EVERY ABLATION IS A DOSE: COUNTERWEIGHTS AND THE SEMBLANCE OF SELF-REPAIR

Anonymous authors
Paper under double-blind review

ABSTRACT

Ablate a component of the language model, and it's often observed that other components adjust and appear to compensate. This phenomenon, termed self-repair, has been observed repeatedly, but its mechanism remains unclear. The most systematic study to date concluded that self-repair is noisy and unlikely to have a single explanation. We argue that it has one, a gain already present before any ablation. Any intervention on a causally important component can be viewed as a point on a coordinate axis λ, the signed strength of a counterfactual contrast. Hence, any conventional ablation methods are uncalibrated points on this axis. We show that the causal repair response for a fine-grained unit r is governed by an affine law $E_r(\lambda) = own_r + \gamma_r\lambda$. The slope γ_r is a fixed coefficient that consistently influences the model, with or without ablation, and its sign determines whether the unit counteracts or reinforces the removed signal. On a factual-verdict task across four models from distinct families (Gemma, Qwen, LLaMA, and Mistral), we identify components (MLP neurons, OV neurons, and singular directions) that follow this affine law, 68 of 81 downstream directions in all. Moreover, we can anticipate the magnitude of γ_r from the fixed weights. On the IOI circuit of GPT-2 Small, seven of the ten heads the intervention can reach follow the law, and all seven are counterweights. From this perspective, what may appear as self-repair is merely a counterweight doing its usual operation when the contrastive signal emerges at the core.

1 INTRODUCTION

Ablation, deleting a component and observing its effect, is a standard tool for understanding model internals. Circuit discovery, attribution, and unlearning evaluation all infer a component's contribution from ablations (Wang et al., 2023; Conmy et al., 2023; Dai et al., 2022). However, self-repair undermines this inference. Ablating a component causes downstream components to adjust, so the behavior partly survives, and ablation can underestimate the component's importance while also misattributing what it did.

Backup name movers (Wang et al., 2023), the Hydra effect (McGrath et al., 2023), and anti-erasure neurons (Rushing & Nanda, 2024) were each found the same way, by ablating first and sorting components by what moved, and the most systematic study concluded that the process is noisy and unlikely to have a single mechanism (Rushing & Nanda, 2024). It remains open why a model possesses this capacity and which components are predisposed to compensate.

Prior work has documented instances without identifying the rule. McDougall et al. (2024) showed that one head in GPT-2 Small suppresses whatever token the model is currently predicting, so ablating it lifts the suppression, and Rushing & Nanda (2024) found late-layer MLP neurons that oppose an upstream head and relax when it is ablated. Neither could tell the rule from the instance, because a deletion is a single point and cannot distinguish a component that switches on from one that responds in proportion to the displacement of its input. Direct logit attribution, the readout of every prior self-repair study, compounds this, since it degrades with distance from the unembedding (nostalgebraist, 2020; Belrose et al., 2023), which is why every compensator found so far is a late-layer component.

We claim that *compensation is a standing coefficient*. A downstream direction r (a computational unit in weights, e.g., a neuron) exerts a fixed opposition in the clean model before the upstream

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

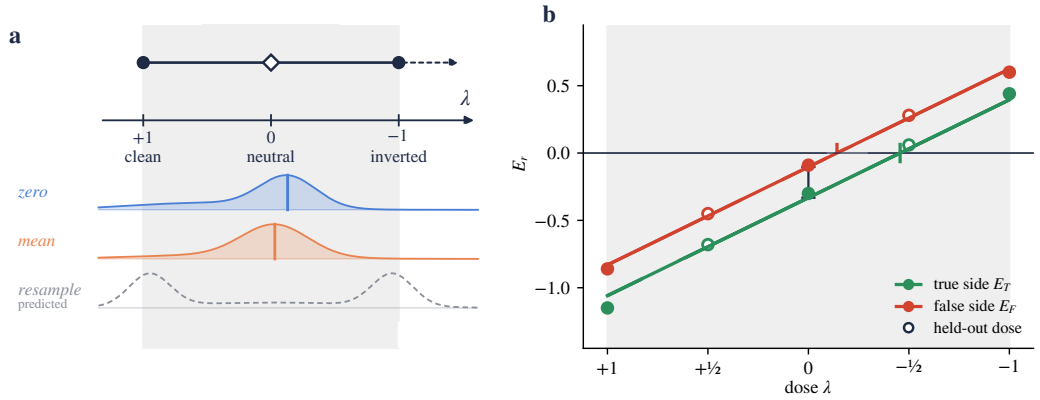

Figure 1: **(a)** Any intervention on the core `L28.n2046` (Gemma-2-9B-it) is a dose λ in units of the pair's class contrast, with $+1$ clean, 0 class signal removed, and -1 inverted. Zero and dataset-mean ablation land at per-prompt doses set by the data (true prompts shown), while resample is predicted to land at the donor's dose. **(b)** Its counterweight `L38.n7088` ($\gamma_r < 0$) has a contribution affine in λ, fitted on three doses (filled), predicting held-out doses (hollow).

direction c is deleted, and deletion records the release of that opposition. To see this, the intervention must be a continuous dose. We study a True/False verdict task in which every prompt has a counterfactual partner of the opposite class, and we displace the signal in c by a signed dose λ in units of the pair's own class contrast. On this axis, every zero, mean, or resample ablation (Miller et al., 2024; Zhang & Nanda, 2024; Heimersheim & Nanda, 2024) is a point at a dose set by the data (Figure 1a). Within the window of naturally occurring activations $|\lambda| \leq 1$, r's contribution to the verdict is affine in the dose,

$$E_r(\lambda) = \text{own}_r + \gamma_r \lambda, \tag{1}$$

with a fixed gain γ_r, which makes r a *counterweight* if $\gamma_r < 0$ and a *relay* if $\gamma_r > 0$ (Figure 1b). The capacity is therefore standing, and compensation is its release as seen through a coarse intervention. The predisposed directions are the ones the weights already couple to the core, and the magnitude of γ_r is anticipated by the alignment of the core's write vector with r's read weights in the static checkpoint (§6).

We test this with a nested counterfactual intervention (§3) in four instruction-tuned models of different families and on the IOI circuit of GPT-2 Small. Scope conditions are stated in §9.

Contributions.

1. **Compensation is standing.** The directions that self-repair already oppose the core in the clean model, so deletion releases a gain rather than recruiting one. This holds across four instruction-tuned models and for the seven IOI heads that follow the law (§4, §7).
2. **A dose axis and a nested intervention.** Every conventional ablation is an uncontrolled point on a single axis, and a nested counterfactual intervention, read by total effect inside the dosed model, makes it observable (§3, §7).
3. **The coupling has an affine form.** Within the on-manifold window $|\lambda| \leq 1$, a downstream direction's contribution to the verdict varies nearly linearly with the dose applied to the core, Eq. 1. Lines fitted on three doses predict held-out intermediate doses with out-of-sample R^2 of 0.65 to 0.91 by family (§5).
4. **A population of named counterweights and relays.** Directions named from the weights alone are legible at the unembedding yet inert or opposing in the clean model, which is why deletion never surfaces them. They are certified by intervention rather than assumed from their labels, and their coupling magnitude is anticipated from the static checkpoint (§4, §6).

2 SETUP

Weight-space bases. Decoder-only transformers add writes $o_c(x) \in \mathbb{R}^d$ from each attention head and MLP into a residual stream (Elhage et al., 2021). We decompose each write along two fixed sets of weight-space directions. (I) **Neuron directions** give $o_c = \sum_j z_{cj}\, w_{cj}$, where w_{cj} is a column of the component's write matrix (W_c^d for MLPs, W_O for heads) and z_{cj} is its scalar activation. (II) **SVD basis** gives $o_c = \sum_k \alpha_{ck}\, v_{ck}$, where $\{v_{ck}\}$ are the left-singular vectors of W_c^d for MLPs and of the OV circuit $W_O W_V$ for heads, and α_{ck} is an input dependent scalar (Ahmad et al., 2025; Xue & Andrzejak, 2026).

A single interface. We unify these as a single *direction* r, parameterized by a fixed write vector $\omega_r \in \mathbb{R}^d$ and an activation coefficient $a_r(x) \in \mathbb{R}$. This factorization isolates single-direction interventions, where $\mathrm{do}(a_r := v)$ overwrites the coefficient of a specific direction ω_r. A direction's pole $\omega_r^\top W_U$ at the unembedding W_U sets its *direct effect*. The write moves the logit of token t by $a_r(x)\langle W_U[:,t], \omega_r\rangle$ before the final norm. We read these poles from the input-independent weights (nostalgebraist, 2020; Geva et al., 2021; Dar et al., 2023; Millidge & Black, 2022; Ahmad et al., 2025; Hong et al., 2025; Xue & Andrzejak, 2026).

Interchange interventions. We treat the forward pass as a deterministic structural causal model (Pearl, 2009) in which the prompt is the exogenous variable and the basis coefficients $a_r(x)$ are the endogenous variables. On this model, activation patching is the operation $\mathrm{do}(a_r := a_r(x'))$, which pins a coefficient to the value it takes on a source input x' (Vig et al., 2020; Chan et al., 2022; Geiger et al., 2020; 2021). Every intervention in this paper, including the dose ladder of §3, is built from this primitive.

Verdict margin. We evaluate the model's factual verdict with the *verdict margin*, the gap in mean logits between faithful (X) and contrastive (Y) answer tokens,

$$D_{XY}(x) = \frac{1}{|X|}\sum_{t\in X} \ell_t(x) - \frac{1}{|Y|}\sum_{t\in Y} \ell_t(x). \tag{2}$$

Evaluation is conditioned on clean-correct pairs, so $D_{XY} > 0$ on every clean run, and any intervention that damages the verdict lowers D_{XY}.

Models, task, and data. We study four instruction-tuned models (Llama-3-8B-Instruct, Gemma-2-9B-it, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct-v0.3). The task is a zero-shot factual verdict (true/false) on declarative statements (App. C.1). Contrastive pairs are synthesized from the Known-Facts corpus (Meng et al., 2022) so that the true and false variants of a pair differ only in the asserted attribute, leaving all other phrasing identical (App. C.2).

3 THE DOSE AXIS

Every experiment in this paper measures what a downstream direction does when the core's class signal is disturbed. This requires two interventions composed in sequence, a *primary* that disturbs the core and a *secondary* that measures a direction in the disturbed model. This section defines both, puts every disturbance on a single axis, and states the estimand.

A single coordinate for the core direction. Let (x, x') denote a contrastive prompt pair (e.g., True vs. False), and let $a(x)$ and $a(x')$ be the corresponding activations along a core direction c. Decomposing these activations yields

$$m \;=\; \frac{a(x)+a(x')}{2}, \qquad \Delta \;=\; a(x') - a(x),$$

the *mid-point* (m), which isolates the shared context across the prompt pair, and the *contrast* (Δ), which isolates the class-differential signal. Any clamp value v applied to the core lies on the line connecting $a(x)$ and $a(x')$, parameterized by a single scalar,

$$a(\lambda) \;=\; m - \frac{\lambda}{2}\,\Delta, \qquad \lambda(v) \;=\; \frac{2\,(m-v)}{\Delta}. \tag{3}$$

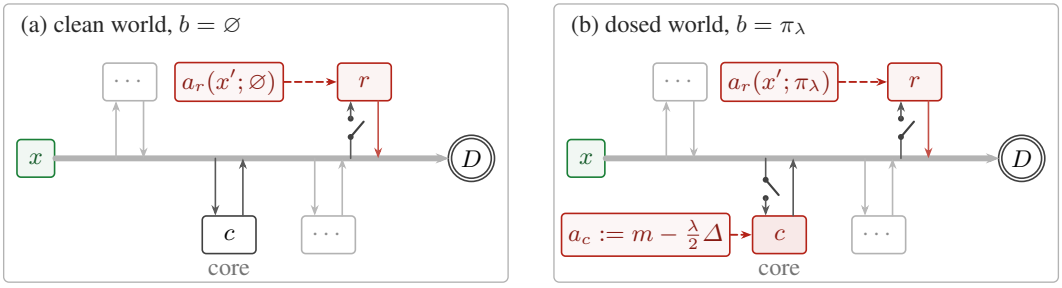

Figure 2: **The nested counterfactual intervention**. Open switches cut a read from the residual stream, and dashed arrows inject the replacement value. *(a) Clean world,* $b = \varnothing$. The downstream direction r is patched with $a_r(x'; \varnothing)$, its value on the partner prompt x' in the unmodified model. *(b) Dosed world,* $b = \pi_\lambda$. The core c is first clamped to dose λ (Eq. 4), and r is then patched with $a_r(x'; \pi_\lambda)$, computed on x' through this *same* mirror-dosed graph.

The continuous axis has three reference points that are intuitively conventional. At $\lambda = +1$ the core holds its clean activation for the given class. At $\lambda = -1$, it assumes the opposite-class activation, that is, the standard activation swap (Vig et al., 2020; Geiger et al., 2020; 2021). At $\lambda = 0$, the core rests at the class-neutral midpoint m, the unique point with no class differential. Doses $\lambda < -1$ overshoot the counterfactual, injecting more than one unit of the pair's own contrast, as in graded steering with a multiplier above one (Zou et al., 2023; Rimsky et al., 2024; Turner et al., 2024). We refer to $k = (1 - \lambda)/2$ as the corruption fraction, where it aids reading, and k runs $0 \to 1$ as λ runs $+1 \to -1$.

The primary intervention (clean world). The primary is a dosed swap of the core toward its partner value,

$$\pi_\lambda \;:=\; \mathrm{do}\Big(a_c := m - \tfrac{\lambda}{2}\Delta\Big) \;=\; \mathrm{do}\Big(a_c := a_c(x) + k\,[a_c(x') - a_c(x)]\Big), \tag{4}$$

applied to the core direction(s) c certified in §4, and to all of them jointly, where a family has more than one. The clean model is the point $\lambda = +1$. We write $b = \varnothing$ for it and $b = \pi_\lambda$ for the dosed world. The ladder is $\lambda \in \{+1, +\frac{1}{2}, 0, -\frac{1}{2}, -1, -3, -5, -7\}$.

The secondary intervention (dosed world). Within a world $b \in \{\varnothing, \pi_\lambda\}$, we read a downstream direction r by patching its coefficient with the value it takes on the partner prompt in that same world ,

$$\mathrm{sec}_r^b \;:=\; \mathrm{do}\big(a_r := a_r(x';\, b)\big). \tag{5}$$

Here π_λ is a rule applied per prompt, so on x' it clamps the core to $m + \frac{\lambda}{2}\Delta$, the mirror image of Eq. 4 about m. At every dose the two prompts' cores sit equidistant from the midpoint on opposite sides, coinciding only at $\lambda = 0$. The coefficient $a_r(x'; \pi_\lambda)$ is then r's activation on the partner prompt inside this mirror-dosed graph (Figure 2).

The estimand. Let $D(x; b)$ be the verdict margin on x in world b. The contribution of r in world b is the margin erased when its natural reading is intervened on,

$$E_r^b(x) \;=\; D(x;\, b) \;-\; D(x;\, b,\, \mathrm{sec}_r^b). \tag{6}$$

$E > 0$ means r supports the verdict in that world, $E < 0$ that it opposes. It is a total effect. Both passes run in world b, so a shift of the primary adds to both margins, which cancels in E. In the clean world (with $b = \varnothing$) we write E_r^0.The point contrast

$$R_r(\lambda) \;=\; E_r^b - E_r^0 \tag{7}$$

measures how much r's contribution changes between the clean and dosed worlds at a single dose.

Table 1: **The core carries roughly half the total effect.** Total effect $E = E_T + E_F$ of the whole model on the verdict, where E_T is the True−False logit margin on the true statement and E_F the False−True margin on the false statement, at the answer token. E^0 is the clean model, and E^π is the same with the core swapped to its full counterfactual ($k = 1$, $\lambda = -1$). The last column is the share of the clean total effect the core is responsible for, $(E^0 - E^\pi)/E^0$. Standard errors in parentheses.

Model	Core(s)	E^0	E^π	core share
Llama-3-8B-Instruct	`L18.n11065`	7.78 (0.11)	3.67 (0.05)	52.7%
Gemma-2-9B-it	`L28.n2046` + `L33.n4294`	9.50 (0.11)	4.38 (0.05)	53.9%
Qwen2.5-7B-Instruct	`L22.n13149` + `L24.n14758`	18.76 (0.17)	8.46 (0.07)	54.9%
Mistral-7B-Instruct-v0.3	`L20.n14286` + `L19.n8228`	11.59 (0.15)	5.03 (0.06)	56.6%

4 Census of the Core and Its Counterweights

A direction's label in the weights says what it writes, the tokens its output promotes or suppresses at the unembedding. It says nothing about the direction's causal role (Bolukbasi et al., 2021; Huang et al., 2023). This section separates the two. It certifies the core, the one or two directions per family that carry the verdict, and identifies the legible directions that do not carry it at rest yet respond once the core is intervened on. Both come from one bench, named from the weights alone.

A bench named from the weights. We scan the four families (§2) and label each direction r by the tokens it promotes most and suppresses most on the unembedding. A direction enters the bench if a truth or falsity token, in English or another language, appears at either pole, and the suppressing pole counts as much as the promoting one (App. D.2).[1] Selection uses only the checkpoint. The bench was fixed before any dose was run and not revised after. Whether a direction affects the verdict is the causal question the census answers.

One or two directions carry the verdict. In the clean world ($\lambda = +1$,§3), we measure each direction's contribution E_r^0 (Eq. 6) and certify it as a core only if $E_r^0 > 0$, reliably, on both classes. The core is small in every family, a single neuron in Llama (`L18.n11065`) and a pair in each of Gemma(`L28.n2046`, `L33.n4294`), Qwen (`L22.n13149`, `L24.n14758`), and Mistral (`L20.n14286`, `L19.n8228`), and it carries more than half of the clean total effect, 52.7 to 56.6% (Table 1). Of the other directions, some are inert, others oppose the verdict on both classes despite legible truth or falsity poles.

The dismissed directions move against the core. Now intervene on the core (Eq. 4) and re-measure Eq. 6 inside the dosed world. Most of these directions move against it. As the core's class signal is weakened, their opposition to the verdict shrinks. At some dose it reaches zero, and past that dose they support the verdict instead. We call such a direction a *counterweight* (formally, $\gamma_r < 0$ in Eq. 1). Figure 3 follows one *counterweight* per family. Llama's `L23.n8972` goes from -0.10 to $+0.25$ nats, Gemma's `L38.n7088` from -2.1 to $+3.9$, Qwen's `L27.n13033` from -5.7 to $+8.7$, and Mistral's `L31.n8773` from -1.5 to $+3.2$, each crossing zero on both classes. Together with the census, this shows that these directions write the truth vocabulary and pull against the core. A smaller set moves with the core, and §5 treats these *relays* on the same axis. The rest of the bench is tabulated in App. F.1. The base checkpoints show the same response (App. H). Renormalization, which Rushing & Nanda (2024) credit with part of self-repair, explains at most 4.5% of the audited recruitment (App. I).

The response tracks the amount of signal removed. In families with two cores, dosing either core alone moves part of the bench, and dosing both moves more of it than either alone, and at low dose by more than the sum of the two (App. F.3). Two independent couplings would add exactly. The excess is the share of the signal that the free core carries when only one is clamped.

[1] The core neurons were first noticed as the sites at which the two statement classes separate most sharply in the clean forward pass, before the bench was constructed. The scan uses no activations and recovers them.

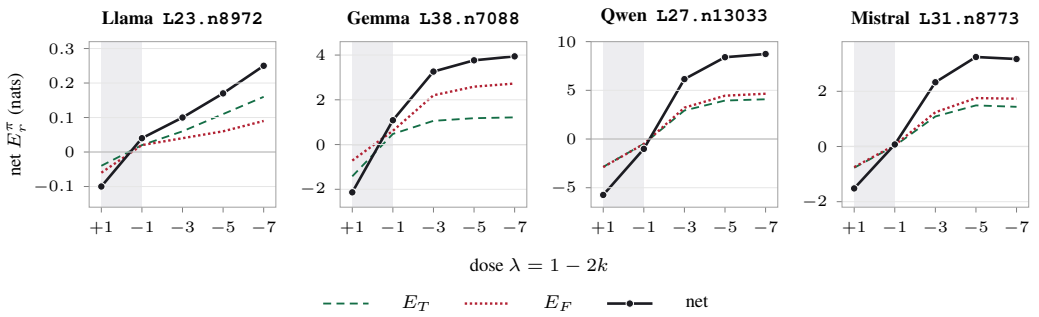

Figure 3: **The featured counterweight in each family moves against the core.** Net contribution $E_r^\pi = E_T + E_F$ of one direction per family against the dose λ on that family's certified core(s). The shaded band is the on-manifold window $|\lambda| \leq 1$, where $+1$ is the clean model and -1 the counterfactual. Each opposes the verdict at rest and crosses zero on both classes at a sufficient dose. Points are per-dose means, and standard errors are smaller than the markers (paired t, App. F.1)

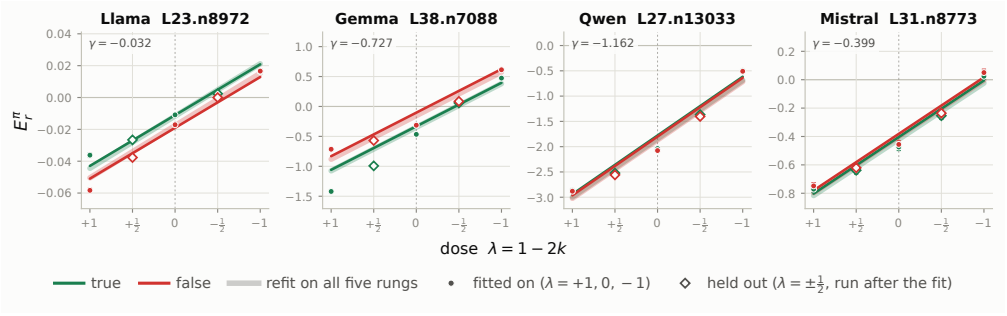

Figure 4: **The Affine Law**, four model families. For one exemplar direction per model, the erased margin E_r^π against the dose λ. Lines are fit on the three filled rungs, with the hollow rungs run afterward and held out. The band is the refit on all five.

The response follows the signal across interventions. The same directions respond to a procedurally unrelated primary intervention, a learned label-free perturbation of the singular values, and to its conjunction with the swap (App. G). We can see that recruitment/compensation is robustly present in the model, irrespective of the disruption method.

5 The Coupling Law

§4 fixed the sign of the response. As the core's signal is withdrawn, most of the bench moves toward supporting the verdict. It could not fix the shape, because only two of its doses, the clean model and full inversion, lie inside the on-manifold window ($|\lambda| \leq 1$, where the natural activation values lie). Shape is what separates the two readings. A component that switches on when the core's signal is lost, sits at its resting value until some level, and rises after it, one that responds in proportion to the displacement gives a line.Across the dose ladder, the response is a line. For each bench direction r and class $c \in \{True, False\}$, on $|\lambda| \leq 1$,

$$E_r^c(\lambda) \;=\; \text{own}_r^c \;+\; \gamma_r\,\lambda, \tag{8}$$

where $E_r^c(\lambda)$ is the contribution of Eq. 6 on class c in the world π_λ, in which the core is held at dose λ in both passes the contrast compares, the measured pass on x and the partner pass on x' that supplies r's counterfactual value (Figure 2). We fit Eq. 8 by weighted least squares on both classes at once, with weights from the per-rung standard errors, on the three rungs $\lambda \in \{+1, 0, -1\}$, and we hold out the rungs at $\pm\frac{1}{2}$ and beyond -1 to test the line.

What the two parameters are. γ_r is the coupling, how much of the core's signal reaches the verdict through r and with what sign. A negative slope marks a *counterweight*, a positive one a

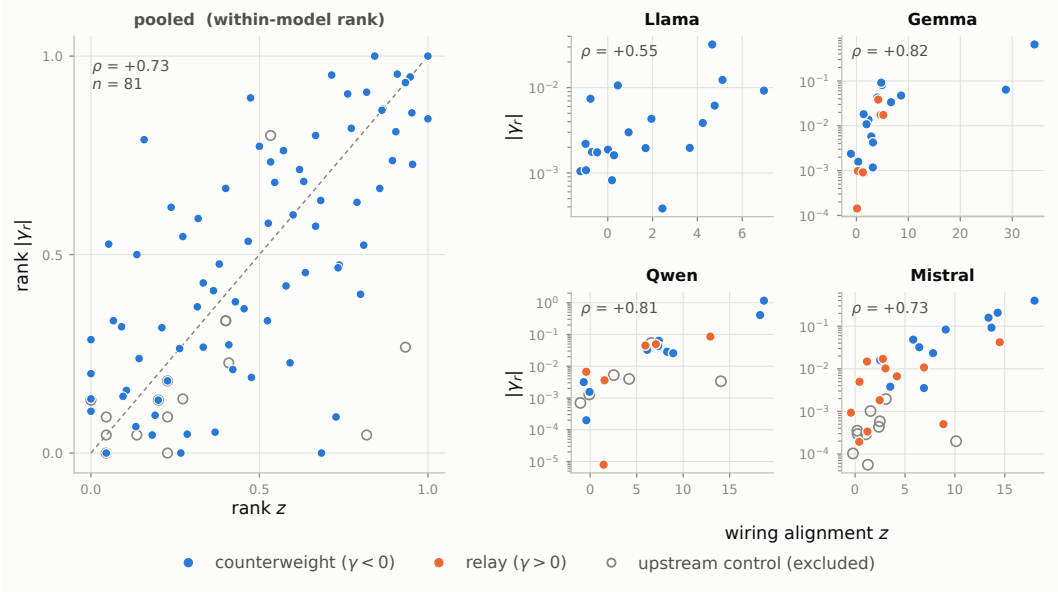

Figure 5: **Coupling strength is legible in the weights**. Weight-space alignment z between the core's write and each direction's read, computed from the checkpoint alone, against the intervention-measured coupling $|\gamma_r|$, pooled on a within-model rank scale (left, $\rho = +0.73$, $n = 81$) and per model (right). Hollow markers are upstream components with no causal path to the core.

relay. One slope serves both classes (fitted separately, they agree at $\rho = 0.83$) and one intercept serves each. Their mean is r's baseline, and their difference Δown_r is the class information r gathers through routes other than the core, constant in dose, so the two class lines run parallel. The split raises the median in-sample R^2 in every family, from $[0.50, 0.76]$ to $[0.82, 0.91]$, and without it the law fails Mistral out of sample ($R^2 = -0.41$, App. K).

Most of the bench couples to the core. Of the 81 bench directions the core can reach, 68 have a slope reliably different from zero under both fits, with the false discovery rate controlled at $q = 0.05$ across each family's full bench, including upstream controls (App. K). Of these, 52 are counterweights, and 16 are relays (Table 33), a ratio specific to a bench selected for legible poles. The remaining 13 fail the significance gate. Seven have near-zero slopes, and six sit narrowly outside the correction. Figure 4 shows one direction per family with its fitted lines.

The lines hold on doses they never saw. After freezing the fits we ran the half rungs at $\lambda = \pm\frac{1}{2}$. The line predicts them to within a median 9% of each direction's response range, with out-of-sample R^2 between 0.65 and 0.91 by family. These doses lie inside the window, so the test shows that the line tracks a smooth response, not that no other smooth curve would. A quadratic term changes γ_r by less than its standard error (App. K). Past full inversion the response saturates, so we certify Eq. 8 on $|\lambda| \leq 1$ only. The same law resolves on the IOI circuit of GPT-2 Small, where every certified head is a counterweight (App. L).

6 The Coupling Magnitude Is Grounded in the Weights

γ_r could be viewed as an artifact of the experimental design rather than an intrinsic network property. To address this, we define a static measure computed from the model weights. For downstream direction r, we compute the wiring alignment z_r, the cosine similarity between the core's write vector ω_c and the rows where r reads the residual stream, z-scored against 2,000 random readers in the same layer (App. M). Figure 5 shows that $|\gamma_r|$ correlates with z_r across all model families (Spearman $\rho \in [0.55, 0.82]$, one-sided permutation $p \leq 0.01$ in every model, Table 32). Given that four models were trained independently, Fisher's method combines them to $p \approx 10^{-10}$, an upper bound because Gemma's p sits at the permutation floor. This relationship remains robust when

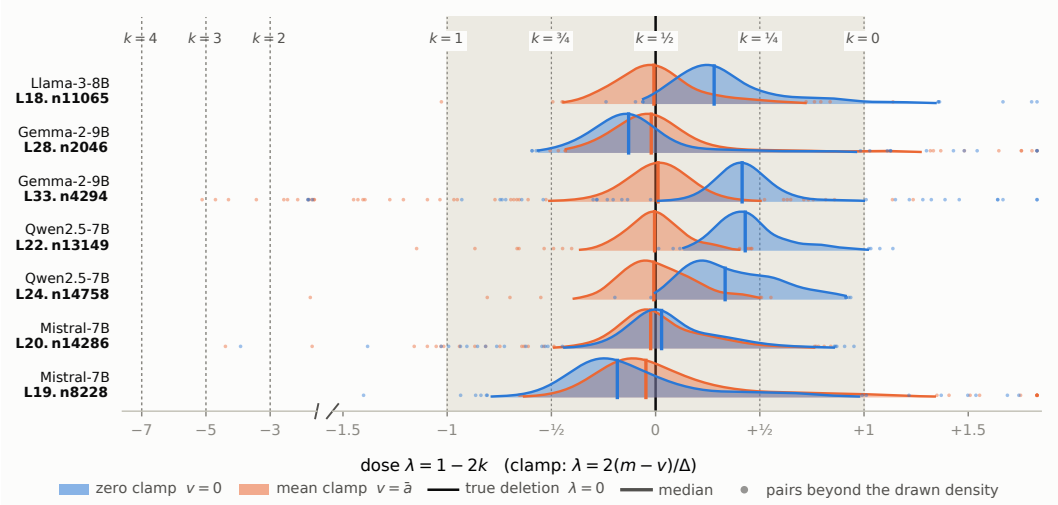

Figure 6: **Deletion on the dose axis.** For each certified core, the distribution over pairs of the dose $\lambda(v) = 2(m - v)/\Delta$ administered by a zero clamp ($v = 0$, blue) and a dataset-mean clamp ($v = \bar{a}$, orange), with medians marked. The vertical line is $\lambda = 0$, the shaded band is the on-manifold window, the broken axis is the off-manifold tail.

γ_r is refitted exclusively on the intervened rungs $\lambda \in \{0, -1\}$, excluding clean-run data entirely ($\rho \in [0.43, 0.83]$, combined $p < 10^{-9}$). This statistic is ordinal and unsigned, accounts only for the direct value pathway, and tracks the clean effect $|E_r^0|$ ($\rho \in [0.68, 0.86]$) about as strongly as $|\gamma_r|$, since $E_r^0 = \mathrm{own}_r + \gamma_r$ at $\lambda = +1$ (Eq. 8), a link the refit does not remove. Consequently, we claim only that the checkpoint anticipates the *magnitude* of the coupling, not the exact coefficient (App. M.3).

