagger$: saturates only far past k_{repr}^* , destroying identity (Fig. 3c). PaliGemma cells are caption-probability movement as a fraction of the prompted gap ($\ddagger$: absolute, since orientation’s prompted gap is ≈ 0 ; the LM cannot report orientation even when asked). $\S$: role-criterion head count, not a reliable width; the k_{eff}^* and width-1 entries for this row come from criterion-free joint greedy selection (Table 3).

granularity / system	attribute	k_{repr}^*	eff. @ width 1	k_{eff}^*	max eff.	verdict
dinov2 L11	size	16	0.23	—	0.40	basis-limited
dinov2 L11	orientation	49	0.11	53	0.96	width-limited
clip L11	size	15	0.33	—	0.64	basis-limited
clip L11	orientation	65	-0.03	86	1.72	width-limited
siglip L11	size	12	0.17	76	1.07	width-limited [†]
siglip L11	orientation	50	0.01	53	1.30	width-limited
7 LMs (Fig. 2)	induction writers	2–9 [§]	0.24 (median)	2–22	1.00	width-limited
GPT-2 IOI	head coalition	9	0.09	9	1.00	width-limited
Llama-3.1-8B	layer band	15 (transfer)	0.10	13	0.95	width-limited
PaliGemma readout	size	13	0.02	none	0.02	readout-limited
PaliGemma readout	orientation	54	0.00	none	0.01 [‡]	readout-limited

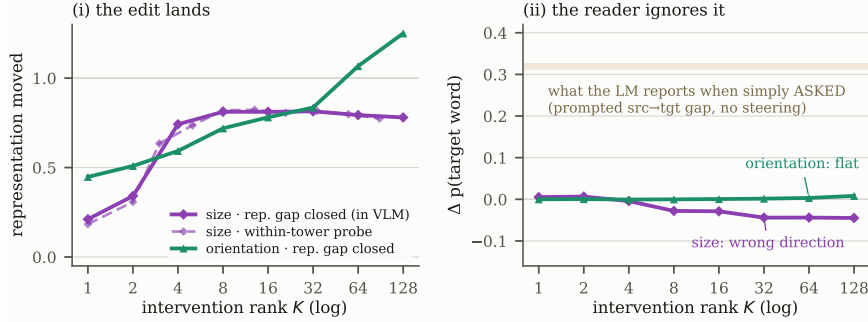


Figure 4: **The readout boundary** (PaliGemma, tower L24). (i) rank- K steering closes up to 81% of the representation gap, confirmed by a within-tower probe. (ii) the frozen LM’s answer is flat at every rank (size: *wrong direction*, -0.045), at three injection sites, prompted or not. Prompting alone shows a +0.32 gap. The signal moved; this handle’s content is not what the consumer reads (§5).

What width sweeps cannot catch. Matching width does not certify a circuit *complete*. Our GPT-2 analysis initially missed three heads, two of which no ablation threshold could reach: the analysis schema had three role fields, and the missing heads performed roles outside the taxonomy. A sweep explores the space the analyst wrote down; errors encoded in structure do not look like decisions when the code is read back.

B Threshold sensitivity and criterion-free head widths

Thresholds. Four constants enter the diagnostic: τ_{repr} (relevance fraction defining k_{repr}^*), τ_{eff} (effect fraction defining k_{eff}^*), the saturation cutoff below which no k_{eff}^* exists, and the ratio r for $k_{\text{eff}}^* \leq r k_{\text{repr}}^*$. We evaluated every Table 1 row on all $3^4 = 81$ combinations (Table 2). The class (width-limited vs. basis- or readout-limited) never changes: size rows in DINOv2 and CLIP peak at 0.40 and 0.64, below every saturation cutoff, and PaliGemma’s readout moves 0.02. What does move is the finer claim that k_{eff}^* lands within $r k_{\text{repr}}^*$: it holds for orientation in 54–66 of 81 settings, and every failure has $\tau_{\text{repr}} = 0.90$ with $r \leq 1.4$ (plus CLIP at $r = 1.2$): a looser relevance cutoff shrinks k_{repr}^* faster than the effect cutoff shrinks k_{eff}^* . SigLIP size saturates only far past k_{repr}^* in 78

of 81 settings, as its $\dagger$ already marks. k_{eff}^* in the three Fig. 1a testbeds moves by at most 1.5 units over $\tau_{\text{eff}} \in [0.90, 0.99]$. The layer-side k_{repr}^* (probe-transfer band, 0.9 of self-accuracy) was not varied.