7 Deletion on the Dose Axis

Self-repair has only ever been measured by deletion, whether by mean ablation (Wang et al., 2023), task-matched resampling (McGrath et al., 2023), or open-distribution resampling (Rushing & Nanda, 2024). All three assume that the clamp strictly removes the component's contribution, interpreting whatever follows as the network's response to the loss. In our perspective, any clamp $a_c := v$ is a point on the axis of §3. It lands at dose $\lambda(v) = 2(m - v)/\Delta$, where Eq. 1 governs the behavior of every certified counterweight and relay. Figure 6 shows how each clamp v distributes on the dose axis. Zero ablation lands at $2m/\Delta$. Mean ablation is neutral on average because a balanced reference set puts its mean $\bar{a}$ near m, but not per pair. Since the dose divides by the pair's contrast, pairs with a small contrast receive a large dose. Resample ablation lands wherever its donor pool places it (App. N and Table 20). In none of these three methods does the experimenter choose the dose. The data chooses it, assigning a different dose to every pair.

What the founding literature's observations saw. A deletion study records the rise of the counterweights it evaluates, scaling with $|\gamma_r|$ and an uncalibrated dose that varies across studies and random draws. For IOI, the heads of Wang et al. (2023) that obey Eq. 8 are all counterweights ($\gamma_r < 0$) (App. L). The reported backup is the release of a coupling already active in the clean model.

8 Related Work

Self-repair. In the IOI circuit, the backup name mover head engages, and the negative name movers' effect weakens when the name movers are ablated (Wang et al., 2023). Layer-level compensation in Chinchilla 7B was referred to as the Hydra effect by McGrath et al. (2023), where late MLPs erase the top token. Rushing & Nanda (2024) attributes the self-repair partly to LayerNorm

and partly to sparse anti-erasure neurons. Patrawala et al. (2025) shows that adjacent layers reduce the effect of predecessors. The OR-gate redundancy hides components from noising-based discovery (Chen et al., 2025). Concurrently, Gong et al. (2026) retrieves backup components by measuring the increase in ablation after the primary set is removed. These works mostly locate compensators after deletion. We show the opposition is a standing coupling in the clean model, and that on IOI the seven reachable heads that obey our dose law are all counterweights.

Suppressive components. Copy suppression (McDougall et al., 2024), suppression neurons (Gurnee et al., 2024) and function-vector cancellers (Wang, 2026) write against the output, and entropy neurons damp confidence (Gurnee et al., 2024; Stolfo et al., 2024). Concurrently, Oskin (2026) reports that across 18 models, the signed logit mass opposing the predicted token is a median of seven times the mass carrying it. These are characterized by their clean-pass sign, often via direct logit attribution. We measure how a direction's *total* effect varies with the dose of a specific core.

Interventions and weights. The choice of ablation changes localization and faithfulness (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024; Miller et al., 2024), motivating optimal ablation (Li & Janson, 2024). Subspace patching may act through dormant parallel pathways (Makelov et al., 2024), a claim contested by (Wu et al., 2024). Concurrently, Vaidyanathan et al. (2026) show that patching effects contain interaction terms that, on IOI, grow linearly with patch distance with role-dependent sign, and γ_r is a related, per-direction coupling along a controlled dose. Composition scores (Elhage et al., 2021), SVD communication channels (Merullo et al., 2024), and a concurrent weights-only map (Oskin, 2026) chart which components *can* interact. We place every clamp of the core, including standard ablations, on one dose axis, and show that a weight-only statistic anticipates the rank of $|\gamma_r|$.

9 Conclusion

Findings. In this work, we investigated the phenomenon of self-repair that confounds the interpretation of ablations. We uncovered weight-space components that are finer-grained than heads or MLPs and are coupled to one another along a class-contrast dose axis, thereby responsible for the compensatory effect. Conventional ablation methods, such as zero, mean, or resampling, are unpicked points on this axis. Therefore, what is regarded as self-repair is the passive release of a pre-existing coupling rather than a dynamic, reactive process. Furthermore, this interaction is governed by an approximate affine law, whose slope can be anticipated from the alignment between downstream read and upstream write vectors. Finally, we showed that this coupling is robust, present in the base counterparts of the instruction models, and active under a procedurally unrelated, learned perturbation.

Limitations. We studied one task at one answer position, a True/False verdict read at a single token. How it may translate to other tasks, and especially generative ones, remains an open question. We found the cores, counterweights and relays among legible monosemantic weight-space directions, although nothing requires them to be legible. Our experiments also suggest that a counterweight may respond to coupling from distributed components. In models with two cores, the downstream counterweight responded more strongly to joint intervention than either core (App. F.3). Mapping these interactions exhaustively is left for future work.

Reflection The methodology underlying the critique our findings motivate generalizes further than the findings themselves. The meta-conclusion of this work focuses on our instruments, which can become our own Eddington's net (Eddington, 1939). Our very tools and methods may be entangled with an existing computational structure within the model that no one yet knows, which eventually gives rise to a semblance of a phenomenon, as ablations did with self-repair. This applies to this work, too. We have used a tool to uncover what artifacts another tool can produce, and our findings are subject to the same scrutiny. We offer this work as a ladder to be climbed, for what it lets you see, and once you have seen, it should be thrown away (Wittgenstein, 1922).

AI USE STATEMENT

Language models are the object of study in this paper, and the models we analyse are described in §2. Separately, the authors used AI assistants to copy-edit and redraft parts of the manuscript, to search for related work and check references, and to draft plotting and TikZ code for two figures. The authors also used AI coding assistants to help write the experiment and analysis code and to audit the reported statistics against that code, and used AI to generate the synthetic false counterparts of the true statements (App. C.2). The authors ran and verified all AI-assisted code and reviewed all AI-assisted text. Every reported number comes from the authors' own runs, and the authors take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies open-weight models on synthetic statements from a public corpus and involves no human subjects or personal data.

REPRODUCIBILITY STATEMENT

All models and the fact corpus are public. Setup and data are in §2 and App. C.1, the intervention in §3, and fitting and tests in App. K. We will release code and per-run outputs with the camera-ready version.

REFERENCES

Areeb Ahmad, Abhinav Joshi, and Ashutosh Modi. Beyond components: Singular vector-based interpretability of transformer circuits. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2511.20273.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

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

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: a method for rigorously testing interpretability hypotheses. AI Alignment Forum, https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing, 2022.

Hang Chen, Jiaying Zhu, Xinyu Yang, and Wenya Wang. Rethinking circuit completeness in language models: AND, OR, and ADDER gates. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.10039.

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

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16124–16170, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.893. URL https://aclanthology.org/2023.acl-long.893/.

Arthur Stanley Eddington. *The Philosophy of Physical Science*. Cambridge University Press, 1939.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Atticus Geiger, Kyle Richardson, and Christopher Potts. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2020. URL https://arxiv.org/abs/2004.14623.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2106.02997.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. URL https://arxiv.org/abs/2012.14913.

Zhiren Gong, He Lu, Tiantong Wang, Yichi Zhang, Yixin Wang, Zihao Zeng, Ming Xiao, Chau Yuen, and Wei Yang Bryan Lim. Conditional co-ablation: Recovering self-repair backups in transformer circuits. *arXiv preprint arXiv:2607.01940*, 2026. URL https://arxiv.org/abs/2607.01940.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal neurons in GPT2 language models. *Transactions on Machine Learning Research*, 2024.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Yihuai Hong, Lei Yu, Haiqin Yang, Shauli Ravfogel, and Mor Geva. Intrinsic test of unlearning using parametric knowledge traces. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19513–19535, 2025.

Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. Rigorously assessing natural language explanations of neurons. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 317–331, 2023. URL https://aclanthology.org/2023.blackboxnlp-1.24/.

Maximilian Li and Lucas Janson. Optimal ablation for interpretability. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2409.09951.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through l_0 regularization. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1712.01312.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024.

Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 337–363, 2024. URL https://aclanthology.org/2024.blackboxnlp-1.22/.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The Hydra effect: Emergent self-repair in language model computations, 2023. URL https://arxiv.org/abs/2307.15771.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Talking heads: Understanding inter-layer communication in transformer language models. In *Advances in Neural Information Processing Systems*, 2024.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 14014–14024. Curran Associates, Inc., 2019.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust. In *First Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2407.08734.

Beren Millidge and Sid Black. The singular value decompositions of transformer weight matrices are highly interpretable. AI Alignment Forum, November 2022.

nostalgebraist. Interpreting GPT: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020.

Mark Oskin. Through the looking glass: Directly reading and writing transformers, 2026. URL https://arxiv.org/abs/2609.10210.

Arjun Patrawala, Jiahai Feng, Erik Jones, and Jacob Steinhardt. LLM layers immediately correct each other. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=7DY7kB8wyZ.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2nd edition, 2009. ISBN 978-0521895606.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.15390.

Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. Confidence regulation neurons in language models. 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. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The curse of multiple mediators: Hidden interaction effects in activation patching. *arXiv preprint arXiv:2606.27510*, 2026.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias, 2020. URL https://arxiv.org/abs/2004.12265.

Han-yu Wang. Function-vector heads are two populations: Writers and cancellers in in-context learning. *arXiv preprint arXiv:2606.07560*, 2026.

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.

Ludwig Wittgenstein. *Tractatus Logico-Philosophicus*. Kegan Paul, 1922. Translated by C. K. Ogden.

Zhengxuan Wu, Atticus Geiger, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, and Noah D. Goodman. A reply to Makelov et al. (2023)'s "interpretability illusion" arguments. *arXiv preprint arXiv:2401.12631*, 2024.

Min Xue and Artur Andrzejak. SVD as a fast interpretability method for transformers. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=7tt8TwMjdJ.

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

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

CONTENTS

§A Reader's Guide and Notation . . . 15
§A.1 What each section is evidence for . . . 15
§A.2 Notation . . . 15
§B Models and Weight-Space Bases . . . 16
§C Task, Prompts, and Data . . . 16
§C.1 Template and answer tokens . . . 16
§C.2 Data . . . 17
§C.3 Sample sizes across experiments . . . 17
§D Weight-Space Census of the Bench . . . 17
§D.2 Catalogues . . . 18
§E Regularity and Certification of the Second-Order Intervention . . . 22
§E.1 The dose axis is smooth across the full ladder . . . 22
§E.2 Certification statistics . . . 22
§F Recruitment Under the Dosed Swap . . . 22
§F.1 Per-model recruitment tables . . . 22
§F.2 Dose-resolved recruitment . . . 22
§F.3 Single-core versus joint-core recruitment . . . 22
§G The Learned Perturbation as an Independent Primary . . . 32
§G.1 Objective and training . . . 32
§G.2 Results . . . 34
§G.3 Three-route convergence . . . 34
§H Pretraining-Native Recruitment in Base versus Instruct . . . 37
§I Recruitment Is Not Renormalization . . . 41
§I.1 The objection . . . 41
§I.2 What the objection predicts . . . 41
§I.3 Measurement . . . 41
§I.4 Class-resolved margins in the worlds the audit indexes . . . 42
§I.5 The refutation . . . 43
§I.6 Why an inversion-type primary need not shrink the stream . . . 44
§I.7 What this establishes, and what it leaves open . . . 44
§J Fits and Diagnostics of the Coupling Law . . . 45
§J.1 Estimation procedure . . . 45
§J.2 Family-level summary . . . 45
§J.3 Per-direction fits . . . 45
§J.4 Fit diagnostic figures . . . 45
§J.5 Exemplar fits and the half-rung holdout . . . 50
§K Tests of the Coupling Law . . . 50
§L Out-of-Domain Test on GPT-2 Small and Indirect-Object Identification . . . 54
§M The Wiring Statistic as a Weight-Space Predictor of Coupling . . . 59
§M.1 Family breakdown . . . 59
§M.2 The placement control on directions the core cannot reach . . . 59
§N Derivations Showing That Every Clamp Is a Dose . . . 62
§O Beyond a Binary Contrast . . . 62
§P A Conjecture That Every Design Is a Point in the (u, v) Plane . . . 63

A READER'S GUIDE AND NOTATION

A.1 WHAT EACH SECTION IS EVIDENCE FOR

Main-text claim	Supporting appendix section
Contribution 1 (compensation is standing)	§F (recruitment), §G (other primaries), §H (base checkpoints), §I (not a LayerNorm artefact)
Contribution 2 (a dose axis and a nested intervention)	§E, §N
Contribution 3 (the coupling has an affine form)	§J and §K, with IOI in §L
Contribution 4 (named counterweights and relays, coupling legible in the weights)	§D (catalogues, clean census), §M

Table 2: Claim-to-evidence map. Each row names the section that carries the primary burden of proof. Supporting figures/tables for a claim may also appear incidentally elsewhere.

A.2 NOTATION

Symbol	Meaning
x	The prompt (SCM exogenous variable).
x'	The matched counterfactual partner of x (opposite class, identical phrasing).
b	World index: $b = \varnothing$ (clean), $b = \pi_k$ (core dosed to k; the main text writes π_λ for the same world), $b = \delta$ (learned perturbation).
r	A weight-space direction (an `mlp_neuron`, `ov_neuron`, `ov_svd`, or `mlp_svd` unit; §B).
ω_r	r's fixed write vector.
$a_r(x; b)$	r's activation coefficient on prompt x in world b.
m, Δ	Pair mean $\frac{1}{2}(a_T + a_F)$ and contrast $a_F - a_T$ of the core's clean activations on a matched pair.
λ	**The dose coordinate.** $\lambda(v) = 2(m - v)/\Delta$; $\lambda = +1$ clean, $\lambda = -1$ full counterfactual inversion, $\lambda = 0$ class-neutral midpoint.
k	Corruption fraction, $k = (1-\lambda)/2$. The ladder of §3 is $k \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, 2, 3, 4\}$; recruitment tables label $k = n$ as $\times n$. $k \leq 1$ is on-manifold, $k > 1$ is overshoot/extrapolation.
$E_r^b(x)$	r's causal contribution in world b: the verdict margin erased by overriding r's activation with its regime-consistent counterfactual (main text Eq. 5).
R_r	Recruitment, $E_r^\pi - E_r^0$.
D_{TF}	Class-fixed verdict margin, True minus False tokens; D_{XY} of Eq. 2 equals D_{TF} on true prompts and $-D_{\mathrm{TF}}$ on false ones.
own_r, γ_r	Intercept and slope of the affine law $E_r^c(\lambda) = \mathrm{own}_r^c + \gamma_r \lambda$ (main text Eq. 8). $\gamma_r < 0$: counterweight; $\gamma_r > 0$: relay.
z_r	Weight-only wiring-alignment statistic (§M) predicting $\lvert\gamma_r\rvert$ from the static checkpoint.
$\rho_b(x)$	Answer-position RMS of the pre-unembedding stream in world b (§I).
g_k	Final-norm gain, $\rho_\varnothing / \rho_{\pi_k}$.

Table 3: Notation used throughout the appendix. The dose is always λ, and x always denotes a prompt.

B MODELS AND WEIGHT-SPACE BASES

Gated MLP, neuron j of layer ℓ: $g_j = \mathbf{w}_j^{\text{gate}} \cdot x \quad u_j = \mathbf{w}_j^{\text{up}} \cdot x \quad \text{MLP}(x) = \sum_j a_j(x)\, W_{\text{down}}[:, j]$

Model	MLP input x	Neuron scalar $a_j(x)$	Residual write
Llama-3-8B-Instruct	$\text{RMSNorm}_{\text{post-attn}}(h)$	$\text{SiLU}(g_j)\, u_j$	$W_{\text{down}}[:, j]$
Gemma-2-9b-it	$\text{RMSNorm}_{\text{pre-ffw}}(h)$	$\text{GELU}_{\text{tanh}}(g_j)\, u_j$	$\boldsymbol{\gamma}_{\text{post}} \odot W_{\text{down}}[:, j]$
Qwen2.5-7B-Instruct	$\text{RMSNorm}_{\text{post-attn}}(h)$	$\text{SiLU}(g_j)\, u_j$	$W_{\text{down}}[:, j]$
Mistral-7B-Instruct-v0.3	$\text{RMSNorm}_{\text{post-attn}}(h)$	$\text{SiLU}(g_j)\, u_j$	$W_{\text{down}}[:, j]$

Table 4: **The `mlp_neuron` scalar, per model.** Every intervention in this paper acts on a_j, which is coordinate j of the `down_proj` input and the scalar multiplying that neuron's fixed write vector. All four models use a gated MLP, and in all four, the up branch is *linear*. Only the gate passes through the nonlinearity, so the scalar is $\sigma(g_j)\, u_j$ and never $\sigma(g_j)\, \sigma(u_j)$.

The other three weight-space bases used to assemble the bench (`ov_neuron` is a single output channel of an attention head's W_O, `ov_svd` a left-singular direction of a head's OV circuit, and `mlp_svd` a left-singular direction of an MLP's down-projection) follow the same interface (a fixed write vector ω_r and a scalar activation coefficient $a_r(x)$) with the coefficient read from the corresponding decomposition rather than a single `down_proj` coordinate. Main text §2 gives the shared interface, and Table 4 exists because the neuron basis is the one whose scalar is model-family-dependent, per the divergence above.

C TASK, PROMPTS, AND DATA

C.1 TEMPLATE AND ANSWER TOKENS

Every prompt is built by substituting the statement into the fixed template below.

```
Given the following statement, answer if it is True or
False.
Statement: {q}
Answer:
```

where {q} is the statement and the literal newlines are as shown (a blank line after the instruction, none before `Answer:`). The prompt is tokenised with the model's own tokenizer, with its BOS token but *without* a chat template. The instruct models are used as plain conditional language models, so that the evaluation regime and the intervention regime coincide. All readouts are taken at the final position, i.e. immediately after `Answer:`, so the next token is the verdict word.

The verdict is read from two sets of vocabulary items rather than a single token, so that the measurement does not depend on which capitalisation or leading-space form a given tokenizer happens to prefer. The two sets are

$$\mathcal{T} = \{\text{"_true"}, \text{"_True"}, \text{"_TRUE"}, \text{"true"}, \text{"True"}\},$$
$$\mathcal{F} = \{\text{"_false"}, \text{"_False"}, \text{"_FALSE"}, \text{"false"}, \text{"False"}\},$$

writing _ for a leading space. Two conventions matter for reproducibility. **Single-token filtering.** A surface form is retained only if the model's tokenizer encodes it as exactly one token. Multi-token forms are dropped, since a form with no column in the unembedding cannot contribute a logit. The retained subset is therefore tokenizer-dependent and recorded in each run's metadata. **Means rather than sums.** D_{TF} averages within each set before subtracting, so the scale is comparable across models even when the two sets survive filtering at different sizes. Logits are used raw (no softmax, no temperature) because the interventions are linear perturbations of the residual stream and a normalisation would make the measured effect depend on the rest of the vocabulary.

C.2 DATA

Source. Statements derive from the **Known-Facts** corpus released with ROME (Meng et al., 2022), obtained from `https://rome.baulab.info/data/dsets/known_1000.json` (1209 records, 1208 usable). Each record is a (subject, relation, attribute) triple with a cloze template, and we render a claim as `template.format(subject) + " " + attribute + "."`, i.e. a declarative sentence with a terminal period. The corpus ships a pre-curated `prompt` field that extends some templates with additional text. We do not use it, because those extensions are not well-formed for every record.

Synthesized false counterparts. Known-Facts contains *only true* facts, so it supplies no negative side. For each record, we synthesise a false claim by drawing a type-matched alternative attribute from the pool of attributes attested for that same relation, excluding the record's own attribute. The true and false members of a pair therefore share subject, relation, and surface template, and differ only in the asserted attribute, the minimal edit the contrastive interventions require. Sampling is deterministic given the build seed, so the pairing is reproducible. Only the true side is attributable to Meng et al. (2022). The counterfactual side is constructed here, and any error in it is ours.

Splits. Records are partitioned 70/10/20 into train/validation/test, stratified by relation so that every split contains every relation, and with no record appearing in more than one split (Splits are only used in learned steering perturbation G).

Filtering to clean-correct pairs. The corpus is not filtered by model behaviour at build time. The restriction used throughout the paper is applied afterward, and per model. A pair is retained iff $D_{\mathrm{TF}} > 0$ on its true member *and* $D_{\mathrm{TF}} < 0$ on its false member, evaluated zero-shot with the template of §C.1. Because this depends on the model, the retained subset differs across the four models.

C.3 SAMPLE SIZES ACROSS EXPERIMENTS

The source appendix reported n in each table/figure caption individually, which makes it hard to see at a glance whether a given comparison used a matched subsample or the full retained pool. Table 5 consolidates every n that appears in this appendix.

Experiment	n	Section
Llama single-core swap (`L18.n11065`)	532	§F
Gemma joint-core swap (`L28.n2046+L33.n4294`)	583	§F
Qwen joint-core swap (`L22.n13149+L24.n14758`)	478	§F
Mistral joint-core swap (`L20.n14286+L19.n8228`)	598	§F
Gemma / Qwen dose-resolved single-core comparisons	100	§F
Learned perturbation (`perfect_on_pairs`, all models)	– (full pool)	§G
Base-vs-instruct, Llama	240 (shared, both checkpoints correct)	§H
Base-vs-instruct, Gemma	535	§H
Base-vs-instruct, Qwen	470	§H
Base-vs-instruct, Mistral	423	§H
Final-norm gain audit, per family	50 matched pairs	§I
Half-rung holdout ($\lambda = \pm\frac{1}{2}$), per family	478–598 (same pop. as ladder)	§J

Table 5: Sample sizes referenced in this appendix, consolidated from individual table/figure captions.

D WEIGHT-SPACE CENSUS OF THE BENCH

D.1 SCAN PROTOCOL

Directions are named from the weights before any intervention. Scanning the four model families (main text §2), each direction r is labelled by the top and bottom tokens of $\omega_r^\top W_U$ and retained if its

pole represents words corresponding to truth or falsity (*lie, truth, wrong, sure, illusion*), including cross-lingual matches pairing English true/false with other counterparts. Selection relies solely on static weights. Whether a retained direction contributes, opposes, or is dormant for the realised verdict is a separate, causal question, answered in §F.

D.2 CATALOGUES

Tables 6 to 9 give the weight-fixed, intervention-independent poles for the benched directions in each model.

Table 6: **Vocabulary poles of the recruited directions in Llama-3-8B-Instruct** ($\omega_r^\top W_U$, weight-fixed, intervention-independent, with the top-4 promoted and the bottom-4 suppressed). **Green** marks truth-coding tokens, **red** falsity-coding tokens, gray non-interpretable subword fragments.

Axis	Promoted (top-4)	Suppressed (bottom-4)
L23.n8972	iller, rello, LC, etz	**lie**, **lies**, **lying**, **Lie**
L19.n2738	ledged, Alo, ADDE, くだ (Jp.)	**illusion**, **illusions**, **false**, **illusion**
L23.n9811	**truth**, **Truth**, **truth**, **Truth**	oby, imore, Tell, tell
L19.n3754	Edition, onte, フ (Jp.), rale	**sure**, **yes**, **yes**, **Sure**
L22.h1.svd117	miss, missing, chances, misses	**true**, **true**, Important, **TRUE**
L24.n4543	**yes**, **true**, **True**, **Yes**	OrNil, uelle, illard, izh (Cyr.)
L21.n798	Arena, quer, KER, .BorderFactory	**true**, is, **True**, **true**
L21.n4049	642, upe, airo, 立て (Jp.)	**wrong**, **wrong**, **Wrong**, **WRONG**
L20.h11.c95	reversal, reversed, loss, ercul	**false**, **false**, **False**, (explain)
L20.n14044	lagod (Cyr.), IGGER, /generated	**truly**, **true**, **really**, 真正 **(genuine)**
L22.h1.svd116	fun, Needs, needs, hopes	**correct**, **legitimate**, 正确 **(correct)**, **legitimacy**
L28.h27.c34	total, total, Total, TOTAL	**True**, **true**, fully, **True**
L28.h27.c120	about, about, **True**, truly	**real**, near, Near, **Real**
L21.n14165	**Genuine**, **authentic**, **genuine**, **pure**	atorio, čet, strcasecmp, yh
L18.n11065†	**true**, **yes**, **YES**, .YES	**false**, **False**, **false**, **False**

†Core. Its full $k=6$ lists are `true, yes, YES, .YES, Yes, TRUE` (promoted) and `false, False, false, False, FALSE, (false` (suppressed).

Table 7: **Logit-lens poles of the recruited Gemma-2-9b-it directions and the two cores** ($\omega_r \cdot W_U$, weight-fixed, intervention-independent, with the top-4 promoted and the bottom-4 suppressed). **Green**=truth-coding, **red**=falsity-coding, gray=non-interpretable subword. Most recruited anti-directions carry a clean English truth/falsity pole on one side against subword noise on the other. The core `L28.n2046` carries the cross-lingual 否 (no) beside English *yes/true*.

Axis	Promoted (top-4)	Suppressed (bottom-4)
L38.n7088	**wrong**, **wrong**, **Wrong**, **Wrong**	__)., BarStyle, preferably, AutoSize
L39.n10085	**fake**, **Fake**, **pseudo**, **pseud**	argout, cyd, MessageOf, MigrationBuilder
L38.n854	ModelExpression, _":, ituary, orghini	**true**, **true**, **True**, **True**
L41.n8771	SOUNDBITE, estekak, protoimpl, ChrTalk	**True**, **true**, **True**, **true**
L39.n10210	**incorrect**, **incorrect**, **Incorrect**, **WRONG**	IBOutlet, setopt, openzeppelin, Geos
L41.n3789	haikusbot, writeFieldEnd, complexType, getDescription	**real**, skuto, **truly**, **genuine**
L32.n8200	LookAnd, GEBURTS, ItemBackground, <bos>	**False**, **false**, **False**, **True**
L33.n1466	**deception**, **fake**, **deceiving**, **deceptive**	rungsseite, createSlice, surate, photobucket
L34.n4784	Хьажоргаш (Cyr.), TypedDataSet, RegressionTest, pasto	tagHelperRunner, nahilalakip, **truth**, **truths**
L40.h11.svd18	**truth**, tissue, traffic, bishop	GEBURTSDATUM, StoryboardSegue, ModelExpression, SourceChecksum
L40.h7.c177	brekdaun (Dev.), Upper, Upper, Continuous	DebuggerNonUser, ∅, StoryboardSegue, perfect
L40.h7.c105	**real**, **Real**, **Real**, **real**	principalColumn, déf, PerformLayout, <>",
L33.n4294	InputDecoration, **true**, jMenuItem, **true**	**wrong**, **wrong**, **Wrong**, **Wrong**
L28.n2046	否 (no), **Nope**, **nope**, **NOPE**	**yes**, **true**, **TRUE**, **Yes**

Table 8: **A catalogue of interpretable truth/falsity directions in Qwen2.5-7B-Instruct** ($\omega_r \cdot W_U$, weight-fixed, intervention-independent, with the top-4 promoted and the bottom-4 suppressed). **Green**=truth-coding, **red**=falsity-coding, gray=non-interpretable subword. As in Gemma-2-9b and Llama-3-8b, Qwen carries a large population of directions whose unembedding poles encode the True/False verdict for this task (6,310 marked across the four component families). A representative sample spanning MLP neurons (layers L20 to L27) and attention (OV) axes is listed by $|\cos(\omega_r, w_{TF})|$. Coding is markedly cross-lingual, pairing Chinese 真 (true)/ (wrong)/假 (fake)/虚假 (false) with English *true/false/fake*. Whether each contributes to, opposes, or is dormant for the realized verdict is measured separately (§F).

Axis	Promoted (top-4)	Suppressed (bottom-4)
L27.n4228	∅, -Headers, ómo, '}}>	**truly**, 真正 (genuine), **true**, 真正的 (genuine)
L22.n13149	**true**, **True**, _true, **true**	**false**, **False**, **false**, **False**
L20.n2073	**False**, **false**, **False**, **false**	**true**, **True**, **true**, **True**
L22.n4538	**false**, **false**, **False**, **False**	**true**, **True**, **True**, **true**
L27.n13033	fa, ase, Л (Cyr.), fa	**false**, **False**, **False**, **false**
L24.n3278	**true**, **true**, 真 (real), **True**	ulfill, .Gray, naz, 泄露 (Zh.)
L25.n4929	assertNotNull, seedu, reeNode, addtogroup	**false**, **false**, **False**, **False**
L25.n11565	**true**, **true**, 真 (true), **True**	commodo, 情人 (Zh.), ollah, 蘑 (Zh.)
L22.n9820	**yes**, **sure**, **Yes**, **Sure**	YO, áll, ublish, .Itoa
L24.n14758	**right**, **right**, **Right**, **Right**	(wrong), 錯 (wrong), ∅, (error)
L27.n6739	**none**, **none**, **None**, **None**	;line, 个 (Zh.), oriously, odox
L23.n3002	,},, ('','$, akest, TMPro	**true**, **true**, (true, **True**
L24.n9321	ORITY, twor, kowski, akthar (Ar.)	真正的 (genuine), 真正 (genuine), **true**, **genuine**
L22.n9609	**False**, **false**, **false**, **False**	glEnable, 阳性 (positive), 正品 (genuine), (Zh.)
L23.n7028	.Itoa, 机 (Zh.), 싼, StringBuffer	**yes**, **Yes**, **Yes**, **yes**
L25.n8444	seedu, Decompiled, anes, .uf	假 (fake), **fake**, (pseudo), pseud
L23.n14476	ournemouth, ulfill, indsight, になれ (Jp.)	真正的 (genuine), 真 (real), **real**, 真的 (Zh.)
L22.n13025	:".$, km, .IsNullOr, :normal	真正 (genuine), 真正的 (genuine), **truly**, **true**
L23.n8341	.getOwnProperty, Hath, 眉 (Zh.), 렵	**illusion**, Ill, **illusions**, 幻 (Zh.)
L24.n8308	**yes**, **Yes**, **yes**, **Yes**	uk, @$, 究竟 (Zh.), 到底 (Zh.)
L24.n12848	虚假 (false), **false**, 假 (fake), **fake**	苡 (Zh.), nieu, , [Byte
L24.n5510	**none**, **none**, **None**, **None**	个 (Zh.), **no**, ,no, 不再 (Zh.)
L23.h27.a9	**correct**, 正确 (correct), ∅, 正确的 (correct)	.Empty, 无 (unrelated), empty, Empty
L23.h27.c88	正确 (correct), **correct**, :[[, **Correct**	不可以 (Zh.), 可以直接 (Zh.), egot, 不可 (Zh.)
L26.h17.svd117	.Active, active, 真 (true), active	base, internal, 内部 (internal), 基 (Zh.)
L21.n17369	"Yes, **Yes**, “Yes, **Yes**	Nor, **Nope**, -not, nor
L23.h27.c82	**correct**, **true**, 正确 (correct), know	al-ithn (Ar.), .Undef, "nil, _TypeInfo

Table 9: **A catalogue of interpretable truth/falsity directions in Mistral-7B-Instruct-v0.3** ($\omega_r \cdot W_U$, weight-fixed, intervention-independent, with the top-4 promoted and the bottom-4 suppressed). **Green**=truth-coding, **red**=falsity-coding, gray=non-interpretable subword. All 35 bench directions are listed, ordered by layer. Of these, 14 carry a clean bipolar verdict axis and the rest carry a clean pole on one side against subword noise on the other. † marks the core(s) used as the primary intervention.

Axis	Promoted (top-4)	Suppressed (bottom-4)
L15.n11835	**fake**, **invalid**, **false**, **False**	riel, lib, trends, HAL
L16.n1170	ensures, ensuring, confirmed, ingo	STO, **FALSE**, **false**, ∅
L16.n5090	ugno, iten, Graf, rolog	pret, pretend, **fake**, nomin
L16.n9991	**false**, fals, **False**, **false**	∅, amon, bert, bis
L17.n11073	integrity, respons, **genuine**, proper	artificial, stere, unsafe, **false**
L17.n13852	**true**, **real**, **True**, **TRUE**	asp, ape, dom, Mend
L17.n8203	**fake**, fals, fraud, **false**	aju, Nu, ∅, pn
L18.n10672	odi, Roose, berger, **Verify**	mock, **fake**, Mock, mock
L18.n8574	anymore, excessive, too, unnecessary	properly, rather, proper, hopefully
L19.n11056	**genuinely**, **truly**, **genuine**, **actual**	∅, indirect, ester, oll
L19.n8228†	**true**, **truth**, **True**, **TRUE**	**false**, **False**, **False**, **false**
L20.n14286†	shake, shook, **False**, shaking	**yes**, **Yes**, **true**, **Yes**
L20.n4786	**right**, achuset, **TRUE**, Seth	**wrong**, mistaken, mis, **incorrect**
L22.h20.c93	perfection, exact, exactly, mainstream	unexpected, new, **false**, **incorrect**
L22.h20.svd0	poorly, **wrong**, poor, incons	**correctly**, properly, **correct**, proper
L22.h20.svd40	slightly, slight, missing, remaining	various, increasingly, **actual**, **real**
L22.h23.svd2	unknown, **wrong**, absent, isEmpty	**correctly**, **valid**, viable, **legitimate**
L22.n1070	reliable, properly, efficient, **accurate**	unsafe, poorly, unfair, undefined
L22.n5778	**wrong**, wr, **incorrect**, Wr	properly, proper, gc, olia
L22.n8735	sek, pred, ork, otal	fals, **false**, **incorrect**, **False**
L24.n1021	Strip, Tib, endo, strip	**false**, **fake**, **False**, **false**
L24.n4907	poorly, poor, **incorrect**, **wrong**	**correctly**, ben, adequate, **correct**
L24.n5072	**true**, **True**, **TRUE**, **true**	zero, Zero, Zero, zero
L25.h8.svd45	wild, **false**, artificial, amateur	**real**, normal, Private, private
L27.n11713	**truly**, **true**, **genuinely**, **true**	oft, 宮 (palace), uvud, ço
L28.n4079	inet, stract, èse, teger	**true**, **True**, **true**, **True**
L28.n9060	**truly**, **true**, **True**, Tru	**accurate**, accur, accurately, accuracy
L29.n3831	**real**, **real**, **Real**, **Real**	esser, gorith, otta, **False**
L29.n8705	**true**, **True**, **true**, **True**	illet, agr, ícula, þl
L30.h3.c8	Complete, complete, Complete, formal	exact, extract, aver, Extract
L30.n11434	**false**, **fake**, **False**, **false**	F, Bog, chen, icc
L30.n11755	**real**, **real**, **actual**, **true**	ally, ALL, benfalls, iously
L30.n498	**genuine**, **authentic**, **accurate**, **true**	arbitrary, **False**, idle, lazy
L31.n13669	xF, stract, icer, kör	**true**, **True**, **True**, **true**
L31.n8773	**false**, **False**, **False**, fals	ulo, heng, Ä, kennis

E Regularity and Certification of the Second-Order Intervention

E.1 The dose axis is smooth across the full ladder

Certification, the coupling fits of §J, and every claim anchor at $k \leq 1$. The regime $k > 1$ ($\lambda = 1 - 2k$) is read as extrapolation along a regular response surface rather than as an artefact of an off-manifold $do(\cdot)$, and that reading rests on the surface being smooth through the transition. Table 18 (§I) gives the evidence. At $\times 1$ every family still answers correctly on both classes, every family-class pair crosses zero between $\times 1$ and $\times 2$, and no family shows a discontinuity there. Llama's per-dose steps on the true side are near-constant ($-2.07, -2.27, -2.48, -2.70$), while the two joint-core families decelerate sharply at depth (Gemma $-2.85, -2.38, -1.09, -0.54$ and Qwen $-5.26, -5.07, -3.56, -2.17$). Compensatory recruitment predicts a marginal effect that flattens as the intervention deepens, because the reserve's contribution grows with dose (§F) as the primary's marginal effect shrinks. Llama, whose reserve is by far the smallest of the three (§H), is the family whose steps do not flatten. This co-occurs with the true-side norm growth documented in §I (Table 17), consistent with Eq. 12.

Magnitudes at $k > 1$ are nonetheless read as ordinal rather than calibrated, because the per-direction fits place the linear boundary at or near full inversion for most directions (§J), and §I locates the dose at which the intervention itself leaves the linear regime of Eq. 12.

E.2 Certification statistics

A direction is scored per side (true-statement, false-statement) with a paired $t = |\text{mean}|/\text{sem}$ against a threshold of $t \geq 2$. Cells below threshold are treated as not reliably different from zero. ***Certified contributor*** (main text §4) requires $E_r^0 > 0$, reliably, on both sides in the clean model. ***Recruited*** requires the same sign-and-reliability test to hold under the primary intervention for a direction that was clean-null or clean-anti. `flip@` in Tables 10 to 13 records the first dose at which that happens. Evaluation throughout uses pairs on which the model is correct, zero-shot, on both the true and false members (§C.2), so that recruitment is measured against a background where the clean verdict holds.

F Recruitment Under the Dosed Swap

F.1 Per-model recruitment tables

Tables 10 to 13 give, per model, each benched direction's clean and dosed contributions on both classes and the dose at which it flips.

F.2 Dose-resolved recruitment

Figures 7 to 10 plot the same contributions against dose.

F.3 Single-core versus joint-core recruitment

Gemma, Qwen, and Mistral each carry two cores, while Llama carries one, so no single-vs-joint comparison exists for it. Figures 11, 12, and 13 give the aggregate recruitment (mean R and count crossing into contribution) under each core alone versus both jointly. The per-detector breakdowns (Figs. 14, 15, 16) are kept as a distinct, lower-priority subsection rather than cut, because they show which specific detector each single core recruits, information the aggregates lack (e.g. Mistral's `L19.n8228` alone never recruits `L31.n13669` at any dose, a complementarity invisible in the aggregate count).

Table 10: **Backup recruitment across dose under ablation of the Llama-3-8B-Instruct verdict core** (`L18.n11065`, contrastive swap, `sec_ref=regime`, n=532). Per side, the clean erased margin $E^0_T = D_{TF}(T) - D_{TF}(T, \text{sec})$ and $E^0_F = D_{TF}(F, \text{sec}) - D_{TF}(F)$ (+ supports the verdict, with colour giving the clean class as anti/contributor/null), then E^π_T and E^π_F under the ablation at each swap dose. Green=reliably > 0 (now a contributor on that side), red=reliably < 0. The superscript is the paired $t = |\text{mean}|/\text{sem}$. Uncoloured cells are below threshold ($t < 2$). Row shade = clean class, and 'flip@'=first dose a clean null/anti axis is reliably > 0 on both sides. Across the coloured entries the median $t \approx 8$ (up to $t \approx 56$). The pair-averaged verdict gap runs -7.8 (clean) $\to -3.7, +1.4, +6.8, +12.3$ across $\times 1$.. $\times 4$, so the verdict gives way between $\times 1$ and $\times 2$.

			E^0 (clean)		$E^\pi \times 1$		$E^\pi \times 2$		$E^\pi \times 3$		$E^\pi \times 4$		
secondary	family	detects	T	F	T	F	T	F	T	F	T	F	flip@
`L23.n8972`	mlp neuron	lie	-0.04	-0.06	$+0.02^{13}$	$+0.02^{13}$	$+0.06^{34}$	$+0.04^{32}$	$+0.11^{47}$	$+0.06^{41}$	$+0.16^{56}$	$+0.09^{52}$	$\times 1$
`L23.n9811`	mlp neuron	truth	-0.02	-0.02	$+0.00^{3}$	$+0.00^{4}$	$+0.05^{25}$	$+0.03^{22}$	$+0.10^{37}$	$+0.05^{34}$	$+0.11^{48}$	$+0.02^{8}$	$\times 1$
`L19.n3754`	mlp neuron	sure	-0.01	-0.02	$+0.00^{3}$	$+0.00^{4}$	$+0.01^{11}$	$+0.02^{15}$	$+0.02^{18}$	$+0.02^{13}$	$+0.02^{14}$	$+0.01^{9}$	$\times 1$
`L21.n4049`	mlp neuron	wrong	-0.01	-0.01	$+0.01^{8}$	$+0.01^{10}$	$+0.03^{21}$	$+0.03^{28}$	$+0.04^{29}$	$+0.05^{37}$	$+0.03^{19}$	$+0.04^{30}$	$\times 1$
`L19.n2738`	mlp neuron	illusion	-0.01	-0.01	$+0.00$	$+0.00^{2}$	$+0.01^{5}$	$+0.02^{13}$	$+0.01^{6}$	$+0.04^{30}$	$+0.01^{4}$	$+0.06^{42}$	$\times 2$
`L22.h1.svd117`	ov svd	true	-0.00	-0.01	$+0.01^{10}$	$+0.01^{9}$	$+0.02^{16}$	$+0.01^{12}$	$+0.03^{21}$	$+0.02^{18}$	$+0.03^{18}$	$+0.02^{18}$	$\times 1$
`L21.n798`	mlp neuron	true	-0.00	-0.00	$+0.00^{2}$	$+0.00^{2}$	$+0.00$	$+0.00^{2}$	$+0.00^{3}$	$+0.00^{2}$	-0.01^{10}	-0.01^{8}	$\times 1$
`L22.svd323`	mlp svd	–	-0.00	-0.00	$+0.00$	$+0.00^{2}$	$+0.00$	$+0.00$	$+0.00^{4}$	$+0.00$	$+0.00^{2}$	$+0.00^{3}$	$\times 4$
`L21.n517`	mlp neuron	yes	-0.00	-0.00	$+0.00$	$+0.00^{3}$	$+0.00^{4}$	$+0.01^{4}$	$+0.01^{8}$	$+0.01^{10}$	$+0.01^{4}$	$+0.02^{13}$	$\times 2$
`L28.h27.c120`	ov neuron	True	-0.00	-0.00	$+0.00$	$+0.00^{3}$	$+0.00^{2}$	$+0.00$	$+0.01^{12}$	$+0.00$	$+0.02^{16}$	$+0.00$	–
`L20.h11.c95`	ov neuron	false	$+0.00$	-0.00	-0.00	$+0.00$	$+0.00$	$+0.00$	$+0.00^{3}$	$+0.01^{10}$	$+0.00$	$+0.03^{19}$	$\times 3$
`L21.n14165`	mlp neuron	Genuine	-0.00	-0.00	$+0.00^{4}$	$+0.01^{6}$	$+0.01^{8}$	$+0.01^{7}$	$+0.02^{13}$	$+0.01^{6}$	$+0.01^{11}$	$+0.00^{2}$	$\times 1$
`L28.h27.c34`	ov neuron	True	$+0.00$	$+0.00$	$+0.00^{2}$	$+0.00^{3}$	$+0.00^{5}$	$+0.00$	$+0.01^{10}$	$+0.00^{5}$	$+0.02^{15}$	$+0.01^{5}$	$\times 1$
`L22.h1.svd116`	ov svd	correct	-0.00	$+0.00$	$+0.00^{3}$	$+0.00^{4}$	$+0.01^{6}$	$+0.01^{6}$	$+0.01^{7}$	$+0.01^{8}$	$+0.01^{5}$	$+0.01^{8}$	$\times 1$
`L28.h27.c0`	ov neuron	real	$+0.00$	-0.00	$+0.00$	$+0.00$	$+0.00$	-0.00	$+0.01^{5}$	$+0.00^{4}$	$+0.00^{3}$	$+0.01^{9}$	$\times 3$
`L22.h1.c60`	ov neuron	valid	$+0.00$	-0.00	$+0.01^{5}$	$+0.00^{3}$	$+0.01^{6}$	$+0.00^{3}$	$+0.01^{9}$	$+0.01^{8}$	$+0.01^{8}$	$+0.01^{9}$	$\times 1$
`L28.h27.c43`	ov neuron	–	$+0.00$	-0.00	$+0.00$	$+0.00^{3}$	$+0.00^{3}$	$+0.00^{2}$	$+0.01^{11}$	$+0.00^{3}$	$+0.02^{17}$	$+0.01^{5}$	$\times 2$
`L24.n4543`	mlp neuron	yes	-0.00	-0.01	$+0.00$	$+0.01^{8}$	$+0.03^{18}$	$+0.02^{13}$	$+0.06^{29}$	-0.01^{9}	$+0.08^{36}$	-0.05^{23}	–
`L20.n14044`	mlp neuron	truly	-0.00	-0.00	$+0.00$	$+0.00^{4}$	$+0.00^{3}$	$+0.00^{4}$	$+0.01^{10}$	$+0.01^{6}$	$+0.02^{12}$	$+0.01^{5}$	–
`L20.n1390`	mlp neuron	wrong	$+0.00$	$+0.00$	$+0.01^{7}$	$+0.01^{11}$	$+0.01^{7}$	$+0.02^{11}$	$+0.02^{14}$	$+0.03^{19}$	$+0.04^{22}$	$+0.05^{27}$	–

Table 11: **Backup recruitment across dose under JOINT ablation of both Qwen2.5-7B-Instruct verdict cores** (`L22.n13149+L24.n14758`, contrastive swap, `sec_ref=regime`, n=478). Columns as in Table 10. Across the coloured entries the median $t \approx 21$ (up to $t \approx 201$). $\kappa_{\text{base}} = -18.8$, and the joint κ_{primary} at $\times 1$.. $\times 4 = -8.5, +4.7, +15.8, +23.4$ crosses 0 at $\times 2$, where the verdict flips. Null-class rows stay ~ 0 across all doses (specificity).

			E^0 (clean)		$E^\pi \times 1$		$E^\pi \times 2$		$E^\pi \times 3$		$E^\pi \times 4$		
secondary	family	detects	T	F	T	F	T	F	T	F	T	F	flip@
`L27.n13033`	mlp neuron	false	-2.86	-2.88	-0.50^{15}	-0.51^{15}	$+2.93^{116}$	$+3.22^{109}$	$+3.95^{189}$	$+4.45^{178}$	$+4.08^{199}$	$+4.65^{201}$	$\times 2$
`L27.n4228`	mlp neuron	truly	-0.84	-0.87	$+0.03^{4}$	$+0.03^{4}$	$+0.37^{29}$	$+0.41^{29}$	$+1.43^{47}$	$+1.58^{47}$	$+2.62^{63}$	$+2.96^{63}$	$\times 1$
`L25.n4929`	mlp neuron	false	-0.25	-0.07	-0.01^{3}	-0.00^{2}	$+0.02^{12}$	$+0.05^{17}$	$+0.05^{21}$	$+0.21^{35}$	$+0.05^{21}$	$+0.41^{45}$	$\times 2$
`L23.n8341`	mlp neuron	illusion	-0.18	-0.11	$+0.00^{2}$	$+0.00$	$+0.04^{18}$	$+0.15^{29}$	$+0.08^{28}$	$+0.49^{52}$	$+0.12^{37}$	$+0.84^{66}$	$\times 2$
`L24.n12848`	mlp neuron	false	-0.21	-0.04	$+0.01$	$+0.00$	$+0.11^{25}$	$+0.29^{33}$	$+0.04^{15}$	$+0.32^{41}$	-0.01^{3}	$+0.26^{35}$	$\times 2$
`L23.n14476`	mlp neuron	real	-0.05	-0.09	-0.00^{3}	-0.01^{5}	$+0.02^{10}$	$+0.04^{14}$	$+0.13^{34}$	$+0.16^{38}$	$+0.29^{44}$	$+0.13^{28}$	$\times 2$
`L25.n8444`	mlp neuron	fake	-0.06	-0.03	$+0.01^{5}$	$+0.02^{7}$	$+0.07^{13}$	$+0.14^{14}$	$+0.21^{49}$	$+0.64^{44}$	$+0.26^{71}$	$+1.09^{70}$	$\times 1$
`L22.n13025`	mlp neuron	真正 (genuine)	-0.02	-0.04	-0.03^{17}	-0.06^{29}	-0.03^{16}	-0.04^{22}	-0.02^{10}	-0.01^{6}	-0.02^{11}	-0.00	–
`L22.n4538`	mlp neuron	false	-0.03	-0.03	-0.03^{18}	-0.04^{21}	-0.02^{13}	-0.07^{27}	-0.01^{7}	-0.06^{26}	-0.01^{4}	-0.05^{20}	–
`L26.h20.svd17`	ov svd	–	-0.01	-0.01	-0.00^{4}	-0.01^{5}	$+0.00$	$+0.00$	$+0.00^{2}$	$+0.01^{4}$	$+0.00$	$+0.00$	$\times 3$
`L11.svd690`	mlp svd	true	-0.00	-0.00	-0.00	-0.00	-0.01^{2}	-0.00	-0.00	$+0.00$	-0.00	$+0.00$	–
`L24.n5510`	mlp neuron	none	-0.00	$+0.00$	-0.00	-0.00	-0.01^{3}	-0.00	-0.00	$+0.00^{3}$	-0.00	$+0.00$	–
`L21.svd2451`	mlp svd	–	$+0.00$	$+0.00$	$+0.00^{3}$	-0.00	-0.00^{2}	$+0.00^{2}$	-0.00^{2}	$+0.00$	$+0.00$	-0.00	–
`L27.n6739`	mlp neuron	none	$+0.00$	$+0.00$	$+0.00^{3}$	$+0.00^{3}$	$+0.00^{3}$	$+0.00^{4}$	$+0.00^{3}$	$+0.00^{5}$	$+0.00^{2}$	$+0.00^{3}$	–
`L24.n8308`	mlp neuron	yes	$+0.16$	$+0.49$	$+0.06^{27}$	$+0.10^{32}$	-0.25^{44}	-0.17^{52}	-0.46^{68}	-0.06^{15}	-0.58^{85}	$+0.16^{22}$	–
`L22.n9609`	mlp neuron	False	$+0.29$	$+0.28$	$+0.39^{46}$	$+0.41^{45}$	$+0.40^{43}$	$+0.42^{50}$	$+0.24^{44}$	$+0.28^{53}$	$+0.17^{42}$	$+0.19^{45}$	–
`L27.n15192`	mlp neuron	actual	$+0.19$	$+0.20$	$+0.09^{52}$	$+0.10^{55}$	-0.09^{41}	-0.10^{41}	-0.22^{78}	-0.24^{77}	-0.29^{86}	-0.33^{85}	–
`L25.n11824`	mlp neuron	真正 (genuine)	$+0.10$	$+0.12$	$+0.01^{10}$	$+0.04^{15}$	-0.04^{20}	-0.11^{52}	-0.07^{32}	-0.09^{43}	-0.08^{33}	-0.03^{11}	–
`L26.h17.svd117`	ov svd	真 (true)	$+0.11$	$+0.11$	$+0.07^{38}$	$+0.13^{57}$	$+0.04^{27}$	$+0.13^{61}$	$+0.04^{24}$	$+0.08^{52}$	$+0.05^{27}$	$+0.03^{18}$	–
`L20.h1.svd6`	ov svd	YES	$+0.06$	$+0.14$	$+0.03^{7}$	$+0.17^{48}$	$+0.01^{3}$	$+0.10^{33}$	$+0.01^{4}$	$+0.03^{8}$	$+0.01^{4}$	-0.01	–
`L23.h27.c82`	ov neuron	correct	$+0.01$	$+0.02$	$+0.02^{12}$	$+0.02^{12}$	$+0.03^{14}$	$+0.01^{5}$	$+0.02^{11}$	$+0.00$	$+0.02^{11}$	-0.00	–

Table 12: **Backup recruitment across dose under JOINT ablation of both Gemma-2-9b-it verdict cores** (`L28.n2046+L33.n4294`, contrastive swap, `sec_ref=regime`, n=583). Columns as in Table 10. Across the coloured entries the median $t \approx 20$ (up to $t \approx 252$). $\kappa_{\text{base}} = -9.50$, and joint κ_{primary} at $\times 1.. \times 4 = -4.38, +2.30, +5.59, +7.14$ (crosses 0 at $\times 2$). The one null-class row, `L36.n1119`, turns reliably negative from $\times 2$ (E^{π} between -0.10 and -0.22) and never supports the verdict.

			E^0 (clean)		$E^{\pi} \times 1$		$E^{\pi} \times 2$		$E^{\pi} \times 3$		$E^{\pi} \times 4$		
secondary	family	detects	T	F	T	F	T	F	T	F	T	F	flip@
`L38.n7088`	mlp neuron	wrong	-1.420	-0.715	$+0.474^{72}$	$+0.613^{54}$	$+1.060^{194}$	$+2.202^{122}$	$+1.178^{228}$	$+2.587^{163}$	$+1.215^{252}$	$+2.727^{197}$	×1
`L39.n10085`	mlp neuron	fake	-0.504	-0.418	-0.364^{88}	-0.342^{94}	-0.179^{44}	-0.181^{43}	-0.017^{4}	-0.022^{4}	$+0.071^{15}$	$+0.077^{15}$	×4
`L38.n854`	mlp neuron	true	-0.179	-0.281	-0.044^{26}	-0.055^{26}	$+0.075^{35}$	$+0.072^{40}$	$+0.177^{55}$	$+0.146^{66}$	$+0.247^{69}$	$+0.192^{78}$	×2
`L41.n8771`	mlp neuron	True	-0.107	-0.202	$+0.047^{16}$	$+0.057^{16}$	$+0.261^{69}$	$+0.221^{66}$	$+0.339^{80}$	$+0.236^{80}$	$+0.354^{90}$	$+0.232^{86}$	×1
`L39.n10210`	mlp neuron	incorrect	-0.167	-0.127	-0.019^{13}	-0.018^{13}	$+0.046^{30}$	$+0.048^{31}$	$+0.019^{15}$	$+0.021^{17}$	$+0.007^{5}$	$+0.006^{5}$	×2
`L41.n3789`	mlp neuron	real	-0.100	-0.182	-0.051^{37}	-0.063^{40}	$+0.064^{40}$	$+0.054^{38}$	$+0.141^{69}$	$+0.099^{62}$	$+0.177^{88}$	$+0.117^{74}$	×2
`L32.n8200`	mlp neuron	False	-0.067	-0.175	$+0.028^{12}$	-0.082^{36}	-0.002	-0.003^{2}	$+0.032^{17}$	-0.048^{22}	$+0.055^{26}$	-0.080^{32}	–
`L33.n1466`	mlp neuron	deception	-0.147	-0.039	-0.035^{22}	-0.032^{21}	$+0.011^{9}$	$+0.029^{19}$	$+0.016^{13}$	$+0.038^{30}$	$+0.013^{12}$	$+0.037^{29}$	×2
`L34.n4784`	mlp neuron	truth	-0.044	-0.037	-0.005^{5}	-0.004^{4}	$+0.021^{20}$	$+0.039^{37}$	$+0.035^{30}$	$+0.057^{48}$	$+0.044^{36}$	$+0.065^{51}$	×2
`L40.h11.svd18`	ov svd	truth	-0.027	-0.044	-0.012^{15}	-0.015^{18}	$+0.009^{10}$	$+0.008^{10}$	$+0.026^{26}$	$+0.020^{22}$	$+0.037^{36}$	$+0.027^{29}$	×2
`L40.h7.c105`	ov neuron	real	-0.015	-0.014	-0.003^{4}	-0.003^{4}	$+0.006^{8}$	$+0.005^{8}$	$+0.012^{15}$	$+0.013^{16}$	$+0.017^{19}$	$+0.017^{20}$	×2
`L40.h7.c177`	ov neuron	genuine	-0.011	-0.014	-0.005^{6}	-0.004^{6}	$+0.003^{5}$	$+0.002^{2}$	$+0.008^{10}$	$+0.007^{9}$	$+0.011^{13}$	$+0.010^{12}$	×2
`L36.n1119`	mlp neuron	truth	$+0.041$	$+0.207$	$+0.002$	$+0.001$	-0.122^{46}	-0.222^{62}	-0.176^{57}	-0.166^{42}	-0.210^{65}	-0.104^{31}	–

Table 13: **Backup recruitment across dose under JOINT ablation of both Mistral-7B-Instruct-v0.3 verdict cores** (`L20.n14286+L19.n8228`, contrastive swap, `sec_ref=regime`, n=598). **Rows are the clean anti/null axes that become verdict contributors under the ablation** (the recruitment set). Clean contributors and never-flipping null controls are omitted (they are in the total-effect scan). Columns as in Table 10. Across the coloured entries the median $t \approx 20$ (up to $t \approx 129$). $\kappa_{\text{base}} = -11.6$, and joint κ_{primary} at $\times 1.. \times 4 = -5.0, +3.9, +10.5, +15.9$ (crosses 0 at $\times 2$, where the verdict flips).

			E^0 (clean)		$E^{\pi} \times 1$		$E^{\pi} \times 2$		$E^{\pi} \times 3$		$E^{\pi} \times 4$		
secondary	family	detects	T	F	T	F	T	F	T	F	T	F	flip@
`L31.n13669`	mlp neuron	true	-1.929	-2.083	-1.470^{88}	-1.624^{79}	$+0.227^{7}$	$+0.209^{6}$	$+0.736^{43}$	$+0.793^{44}$	$+0.522^{40}$	$+0.583^{40}$	×2
`L31.n8773`	mlp neuron	false	-0.771	-0.748	$+0.024$	$+0.050^{2}$	$+1.088^{61}$	$+1.240^{56}$	$+1.489^{78}$	$+1.754^{72}$	$+1.438^{58}$	$+1.732^{56}$	×2
`L28.n4079`	mlp neuron	true	-0.254	-0.757	-0.062^{33}	-0.149^{38}	$+0.415^{37}$	$+0.325^{56}$	$+0.958^{55}$	$+0.511^{95}$	$+1.153^{88}$	$+0.569^{129}$	×2
`L22.h23.svd2`	ov svd	wrong	-0.092	-0.132	-0.049^{29}	-0.092^{37}	-0.018^{12}	-0.017^{10}	$+0.010^{8}$	$+0.018^{11}$	$+0.032^{25}$	$+0.053^{36}$	×3
`L22.n8735`	mlp neuron	false	-0.123	-0.083	-0.045^{16}	-0.096^{23}	$+0.018^{5}$	-0.034^{6}	$+0.025^{11}$	$+0.058^{8}$	$+0.008^{5}$	$+0.112^{18}$	×3
`L30.n11434`	mlp neuron	false	-0.111	-0.068	-0.013^{11}	-0.017^{14}	$+0.010^{11}$	$+0.024^{19}$	$+0.030^{24}$	$+0.103^{51}$	$+0.044^{34}$	$+0.151^{68}$	×2
`L22.n1070`	mlp neuron	accurate	-0.024	-0.040	-0.017^{13}	-0.032^{21}	$+0.001$	-0.002	$+0.008^{6}$	$+0.008^{7}$	$+0.013^{11}$	$+0.014^{12}$	×3
`L20.n4786`	mlp neuron	right	-0.013	-0.003	-0.008^{6}	-0.009^{7}	$+0.002$	$+0.005^{5}$	$+0.006^{5}$	$+0.016^{12}$	$+0.005^{4}$	$+0.025^{18}$	×3

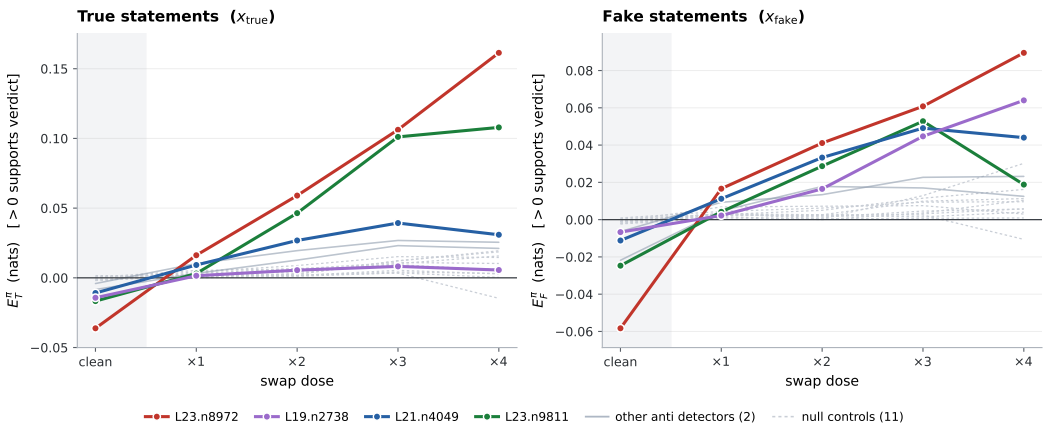

Figure 7: **Interpretable truth detectors that are dormant or opposing in the clean model become a verdict contributor when the core neuron is ablated, in proportion to dose and on both prompt classes.** For each named secondary r we exchange its activation with the value it takes on the contrastive partner *within the ablated regime* (sec_r in Eq. 5), the SCM-faithful counterfactual, and read off E_r^π (Eq. 6), the faithful True−False margin the swap erases in the ablated world π ($E > 0$ *supports* the verdict, $E < 0$ *opposes* it). The horizontal axis sweeps the dose from the clean baseline E^0 (leftmost) through swap scales $k \in \{1, 2, 3, 4\}$, driving the primary neuron `L18.n11065` progressively past its contrastive value. The **left** panel shows true statements, the **right** fake statements. Each detector starts in the dormant/opposing band ($E \leq 0$) and rises into the contributor band ($E > 0$) as the ablation deepens, monotonically for the core detectors, led by the lie-writer `L23.n8972`. A contributor must be positive on *both* panels, and the both-classes gate also exposes the lone one-sided case, `L24.n4543` (*yes/true*), which rises on true statements but reverses on fake statements at high dose. Junk-poled directions (grey dashed) stay pinned at zero. LLAMA-3-8B-INSTRUCT.

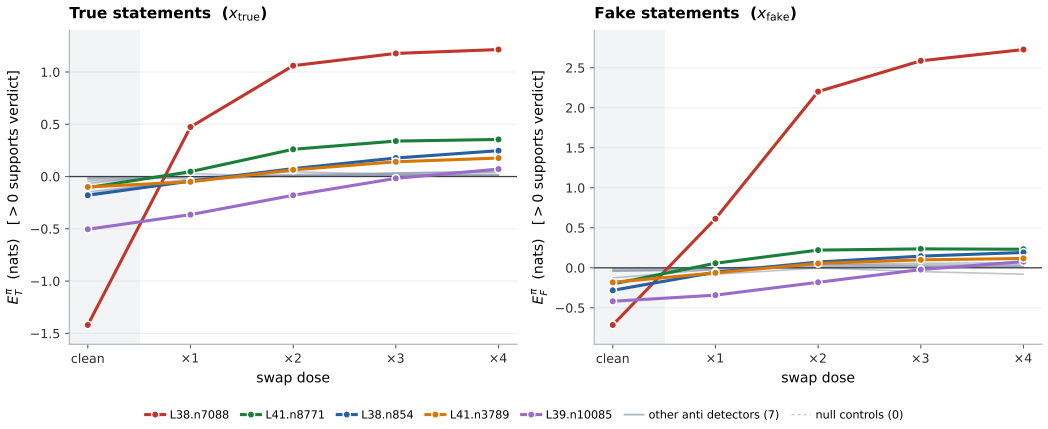

Figure 8: **Dose-resolved recruitment under joint ablation of both Gemma-2-9b-it verdict cores** (`L28.n2046+L33.n4294`). Per-side erased margin E^π vs. swap dose (*clean*$= E^0$, then $\times 1.. \times 4$), with true statements on the left and fake statements on the right. Each coloured curve is one recruited anti-detector, and the dotted grey curves are the null controls. As the joint swap strengthens, the opposing detectors are pulled from $E^\pi < 0$ up through zero into positive (verdict-supporting) territory, while the null controls stay flat, so recruitment is specific to truth axes, including opposing ones.