Table 2: **Threshold sensitivity.** k_{eff}^* at saturation 0.90 (— = never saturates). *Class stable*: settings (of 81) giving the canonical class. Last column: settings where $k_{\text{eff}}^* \leq r k_{\text{repr}}^*$ (n/a for basis-limited rows). Generated by code/e11_threshold_sensitivity.py.

row	k_{repr}^* at τ			k_{eff}^* at τ			class stable (of 81)	$k_{\text{eff}}^* \leq r k_{\text{repr}}^*$ (of 81)
	.90	.95	.99	.90	.95	.99		
dinov2 size	7	16	42	—	—	—	81	n/a
dinov2 orientation	36	49	75	51	53	55	81	63
clip size	8	15	42	—	—	—	81	n/a
clip orientation	47	65	91	82	86	88	81	54
siglip size	7	12	36	63	76	86	81	3
siglip orientation	38	50	75	52	53	55	81	66
GPT-2 IOI heads	—			8.2	8.6	8.9	81	—
Llama layers	—			12.3	13.1	13.8	81	—
DINOv2 orient. (Fig. 1)	—			51.4	53.2	54.6	81	—
PaliGemma size (readout)	max 0.02 < all cutoffs						81	—

Criterion-free head widths. The Fig. 2 sweep ablates only heads passing the previous-token criterion, so it cannot test whether the criterion counts the mechanism correctly. For the seven models with full ladders we reran selection with no role score: at each of 30 steps, add the upstream head (all heads in layers before the reader) whose ablation *jointly* with those already chosen most reduces the reader’s induction score; k_{eff}^* is the smallest set reaching 95% of the maximum over the sweep. Every curve is flat by step 30 (last five steps ≥ 0.98 of the maximum). Same prompts, seed, reader and mean references as Fig. 2; candidates are batched and the forward pass stops after the reader’s layer, which we verified reproduces the unbatched search exactly on GPT-2 small and medium. Ranking heads by their *individual* ablation effect instead is itself a width-1 measurement and misses jointly necessary heads (GPT-2 medium: the six criterion heads together outdo that ranking’s best 30); we report it only for comparison (Table 3, “single-rank”).

Table 3: **Criterion-free head widths.** n_w : heads passing the role criterion. k_{eff}^* joint: joint greedy selection with no role score. Overlap: criterion heads among the first n_w greedy picks. Width 1: first greedy head’s effect over the maximum. Criterion set: joint effect of the n_w criterion heads over the greedy maximum. Generated by code/e12_head_width_criterion_free.py.

model	n_w (criterion)	k_{eff}^* single-rank	k_{eff}^* joint	overlap	width 1	criterion set
GPT-2 small	2	11	10	2/2	0.34	0.51
GPT-2 medium	6	22	21	6/6	0.03	0.43
GPT-2 large	7	25	14	5/7	0.03	0.32
GPT-2 XL	9	23	22	0/9	0.01	0.20
Pythia-410m	5	3	2	4/5	0.80	0.98
Pythia-1.4b	4	5	8	4/4	0.24	0.85
Llama-3.2-1B	7	23	8	3/7	0.33	0.70

C Reproducibility notes

Normalization anchors and the affine conversion between coalition-fraction conventions; band definitions (11-centered k -curves vs. 16-centered scale comparisons); the canonical k^* definition (centered SVD, $n_{pc} = 120$, 95% cumulative corr^2 relevance) and its audit against the four variant definitions used across source scripts; per-model writer ladders; NDIF remote-ablation mechanics (arbitrary-subset rebinding); the threshold grid (e11) and criterion-free head selection (e12, with the joint-greedy search in `m5y_joint_greedy_batched.py`). Every number in the text is generated from committed artifacts by `make_stats.py`.