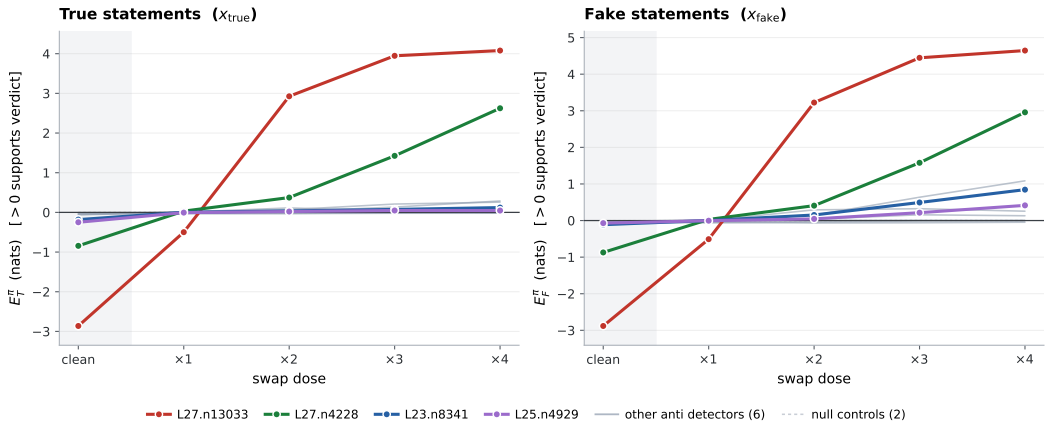

Figure 9: **Dose-resolved recruitment under joint ablation of both Qwen2.5-7B-Instruct verdict cores** (`L22.n13149+L24.n14758`, contrastive swap, `sec_ref=regime`, n=100). Per-detector erased margin E^π on true statements (left) and fake statements (right) as the joint swap dose increases from clean to ×4, where $E^\pi > 0$ means the detector now supports the verdict. The four identified detectors (colour) start as clean opponents ($E^0 < 0$, shaded clean band) and are progressively recruited. `L27.n13033` (*false*) rockets from -2 to $+3.7/+4.2$ nats, flipping at ×2, and `L27.n4228` (suppresses *true*) flips at ×1. The remaining anti detectors (slate) follow the same upward drift at smaller magnitude, while the null controls (dotted) stay pinned at zero across all doses. The joint ablation flips the verdict at ×2 (κ_{primary} crosses 0), so beyond that the panels report reconstitution of the erased fact/fake signal by the backup detectors.

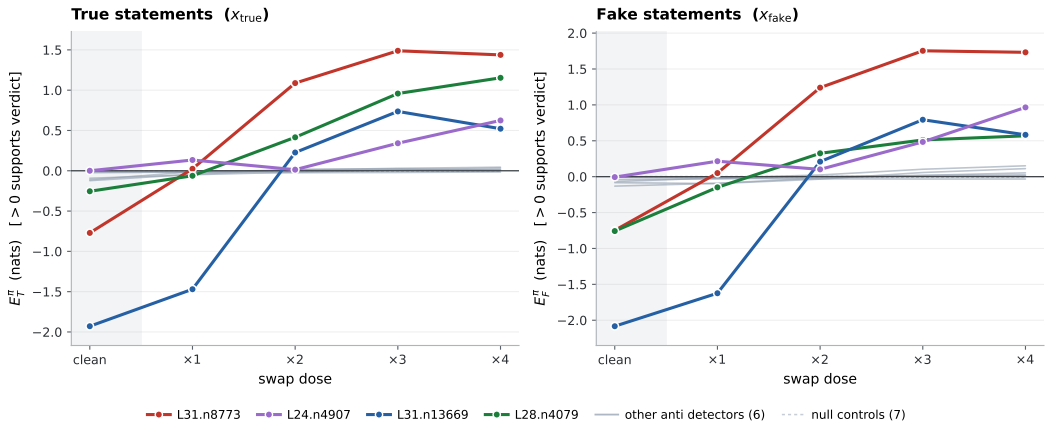

Figure 10: **Dose-resolved recruitment under joint ablation of both Mistral-7B-Instruct-v0.3 verdict cores** (`L20.n14286+L19.n8228`, contrastive swap, `sec_ref=regime`, n=598). Per-side erased margin E^π vs. swap dose (*clean*= E^0, then $\times 1..\times 4$), with true statements on the left and fake statements on the right, following the layout of Figs. 8 to 9. The eight recruited anti-detectors of Table 13 are pulled from $E^\pi < 0$ through zero as the joint swap strengthens. `L31.n8773` and `L31.n13669` show the largest swings (clean $E^0 \approx -0.8$ and -1.9 to -2.1, rising to $E^\pi \approx +1.4$ to 1.8 and $+0.5$ to 0.8 by ×3 to ×4), and `L28.n4079` reaches $+1.15$ on the true side by ×4, while `L22.h23.svd2`, `L22.n8735`, `L22.n1070` and `L20.n4786` recruit later and more weakly, consistent with the `flip@` column of Table 13.

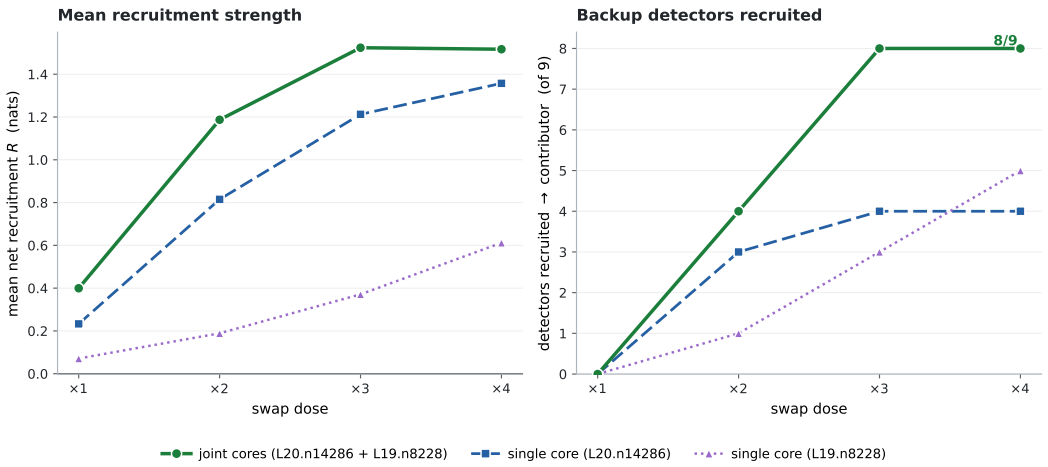

Figure 11: **Removing the verdict as a whole recruits more backup than removing either core separately** (Mistral-7B-Instruct-v0.3, n=598). Over the 9 axes that oppose the verdict in the clean model, the panels show mean net recruitment R (left) and the number crossing into net contribution (right), with the cores `L20.n14286` and `L19.n8228` ablated jointly (solid) or each alone (dashed, dotted). Joint ablation dominates at every dose (R=1.52, 8/9 recruited at $\times$4, vs. 1.36/4 and 0.61/5 for the single cores) and is super-additive at low dose, exceeding the sum of the two single-core effects at $\times$1 and $\times$2 before saturating.

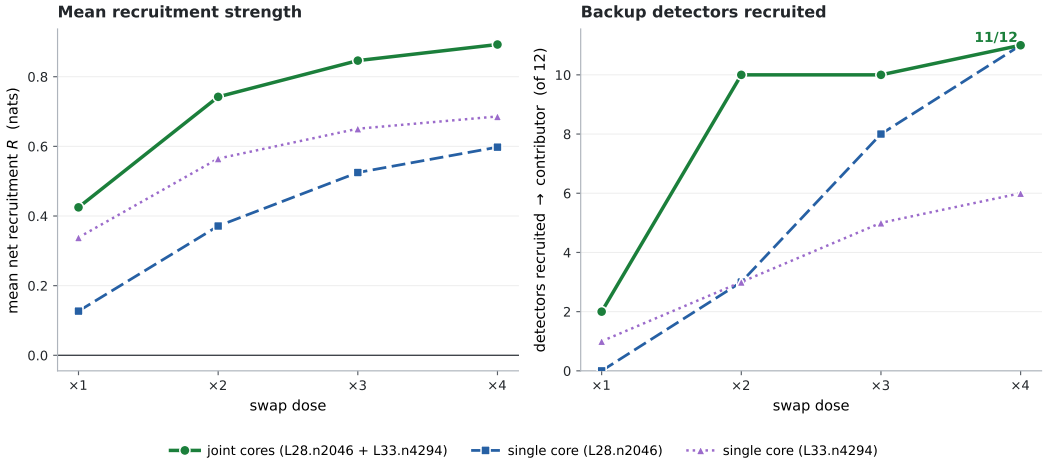

Figure 12: **Single-core vs. joint-core recruitment across dose** (Gemma-2-9b-it, contrastive swap, `sec_ref=regime`, n=100). Recruitment of the 12 clean *anti* detectors when we ablate each verdict core *alone* (`L28.n2046` dashed, `L33.n4294` dotted) versus *jointly* (both cores, solid), as a function of the swap dose $\times$1.. $\times$ 4. The **left** panel shows mean net recruitment $R = R_T + R_F$ averaged over the 12 detectors. The joint ablation dominates either core alone at every dose (0.43/0.74/0.85/0.89 vs. `L33` 0.34/0.57/0.65/0.69 vs. `L28` 0.13/0.37/0.53/0.60). The **right** panel shows the number of the 12 that cross into net contribution ($E^\pi > 0$ on both sides). The joint ablation recruits 10/12 by $\times$2 and 11/12 at $\times$4. `L33` leads at $\times$1 (1 vs. 0), the two single cores tie at $\times$2 (3 each), and `L28` then overtakes, 8 vs. 5 at $\times$3 and 11 vs. 6 at $\times$4, where it matches the joint count. No single core reproduces the joint mean recruitment.

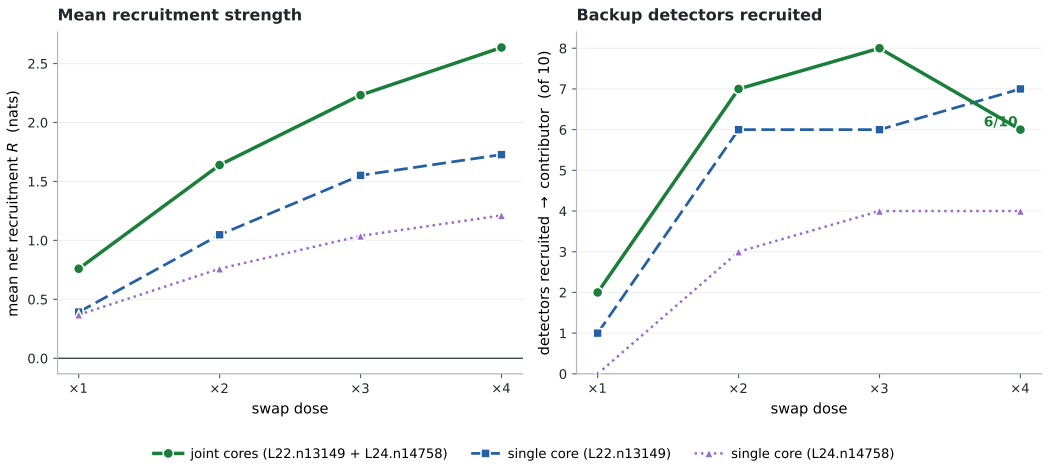

Figure 13: **Single-core vs. joint-core recruitment across dose** (Qwen2.5-7B-Instruct, contrastive swap, `sec_ref=regime`, n=100). Recruitment of the 10 clean *anti* detectors when we ablate the dominant truth core `L22.n13149` alone (dashed), the correctness co-core `L24.n14758` alone (dotted), or both *jointly* (solid), vs. swap dose $\times 1..\times 4$. The **left** panel shows mean net recruitment $R = R_T + R_F$ over the anti detectors. The joint ablation dominates either core alone at every dose, but `L22.n13149` alone, which carries $\sim 35\%$ of the clean fact/fake margin, recovers most of it, while the correctness co-core `L24.n14758` recruits far less. The **right** panel shows the number crossing into net contribution ($E^\pi > 0$ both sides). The joint ablation peaks at $8/10$ at $\times 3$ and falls to $6/10$ at $\times 4$, while `L22.n13149` alone reaches $7/10$ and `L24.n14758` alone $4/10$. Unlike Gemma-2-9b, whose two co-equal cores each recruit only partially, Qwen's backup recruitment is carried predominantly by the single dominant core.

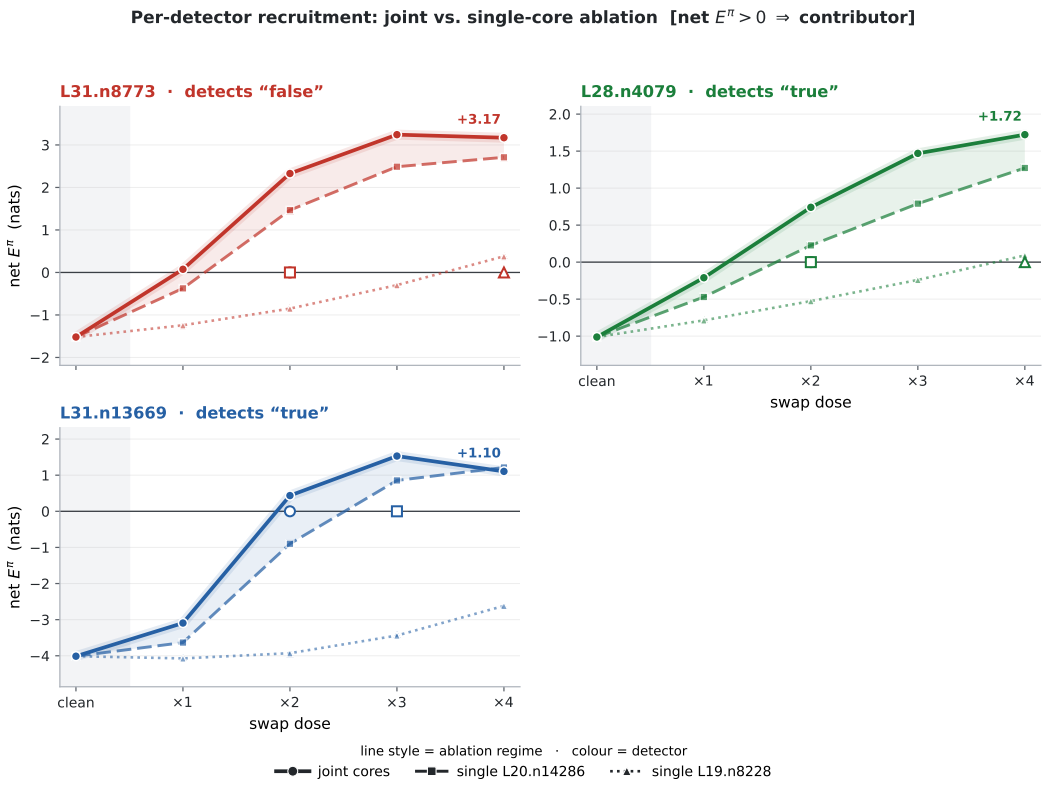

Figure 14: **Backup recruitment is driven by the cores jointly rather than by either core alone** (Mistral-7B-Instruct-v0.3, contrastive swap, `sec_ref=regime`, perfect-on-pairs partition, n=598). One panel per recruited detector, showing its net erased margin $E^{\pi} = E^{\pi}_{T} + E^{\pi}_{F}$ against swap dose when the two verdict cores `L20.n14286` and `L19.n8228` are ablated *jointly* (solid) or each *alone* (dashed, dotted). $E^{\pi} > 0$ means the detector now supports the verdict, and the shaded band is the gain from joint ablation beyond whichever single core does better. Every panel is recruited by the joint ablation at $\times 2$. `L31.n8773` moves $-1.52 \rightarrow +3.17$, `L28.n4079` $-1.01 \rightarrow +1.72$, `L31.n13669` $-4.01 \rightarrow +1.10$ (clean $\rightarrow$ joint at $\times 4$). Single-core ablation is strictly weaker and asymmetric. `L20.n14286` alone reaches $+2.71$, $+1.27$ and $+1.21$ and recruits 4 of the 9 clean-anti axes in the bench, `L19.n8228` alone reaches only $+0.38$, $+0.09$ and -2.62 (it never recruits `L31.n13669` at any dose), and both together as a joint ablation recruit 8 of 9.

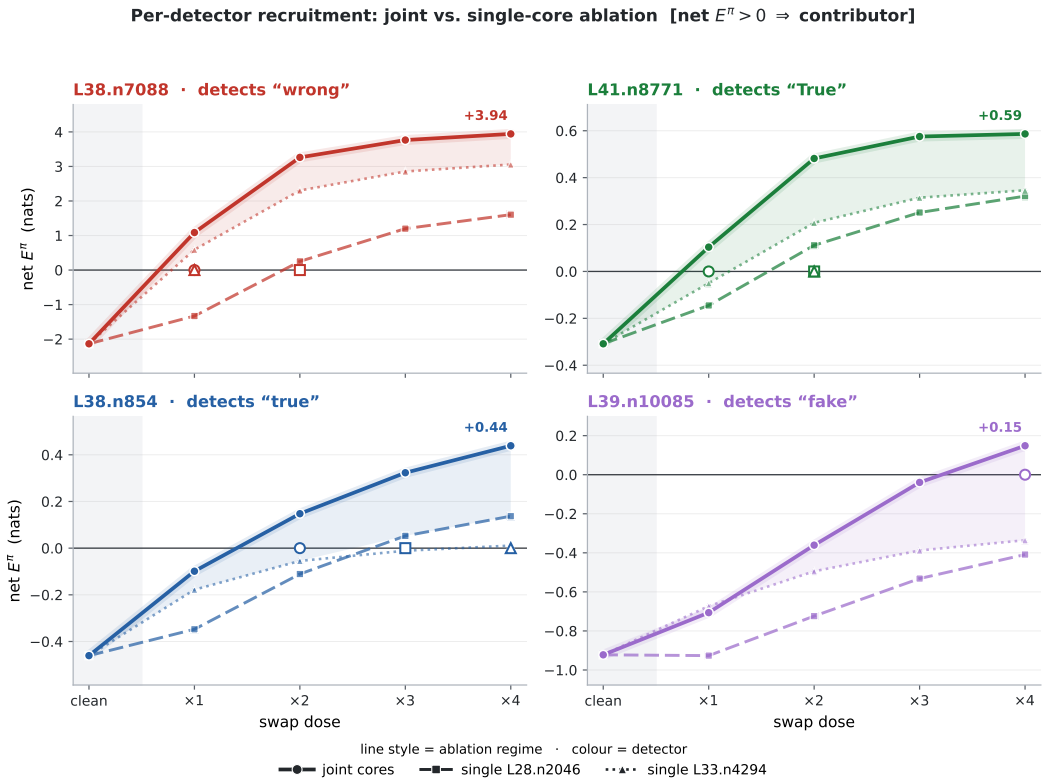

Figure 15: **Backup detectors are recruited into verdict contributors, and more strongly under joint than single-core ablation (Gemma-2-9b-it).** Each panel follows one of the four identified secondary detectors, and the net erased margin $E^\pi = E^\pi_T + E^\pi_F$ is plotted at the clean state (E^0) and at contrastive-swap doses $\times 1$ to $\times 4$, under *joint* ablation of both verdict cores (`L28.n2046+L33.n4294`, solid) versus each core alone (`L28.n2046` dashed, `L33.n4294` dotted). All four are *anti* directions at rest ($E^0 < 0$, with values $-2.14, -0.31, -0.46, -0.92$ for `L38.n7088`, `L41.n8771`, `L38.n854`, `L39.n10085`) yet cross zero under ablation. Joint ablation flips `L38.n7088` and `L41.n8771` already at $\times 1$, `L38.n854` at $\times 2$, and even the initially strongly opposing `L39.n10085` by $\times 4$. n=583 perfect-on-pairs prompts.

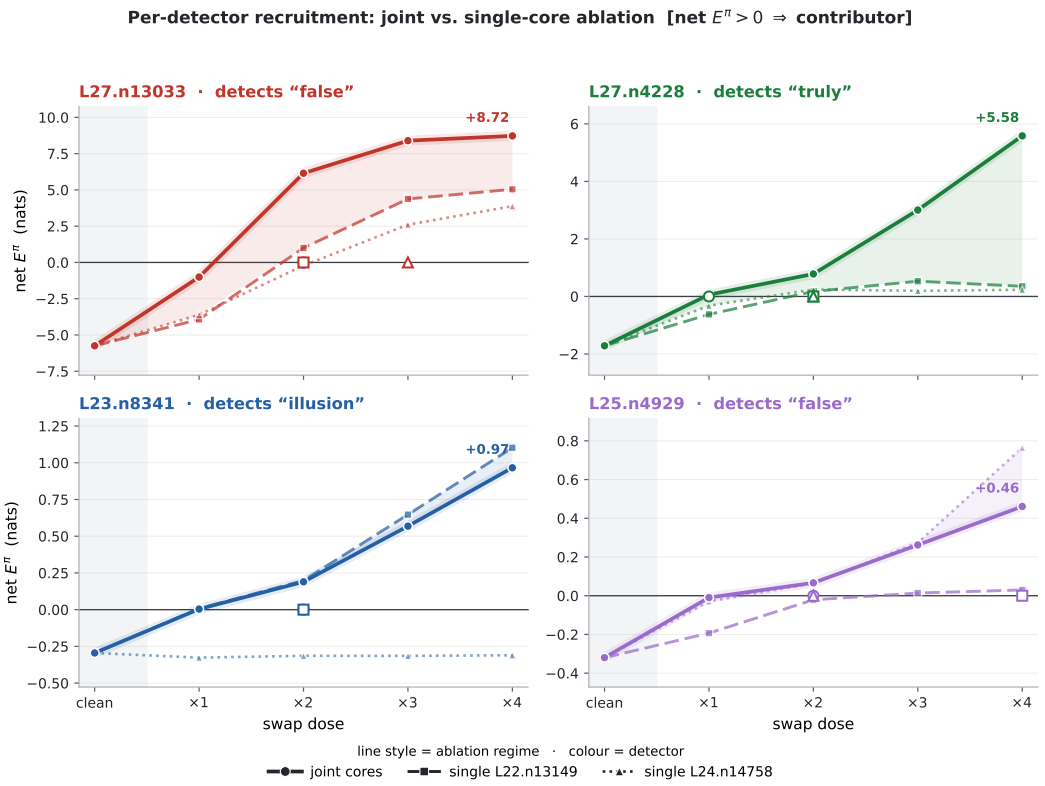

Figure 16: **Backup detectors are recruited into verdict contributors, and more strongly under joint than single-core ablation (Qwen2.5-7B-Instruct).** Each panel follows one of the four identified secondary detectors and plots the net erased margin $E^{\pi} = E^{\pi}_T + E^{\pi}_F$ at the clean state (E^0) and at contrastive-swap doses $\times 1$ to $\times 4$, under *joint* ablation of both verdict cores (`L22.n13149`+`L24.n14758`, solid) versus each core alone (truth core `L22.n13149` dashed, correctness co-core `L24.n14758` dotted). All four are *anti* at rest ($E^0 < 0$, with values $-5.74, -1.71, -0.29, -0.32$ for `L27.n13033`, `L27.n4228`, `L23.n8341`, `L25.n4929`). Joint ablation flips `L27.n4228` at $\times 1$ and the remaining three by $\times 2$. Several directions are recruited by only one of the two cores unless both are removed together. n=478 perfect-on-pairs prompts.

G THE LEARNED PERTURBATION AS AN INDEPENDENT PRIMARY

G.1 OBJECTIVE AND TRAINING

Overview. We localise the verdict mechanism by learning a sparse perturbation δ of the receptor coefficients α_{ck} (§2) that, applied with opposite signs, simultaneously flips and preserves the model's binary verdict across a contrastive dataset. We describe the construction for a generic binary classification. In the main text the classes are True/False, and the same design instantiates other contrastive verdicts (e.g. harmful/harmless).

CONTRASTIVE UNITS (BUNDLES)

The dataset is structured into distinct groups called **bundles**. Each bundle uses a 2×2 factorial design to test how the model responds to two variables, the underlying class of the test query (e.g., true vs. fake) and the arbitrary labels assigned to those classes in the prompt.

Every bundle starts with a matched pair of queries, (q^+, q^-), which are identical except for their class. Instead of using standard labels like "True" and "False," each bundle randomly draws a pair of neutral tokens (x, y), such as random letters or digits with similar baseline probabilities, to serve as the target answers.

The bundle tests all four possible combinations of these queries and label assignments, listed in the table below.

Label Assignment Rule (π)	**Class** $+$ **Query** (q^+)	**Class** $-$ **Query** (q^-)
xy (Class $+$ maps to x)	Prompt: $\mathrm{ICL}^i_{xy}(q^+)$ Expected Output: x	Prompt: $\mathrm{ICL}^i_{xy}(q^-)$ Expected Output: y
yx (Class $+$ maps to y)	Prompt: $\mathrm{ICL}^i_{yx}(q^+)$ Expected Output: y	Prompt: $\mathrm{ICL}^i_{yx}(q^-)$ Expected Output: x

Structuring the data this way guarantees two experimental properties.

- **Label Invariance.** Because the answer tokens (x, y) change from bundle to bundle, the learned perturbation (δ) cannot succeed by memorizing a fixed target word.
- **Factor Decoupling.** Within a single bundle, every class yields both possible answer tokens depending on the context. This breaks any statistical correlation between the query's class, the assignment rule, and the final output token.

Consequently, a successful intervention must act on the combination of the query's actual class and the mapping rules established in the prompt. This forces the model into a "semantically-unrelated-label regime," meaning it must actively read and apply the rules demonstrated in the context rather than relying on its pre-trained biases for standard labels.

Perturbation and asymmetric magnitudes. During the forward pass for a given prompt p, the intervention targets the receptor coefficients α_{ck}, which correspond to the singular directions of the weight matrices as established by Ahmad et al. (2025). Specifically, these coefficients are perturbed according to the mapping $\alpha_{ck} \mapsto \alpha_{ck} + \eta(p, s)\delta_{ck}$. The scalar multiplier $\eta(p, s)$ is determined exclusively by the steering direction s and the prompt's underlying class $q(p) \in \{+, -\}$.

The steering direction s dictates the sign of the perturbation, whereas its magnitude is drawn from two learnable, class-specific parameters, η_+ and η_- (where $\eta_+, \eta_- > 0$). These magnitudes are intentionally kept independent to account for the model's asymmetric commitment, since a network may not require a symmetric application of force to be pushed in opposing directions.

To structure the optimization, cells where the steering direction s aligns with the ground truth are designated as the *preserve* group, while the remaining cells constitute the *flip* group. Within any single class, the preserve and flip cells share the identical magnitude parameter $\eta_{q(p)}$. Coupling the magnitudes in this way prevents the preserve magnitude from trivially collapsing to zero, which would otherwise happen given that the unperturbed model already predicts the correct answer on those instances.

Label-Space Scoring and Objective To evaluate the effect of the perturbation, we compute the probability of a given label string ℓ. Letting $\tau(\ell)$ denote the set of tokenizer-variant IDs for ℓ, this probability is the sum of the softmax probabilities over the perturbed logits $\widetilde{L}(p)$, giving $P(\ell \mid p) = \sum_{v \in \tau(\ell)} \mathrm{softmax}(\widetilde{L}(p))_v$.

We then isolate the binary classification probability over the two available label strings, ℓ_a and ℓ_b, calculating $\widetilde{P}(t^\star \mid p, s) = \frac{P(t^\star|p)}{P(\ell_a|p)+P(\ell_b|p)}$. The target $t^\star$ represents the ground-truth label for cells in the *preserve* group, and the complementary (flipped) label for cells in the *flip* group.

The loss for an individual cell is the negative log-likelihood, $-\log \widetilde{P}(t^\star \mid p, s)$. Because each bundle d_i generates four preserve and four flip prompts, the averaged loss terms across the N bundles are defined as

$$\mathcal{L}_{\text{flip}} = \frac{1}{N} \sum_i \frac{1}{4} \sum_{p \in d_i} -\log \widetilde{P}(t^\star \mid p, \text{flip})$$

$$\mathcal{L}_{\text{pres}} = \frac{1}{N} \sum_i \frac{1}{4} \sum_{p \in d_i} -\log \widetilde{P}(t^\star \mid p, \text{pres})$$

.

These two terms fulfill distinct theoretical roles.

- $\mathcal{L}_{\text{flip}}$ certifies *sufficiency*, ensuring the intervention is strong enough to invert the model's verdict.
- $\mathcal{L}_{\text{pres}}$ certifies *necessity*, ensuring that applying the steering vector with the opposite sign correctly maintains the ground-truth verdict.

Both terms will vanish only if the perturbation δ functions as a directed control.

To enforce sparsity and localize the intervention, the objective incorporates a Hard-Concrete ℓ_0 relaxation (Louizos et al., 2018). This approach gates each coordinate using a stochastic variable on $[0, 1]$ with endpoint mass, yielding $\mathcal{L}_0(\delta)$, a differentiable expectation of the number of open gates.

With the language model's weights completely frozen, the perturbation δ and the magnitudes $\eta = (\eta_+, \eta_-)$ are optimized end-to-end using AdamW to minimize the total objective

$$\mathcal{L}(\delta, \eta) = \mathcal{L}_{\text{flip}} + \mathcal{L}_{\text{pres}} + \lambda_0 \mathcal{L}_0(\delta)$$

.

Here, the hyperparameter λ_0 governs the trade-off between behavioral fidelity (achieving the desired flip/preserve outcomes) and localization (maintaining a sparse perturbation). It follows the fixed schedule given under the training recipe below.

Training recipe. For every layer we take the SVD of each attention head's OV map and of the MLP output map and keep the directions with $\sigma_k > 10^{-5}\sigma_1$ as receptors. The perturbation $\alpha = c\,(z \odot v)$ is added to every attention and MLP output at the final token only. v is initialized at 0, and c is $\pm s_{\text{fl}}$ or $\pm s_{\text{lf}}$ by verdict side, each a learned softplus scale initialized at 1. The gate z is Hard-Concrete with $\log \alpha_0 = 2.5$, fixed temperature $2/3$, $\gamma = -0.1$ and $\zeta = 1.1$, sampled during training and deterministic at evaluation. Training uses 12-shot in-context true/false judgements on known facts, with random single-letter labels from a pool of 13 and no chat template, restricted to bundles the model answers correctly on all four quad prompts and split 70/10/20 with seed 42. The objective sums eight restricted two-way negative log-likelihoods per bundle (four prompts, each preserved and flipped), averages them over a batch of four bundles, and adds $\lambda(t)\,\mathbb{E}\|z\|_0$. The weight λ is 0 in epoch 1 and then ramps linearly from 0 to 9×10^{-4} over epochs 2 to 20, to 10^{-2} over epochs 21 to 25 and to 1 over epochs 26 to 35, resetting to 0 at each phase. We optimize with AdamW at learning rate 5×10^{-3} for all intervention parameters, with no weight decay and a gradient-norm clip of 1.0, for 35 epochs with seed 0 and one run per model. The base model is frozen.

Deployment. We use the epoch-31 checkpoint, where λ is between 0.5 and 0.6, and apply $\delta = \pm 2\, s_{\text{side}}(z \odot v)$, twice the trained magnitude, a scale at which the perturbation flips the verdict in all four models. Evaluation is zero-shot on the True/False template over all statements ($n = 532$,

583, 478 and 598), which includes statements whose in-context bundles were in the perturbation's training split.

G.2 Results

Tables 14 and 15 report the verdict flips and the recruited detectors under the learned perturbation.

Table 14: **The trained receptor-space perturbation as the primary intervention ($\delta = \text{scale} \cdot \eta_{\text{learned}} \cdot v$, scale$= 2$, no dose sweep, epoch 31).** Per-class verdict $D_{TF} = \text{mean}(\text{logit}[\text{True}]) - \text{mean}(\text{logit}[\text{False}])$ at the answer token, on `perfect_on_pairs`. A *strict* verdict flip requires BOTH sides to cross zero, so the true statement must read False (suppress, $D_{TF} < 0$) *and* the fake statement must read True (fabricate, $D_{TF} > 0$). Only Qwen2.5-7B inverts both sides, Llama-3-8B and Gemma-2-9B flip the fabricate side only, and Mistral-7B-v0.3 crosses neither side. "recruited" counts interpretable null/anti axes turned into positive contributors (Table 15).

model	true D_{TF} clean→sup	fake D_{TF} clean→fab	strict flip?	interp. backups recruited
`Llama-3-8B`	$+3.60 \rightarrow +2.40$	$-4.17 \rightarrow +0.08$	×	10
`Gemma-2-9B`	$+4.28 \rightarrow +0.41$	$-5.22 \rightarrow +1.67$	×	5
`Qwen2.5-7B`	$+6.69 \rightarrow -0.99$	$-12.07 \rightarrow +2.28$	✓	6
`Mistral-7B-v0.3`	$+4.60 \rightarrow +4.23$	$-6.99 \rightarrow -0.65$	×	3

G.3 Three-route convergence

Table 16 gives the per-axis, per-side detail behind the convergence claim of main text §4. It shows which interpretable truth/falsity axes, null or opposing in the clean model, are recruited into two-sided verdict defence under each of the three mechanistically distinct primaries (the dosed swap of §F, the learned perturbation of §G.1, and their conjunction). Restricting to the three-primary comparison (excluding rows only tested under the swap), 15 of 31 axes are recruited on both sides under *all three* primaries, and a further 12 axes are recruited under the swap alone (not separately tested under the other two primaries in this table).

Table 15: **Interpretable backup detectors recruited from opposing/inert to contributing by the trained receptor-space perturbation** (scale= 2, epoch 31, `perfect_on_pairs`, with 10 / 5 / 6 / 3 axes for Llama-3-8B/Gemma-2-9B/Qwen2.5-7B/Mistral-7B-v0.3, ranked by recruitment R). Each axis is *null/anti* in the clean model and turns a positive contributor to D_{TF} under the primary intervention, shown separately on the two sides as E_T (suppress side) and E_F (fabricate side). E^0 is the same quantity with no primary, and $R = E^\pi - E^0$ is the recruitment. Row shade = clean class (anti , null). Superscripts are paired t values, which are inflated because a fixed intervention makes each per-prompt shift near-deterministic, so recruitment is read from the $E^0 \to E^\pi$ sign flip and R rather than from t.

			E^0 (clean)		E^π (perturb)		$R = E^\pi - E^0$	
axis	family	detects	E^0_T	E^0_F	E^π_T	E^π_F	R_T	R_F
Llama-3-8B (clean D_{TF}: T+3.60/F−4.17 → prim T+2.40/F+0.08)								
`L22.h1.svd117`	ov svd	true	−0.004	−0.006	$+0.111^{88}$	$+0.177^{117}$	+0.115	+0.183
`L19.n3754`	mlp neuron	sure	−0.008	−0.022	$+0.018^{16}$	$+0.034^{29}$	+0.027	+0.056
`L22.h1.c60`	ov neuron	valid	+0.000	−0.001	$+0.035^{30}$	$+0.030^{27}$	+0.035	+0.031
`L21.n4049`	mlp neuron	wrong	−0.011	−0.011	$+0.018^{16}$	$+0.018^{16}$	+0.029	+0.029
`L21.n798`	mlp neuron	true	−0.002	−0.002	$+0.021^{18}$	$+0.028^{24}$	+0.023	+0.030
`L28.h27.c120`	ov neuron	True	−0.002	−0.001	$+0.017^{20}$	$+0.030^{33}$	+0.019	+0.031
`L28.h27.c34`	ov neuron	True	+0.000	+0.001	$+0.015^{16}$	$+0.023^{29}$	+0.015	+0.023
`L28.h27.c0`	ov neuron	real	+0.000	−0.001	$+0.009^{10}$	$+0.015^{18}$	+0.009	+0.016
`L20.h11.c95`	ov neuron	false	+0.000	−0.002	$+0.009^{8}$	$+0.006^{5}$	+0.009	+0.008
`L22.h1.svd116`	ov svd	correct	−0.000	+0.001	$+0.006^{6}$	$+0.010^{9}$	+0.006	+0.009
Gemma-2-9B (clean D_{TF}: T+4.28/F−5.22 → prim T+0.41/F+1.67)								
`L38.n7088`	mlp neuron	wrong	−1.420	−0.715	$+0.014^{11}$	$+0.024^{16}$	+1.434	+0.738
`L32.n8200`	mlp neuron	False	−0.067	−0.175	$+0.865^{63}$	$+0.105^{50}$	+0.932	+0.280
`L39.n10210`	mlp neuron	incorrect	−0.167	−0.127	$+0.251^{23}$	$+0.216^{23}$	+0.418	+0.343
`L34.n4784`	mlp neuron	truth	−0.044	−0.037	$+0.027^{20}$	$+0.017^{16}$	+0.072	+0.053
`L40.h11.svd18`	ov svd	truth	−0.027	−0.044	$+0.030^{25}$	$+0.007^{8}$	+0.057	+0.051
Qwen2.5-7B (clean D_{TF}: T+6.69/F−12.07 → prim T−0.99/F+2.28)								
`L27.n13033`	mlp neuron	false	−2.863	−2.881	$+0.249^{39}$	$+0.265^{39}$	+3.112	+3.147
`L22.n4538`	mlp neuron	false	−0.035	−0.026	$+0.201^{75}$	$+0.389^{74}$	+0.236	+0.415
`L23.n8341`	mlp neuron	illusion	−0.182	−0.113	$+0.042^{39}$	$+0.046^{27}$	+0.224	+0.159
`L25.n4929`	mlp neuron	false	−0.249	−0.071	$+0.009^{7}$	$+0.012^{6}$	+0.258	+0.083
`L25.n8444`	mlp neuron	假 (fake)	−0.063	−0.031	$+0.027^{14}$	$+0.040^{14}$	+0.091	+0.071
`L22.n13025`	mlp neuron	真正 (genuine)	−0.017	−0.045	$+0.040^{36}$	$+0.044^{23}$	+0.056	+0.089
Mistral-7B-v0.3 (clean D_{TF}: T+4.60/F−6.99 → prim T+4.23/F−0.65)								
`L31.n13669`	mlp neuron	true	−1.929	−2.083	$+1.045^{128}$	$+1.238^{126}$	+2.974	+3.322
`L20.n4786`	mlp neuron	right	−0.013	−0.003	$+0.008^{8}$	$+0.011^{10}$	+0.021	+0.014
`L16.n9991`	mlp neuron	false	+0.002	+0.001	$+0.005^{5}$	$+0.006^{5}$	+0.004	+0.005

Table 16: **Recruitment is consistent across three distinct interventions.** Each row is an interpretable truth/falsity axis that is null/anti in the clean model ($E^0 \leq 0$). The verdict sides are scored separately, **T** (true/suppress, E_T) and **F** (false/fabricate, E_F), and ✓ marks that side becoming a contributor ($E^\pi > 0$, $|t| \geq 2$). **Swap** is an in-distribution contrastive fact↔fake exchange, dose-swept $\times 1 - \times 4$ (✓ = rose monotonically and reliable by $\times 4$). **Perturb** is an out-of-distribution trained steering vector applied at every layer (scale 2), and **Swap+Perturb** stacks both in one pass. The clean class is anti ($E^0 < 0$) or null ($E^0 \approx 0$), and ★ marks recruitment on both sides under all three.

		clean	Swap		Perturb		Swap+Perturb		
axis	detects	E^0	T	F	T	F	T	F	all
Llama-3-8B									
L19.n3754	sure	anti	✓	✓	✓	✓	✓	✓	★
L21.n4049	wrong	anti	✓	✓	✓	✓	✓	✓	★
L22.h1.c60	valid	null	✓	✓	✓	✓	✓	✓	★
L22.h1.svd116	correct	null	✓	✓	✓	✓	✓	✓	★
L22.h1.svd117	true	anti	✓	✓	✓	✓	✓	✓	★
L28.h27.c0	real	null	✓	✓	✓	✓	✓	✓	★
L28.h27.c34	True	null	✓	✓	✓	✓	✓	✓	★
L20.h11.c95	false	null	–	✓	✓	✓	✓	✓	
L28.h27.c120	True	null	✓	–	✓	✓	✓	✓	
L19.n2738	illusion	anti	✓	✓	–	–	✓	✓	
L21.n798	true	null	–	–	✓	✓	✓	✓	
L21.n14165	Genuine	null	✓	✓	–	–	–	–	
L23.n8972	lie	anti	✓	✓	–	–	–	–	
L23.n9811	truth	anti	✓	✓	–	–	–	–	
Gemma-2-9B									
L34.n4784	truth	anti	✓	✓	✓	✓	✓	✓	★
L38.n7088	wrong	anti	✓	✓	✓	✓	✓	✓	★
L39.n10210	incorrect	anti	✓	✓	✓	✓	✓	✓	★
L40.h11.svd18	truth	anti	✓	✓	✓	✓	✓	✓	★
L32.n8200	False	anti	✓	–	✓	✓	✓	✓	
L38.n854	true	anti	✓	✓	–	–	✓	✓	
L40.h7.c177	genuine	anti	✓	✓	–	–	✓	✓	
L41.n3789	real	anti	✓	✓	–	–	✓	✓	
L41.n8771	True	anti	✓	✓	–	–	✓	✓	
L33.n1466	deception	anti	✓	✓	–	–	–	–	
L39.n10085	fake	anti	✓	✓	–	–	–	–	
L40.h7.c105	real	anti	✓	✓	–	–	–	–	
Qwen2.5-7B									
L23.n8341	illusion	anti	✓	✓	✓	✓	✓	✓	★
L27.n13033	false	anti	✓	✓	✓	✓	✓	✓	★
L22.n13025	真正 (genuine)	anti	–	–	✓	✓	✓	✓	
L22.n4538	false	anti	–	–	✓	✓	✓	✓	
L25.n4929	false	anti	✓	✓	✓	✓	–	–	
L25.n8444	假 (fake)	anti	✓	✓	✓	✓	–	–	
L23.n14476	真正的 (genuine)	anti	✓	✓	–	–	–	–	
L27.n4228	truly	anti	✓	✓	–	–	–	–	
Mistral-7B-v0.3									
L20.n4786	right	anti	✓	✓	✓	✓	✓	✓	★
L31.n13669	true	anti	✓	✓	✓	✓	✓	✓	★
L22.n1070	accurate	anti	✓	✓	–	✓	✓	✓	
L16.n9991	false	null	–	–	✓	✓	✓	✓	
L22.h23.svd2	wrong	anti	✓	✓	–	–	✓	✓	
L22.n8735	false	anti	✓	✓	–	–	–	–	
L28.n4079	true	anti	✓	✓	–	–	–	–	
L30.n11434	false	anti	✓	✓	–	–	–	–	
L31.n8773	false	anti	✓	✓	–	–	–	–	

H PRETRAINING-NATIVE RECRUITMENT IN BASE VERSUS INSTRUCT

The identical dose-graded crossing of the bench is present in the base (pretrained) checkpoint for each family, on the *same* pairs each checkpoint answers correctly on both sides zero-shot, so prompt, metric and items are identical and the only difference is post-training.

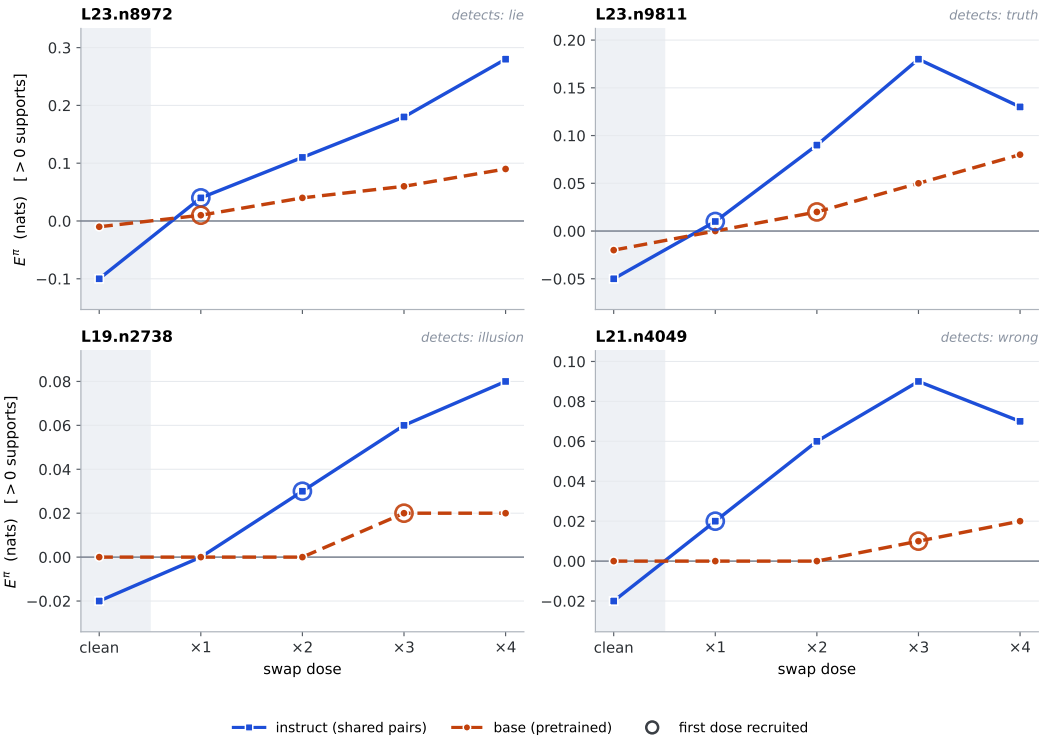

Figure 17: **The same flagship backups are recruited in the pretrained checkpoint** (Llama-3-8B base vs. Llama-3-8B-Instruct, contrastive swap of the single verdict core `L18.n11065`, `sec_ref=regime`, n=240). Each panel shows one flagship detector identified as a recruit in the instruct model, plotting net erased margin $E^\pi = E^\pi_T + E^\pi_F$ against swap dose for instruct (solid) and base (dashed). The ring marks the first dose at which that checkpoint's mean margin turns positive. **Every one of the four is recruited in the base checkpoint as well as in instruct**, monotonically in dose and with the same ordering. `L23.n8972` ('lie') reaches $+0.09$ in base against $+0.28$ in instruct, `L23.n9811` ('truth') $+0.08$ against $+0.13$, and the two weaker axes `L19.n2738` and `L21.n4049` $+0.02$ against $+0.08$ and $+0.07$. The paired t reaches 44 (instruct) and 24 (base) on `L23.n8972`, while base's whole verdict is smaller to begin with ($\kappa_{\text{base}} = -2.0$ against -8.4). The clean-model classification does differ for `L19.n2738` and `L21.n4049`, which oppose the verdict outright in instruct but are only weakly polarised in base, so instruction tuning sharpened them into opponents without changing whether the ablation recruits them.

Together, the four figures make backup recruitment a property of the pretrained network that instruction tuning inherits (and, in absolute terms, amplifies for Qwen and Mistral) rather than one instruction tuning creates (Figures 17 to 20).

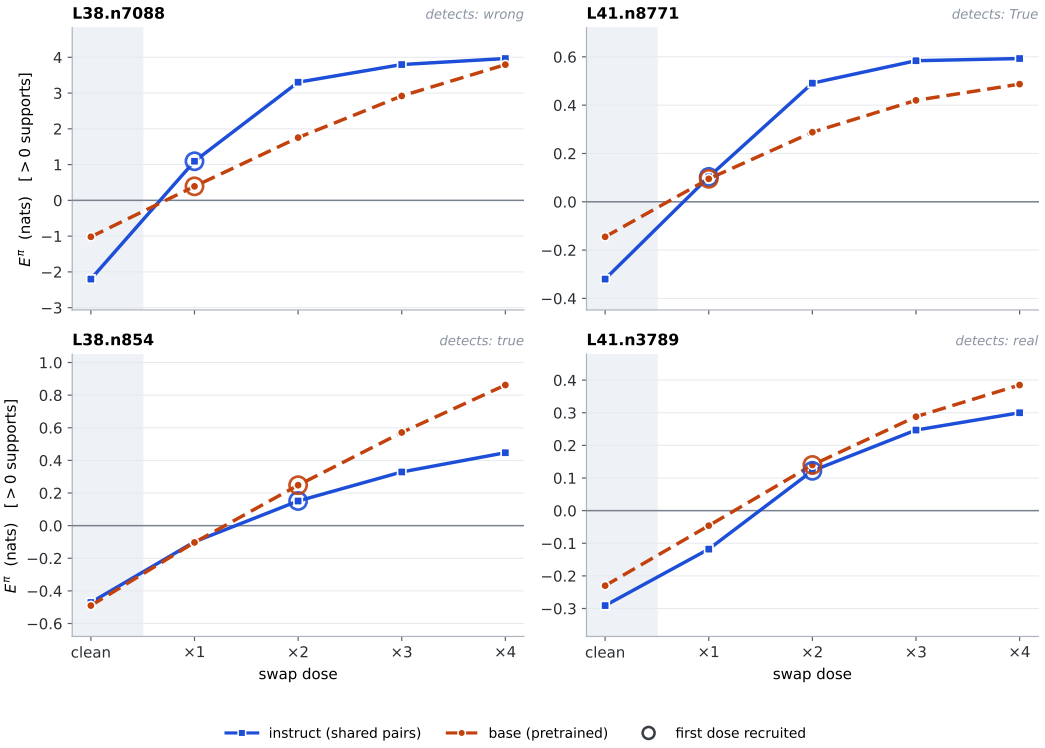

Figure 18: **Backup recruitment is already present in the pretrained checkpoint, in the same components** (Gemma-2-9B base vs. Gemma-2-9b-it, joint contrastive swap of the verdict cores `L28.n2046+L33.n4294`, `sec_ref=regime`, n=535). Each panel shows one flagship recruited detector, plotting net erased margin $E^\pi = E^\pi_T + E^\pi_F$ against swap dose for instruct (solid) and base (dashed). All four oppose the verdict in *both* clean models and are recruited in *both*, at the same dose in three of four panels. `L38.n7088` ('wrong') moves $-1.02 \rightarrow +3.79$ in base against $-2.20 \rightarrow +3.97$ in instruct (the curves nearly coincide at $\times 4$), and on `L38.n854` and `L41.n3789` the base curve finishes *above* instruct. The paired t reaches 120 in base and 262 in instruct. Across the full bench, **all** 11 **detectors that oppose the verdict in both clean models are recruited in both checkpoints** by $\times 4$. Base magnitudes are comparable despite its verdict being far smaller ($\kappa_{\text{base}} = -2.83$ against -9.81). Normalised by the margin the ablation destroys, the base backups restore 11 to 13% of the erased verdict against instruct's 6 to 9%.

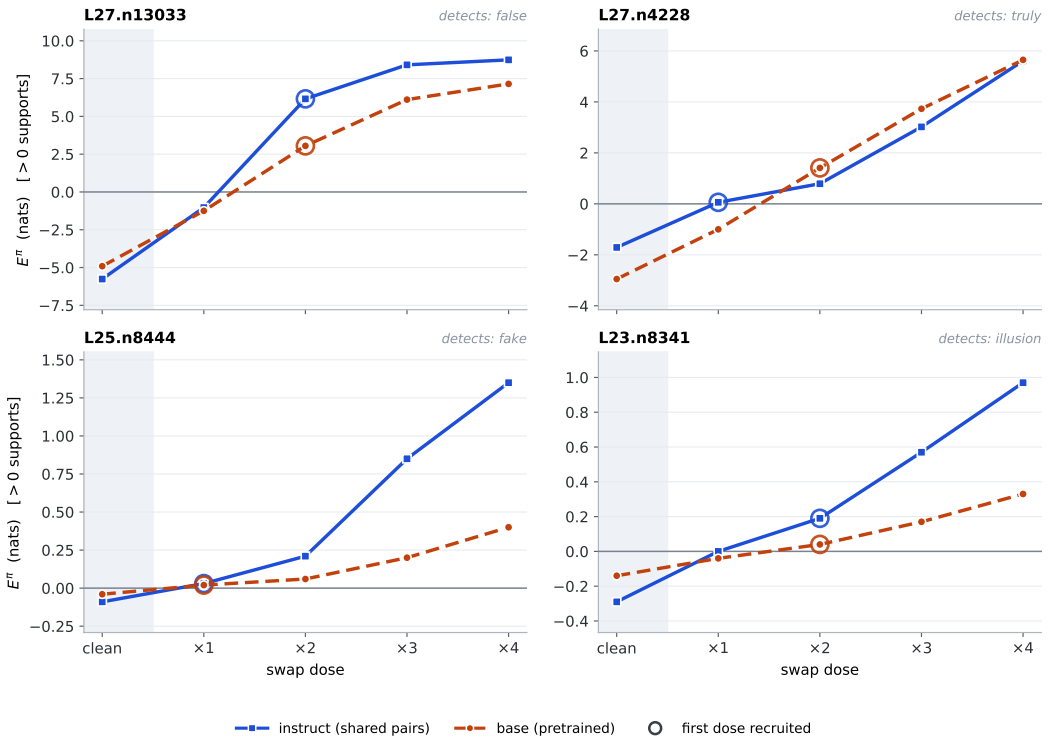

Figure 19: **Backup recruitment is already present in the pretrained checkpoint, in the same components** (Qwen2.5-7B base vs. Qwen2.5-7B-Instruct, joint contrastive swap of the verdict cores `L22.n13149+L24.n14758`, `sec_ref=regime`, n=470). One panel per flagship recruited detector, showing its net erased margin $E^{\pi} = E^{\pi}_{T} + E^{\pi}_{F}$ against swap dose for the instruct (solid) and base (dashed) checkpoint. Every detector opposes the verdict in *both* clean models and is recruited in *both* under ablation, at the same dose in three of four panels. `L27.n13033` moves $-4.91 \rightarrow +7.15$ in base against $-5.76 \rightarrow +8.74$ in instruct, and `L27.n4228` $-2.95 \rightarrow +5.65$ against $-1.71 \rightarrow +5.61$. In that panel the base curve finishes *above* instruct. The recruitment ordering is identical in the two checkpoints (`L27.n13033` > `L27.n4228` > `L25.n8444` > `L23.n8341`), including the multilingual `L25.n8444` (top suppressed token 假(fake)). Absolute magnitudes are smaller in base only because its whole verdict is smaller ($\kappa_{\text{base}} = -6.16$ against -18.87). Normalised by the margin the ablation destroys, the base backups restore 22 to 27% of the erased verdict against instruct's 9 to 10%.

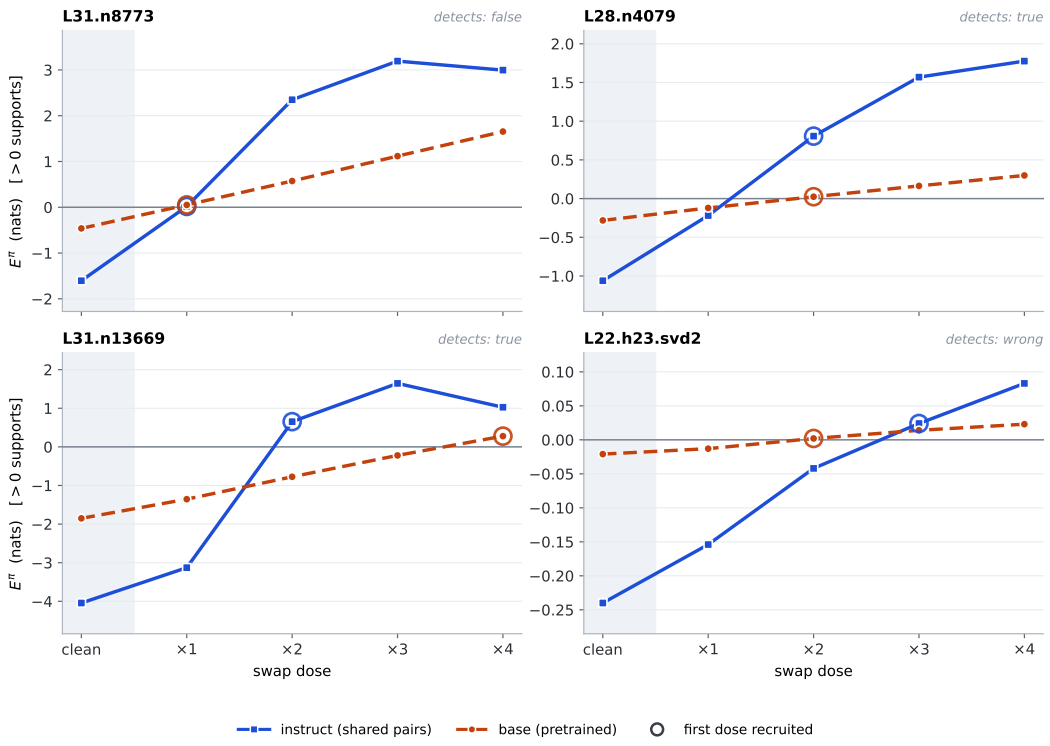

Figure 20: **Backup recruitment is already present in the pretrained checkpoint, in the same components** (Mistral-7B-v0.3 base vs. Mistral-7B-Instruct-v0.3, joint contrastive swap of the verdict cores `L20.n14286+L19.n8228`, `sec_ref=regime`, n=423). All four detectors oppose the verdict in *both* clean models and are recruited in *both*. `L31.n8773` reaches $+1.65$ in base against $+3.00$ in instruct, `L28.n4079` $+0.30$ against $+1.78$, `L31.n13669` $+0.28$ against $+1.03$, and the attention-head axis `L22.h23.svd2` $+0.02$ against $+0.08$, which extends the same recruitment one family beyond MLP neurons. Base magnitudes are smaller only because its whole verdict is smaller ($\kappa_{\text{base}} = -1.80$ against -12.40, a $7\times$ scale gap). Normalised by the margin the ablation destroys, the base backups restore a near-constant 24 to 25% of the erased verdict at every dose, against instruct's 8 to 11%. By $\times 4$ all six shared clean-anti detectors are recruited in both checkpoints.

I RECRUITMENT IS NOT RENORMALIZATION

I.1 THE OBJECTION

Recruitment compares a secondary's contribution across two causal worlds. The score $R_r = E_r^{\pi} - E_r^{0}$ (main text Eq. 7) reads the same direction once in the clean model and once in the dosed one. The two worlds differ in one respect that affects every direction at once. A component's write reaches the logits through the final normalization, which the readout of main text §2 folds into W_U as a per-input gain proportional to $1/\rho_b(x)$, where

$$\rho_b(x) \;=\; \mathrm{RMS}\big(h^L(x;b)\big) \;=\; \Big(\tfrac{1}{d}\textstyle\sum_i h_i^L(x;b)^2 + \varepsilon\Big)^{1/2} \tag{9}$$

is the RMS of the answer-position stream in world b.[2] If the primary shrinks this denominator, as one would expect of an intervention that removes a large write, then every surviving direction's path to the logits is multiplied by

$$g_k \;=\; \rho_{\varnothing} \,/\, \rho_{\pi_k}, \tag{10}$$

and if g_k grows with dose, so does every direction's measured effect, with no change in how the model uses any of them. On this reading, §F documents a change of measurement frame rather than a change of computation. The concern has precedent, since a share of the head-level self-repair observed under ablation has been attributed to exactly this mediator (Rushing & Nanda, 2024), and ρ is the only multiplicative factor between a weight-fixed write and the margin on the direct path.

I.2 WHAT THE OBJECTION PREDICTS

In its strict form, the null says that the two worlds differ only in the denominator. Every logit then carries the same factor, and because D_{XY} is a difference of logits (main text Eq. 2), so does the margin, $D_{XY}(x;\pi_k) = g_k D_{XY}(x;\varnothing)$. For any secondary, therefore,

$$E_r^{\pi} = g_k E_r^{0}, \qquad R_r = (g_k - 1)\, E_r^{0}. \tag{11}$$

Three testable consequences follow.

(P1) No direction can change sign. Since $g_k > 0$ for any nonzero stream, E_r^{π} and E_r^{0} share a sign. A dismissed direction ($E_r^{0} \le 0$) may shrink toward zero or become more negative, but it cannot become a two-sided contributor. Certification (§D) requires $E > 0$ on both classes, each reliably, and is therefore unreachable under the null.

(P2) Amplification predicts the wrong sign of R. On the dismissed bench $E_r^{0} \le 0$, so a gain above unity gives $R_r = (g_k - 1)E_r^{0} \le 0$, and amplification *deepens* opposition. Only attenuation ($g_k < 1$) can produce $R_r > 0$ on these directions, and then only by shrinking opposition toward zero, never by crossing it.

(P3) The gain-attributable share of R is bounded by $|1 - g_k|\,|E_r^{0}|$. This bound holds under the mixed account as well. Whatever genuine recruitment occurs, Eq. 11 fixes the most that renormalization alone can contribute, per direction, class, and dose, from measured quantities alone.

I.3 MEASUREMENT

We record ρ at the answer position, in the clean model and inside each π_k, per statement class, on $n = 50$ evaluation pairs per family, using the primaries of §F (the single core `L18.n11065` for Llama, the joint `L28.n2046`+`L33.n4294` for Gemma, and the joint `L22.n13149`+`L24.n14758` for Qwen). Since $\rho(x;\pi_k)$ is a property of the world rather than of any secondary, the bounds below apply to the entire bench, including units not tabulated. Per-condition dispersion is small (s.e.m. $\le 0.9\%$ of the mean throughout), and the subsample is faithful to the full runs. The class-resolved margins of Table 18 cross zero at the same doses as at full n, and the emblematic contributions they encode reproduce the grids of §F.1 to within 0.03 nats.[3]

[2] Absolute ρ is family-idiosyncratic (Gemma-2's late-stream norms are an order of magnitude larger than Qwen's). Every inference below uses within-family ratios, which are invariant to any fixed rescaling of the measurement hook.

[3] Gemma `L38.n7088` at $\times 1$ gives $E = (+0.473, +0.615)$ here against $(+0.474, +0.613)$ at n=583, Qwen `L27.n13033` at $\times 2$ gives $(+2.95, +3.25)$ against $(+2.93, +3.22)$, and Llama `L23.n8972` at $\times 1$ gives $(+0.024, +0.015)$ against $(+0.02, +0.02)$.

Table 17: **The final-norm gain across dose.** $\rho_\varnothing$ is the clean answer-position RMS of the pre-unembedding stream, and $g_k = \rho_\varnothing/\rho_{\pi_k}$ is the factor the final normalization applies to every fixed write's direct path in the dosed world. Values above unity amplify, values below unity attenuate. The artifact account requires $g_k > 1$ *and increasing in* k. In the data, the gain *decreases* with dose on true statements in both joint-core families, is elevated but dose-flat on Gemma's fake side, and at the on-manifold dose $\times 1$ exceeds 1.03 in no other family-class cell.

Model	Class	$\rho_\varnothing$	$g_{\times 1}$	$g_{\times 2}$	$g_{\times 3}$	$g_{\times 4}$	profile in k
Llama-3-8B-Instruct	T	0.990	1.023	1.060	1.105	1.157	increasing
	F	0.930	0.992	1.000	1.010	1.032	flat/slight
Gemma-2-9b-it	T	28.99	0.997	0.957	0.934	0.926	decreasing
	F	31.07	1.080	1.103	1.089	1.086	flat
Qwen2.5-7B-Instruct	T	6.09	0.996	0.978	0.941	0.890	decreasing
	F	5.92	1.024	1.033	1.017	0.985	flat

I.4 Class-resolved margins in the audited worlds

Table 18 reports the class-resolved verdict margins across dose in the audit's subsample. Three features of it matter for what follows.

First, the margins fix the pre-/post-failure partition. At $\times 1$ every family still answers correctly on both classes, and every family-class pair crosses zero between $\times 1$ and $\times 2$, so the on-manifold dose is pre-failure everywhere.

Second, the primary's dose-response is smooth and family-shaped. On true statements, Llama's single-core margin falls by a near-constant 2.1 to 2.7 nats per dose, whereas both joint-core families decelerate sharply at depth. The per-dose drop shrinks from 2.85 to 0.54 nats in Gemma and from 5.26 to 2.17 in Qwen, and the fake sides mirror the pattern after an initial rise. Compensatory recruitment predicts a marginal effect that flattens as the intervention deepens, and Llama, whose reserve is by far the smallest of the three (§H), is the family whose steps do not flatten. The deceleration coincides with the true-side norm growth in Table 17, for the reason given in §I.6. Gemma's final logit soft-capping cannot account for it on its own, since Qwen, which has none, decelerates as strongly.

Third, the verdict's redundancy is class-asymmetric. Qwen's fake-side margin (-11.86 clean) withstands $\times 1$ at -7.16 and barely crosses at $\times 2$ ($+0.38$), while its true side falls five nats in a single dose.

Table 18: **Class-resolved verdict margins across dose** (n=50 pairs per family, same runs as Table 17). D is the class-fixed margin D_{TF} of main text §2, with correct sign $+$ on true statements and $-$ on fakes. D_π is the margin in the dosed world π_k, and $D_{\pi,\mathrm{sec}}$ additionally overrides the emblematic anti-detector with its regime reference (Llama `L23.n8972`, Gemma `L38.n7088`, Qwen `L27.n13033`), so each column pair regenerates E^π via $E_T = D_\pi - D_{\pi,\mathrm{sec}}$ and $E_F = D_{\pi,\mathrm{sec}} - D_\pi$.

			$\times 1$		$\times 2$		$\times 3$		$\times 4$	
Model	Cls	$D_\varnothing$	D_π	$D_{\pi,\mathrm{sec}}$	D_π	$D_{\pi,\mathrm{sec}}$	D_π	$D_{\pi,\mathrm{sec}}$	D_π	$D_{\pi,\mathrm{sec}}$
Llama-3-8B	T	$+3.51$	$+1.44$	$+1.42$	-0.83	-0.89	-3.31	-3.41	-6.00	-6.17
	F	-4.13	-2.08	-2.07	$+0.70$	$+0.74$	$+3.71$	$+3.77$	$+6.60$	$+6.69$
Gemma-2-9b	T	$+4.40$	$+1.55$	$+1.08$	-0.83	-1.90	-1.92	-3.12	-2.46	-3.68
	F	-5.21	-2.80	-2.19	$+1.73$	$+3.95$	$+3.93$	$+6.53$	$+4.90$	$+7.65$
Qwen2.5-7B	T	$+6.39$	$+1.12$	$+1.50$	-3.95	-6.89	-7.51	-11.50	-9.68	-13.82
	F	-11.86	-7.16	-7.54	$+0.38$	$+3.62$	$+7.65$	$+12.14$	$+13.11$	$+17.80$

I.5 THE REFUTATION

The artifact account requires a gain above unity that also strengthens with dose. Table 17 shows that this conjunction is absent in every joint-core cell. The one cell in which the gain does rise with dose, Llama's true side, is addressed under (P2) and (P3) below.

Sign crossings are outside the null (P1). The central phenomenon of §F is that dismissed directions *change sign*, over and above any growth in magnitude. Llama's `L23.n8972` moves from $(-0.04, -0.06)$ to $(+0.17, +0.10)$, Gemma's `L38.n7088` from $(-1.42, -0.72)$ to $(+1.22, +2.73)$, and Qwen's `L27.n13033` from $(-2.86, -2.88)$ to $(+4.08, +4.65)$. No positive scalar produces a sign change.

Where the effects are largest, the gain moves the wrong way (P2). On true statements in both joint-core families the stream norm *grows* with dose, driving g down to 0.926 (Gemma) and 0.890 (Qwen). Every fixed write's direct path is attenuated as recruitment proceeds, so the reported E_T^π *understate* their fixed-gain counterparts, by roughly 11% for Qwen's strongest recruit at $\times 4$. On the fake sides of Gemma and Qwen, and on Llama's true side, the gain is instead mildly amplifying, and there Eq. 11 predicts $R_r \leq 0$ for every dismissed direction, the opposite of what is observed in every such cell.

Where the gain is elevated, it is flat in dose. Gemma's fake side carries the largest amplification in the study, $g_F \in [1.080, 1.103]$, and it is constant across the sweep, while E_F^π for `L38.n7088` rises from $+0.62$ to $+2.73$ nats, a factor of 4.5. A dose-independent multiplier cannot generate a dose-graded interaction. Dividing the $\times 1$ value by the full gain leaves $+0.613/1.080 \approx +0.57$ nats, still the largest pre-failure effect on the bench.

The gain-attributable share of R is at most a few percent (P3). Table 19 applies the bound of Eq. 11 to the three emblematic recruits. In half the cells the predicted contribution has the wrong sign, and where it has the right sign it accounts for at most 4.5% of the observed recruitment. At the on-manifold dose $\times 1$ the largest correction anywhere is 0.6% of R, and in Llama the implied inflation of E^π is about 0.0005 nats against a per-cell standard error of about 0.0016. Applying the correction unit by unit changes no certification, no sign, and no ordering. Llama's dose trajectory, divided through by g, remains strictly monotone ($0.021 \to 0.060 \to 0.104 \to 0.143$).

Table 19: **The maximum recruitment attributable to renormalization.** For each emblematic recruit, E^0 is the clean contribution, $R = E^\pi - E^0$ the observed recruitment, and $R_{\text{gain}} = (g_k - 1)E^0$ the recruitment predicted by the strict gain null of Eq. 11. "Wrong sign" marks cells in which the gain predicts recruitment opposite in direction to what is observed.

Unit (detects)	Cls	E^0	dose	R	R_{gain}	share of R
`L23.n8972` (*lie*)	T	-0.04	$\times 1$	$+0.06$	-0.001	wrong sign
Llama-3-8B	T	-0.04	$\times 4$	$+0.21$	-0.006	wrong sign
	F	-0.06	$\times 1$	$+0.08$	$+0.000$	0.6%
	F	-0.06	$\times 4$	$+0.16$	-0.002	wrong sign
`L38.n7088` (*wrong*)	T	-1.42	$\times 1$	$+1.89$	$+0.004$	0.2%
Gemma-2-9b	T	-1.42	$\times 4$	$+2.64$	$+0.105$	4.0%
	F	-0.72	$\times 1$	$+1.33$	-0.057	wrong sign
	F	-0.72	$\times 4$	$+3.44$	-0.061	wrong sign
`L27.n13033` (*false*)	T	-2.86	$\times 2$	$+5.79$	$+0.063$	1.1%
Qwen2.5-7B	T	-2.86	$\times 4$	$+6.94$	$+0.315$	4.5%
	F	-2.88	$\times 2$	$+6.10$	-0.095	wrong sign
	F	-2.88	$\times 4$	$+7.53$	$+0.043$	0.6%

The mediator also leaves the estimand's internal comparison untouched. The secondary's own effect on the denominator, $\rho(x; \pi_k, \sec_r)/\rho(x; \pi_k)$, departs from unity by at most 0.28%, 0.94%, and 0.50% (Llama, Gemma, Qwen) at any dose, so the two forward passes differenced inside every E_r^b (main text Eq. 6) are gain-matched to under 1%. Only the cross-world comparison R_r was exposed to the gain, and Tables 17 and 19 bound that comparison.

Information in the margin levels. The paired columns of Table 18 regenerate E^{π}, but the margin levels carry information of their own. At $\times 1$, where the verdict still stands in every family, overriding the emblematic recruit costs Gemma 31% of its surviving true-side margin ($1.55 \to 1.08$) and 22% of its fake-side margin ($-2.80 \to -2.19$). Measured against what the primary erased, that single unit is restoring 17% and 25% of it. The same operation costs Llama 1.7% ($1.44 \to 1.42$), the family asymmetry in reserve size seen again at the level of margins. In Qwen at $\times 1$ the override moves the margin the *other* way ($1.12 \to 1.50$). `L27.n13033` is still depressing the verdict there, and it crosses at $\times 2$, so the opponent-to-recruit transition is legible without differencing. Past the crossing, $D_{\pi,\mathrm{sec}}$ sits farther from zero than D_{π} on the inverted side (for Gemma at $\times 2$, $-0.83 \to -1.90$ on T and $+1.73 \to +3.95$ on F, or 20% and 32% of the erased-and-inverted signal), so the recruit is partially reconstituting the erased class signal.

I.6 Why an inversion-type primary need not shrink the stream

The expectation that suppressing an important component shrinks the residual norm is an expectation about deletion. The dosed swap of main text Eq. 4 deletes nothing. It replaces the core's coefficient with a value drawn from the opposite class and, at $k > 1$, injects a write *larger* in magnitude than the one it displaces. Writing $\Delta a \propto k$ for the induced coefficient shift,

$$\left\|h + \Delta a\, \omega_{\star}\right\|^2 = \|h\|^2 + 2\,\Delta a\,\langle h, \omega_{\star}\rangle + \Delta a^2 \|\omega_{\star}\|^2. \tag{12}$$

The linear term is negative, because the injected signal opposes the stream's prevailing class direction, while the quadratic term grows as k^2. The recruits' own writes then feed back into ρ in the same direction. Beyond a crossover dose the norm must therefore rise. The joint-core primaries, which inject two enlarged writes at once, cross it by $\times 1$ to $\times 2$, while Llama's single, more modest core remains in the linear regime through $\times 4$ on the true side. This locates the LayerNorm-mediated component of self-repair reported for ablation studies (Rushing & Nanda, 2024) rather than contradicting it. Deletion-type interventions sit in the shrinking regime, where the mediator is real and must be controlled. We therefore retain the frozen-denominator control for any deletion-type variant of the primary, to which the audit above does not transfer.[4]

I.7 What this establishes, and what it leaves open

The final normalization is the only multiplicative mediator between a weight-fixed write and the logits on the direct path, because above the write's layer the residual stream is the identity. We have measured that mediator and bounded its contribution to a few percent of R, with the wrong sign in half the cells, the wrong dose profile in the rest, and no capacity in principle to produce the sign crossings that define certification. We regard the renormalization account of recruitment as excluded.

The causal claim itself never depended on this mediator. Every effect reported in this paper is a total effect computed exactly in a deterministic, fully observed structural causal model. Recruitment is the statement that a direction's contribution to the verdict is larger in the dosed world than in the clean one and, for the certified recruits, positive where it was negative. The interventions themselves establish that the verdict comes to depend on these directions once the core is suppressed, without appeal to a model of how the dependence arises.

It remains open whether recruitment is carried by re-expression, the detectors computing different coefficients because their inputs have changed, or by re-routing through downstream components. This concerns the internal route of an effect whose existence and magnitude are already fixed. Both are the model computing, neither is an artifact of measurement, and the choice between them does not bear on whether the reserve is real. We leave the decomposition to future work, noting that emergent self-repair was likewise established causally before any instance of it received an end-to-end account.

[4] The RMSNorm Jacobian is $\rho^{-1}\left(I - \hat{h}\hat{h}^{\top}\right)$, so the scalar g summarizes the isotropic part of the rescaling. The projection term and all finite-swap curvature are included exactly in every reported effect, none of which is linearized, and g serves here only to bound the passive channel.

J FITS AND DIAGNOSTICS OF THE COUPLING LAW

J.1 ESTIMATION PROCEDURE

Intervention	Clamp v	Dose $\lambda(v)$	In expectation	Used by
Counterfactual patch	$a_{x'}$ (partner)	-1	-1	Vig et al. (2020); Geiger et al. (2025)
Neutral clamp	m	0	0	this paper
Dosed swap	$m - \frac{\lambda}{2}\Delta$	chosen	λ	this paper
Zero ablation	0	$2m/\Delta$	unit-set; 0 iff $m = 0$	Michel et al. (2019); Gong et al. (2026)
Mean ablation	$\bar{a}$	$2(m - \bar{a})/\Delta$	$2q - 1$; 0 if balanced	Wang et al. (2023)
Resample ablation	$a_c(x')$, $x' \sim D$	$2(m - a_c(x'))/\Delta$	$\lambda(\mu_D)$, sd ρ_D	Chan et al. (2022); McGrath et al. (2023); Rushing & Nanda (2024)

Table 20: Every clamp is a dose. For any v, Equation 3 gives its position on the axis and Equation 8 its downstream effect. q is the fraction of the presented class in the reference set. μ_D and σ_D are the donor mean and standard deviation, and $\rho_D = 2\sigma_D/|\Delta|$. Measured placements are in Figure 6, and derivations and per-core detail are in Appendix N.

For every bench direction we regress the class-summed contribution E_r on the signed dose $\lambda = 1 - 2k$ of Eq. 3, using the per-pair values rather than the tabulated means so that standard errors propagate. The reported fit is on the on-manifold segment, $\lambda \in \{+1, 0, -1\}$ (clean, class-neutral clamp, full inversion), and the stress doses $\lambda \in \{-3, -5, -7\}$ are held out and used only to locate the boundary below. Slopes are tested against $\gamma_r = 0$ with Benjamini-Hochberg control within each family at $q = 0.05$, and certification uses the slope rather than R^2. We fit each class separately and pooled, and flag directions whose per-class slopes disagree in sign (Table 22). Directions at or above the core's layer cannot read its write and are excluded from all fits (§M, Table 34).

Where the fit stops holding. Figures 21 to 24 plot, per direction, the observed E_r against λ with each side's on-manifold fit extrapolated across the full range. The *lin.* column of Table 22 records the largest dose at which the residual from the on-manifold line stays within twice its combined standard error. For most certified directions this is $\times 1$, and the residuals turn systematically negative beyond it, because the response saturates once the injected signal is amplified past any value an input attains. Ten of the sixty-eight certified directions remain on the line past full inversion, six of them in Llama.

Crossing-dose comparison. $k^\star$ is obtained by setting Eq. 8 to zero, $\lambda^\star = -\mathrm{own}_r/\gamma_r$ and $k^\star = (1 - \lambda^\star)/2$, using only the fitted parameters. The reference k_{obs} is not a measurement. The ladder is coarse, and we locate the crossing by linear interpolation between the two rungs that straddle zero. Where those rungs lie inside the on-manifold segment the interpolation is over a stretch the fit certifies as linear. Beyond full inversion it is not, and because the true response saturates there, the interpolated reference is biased toward earlier crossings. The discrepancies in the *beyond* column of Table 21 are therefore upper bounds on the prediction error rather than estimates of it.

J.2 FAMILY-LEVEL SUMMARY

Table 21 summarises the certified couplings by family.

J.3 PER-DIRECTION FITS

Table 22 lists every certified direction's fit.

J.4 FIT DIAGNOSTIC FIGURES

Figures 21 to 24 plot, for the six strongest-coupling downstream directions per model, the measured erased margin E against the dose regressor $\lambda = 1 - 2k$, on-manifold window shaded, each side's window fit extrapolated as a straight line . Every per-direction statistic (γ_r, its standard error, R^2, the saturation onset, and the BH-FDR certification) is tabulated in Table 22, so the panels are deliberately unannotated.

Table 21: **The coupling law across families.** *ds.* counts bench directions downstream of the certified core(s), and *cert.* those whose coupling is reliably non-zero (BH-FDR within family, q=0.05). R^2 medians are given separately for MLP-neuron and attention-derived (OV channel, singular mode) directions. *inh.* is the median share of the clean contribution inherited from the core, over the certified directions whose clean value is resolved (n in parentheses). *flip err.* is the median discrepancy between the predicted and the interpolated crossing dose, split by whether the crossing falls inside the on-manifold segment or beyond full inversion.

			coupling sign		median R^2			flip err. (doses)	
model	ds.	cert.	$\gamma < 0$	$\gamma > 0$	mlp	ov/svd	inh.	in-window	beyond
Llama-3-8B-Instruct	20	18	18	0	0.84	0.74	0.67 (6)	0.02 (13)	—
Gemma-2-9b-it	22	17	14	3	0.78	0.72	0.38 (15)	0.05 (2)	0.12 (12)
Qwen2.5-7B-Instruct	16	14	9	5	0.79	0.24	0.40 (13)	0.02 (3)	0.22 (9)
Mistral-7B-Instruct-v0.3	23	19	11	8	0.58	0.45	0.23 (13)	0.03 (3)	0.22 (11)
pooled	81	68	52	16	0.76	0.68	0.39 (47)	0.02 (21)	0.19 (32)

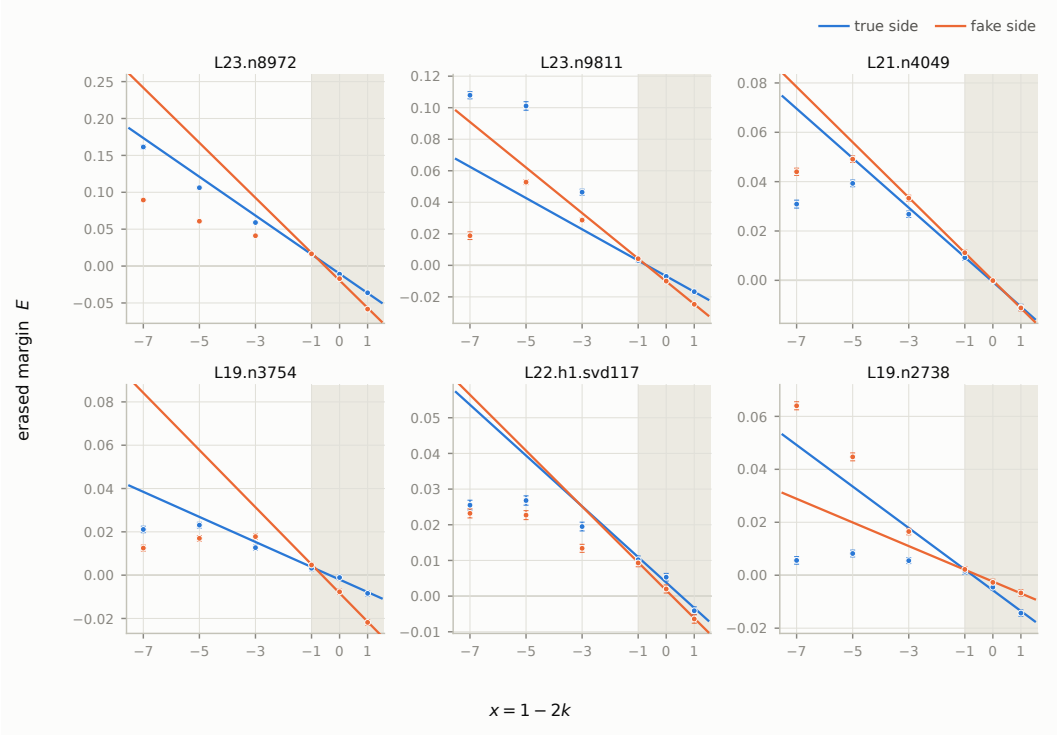

Figure 21: **Affine fits for the six strongest-coupling downstream directions of Llama-3-8B-Instruct.** Each panel plots the measured erased margin E against the dose regressor $\lambda = 1 - 2k$, separately for the true-prompt side (blue) and the fake-prompt side (orange). Markers are the archived per-dose means with their standard errors, so the six markers per colour correspond to $k = 0, \frac{1}{2}, 1, 2, 3, 4$. The shaded band marks the on-manifold window $\lambda \in [-1, +1]$ (equivalently $k \leq 1$) that the headline fit uses. Each side's straight line is fitted on that window and *extrapolated* across the full range. The two lines have their own slopes, so the panels show per-class fits rather than the shared-slope fit of Eq. 8, and drawing them separately makes a sign disagreement between the sides visible. The axis label $x = 1 - 2k$ is λ.

Table 22: **The coupling law, per direction.** For every downstream bench direction with a coupling reliably different from zero (BH-FDR within family, q=0.05), the table gives the on-manifold fit of Eq. 8 (pooled intercept, as certified in App. K.2, while the exemplar tables split it), with intercept own_r, slope γ_r (negative = *counterweight*, positive = *relay*), and fit quality R^2. *inh.* $= \gamma_r/(\mathrm{own}_r+\gamma_r)$. $k^\star$ is the crossing dose predicted by the fit, k_{obs} the crossing located in the ladder by interpolation, and † marks crossings beyond full inversion. *lin.* is the largest dose at which the residual stays on the on-manifold line. z is the weight-only wiring alignment of §M. Rows shaded grey are hybrids, whose clean class and coupling sign disagree. Directions upstream of or level with the cores are omitted because they cannot read the core write. They are listed in Table 34.

direction	own_r	γ_r	R^2	inh.	$k^\star$	k_{obs}	lin.	z
LLAMA-3-8B-INSTRUCT								
L23.n8972	-0.015	-0.032	0.95	0.68	0.74	0.73	×1	+4.7
L23.n9811	-0.009	-0.012	0.95	0.59	0.85	0.85	×1	+5.1
L21.n4049	-0.000	-0.011	0.99	0.97	0.52	0.51	×2	+0.5
L19.n3754	-0.005	-0.009	0.76	0.66	0.76	0.76	×1	+7.0
L22.h1.svd117	+0.003	-0.007	0.95	–	0.31	0.29	×1	-0.8
L19.n2738	-0.004	-0.006	0.84	0.60	0.83	0.83	×1	+4.8
L24.n4543	+0.001	-0.004	0.63	–	0.43	0.51	×1	+1.9
L20.n1390	+0.006	-0.004	0.75	–	–	–	×1	+4.2
L21.n14165	+0.002	-0.003	0.89	–	0.13	0.12	×1	+0.9
L22.h1.c60	+0.003	-0.002	0.74	–	–	–	×4	-1.0
L21.n798	+0.001	-0.002	0.61	–	0.22	0.20	×1	+3.7
L21.n517	+0.001	-0.002	0.90	–	0.31	0.27	×4	+1.7
L28.h27.c120	+0.001	-0.002	0.86	–	0.35	0.33	×1	+0.0
L20.n14044	-0.000	-0.002	0.46	0.93	0.54	0.37	×1	-0.7
L22.h1.svd116	+0.002	-0.002	0.68	–	–	–	×4	-0.5
L22.svd323	+0.000	-0.002	0.68	–	0.39	0.30	×2	+0.3
L28.h27.c34	+0.001	-0.001	0.80	–	–	–	×4	-1.0
L28.h27.c43	+0.001	-0.001	0.37	–	–	–	×1	-1.2
GEMMA-2-9B-IT								
L38.n7088	-0.195	-0.653	0.92	0.77	0.65	0.71	×1	+34.3
L41.n8771	-0.049	-0.092	0.87	0.65	0.77	0.80	×1	+4.8
L38.n854	-0.129	-0.081	0.87	0.39	1.29	1.40†	×1	+4.9
L39.n10210	-0.087	-0.064	0.93	0.42	1.18	1.29†	×1	+28.8
L39.n10085	-0.398	-0.047	0.69	0.11	4.70	3.21†	×1	+8.6
L32.n8200	-0.069	-0.043	0.18	0.38	–	–	×1	+4.0
L36.n1119	+0.048	+0.038	0.34	0.44	1.13	1.01†	×1	+4.2
L41.n3789	-0.091	-0.034	0.59	0.27	1.85	1.49†	×1	+6.7
L34.n4784	-0.023	-0.018	0.98	0.44	1.13	1.13†	×1	+1.4
L39.n12353	+0.048	+0.017	0.96	0.27	1.86	1.70†	×1	+4.6
L30.n761	+0.090	+0.017	0.61	0.16	3.09	1.85†	×1	+5.2
L33.n1466	-0.048	-0.014	0.16	0.22	2.24	1.63†	×1	+2.3
L40.h11.svd18	-0.025	-0.011	0.72	0.30	1.66	1.62†	×1	+1.9
L40.h7.c105	-0.008	-0.006	0.99	0.40	1.24	1.34†	×2	+2.9
L40.h7.c177	-0.009	-0.004	0.82	0.32	1.56	1.65†	×2	+3.2
L34.h15.c22	+0.011	-0.002	0.50	–	–	–	×1	-1.1
L31.h3.c137	+0.011	-0.001	0.18	–	–	–	×1	+3.1
QWEN2.5-7B-INSTRUCT								
L27.n13033	-1.790	-1.162	0.97	0.39	1.27	1.14†	×1	+18.7
L27.n4228	-0.356	-0.409	0.96	0.53	0.93	0.95	×1	+18.3
L24.n8308	+0.171	+0.086	0.36	0.34	1.49	1.28†	×1	+12.9
L23.n8341	-0.058	-0.063	0.90	0.52	0.96	0.98	×1	+7.4
L27.n15192	+0.149	+0.049	0.97	0.25	2.00	1.50†	×1	+7.1
L25.n11824	+0.068	+0.045	0.88	0.40	1.26	1.26†	×1	+5.9
L25.n4929	-0.048	-0.043	0.61	0.47	1.05	1.12†	×1	+7.3
L24.n12848	-0.046	-0.033	0.21	0.42	1.20	0.98†	×1	+6.1
L25.n8444	-0.003	-0.029	0.77	0.90	0.55	0.43	×1	+8.3
L23.n14476	-0.028	-0.026	0.79	0.48	1.05	1.16†	×1	+8.9
L26.h17.svd117	+0.105	+0.007	0.07	0.06	–	–	×1	-0.4
L24.n9321	+0.013	+0.004	0.37	0.21	2.33	1.39†	×1	+1.5
L23.h27.c82	+0.015	-0.003	0.24	–	–	–	×1	-0.7
L26.h20.svd17	-0.007	-0.002	0.72	0.18	2.81	1.97†	×4	-0.1

Table 22: **The coupling law, per direction** (continued).

direction	own$_r$	γ_r	R^2	inh.	$k^\star$	k_{obs}	lin.	z
MISTRAL-7B-INSTRUCT-V0.3								
L31.n8773	-0.393	-0.398	0.98	0.50	0.99	0.96	×1	+18.0
L31.n13669	-1.850	-0.208	0.58	0.10	4.96	1.88†	×1	+14.3
L28.n4079	-0.239	-0.158	0.55	0.40	1.25	1.22†	×1	+13.4
L24.n4907	+0.095	-0.093	0.90	–	–	–	×1	+13.7
L22.h20.svd0	+0.302	-0.083	0.44	–	–	–	×1	+9.1
L29.n8705	+0.251	-0.048	0.11	–	–	–	×1	+5.8
L24.n5072	+0.070	+0.042	0.75	0.38	1.33	1.14†	×1	+14.5
L30.n11434	-0.048	-0.032	0.88	0.40	1.26	1.47†	×1	+6.4
L22.h23.svd2	-0.086	-0.023	0.46	0.21	2.35	2.55†	×1	+7.8
L25.h8.svd45	+0.028	+0.017	0.92	0.38	1.32	1.19†	×1	+2.8
L22.n8735	-0.080	-0.016	0.35	0.17	3.02	2.17†	×1	+2.5
L22.n5778	-0.012	+0.015	0.89	–	0.11	0.07	×1	+1.2
L30.n498	+0.039	+0.011	0.50	0.21	2.33	3.74†	×1	+6.9
L30.n11755	+0.004	+0.010	0.74	0.71	0.70	0.66	×1	+3.0
L24.n1021	-0.013	+0.007	0.93	–	–	–	×1	+4.2
L28.n9060	+0.017	+0.005	0.30	0.22	2.24	2.40†	×3	+0.4
L22.h20.c93	+0.014	-0.004	0.44	–	–	–	×1	+3.5
L22.n1070	-0.028	-0.004	0.11	0.11	4.46	2.04†	×1	+6.9
L27.n11713	+0.006	+0.002	0.08	0.23	2.13	1.59†	×1	+2.5

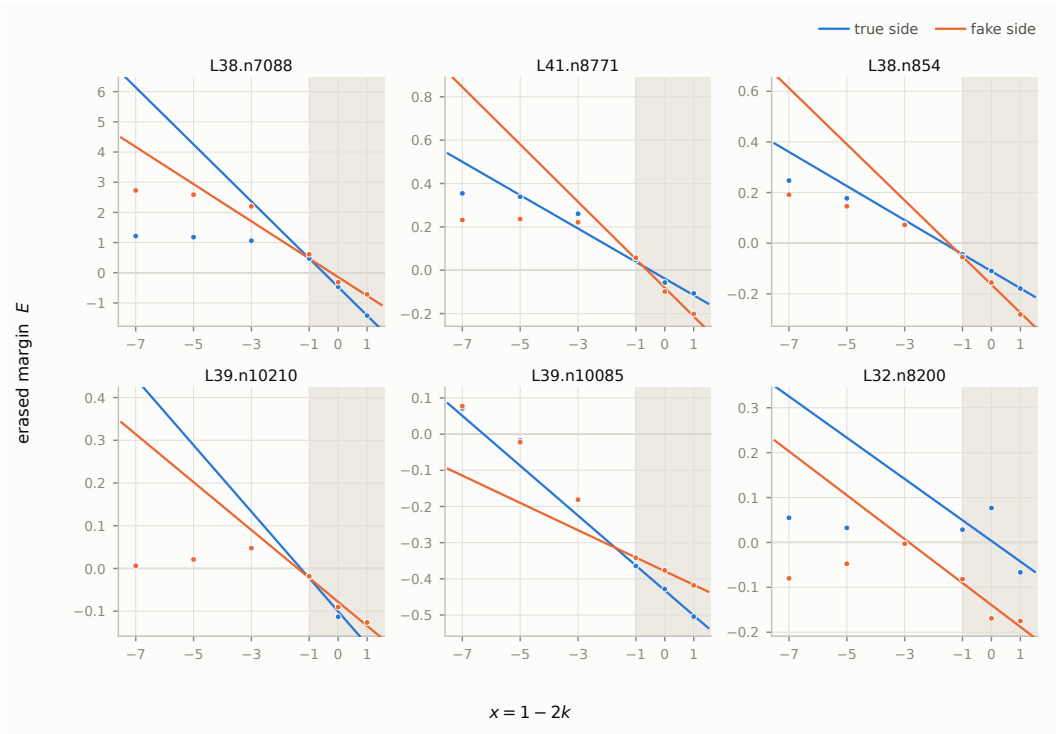

Figure 22: **Affine fits for the six strongest-coupling downstream directions of Gemma-2-9B-it.** Axes, colours, shading and fitted line as in Figure 21 ($\lambda = 1 - 2k$). The vertical scale differs by roughly an order of magnitude between panels, since the coupling strengths $|\gamma_r|$ vary widely within a single model. This spread is why the wiring correlations in the main text are reported on ranks rather than raw values.

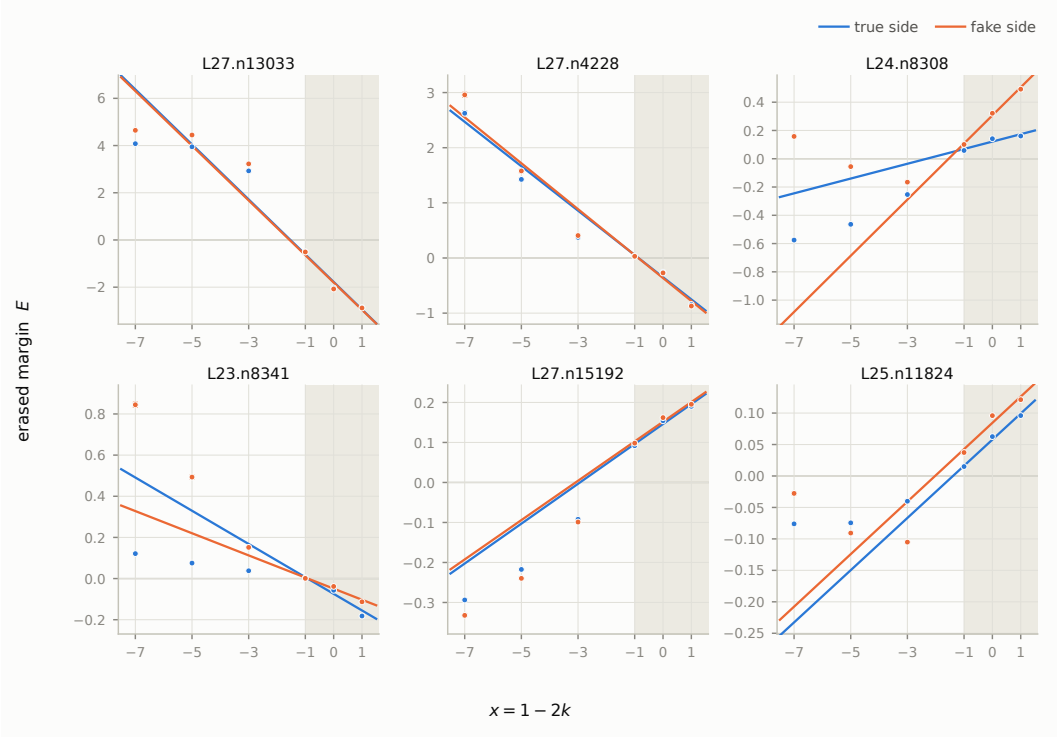

Figure 23: **Affine fits for the six strongest-coupling downstream directions of Qwen2.5-7B-Instruct.** Axes, colours, shading and fitted line as in Figure 21 ($\lambda = 1 - 2k$).

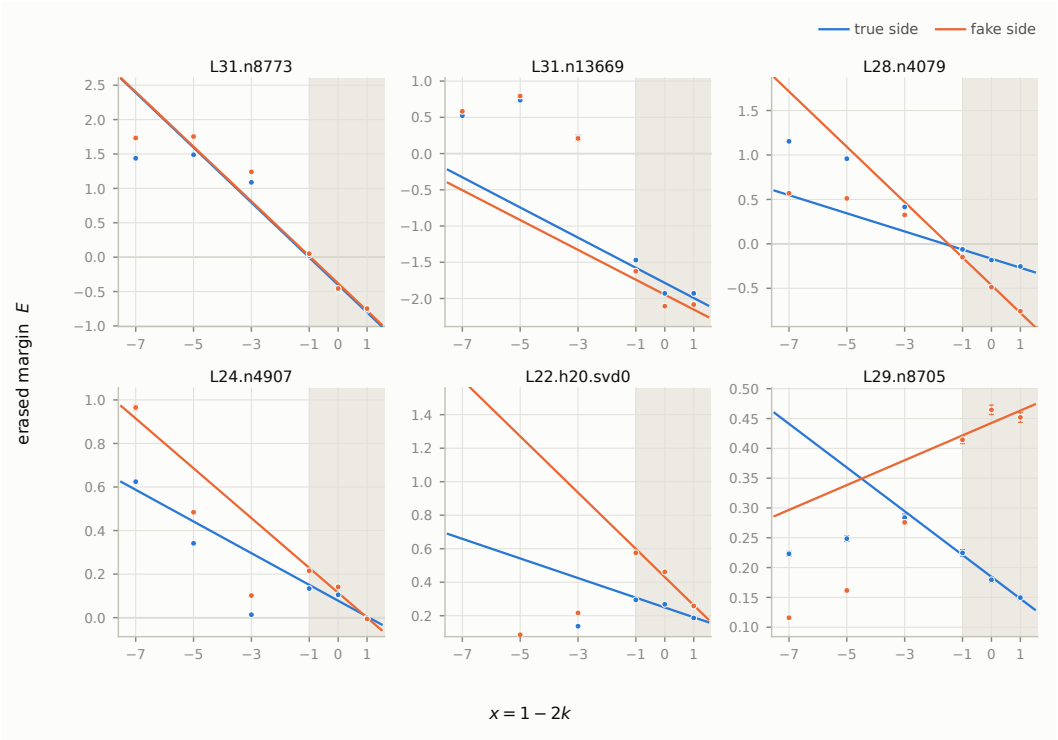

Figure 24: **Affine fits for the six strongest-coupling downstream directions of Mistral-7B-Instruct-v0.3.** Axes, colours, shading and fitted line as in Figure 21 ($\lambda = 1 - 2k$).

J.5 EXEMPLAR FITS AND THE HALF-RUNG HOLDOUT

Tables 23 and 24 give the exemplar fits and the effect of adding the half rungs.

Table 23: **The exemplar fits of main text Figure 4, and where the post-fit rungs landed.** For each exemplar direction the table gives the two per-class intercepts, their difference $\Delta\mathrm{own}_r$, the shared slope with its standard error, the weighted R^2, and the slope refitted with the half rungs included. *track* is how far the two parallel lines sit from their own class's fitted points, as a fraction of the exemplar's full data span. The two half-rung columns give the out-of-sample landing as the largest deviation of the four half-rung means from the line of their own class, in SEs and as a fraction of span. That line was fitted before those rungs existed.

		intercepts			shared slope			half rungs		
Model	Direction	own^T	own^F	Δown	γ_r (SE)	R^2	$\gamma_r^{5\mathrm{pt}}$	max $\lvert z\rvert$	% span	track
Llama-3-8B	`L23.n8972`	-0.0111	-0.0190	$+0.0080$	-0.0319 (0.0007)	0.969	-0.0328	6.1	1.4%	3.3%
Gemma-2-9B	`L38.n7088`	-0.3318	-0.1034	-0.2285	-0.7272 (0.0051)	0.944	-0.7284	17.9	7.2%	8.7%
Qwen2.5-7B	`L27.n13033`	-1.7789	-1.8004	$+0.0215$	-1.1619 (0.0153)	0.973	-1.1542	5.6	2.4%	3.7%
Mistral-7B-v0.3	`L31.n8773`	-0.4038	-0.3815	-0.0223	-0.3986 (0.0119)	0.980	-0.3962	2.1	2.1%	2.9%

Table 24: **The affine window survives doubling its interior sampling.** With the two half rungs $\lambda = \pm\frac{1}{2}$ added (478 to 598 pairs each, same population as the ladder), the columns give agreement between γ_r fitted on five rungs and on the original three, the median slope shift, and two Occam columns (median $|\gamma|$ change and misfit removed when a quadratic term is added). R^2 is the per-class variance in E explained by the affine law over the five rungs, pooled over directions. The last column is a holdout test in which the law's parameters, fitted on the original three rungs alone, are scored against the two half-rung means the fit never saw (324 held-back means). The same test run on the pooled single-intercept form the split intercepts replaced gives 0.86, 0.59, 0.87 and -0.41, so out of sample the pooled form is worse than predicting each direction-class's own mean in Mistral.

		γ: 5 rungs vs 3			shift	+quadratic			
Model	n	r	$\rho_{\lvert\gamma\rvert}$	sign	med. $\lvert\Delta\gamma\rvert$/SE	med. $\lvert\Delta\gamma\rvert/\lvert\gamma\rvert$	misfit rem.	R^2	OOS R^2
Llama-3-8B	20	0.99957	0.982	1.00	0.33	0.6%	5%	0.981	0.905
Gemma-2-9B	22	0.99998	0.998	1.00	0.39	0.9%	4%	0.974	0.803
Qwen2.5-7B	16	0.99999	0.997	1.00	0.51	2.5%	9%	0.965	0.904
Mistral-7B-v0.3	23	0.99973	0.999	1.00	0.26	1.3%	3%	0.913	0.653

K TESTS OF THE COUPLING LAW

This appendix carries the evidence for §5. It establishes that the coupling γ_r is resolved for most of the bench and does not move under any refit we tried, that the affine form with one intercept per class is the right leading-order description on $|\lambda| \leq 1$, predicting doses it never saw, and that the description stops at full inversion. It does not establish that the affine form is selected over other smooth forms, and §K.5 says why. Every number comes from a committed artifact, and none is hand-computed.

K.1 FIT AND POPULATION

The dose is $\lambda = 1 - 2k$, where k is the fraction of the swap toward the counterfactual, with $k = 0$ clean ($\lambda = +1$), $k = \frac{1}{2}$ neutral ($\lambda = 0$), $k = 1$ full inversion ($\lambda = -1$), and $k \in \{2, 3, 4\}$ overshoot. Eq. 8 is fitted by weighted least squares on the three on-manifold rungs $\lambda \in \{+1, 0, -1\}$, both classes at once, one intercept per class and one shared slope, with weights $1/\sigma^2$ from the per-rung standard error of the mean over per-prompt paired differences. The window was fixed before any fit and is the same for every direction, while the half rungs $\lambda = \pm\frac{1}{2}$ and the overshoot rungs are held out. Table 25 gives the population. The class-F margin enters E with its sign flipped so that on both classes a positive E means the direction moved the verdict the way the swap pushes it.

Model	Downstream directions	Correction family	Prompt pairs
Llama-3-8B-Instruct	20	20	532
Gemma-2-9B-it	22	22	583
Qwen2.5-7B-Instruct	16	22	478
Mistral-7B-Instruct-v0.3	23	33	598
Total	81	97	

Table 25: **The fitted population.** Directions downstream of a core enter the fits, and the correction family is the full bench, which in Qwen and Mistral includes 6 and 10 upstream placement controls. The same prompt pairs are measured at every dose.

K.2 Certification

A direction is certified when $|\gamma_r/\mathrm{se}(\gamma_r)|$ survives Benjamini-Hochberg at $q = 0.05$ over its family's full bench. The gate tests the slope alone, and R^2 plays no part. Under the pooled-intercept fit 68 of 81 directions certify, 52 counterweights ($\gamma_r < 0$) and 16 relays ($\gamma_r > 0$). Under the split-intercept fit 69 certify, the same 68 plus Gemma `L31.n4305` ($p = 0.047$ pooled, 0.026 split). §5 quotes the 68, which pass both gates.

The 13 that do not certify fail a significance gate rather than a linearity test (Table 26). Seven have slopes indistinguishable from zero, and six sit narrowly outside the correction. R^2 explains none of them. Llama `L28.h27.c0` (0.59), Gemma `L31.h3.c72` (0.53) and Mistral `L30.h3.c8` (0.55) fit reasonably and still do not resolve a slope, because on three to six points R^2 is a weak diagnostic independent of whether γ_r is resolved. An uncertified direction is not shown to be uncoupled, and its role is unresolved at this sample size.

Status	Count	p (pooled fit)
Slope indistinguishable from zero	7	0.37–0.99
Excluded at the margin of the correction	6	0.031–0.077

Table 26: **The 13 uncertified directions.**

The gate is conservative in two ways. The correction family is the full bench, so in Qwen and Mistral it includes the upstream placement controls (22 and 33 tests rather than 16 and 23). And the borderline decisions are sensitive to the rounding of the archived statistics (§K). Pushing every σ to its rounding extreme moves the certified count to 69 or 70, and every direction that flips (`L30.h3.c8`, `L38.h15.c111`, `L30.h13.c170`) is already among the borderline six. The coefficients are stable, with γ_r over the certified set shifting by a median 0.04%.

K.3 One slope, two intercepts

Fitting each class with its own free slope and correlating the two across the 81 directions gives Pearson $r = 0.954$. Because $|\gamma_r|$ spans orders of magnitude and a few large couplings carry r, the rank correlation, Spearman $\rho = 0.834$, is the scale-free statement.

Splitting the intercept by class makes the law predictive (Table 27). In sample it raises the median per-direction R^2 in every family, from between 0.50 and 0.76 to between 0.82 and 0.91. Out of sample, on the half rungs of §K.4, the single-intercept form predicts Mistral worse than each direction's own mean would ($R^2 = -0.41$), while the split form, fitted on the same three rungs with no new data, turns that into a usable forecast. We report medians across directions because a variance-weighted R^2 is dominated by a few large-coupling directions.

K.4 Doses the fit never saw

We ran $\lambda = \pm\frac{1}{2}$ on the same population after the three-rung fits were frozen and scored each frozen line against the 324 held-back class means (80, 88, 64 and 92 by family). Each score is therefore

	in-sample median R^2		held-out R^2	
Model	split (reported)	single	split (reported)	single
Llama	0.874	0.755	0.905	0.858
Gemma	0.911	0.599	0.803	0.586
Qwen	0.821	0.665	0.904	0.867
Mistral	0.903	0.499	0.653	−0.406

Table 27: **The split intercept, in sample and out of sample.** The left columns give the median per-direction R^2 of the affine law on $\lambda \in \{+1, 0, -1\}$ under the two intercept forms. The right columns give the variance-weighted R^2 of the frozen three-rung line on the 324 held-back half-rung class means, which the line never saw.

a forecast. The variance-weighted held-out R^2 is 0.65 to 0.91 by family (Table 27), and this is the out-of-sample number, rather than the wider-window in-sample figure of 0.91 to 0.98.

The coefficient is stable under the new data. Refitting with the half rungs included shifts γ_r by a median 1.7 to 2.1% of its magnitude in three families and 7.2% in Llama, in every case 0.26 to 0.51 of its own standard error, with the sign preserved on all 81 directions. Adding a free quadratic term moves γ_r by a further 0.6 to 2.5%, again below its standard error.

The weighted R^2 describes the largest couplings, so we screened every certified direction carrying at least 1% of its family's held-out weight for a held-out R^2 below 0.30. Only Mistral `L31.n13669` qualifies. It is a counterweight with $\gamma_r = -0.21$ that fits in sample at $R^2 = 0.72$, carries 12% of Mistral's weight, and predicts its held-out doses no better than its mean ($R^2 = -0.01$).

Table 28 gives the four exemplars of Figure 4, chosen from the top two to six fits by R^2 among each family's certified directions. All four are MLP neurons and counterweights. The hollow markers in the figure are the half-rung means.

		intercepts		shared slope		half rungs	
Model	Direction	own_r^T	own_r^F	γ_r (SE)	R^2	γ_r^{5pt}	% span
Llama-3-8B	`L23.n8972`	−0.011	−0.019	−0.032 (0.0007)	0.97	−0.033	1.4
Gemma-2-9B	`L38.n7088`	−0.332	−0.103	−0.727 (0.005)	0.94	−0.728	7.2
Qwen2.5-7B	`L27.n13033`	−1.779	−1.800	−1.162 (0.015)	0.97	−1.154	2.4
Mistral-7B-v0.3	`L31.n8773`	−0.404	−0.382	−0.399 (0.012)	0.98	−0.396	2.1

Table 28: **The exemplar fits, and where the held-out rungs landed.** Per-class intercepts and shared slope from the three-rung fit with its weighted R^2, the slope refitted with the half rungs included, and the largest deviation of the four half-rung means from their class's line, as a fraction of the direction's data span.

K.5 What the holdout does and does not show

The half rungs lie between the training rungs, so predicting them is interpolation, and a predictor that assumes no law can be scored on the same points (Table 29). Straight interpolation between the bracketing measured rungs beats the affine line by a factor of four in χ^2, and a quadratic through the three training points by six, and the same ordering holds when the neutral rung is withheld (§K.6). Curvature inside the window is therefore resolvable at the precision of the standard errors. It is also small. The quadratic term moves γ_r by 0.6 to 2.5%, below its standard error, and the affine line still captures 65 to 91% of the held-out variance. The holdout establishes that the dose response is smooth and that the affine line tracks it to leading order with a coefficient that does not depend on the form, but it does not select the affine form over its smooth rivals. A test that did would have to extrapolate beyond the window, and §K.7 shows why that is not available.

Withheld	points	affine line χ^2	bracket interpolation χ^2	quadratic χ^2
half rungs $\lambda = \pm\frac{1}{2}$	324	7501	1929	1188
neutral rung $\lambda = 0$	162	3221	530	390

Table 29: **Three predictors on the same held-out points.** The affine line is the frozen fit (three rungs for the half rungs, the four outer rungs for the neutral rung). Bracket interpolation joins the two measured rungs on either side of the withheld one and assumes no law. The quadratic passes exactly through the training rungs. χ^2 is in units of the per-rung standard error, and lower is better.

K.6 Leave one rung out

On the five-rung ladder we withheld each rung in turn, refit on the other four, and scored the prediction at the withheld rung (Table 30, with 162 class means per row). The three interior rungs are predicted to 1.36 to 1.64 standard errors and a median 8 to 9% of each direction's data span, and the two endpoints to about 3.0 SE and 17% of span. Extrapolating past the last anchor costs twice what interpolating does, which is the leverage of any line, and it is independent support for fitting on $|\lambda| \leq 1$ rather than a wider window. The single-intercept form is worse at every rung but -1, where both forms extrapolate.

Withheld rung	median $\lvert z\rvert$	single-intercept $\lvert z\rvert$	median % of span	frac. $\lvert z\rvert > 2$	
$+1$	2.97	4.14	17.9	0.62	endpoint
$+\frac{1}{2}$	1.39	2.36	9.1	0.36	interior
0 (neutral)	1.36	2.19	9.0	0.36	interior
$-\frac{1}{2}$	1.64	1.94	8.4	0.41	interior
-1	3.04	2.73	16.5	0.59	endpoint

Table 30: **Leave-one-rung-out.** Each row refits on the other four rungs and scores the prediction at the withheld rung, in standard errors of the withheld measurement and as a fraction of the direction's data span.

K.7 Where the law stops

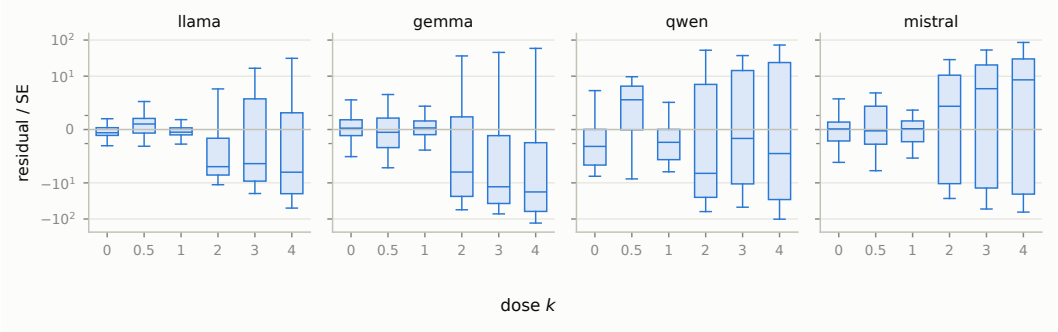

Figure 25: **Standardised residuals from the on-manifold fit, by dose, for all downstream directions in each model.** Each box aggregates the residuals $(E - \hat{E})/\mathrm{SE}$ over every downstream direction and both verdict sides at that dose. Whiskers are the usual $1.5\times$IQR range and outliers are suppressed. The vertical axis is symmetric-logarithmic, linear within ± 2 SE and logarithmic beyond, so that the tightly clustered residuals at $k \leq 1$ and the heavy tails at the stress doses are both resolvable in one frame. Residuals at $k = 0, \frac{1}{2}, 1$ are flat and centred on zero by construction (those are the three points the fit uses), so the informative content is the abrupt fanning at $k = 2, 3, 4$, which is the same saturation seen per direction in Figures 21 to 24 and is why the headline γ_r is estimated on the $k \leq 1$ window rather than the full ladder.

Figure 25 plots the standardised residuals of the on-manifold fit at every dose, including the three overshoot rungs the fit never saw. Inside the window the residuals are small, and beyond it they

grow with dose as the response saturates and the line overpredicts. For each direction we record the last dose at which the measurement stays within two standard errors of its line, and for all four exemplars it is $k = 1$, full inversion. This is why the headline fit uses $k \leq 1$ and not the six-dose ladder, and why no extrapolative test of the affine form is available on it. Gemma `L38.n7088` is shown with its held-out points in Figure 4 because its curvature is the best resolved on the bench (first of 81 by the t-statistic of a freed quadratic term), though not the largest.

L Out-of-Domain Test on GPT-2 Small and Indirect-Object Identification

L.1 Setup

The results so far were obtained on one task family, in four models of comparable size, over a bench of components that we selected ourselves. The natural concern is that the coupling coefficient is an artefact of those choices. Indirect-object identification changes all three at once, and none of the changes is ours to make. The model is GPT-2 small, two orders of magnitude smaller than the models above. The task is syntactic name binding rather than factual verification. The component bench is Wang et al.'s published IOI circuit together with the copy-suppression head `L10.h7` identified by McDougall et al., so the assignment of heads to roles, and in particular which heads the intervention can reach and which it cannot, was settled before this work began.

The intervention and the estimator are the ones used throughout. A prompt pair is a single IOI template instantiated twice with the two names exchanged, so that the margin $D_{AB} = \mathrm{logit}(A) - \mathrm{logit}(B)$ is faithful on one member of the pair and contrastive on the other, exactly as D_{TF} is on the factual task, and the erased margin E is built from it by the same definition. The core is the name-mover pair `L9.h6` + `L9.h9`, swapped at doses $k \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, 2, 3, 4\}$, so that $\lambda = 1 - 2k$ runs over $\{+\frac{1}{2}, 0, -\frac{1}{2}, -1, -3, -5, -7\}$, with 300 prompt pairs at every rung.

The published circuit gives the law two groups to predict on. Ten heads lie in layers 10 and 11, where the swap can reach them, and ten lie in layer 8 and earlier, where it cannot. If γ_r measures what a head inherits from the core, it should resolve in the first group and vanish in the second. The two remaining heads of the circuit, `L9.h7` and `L9.h0`, sit in the core's own layer, where a swap of `L9.h6` and `L9.h9` cannot reach them, and are excluded.

L.2 Seven of the ten reachable heads follow the law

Each reachable head is fitted as on the factual ladders, with one intercept per class and one shared slope γ_r, on the on-manifold window $|\lambda| \leq 1$ (Figure 26a). A head follows the law if γ_r is nonzero after Benjamini-Hochberg correction at $q = 0.05$ and the response is affine, in that a freed quadratic term $\beta_r \lambda^2$ is not required, $|\beta_r/\gamma_r| \leq 1$. Seven of the ten reachable heads do (Table 31, Figure 27), with a median weighted R^2 of 0.971 across the seven. Two of the remaining three, `L10.h1` and `L10.h6`, have $|\gamma_r| \leq 0.012$, within two standard errors of zero, so the swap does not measurably reach them, and their response is a flat line. The third, `L11.h10`, has a curved response, and the next subsection shows why.

Two features of the table stand out. Every head that follows the law is a counterweight, with $\gamma_r < 0$, and there are no relays, while the factual bench divides 52 to 16. A circuit built to move a single name into the output position couples to its own core only in opposition to it. And the coupling is dominated by one head. `L10.h7`, the copy-suppression head, reaches $|\gamma_r| = 0.886$, seven times the next largest value and twenty times the median over the reachable heads. The strongest coupling on the bench falls on the head that prior work independently identified as the circuit's negative feedback element.

The window is read off the data rather than imposed on it. Outside $|\lambda| \leq 1$ the response saturates. Eight of the ten reachable heads leave the two-standard-error band around the fitted line before $\lambda = -7$, four of them by $\lambda = -3$, and the only two that never leave it are the flat heads `L10.h1` and `L10.h6`. This is the same saturation that motivates restricting to $k \leq 1$ on the factual ladders. Inside the window the line predicts. The half rungs at $\lambda = \pm\frac{1}{2}$ were run after the fits had been frozen and never entered them. Over the forty held-out class means the median $|z|$ is 0.23, and only two exceed two standard errors, both at `L10.h7` for $\lambda = -\frac{1}{2}$, the head with by far the steepest slope

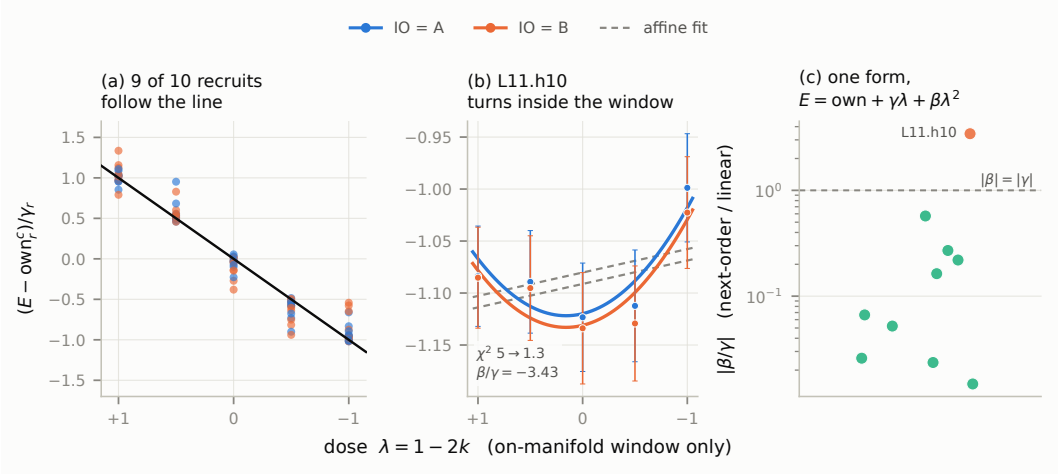

Figure 26: **Which heads require a next-order term.** The panels show only reachable recruits, the heads above the core's layer. Adding $\beta_r\lambda^2$ and taking $|\beta_r/\gamma_r|$ divides them cleanly. Nine of ten sit at or below 0.572 and follow the line *(a)*. One, `L11.h10` at 3.430, sits well above it, a sixfold gap with nothing inside *(c)*. Its parabola turns inside the window *(b)*, and its γ_r does not certify, so the quadratic is free to dominate a ratio whose denominator is unresolved. Built from the clean-reference ladder.

Table 31: **The coupling law on the ten reachable IOI heads.** Shared slope γ_r with its standard error, fitted on the on-manifold window. Roles are Wang et al.'s. Every certified head is a counter-weight, with $\gamma_r < 0$.

Head	Role	γ_r	(SE)	cert.
`L10.h7`	negative name mover	−0.8864	0.0211	yes
`L11.h2`	backup name mover	−0.1269	0.0095	yes
`L10.h2`	backup name mover	−0.0841	0.0055	yes
`L11.h6`	backup name mover	−0.0584	0.0048	yes
`L10.h0`	name mover	−0.0545	0.0120	yes
`L11.h10`	negative name mover	−0.0348	0.0253	–
`L11.h9`	backup name mover	−0.0183	0.0019	yes
`L10.h10`	backup name mover	−0.0175	0.0072	yes
`L10.h6`	backup name mover	−0.0120	0.0070	–
`L10.h1`	backup name mover	−0.0043	0.0063	–

(Figure 28b). Because these rungs lie inside the fitted window this is interpolation. It shows that the line tracks the response but does not show that it beats every smooth alternative.

The law also predicts where the coefficient should vanish, and it does. No head outside the downstream set certifies (Figure 28c). Six of these twelve do not move at all under the intervention, with $\max|E^\pi - E^0| < 10^{-3}$ across the ladder. Among the six that do, whose clean effects reach $E^0 = +0.68$, the median $|\gamma_r|$ is 4.5×10^{-3}, against 4.5×10^{-2} for the downstream heads, a separation of roughly tenfold, on a division of heads that was fixed by the published circuit (Wang et al., 2023).

L.3 The exception

`L11.h10`, the second of the two negative name movers, is the one reachable head with a curved response. Under the two-head core its response falls and then rises inside the window, with $R^2 = 0.413$, a quadratic term larger than the linear one, $|\beta_r/\gamma_r| = 2.608$, and a parabola that turns at $\lambda^* = +0.19$ (Figure 29, left). For the other nine reachable heads the ratio is below one.

The curvature has a mechanism. `L11.h10` reads the core largely through `L10.h7`, the other negative name mover, which is free to move under the two-head swap and is itself strongly curved. If that

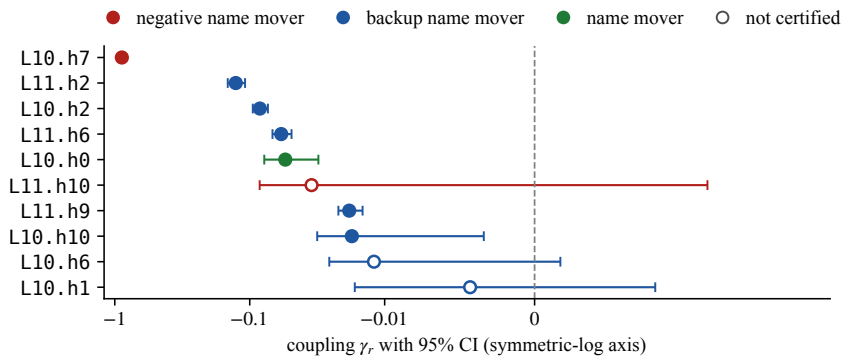

Figure 27: **Coupling of the ten reachable IOI heads.** γ_r with 95% intervals from Table 31. Filled markers are significant after Benjamini-Hochberg correction at $q = 0.05$. Every point estimate is negative, and `L10.h7`, the copy-suppression head, dominates.

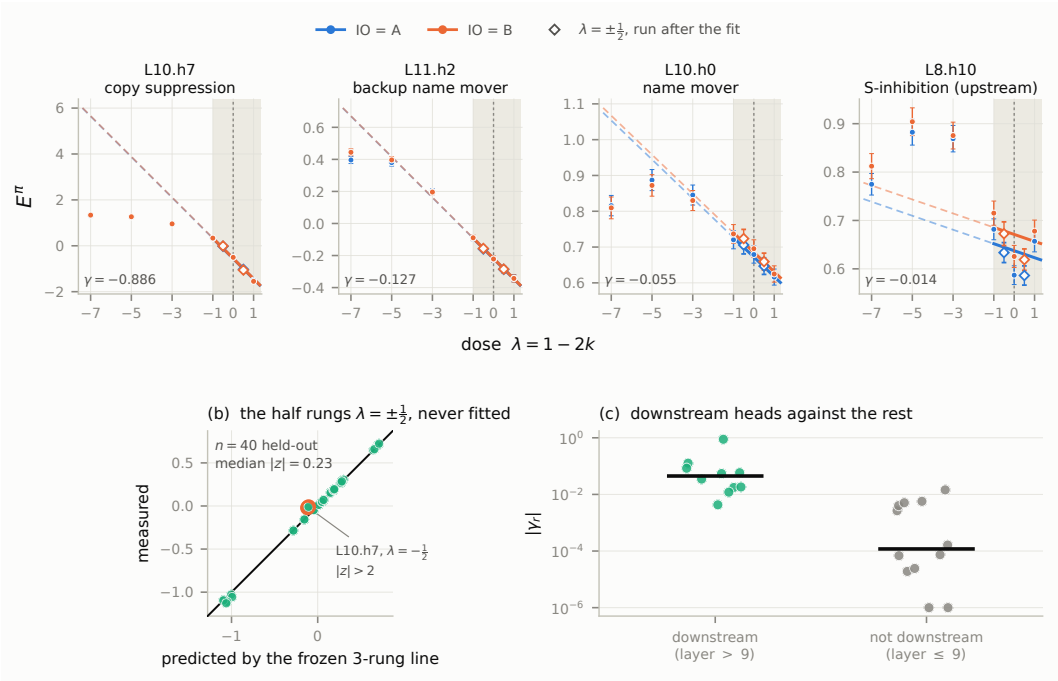

Figure 28: **The coupling law transfers to GPT-2 small on IOI.** *(a)* Dose response of four heads, solid inside the window $|\lambda| \leq 1$ where the law is fitted, dashed outside. *(b)* The 40 held-out means at $\lambda = \pm\frac{1}{2}$ against the frozen three-rung line, with the two beyond 2 SE ringed. *(c)* $|\gamma_r|$ for heads the core swap can reach (above layer 9) and cannot (at or below it).

is the cause, clamping `L10.h7` into the core should remove it. It does (Figure 29, right). The ratio falls from 2.608 to 0.292, R^2 rises from 0.413 to 0.972, the vertex moves from $+0.19$ to $+1.71$, outside the window, the raw means become monotone, and the two held-out points, at $p = 0.070$ and $p = 0.041$ before, are at $p = 0.837$ and $p = 0.459$ after. The coefficient itself nearly quadruples, from -0.0701 to -0.2695. While `L10.h7` was free, it had been absorbing most of `L11.h10`'s response to the core, and holding it at the counterfactual releases that response as a line. The change is specific to `L11.h10`, since every other shared recruit moves by a factor between 0.73 and 1.08, and the two ladders share the same clean effects E^0. The exception therefore traces to a term missing from the core, and the law holds once the mediator is held.

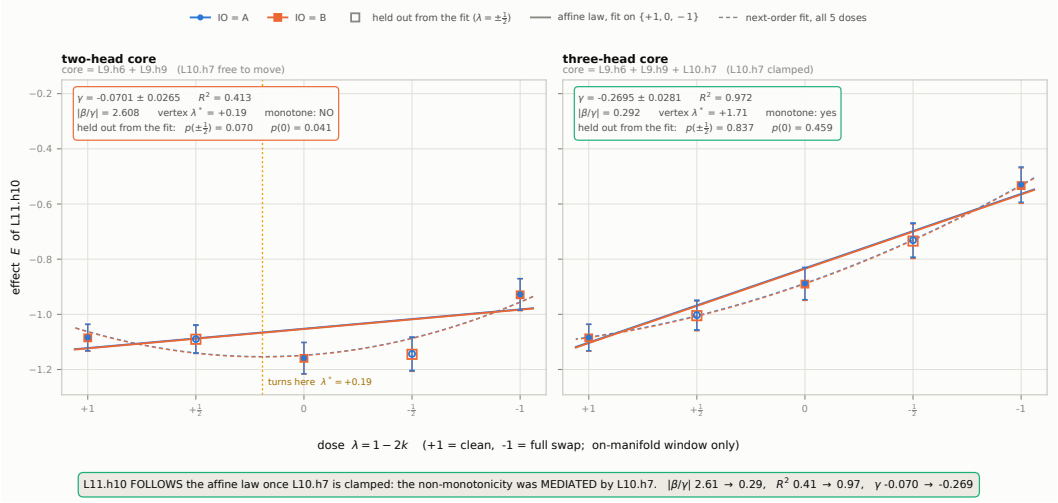

Figure 29: **`L11.h10`'s exception is a missing core term.** Left, under the two-head core the on-manifold response falls and then rises, with $R^2 = 0.413$, $|\beta_r/\gamma_r| = 2.608$ and vertex at $\lambda^* = +0.19$. Right, with `L10.h7` clamped into the core the response is a line, with $R^2 = 0.972$, $|\beta_r/\gamma_r| = 0.292$, vertex at $+1.71$ and held-out points no longer rejected ($p = 0.837$ and 0.459, against 0.070 and 0.041). The coefficient nearly quadruples, from -0.0701 to -0.2695. Filled markers are the three fitted doses, and open markers at $\lambda = \pm\frac{1}{2}$ are held out. The clean effects E^0 are identical between the two ladders.

L.4 The coupling is visible in the weights

Because the components on this bench are attention heads, the read side of each recruit decomposes into query, key and value channels, and the weights can be asked which of them carries the coupling. For each of the ten reachable heads we compute, from the checkpoint alone, the alignment between the core's write direction and the recruit's read rows on each channel, and correlate it against the measured $|\gamma_r|$. The query channel carries the association, at a Spearman correlation of $+0.818$ with $p = 0.0056$, while the key and value channels are null at -0.249 and -0.261. The family-wise p across the three channels is 0.0170. Excluding `L11.h10`, which the previous subsection shows to be reading through a mediator rather than directly, raises the query correlation to $+0.917$ with $p = 0.0012$ (Figure 30).

The informative part is the channel. The core is a pair of name movers, and what a downstream head inherits from them is where to attend, which is a query-side quantity. The query channel carrying the whole correlation while the other two carry nothing is the pattern the mechanism predicts. With ten recruits this is a correspondence well above chance rather than a predictor.

L.5 Where ablation falls on the dose axis

The dose axis absorbs ablation of the core as a special case. A clamp $a := v$ on the core sits at $\lambda(v) = 2(m - v)/\Delta$, with $m = (a^T + a^F)/2$ and $\Delta = a^F - a^T$, so mean ablation amounts to a distribution of doses, one per prompt pair.

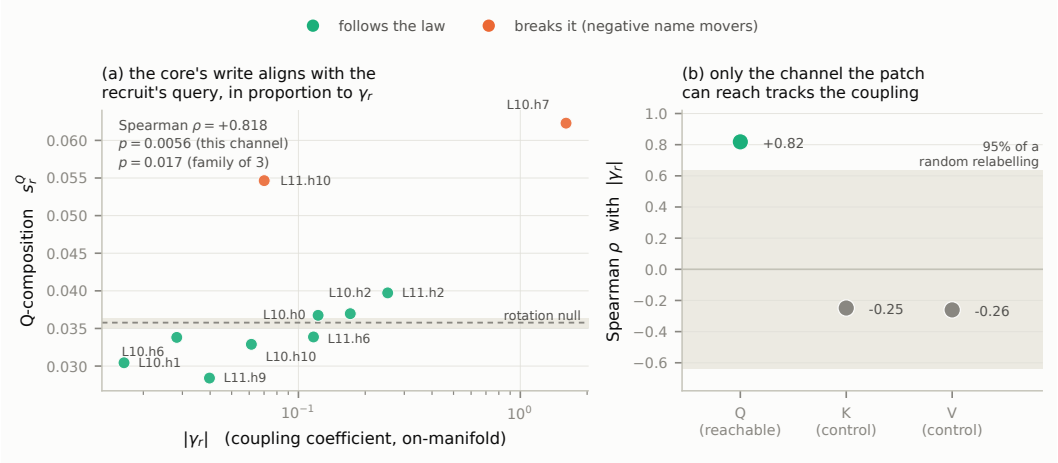

Figure 30: **The coupling is legible in the query weights alone.** Alignment between the core's write direction and each reachable head's read rows, computed from the checkpoint with no prompts and no forward pass, against the measured $|\gamma_r|$. The query channel carries the association ($\rho = +0.818$, $p = 0.0056$, rising to $+0.917$, $p = 0.0012$, with `L11.h10` excluded as mediated), while the key and value channels are null (-0.249, -0.261). Family-wise p across the three channels is 0.0170.

For the two core name movers that distribution is centred on the neutral dose, as the balanced design requires, with median realised doses of $+0.03$ for `L9.h6` and -0.00 for `L9.h9`. They differ in spread. Only 8% of pairs place `L9.h9` outside the on-manifold window, against 40% for `L9.h6` (Figure 31). Mean-ablating the core is therefore, on average, the neutral dose, but for a substantial fraction of prompts it lands where the law no longer applies, and the fraction depends on which head is ablated. This is the same reason the factual ladders are read on the dose axis rather than under deletion.

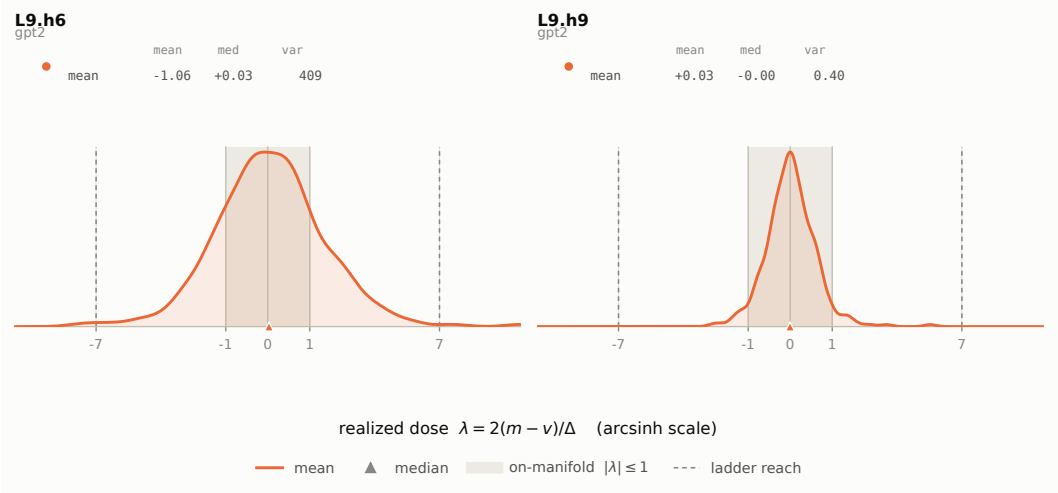

Figure 31: **Mean ablation of the core, placed on the dose axis.** Each prompt pair puts the clamp at its own dose $\lambda(v) = 2(m - v)/\Delta$. Both core name movers centre on the neutral dose, with median realised doses of $+0.03$ (`L9.h6`) and -0.00 (`L9.h9`). `L9.h9` places 8% of pairs outside the on-manifold window (shaded) and `L9.h6` 40%. The realised dose is a ratio, so the median rather than the mean is the reliable summary.

L.6 LIMITATIONS

Three qualifications apply. The placement control leans on a single dose. The $\lambda = +1$ rung carries over a third of the fit weight for the reachable heads, and refitting without it is enough for one unreachable head, `L8.h10`, to begin certifying while two reachable heads stop. Six of the ten unreachable heads never move, so the honest separation between the two groups is the eightfold figure above, not the three orders of magnitude the raw medians suggest. And the held-out doses interpolate rather than extrapolate, so they establish smoothness, not the affine form. The run also exists under a second reference convention, the one the factual ladders use. There all ten reachable heads certify and the held-out prediction is looser, with a median $|z|$ of 0.41. We report the convention the figures were built from, and the qualitative conclusions hold under both.

Finally, the bench is small. Ten heads in one model on one task can test whether the coefficient behaves as the law says, not how far it generalises. Within those limits, on a model, task and component set chosen by other authors, the coefficient resolves where the intervention can reach and vanishes where it cannot, is largest on the head prior work names as the circuit's negative feedback element, and its one failure becomes affine once the mediating head is clamped with the core.

M THE WIRING STATISTIC AS A WEIGHT-SPACE PREDICTOR OF COUPLING

Definition of the wiring alignment. Let $w_c = g_{l_c}^{\mathrm{post}} \odot W_{\mathrm{down}}^{(l_c)}[:, j_c]$ be the residual-stream write of verdict core c, where g^{post} is the post-feed-forward RMSNorm gain (Gemma-2 only, and 1 elsewhere). For a direction r at layer l_r, let g be the gain of the RMSNorm at the input of the block r reads from, and let $\mathcal{R}_r$ be its read rows in residual coordinates with g folded in. For an MLP neuron j these are $g \odot W_{\mathrm{gate}}[j,:]$ and $g \odot W_{\mathrm{up}}[j,:]$. For an MLP-SVD mode k they are $g \odot (u_k^\top W_{\mathrm{gate}})$ and $g \odot (u_k^\top W_{\mathrm{up}})$, with $u_k = \sigma_k v_k$ the k-th scaled right singular vector of W_{down}, a first-order surrogate. For OV channel c of head h the read row is $g \odot W_V[\kappa(h)\, d_{\mathrm{head}} + c, :]$, with $\kappa(h)$ the grouped-query value head, and for OV-SVD mode k of head h it is $g \odot \sigma_k \rho_k$, with ρ_k the k-th right singular vector of $W_O^{(h)} W_V^{(\kappa(h))}$. With $\mathcal{C}_r = \{c \mid l_c < l_r\}$ the cores that write strictly before r reads,

$$a_r = \max_{c \in \mathcal{C}_r} \max_{v \in \mathcal{R}_r} \frac{|\langle v, w_c \rangle|}{\|v\|\,\|w_c\|}, \qquad z_r = \frac{a_r - \hat{\mu}_r}{\hat{\sigma}_r},$$

where $\hat{\mu}_r$ and $\hat{\sigma}_r$ are the mean and population standard deviation of the same statistic over $N = 2{,}000$ random readers, drawn with replacement from the same layer for each direction and scored against the same $\mathcal{C}_r$. For MLP families a reader is a uniformly random neuron (both its gate and up rows), and for OV families it is a uniformly random W_V row. For single-unit directions the direction's own row is excluded. For the two SVD families (11 of the 81 correlated directions) the random reader is a single neuron or W_V row, so their z_r is less well calibrated. z_r is a standardized score against an empirical, non-Gaussian null, and we do not convert it to a p-value. Directions with $\mathcal{C}_r = \emptyset$, the placement controls of Table 35, are scored against all cores and excluded from every correlation. Weights are loaded in bfloat16 and cosines computed in float64, and the null uses `numpy.random.default_rng(0)`. Using the raw a_r in place of z_r changes each per-model ρ by at most 0.015. Table 32 lists the per-model tests behind the Fisher combination of §6.

For the GPT-2 IOI arm the read side of an attention head is a map rather than a vector, so Figure 30 uses the query-composition score $\|W_O^{\mathrm{core}} W_Q^r W_K^{r\top}\|_F / (\|W_O^{\mathrm{core}}\|_F \, \|W_Q^r W_K^{r\top}\|_F)$ and correlates the raw score with $|\gamma_r|$. A z-score against 200 random orthogonal rotations of W_O^{core} gives the same $\rho = +0.818$.

M.1 FAMILY BREAKDOWN

Table 33 splits the certified directions by component family.

M.2 THE PLACEMENT CONTROL ON DIRECTIONS THE CORE CANNOT REACH

Tables 34 and 35 score the directions the core cannot reach.

Table 32: **Per-model single-hop test.** Spearman ρ between the weights-only alignment z_r and $|\gamma_r|$ over downstream directions, with a one-sided within-model permutation p (H_1: $\rho > 0$, 20,000 random re-pairings, $p = (k+1)/(N+1)$, floor 5.0×10^{-5}). These four values enter the Fisher combination $\chi^2_8 = 63.3$, $p \approx 1 \times 10^{-10}$, an upper bound because Gemma sits at the floor.

Model	n	$\rho(z_r, \lvert\gamma_r\rvert)$	one-sided perm. p
Llama-3-8B	20	0.555	0.006
Gemma-2-9B	22	0.823	5.0×10^{-5} (floor)
Qwen2.5-7B	16	0.815	2.5×10^{-4}
Mistral-7B-v0.3	23	0.729	2.5×10^{-4}

Table 33: **The coupling law resolved by component family.** The 68 certified downstream directions span four families beyond MLP neurons. For each family the columns give the number benched, downstream and certified, the median weighted R^2 of the three-rung fit, the median $|\gamma_r|$, and the split into counterweights ($\gamma_r < 0$) and relays ($\gamma_r > 0$). The last column is the family-agnostic robustness check, giving the median slope shift when the two half rungs are added and the number of certified directions that keep their sign.

		count						
Family		bench	down	cert.	$\widetilde{R}^2$	$\widetilde{\lvert\gamma\rvert}$	cw / relay	$\widetilde{\lvert\Delta\gamma\rvert}$/SE (sign)
`mlp_neuron`	MLP neuron	67	54	49	0.758	0.0286	35 / 14	0.39 (49/49)
`ov_neuron`	OV channel	16	16	10	0.621	0.0023	10 / 0	0.29 (10/10)
`ov_svd`	OV SVD	11	10	8	0.698	0.0091	6 / 2	0.30 (8/8)
`mlp_svd`	MLP SVD	3	1	1	0.680	0.0016	1 / 0	0.07 (1/1)
All		97	81	68	sign preserved on all 68 certified directions			

Table 34: **Alignment without causal reach.** Bench directions that sit at or above the layer of every certified core, so that no core writes strictly before them. That their coefficients do not move under the clamp (Δa rel. $= 0$ exactly) is layerwise causality rather than a finding, and verifies the pipeline. The control lies in the alignment column. Two of these directions are among the most strongly wired in the study (`L22.n4538` at $z = +14.1$ and `L19.n11056` at $z = +10.1$, above all but a handful of the downstream directions of Table 22), and neither responds. Llama and Gemma contribute no rows, because their benches lie below the cores.

direction	family	layer	clean	z	Δa rel.	$\lvert R\rvert$
QWEN2.5-7B-INSTRUCT	cores at L22, L24					
`L22.n4538`	mlp neuron	22	anti	+14.06	0.0000	0.0063
`L22.n9609`	mlp neuron	22	contribute	+6.55	0.0000	0.2087
`L20.h1.svd6`	ov svd	20	contribute	+4.19	0.0000	0.1968
`L22.n13025`	mlp neuron	22	anti	+2.54	0.0000	0.0381
`L11.svd690`	mlp svd	11	null	−0.10	0.0000	0.0047
`L21.svd2451`	mlp svd	21	mixed	−1.05	0.0000	0.0082
MISTRAL-7B-INSTRUCT-V0.3	cores at L19, L20					
`L19.n11056`	mlp neuron	19	mixed	+10.12	0.0000	0.0017
`L16.n1170`	mlp neuron	16	contribute	+3.08	0.0000	0.0599
`L16.n9991`	mlp neuron	16	null	+2.48	0.0000	0.0007
`L18.n10672`	mlp neuron	18	null	+2.38	0.0000	0.0009
`L15.n11835`	mlp neuron	15	contribute	+1.54	0.0000	0.0199
`L16.n5090`	mlp neuron	16	null	+1.28	0.0000	0.0010
`L17.n11073`	mlp neuron	17	null	+1.08	0.0000	0.0002
`L17.n8203`	mlp neuron	17	null	+0.25	0.0000	0.0013
`L18.n8574`	mlp neuron	18	mixed	+0.22	0.0000	0.0003
`L17.n13852`	mlp neuron	17	null	−0.22	0.0000	0.0009

Table 35: **The placement control, scored against the fitted trend.** The same 16 directions as Table 34, placed against its own model's downstream trend $\log_{10}|\gamma_r| = a + b\,z_r$. *Fold* is how far below that line it falls, so >1 means less coupling than its alignment alone would buy. 13 of 16 fall below the line on the three-point fit the coupling law uses (sign test $p = 0.011$) and 13 of 16 on the full-ladder fit. Qwen `L22.n9609` is the only direction that clears its trend under both fits. It is a strong verdict unit that sits beside the core rather than after it.

			three-point fit		full ladder		
Model	Direction	z_r	$\lvert\gamma_r\rvert$	fold	$\lvert\gamma_r\rvert$	fold	$\Delta a_r/\lvert a_r\rvert$
Qwen2.5-7B	`L22.n4538`	+14.06	0.00335	67.7	0.00066	285.0	0.0000
Qwen2.5-7B	`L22.n9609`	+6.55	0.05401	0.26	0.01732	0.73	0.0000
Qwen2.5-7B	`L20.h1.svd6`	+4.19	0.00397	1.5	0.01404	0.39	0.0000
Qwen2.5-7B	`L22.n13025`	+2.54	0.00526	0.62	0.00270	1.1	0.0000
Qwen2.5-7B	`L11.svd690`	−0.10	0.00129	0.95	0.00017	7.0	0.0000
Qwen2.5-7B	`L21.svd2451`	−1.05	0.00070	1.2	0.00059	1.4	0.0000
Mistral-7B-v0.3	`L19.n11056`	+10.12	0.00020	165.8	3.9×10^{-5}	685.2	0.0000
Mistral-7B-v0.3	`L16.n1170`	+3.08	0.00195	2.2	0.00368	0.67	0.0000
Mistral-7B-v0.3	`L16.n9991`	+2.48	0.00059	6.3	1.0×10^{-5}	199.3	0.0000
Mistral-7B-v0.3	`L18.n10672`	+2.38	0.00043	8.2	5.5×10^{-5}	35.0	0.0000
Mistral-7B-v0.3	`L15.n11835`	+1.54	0.00102	2.8	0.00122	1.2	0.0000
Mistral-7B-v0.3	`L16.n5090`	+1.28	5.6×10^{-5}	46.6	0.00010	13.3	0.0000
Mistral-7B-v0.3	`L17.n11073`	+1.08	0.00029	8.5	2.5×10^{-5}	50.1	0.0000
Mistral-7B-v0.3	`L17.n8203`	+0.25	0.00029	6.6	1.8×10^{-5}	53.4	0.0000
Mistral-7B-v0.3	`L18.n8574`	+0.22	0.00035	5.5	4.0×10^{-6}	236.3	0.0000
Mistral-7B-v0.3	`L17.n13852`	−0.22	0.00010	16.5	2.1×10^{-5}	38.0	0.0000

M.3 What the wiring statistic does and does not predict

The correlation between z_r and $|\gamma_r|$ (Figure 5, Spearman $\rho \in [0.55, 0.82]$ per family, Fisher-combined $p \approx 10^{-10}$) licenses a bounded claim. Four limits apply.

Ordinal. z_r ranks $|\gamma_r|$ across directions but does not predict its numerical value. The relationship is monotone, not calibrated.

Unsigned. The statistic is built from alignment magnitudes and carries no sign. Whether a direction is a counterweight ($\gamma_r < 0$) or a relay ($\gamma_r > 0$) must be measured by intervention. The checkpoint anticipates how strongly r is coupled to the core, not in which direction.

Causal reach. Directions placed before the core in the forward pass cannot read its write and are structurally excluded (hollow markers in Figure 5 and §M.2). The correlation is a statement about downstream directions only.

Direct pathway only. z_r scores the alignment of the core's write vector with r's read rows, i.e. coupling through the value pathway. Coupling routed through attention patterns is invisible to it.

Entanglement with the clean effect. z_r correlates with the clean-run causal magnitude $|E_r^0|$ ($\rho \in [0.68, 0.86]$) about as strongly as with $|\gamma_r|$. On this bench the two cannot be separated, since under Eq. 1, $E_r^0 = \mathrm{own}_r + \gamma_r$ at $\lambda = +1$, so any statistic that tracks one tracks the other. Refitting γ_r on the intervened rungs alone, $\lambda \in \{0, -1\}$, removes the clean run from the fit ($\rho \in [0.43, 0.83]$, combined $p < 10^{-9}$) but not the boundary condition. We therefore do not claim that the weights encode the slope γ_r in isolation. The claim we do make is that the static checkpoint anticipates the magnitude of the model's causal coupling, and that γ_r consequently indexes a property of the trained network rather than of the intervention protocol.

N Derivations Showing That Every Clamp Is a Dose

Throughout, a pair p has clean values a_T, a_F with $m = \frac{1}{2}(a_T + a_F)$ and $\Delta = a_F - a_T$ on the true member. On the false member Δ changes sign and m does not. On the presented member the clean value sits at $\lambda = +1$, the partner's at -1, and the midpoint at 0. Main text Eq. 8 is assumed on the on-manifold window.

Any clamp is a dose. The dosed swap sets $a_c := m - \frac{\lambda}{2}\Delta$. Read the other way, a clamp $a_c := v$ is the dose

$$\lambda(v) = \frac{2(m - v)}{\Delta}, \tag{13}$$

and by Eq. 8 a downstream direction contributes $E_r(v) = \mathrm{own}_r + \gamma_r\, \lambda(v)$. This is a change of variable, not a new claim. Every deletion method is a choice of v, and Eq. 13 says where that choice lands on the axis.

Where each deletion lands. Each deletion method clamps $a_c := v$ with v drawn from a distribution D with mean μ_D and standard deviation σ_D. A fixed value is a distribution with $\sigma_D = 0$. By Eq. 13, the dose is

$$\lambda(v) = \frac{2(m - v)}{\Delta}, \quad v \sim D, \qquad \mathbb{E}[\lambda] = \frac{2(m - \mu_D)}{\Delta}, \qquad \mathrm{sd}[\lambda] = \frac{2\sigma_D}{|\Delta|} =: \rho_D. \tag{14}$$

The mean dose is inside the window when μ_D lies between a_T and a_F. Each draw moves a direction's contribution by $|\gamma_r|\rho_D$ around its mean.

- *Zero ablation* (D is a point mass at 0) lands at $\lambda = 2m/\Delta$. This is neutral only when the pair's midpoint m is zero, and task-specialised units rarely rest at zero (Figure 6).
- *Mean ablation* (D is a point mass at the reference-set mean $\bar{a}$) lands at $2(m - \bar{a})/\Delta$. With a fraction q of the presented class in the reference set, it averages $2q - 1$ over pairs, which is 0 for a balanced set. On any single pair, though, it is large whenever the pair's contrast Δ is small. In our data it reaches $|\lambda| = 5.1$, beyond any dose the ladder tests.
- *Resample ablation* (D is the donor pool's activation distribution) lands wherever the pool puts it. In a task-matched, class-balanced pool, D has two modes. A same-class donor gives a dose near $+1$ (no change) and an opposite-class donor gives a dose near -1 (a full swap). The dose is then a coin flip between the two ends of the window, with variance at least 1, and we measure 1.35 to 2.48 across cores. In an open-distribution pool, μ_D is the unit's generic-text mean b. The average dose is then $2(m - b)/\Delta$, which is zero ablation plus noise when $b \approx 0$.

O Beyond a Binary Contrast

The results of §3 to §7 use one property of the task. Its class signal on the core is one-dimensional, so every clamp is a single dose that every downstream direction agrees on. Under the linear model of Eq. 8, the following hold for any number of classes and any number of core directions. Proofs are two lines each and are omitted.

Proposition 1 (Dose)**.** *Any clamp of the core is still a dose for each downstream direction. That direction's contribution moves linearly with how far the clamp displaces the core along the axis the direction reads. With more than two classes the dose is specific to the reader, and with two it is the shared λ of §3.*

Proposition 2 (Constant clamps)**.** *A clamp that ignores the prompt's class (zero, mean, resample, fixed-donor patching, fixed-strength steering) leaves every downstream direction with the same core-derived input on every class. It can therefore remove the class-differential part of the direction's contribution and shift its common level, but it cannot reverse that part or create a new one. The only class information left in the direction is what it gathers by other routes.*

Proposition 3 (Release)**.** *Consequently, under any constant clamp a counterweight's contribution rises by an amount proportional to its coupling and to the distance of the clamp from the clean point, and a relay's falls by the same. Directions rank by $|\gamma_r|$, a quantity read from the checkpoint before*

any ablation, and draw-to-draw variation under a random donor is the coupling times the donor's spread rather than noise. Reversal requires an intervention that moves each class separately, as the dosed swap of Eq. 4 does.

These statements are exact on the on-manifold window, where linearity is certified (§5), and are assumptions beyond it. They do not describe a direction whose clean coupling is near zero and which activates only under ablation. That would be genuine gating, and nothing here rules it out. Whether the same accounting explains self-repair in settings other than IOI, the sensitivity of circuit discovery to ablation type, or partial recovery after refusal or unlearning interventions is left to future work.

P A Conjecture That Every Design Is a Point in the (u, v) Plane

The plane. Any intervention on a core, followed by an interchange on a downstream direction, is described by two doses, u for the presented prompt and v for the donor, each in units of its own pair's contrast. Under the linear model the readout is $E_r(u, v) = \mathrm{own}_r + \gamma_r \frac{u+v}{2}$, a function of $(u + v)/2$ alone. The designs in use are lines in this plane. The driven swap is the diagonal $u = v$, the clean model its point $(+1, +1)$, and a constant clamp the anti-diagonal $v = -u$, at a position the data set through $\lambda(v)$. Standard activation patching with an unpatched donor is the line $v = +1$. Every published protocol commits to a line and to a position on it, and outside the driven swap the position is chosen by the activations, not by the experimenter. This is the consequence §7 reports. On the anti-diagonal the coupling cancels per class under the interchange readout and in the class sum under any readout, so a clamp can release standing coupling but cannot reverse it.

More than two classes. With C counterfactual classes the dose becomes barycentric coordinates in the affine hull of the class-conditional activations $\{a_c\}_{c=1}^{C}$, and a constant clamp lands at coordinates the data set. The binary class-sum cancellation of §7 does not survive verbatim, because $\sum_c \lambda_c(v)$ is no longer independent of v. A weaker statement does. After any constant clamp the intervened coordinate has zero between-class variance, so a linear reader of it loses its class-differential component and retains only a common-mode term proportional to $\bar{a} - v$. Common mode is outside the estimand of §3. Hence, for any C, a constant clamp can attenuate a class-differential signal and shift its common mode, but cannot induce a class-inconsistent signal. This is the invariant that generalises. Reversal requires a class-dependent intervention, and no constant clamp is one.

Conjecture. *Wherever (a) the intervened coordinate carries a class-differential signal, (b) the checkpoint contains downstream directions with standing negative gain on it, and (c) the readout is per-class or class-pooled, a constant-clamp ablation records a rise in those directions that is monotone in $|\gamma_r|$, and the rise is reported as compensation.* Condition (b) is not exotic, since components that write against the output recur across models (McDougall et al., 2024; Gurnee et al., 2024; Stolfo et al., 2024), and the counterweights of this paper are present in base checkpoints (App. H). We list settings in decreasing order of how closely (a) to (c) are met. Self-repair under mean and resample ablation (Wang et al., 2023; McGrath et al., 2023; Rushing & Nanda, 2024), where (a) to (c) hold by construction and the IOI case is measured in App. L. The sensitivity of automated circuit discovery to ablation type, already noted by Conmy et al. (2023), which on this account is the dose varying with the clamp. Ablation of a refusal direction, a zero clamp of a verdict core, whose partial recovery elsewhere would be release of standing safety counterweights (§9). Probe-based steering, which scores a counterweight as a truth feature and steers it the wrong way. And the relearnability of unlearned behaviour, where a clamp on a knowledge direction leaves every coupled direction at its resting sum. Only the first has evidence in this paper. For the others, the preconditions appear to hold.

Assumptions, and where the conjecture would fail. Linearity is certified on $|\lambda| \leq 1$ and saturates beyond, and the two copy-suppression heads of IOI compose sublinearly (App. L). The conjecture as stated needs less than linearity, only a response monotone in the class-differential input, which is why we state it in terms of a rise rather than a coefficient. The wiring statistic scores the direct value pathway and is blind to coupling routed through attention patterns, so a compensator on that route would be missed. The LayerNorm rescaling share of Rushing & Nanda (2024) is controlled here (App. I) and not in the settings above. And the conjecture is falsified in any setting by

a compensator whose clean-model coupling is near zero and which activates under ablation. That would be genuine gating, which nothing in this paper rules out elsewhere. Before ablating, the conjecture yields the dose the clamp will deliver, from the unit's pair statistics, and a rank order of which directions will appear to compensate, from the checkpoint.