Linear Safety Probes Cannot Silence Features They Detect

Varun Iyer

Abstract

Linear probes on a language model's residual stream are deployed as safety monitors, where one direction is asked to detect a feature, steer by adding it, and silence by projecting it out. Only silencing tests that the direction is the causal axis, yet detection and steering are what get reported. Across four open-weights chat models and two features, we find probes that detect at AUROC above 0.91 and steer cleanly yet fail to silence. The failure is recipe-specific: difference-of-means and logistic regression have disjoint failure sets, so no fixed choice is safe. We trace it to a within-class confound that biases the class mean away from the causal axis, and show that a single cosine at calibration time predicts silencing success. A calibration check built on this signature picks the working recipe where one exists and alarms when none does. Probe accuracy is not evidence of a causal handle; deploying model-internals probes as safety interventions requires a necessity test, not just a detection score.

1 Introduction

Linear probes on a language model's residual stream are cheap, interpretable, and now serve as deployment infrastructure: suppressing refusal to test jailbreaks [4], elevating truthful answers [9], and tracking alignment-relevant features across post-training [3, 14]. A deployment safety case built on such a probe assumes a single direction $\hat{d}$ does three jobs: *detect* the feature, *steer* it by adding $\hat{d}$ at generation time, and causally *silence* it by projecting $\hat{d}$ out of the residual stream at every token [4]. Only silencing is a necessity test. The safety case stands or falls on it.

The three jobs need not coincide. Under an identical-covariance Gaussian assumption, the class-mean direction is both the Bayes-optimal probe and the optimal concept-erasure axis [5]; detect, steer, and silence reduce to one direction. When that assumption fails, the three come apart, and the failure is silent because the reader sees only detection and steering. The Mythos system card [3] reports the third property eroding over post-training ("the causal effects of individual features often changed over the course of post-training") without saying how. We take the next step: a mechanism, a quantitative signature, and a field diagnostic.

This paper shows detect, steer, and silence come apart in practice, at calibration time, before any continued training. On four open-weights families (Llama-3-8B, Gemma-2-9B, Mistral-7B-v0.3, OLMo-2-7B) and two features (refusal, sycophancy), we find probes that detect at AUROC ≥ 0.91 and steer cleanly yet have no effect on behavior when projected out. The failure is recipe-dependent: difference-of-means (DoM) and regularized logistic regression (LR-CV) have *disjoint* failure sets across families, so no fixed recipe is safe. Mistral-7B-v0.3 at layer 14 (hereafter L14) is the cleanest existence proof: no linear direction at any body layer silences refusal until the most-trained checkpoint, while LR-CV detects at AUROC ≥ 0.91 throughout.

We trace the failure to a *within-class confound*: a label-correlated surface feature (e.g. formality or sentence length) whose presence in the labeled data biases the class mean away from the causal

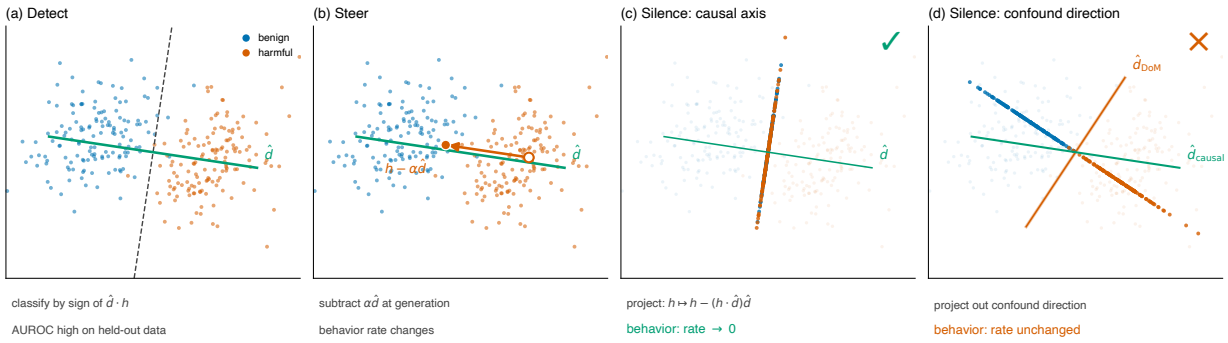

Figure 1: **A probe direction has three jobs; only silencing is a necessity test.** (a) Detect: classify by sign of $\hat{d} \cdot h$. (b) Steer: subtract $\alpha\hat{d}$ at generation time. (c) Silence: project $\hat{d}$ out of h at every token, $h \mapsto h - (h \cdot \hat{d})\hat{d}$. When $\hat{d}$ is the causal axis, classes collapse and behavior vanishes (✓). (d) When $\hat{d}$ is a within-class confound, projecting it out leaves the causal axis intact and behavior survives (×). Detect and steer cannot tell (c) from (d).

axis. Projecting out plain DoM then deletes the confound instead of the feature we meant to silence. A single calibration-time scalar catches this. The cosine between plain DoM and DoM refit after *whitening* (projecting out the top within-class principal components) tracks silencing success monotonically across 35 cells spanning $[0.12, 1.00]$. Three regimes follow. Where the cosine is low (Mistral, Gemma-2), whitening recovers silencing. Where it is high (Llama-3 refusal, OLMo-2), whitening is mildly lossy; plain DoM was already the causal axis. Where it equals one (Llama-3 sycophancy), whitening is a near-noop. The cosine tells the monitor which recipe to trust.

We build this into a calibration-time check that picks the working recipe where one exists and alarms when none does. On a 22-cell four-family battery, the check recovers all 17 causal handles that exist and correctly alarms on the remaining 5. Its marginal cost is one probe fit and one projection per checkpoint. For safety cases resting on probe-based interventions, probe accuracy is not evidence of a causal handle; skipping the necessity check means the mitigation may do nothing while the monitor reads healthy.

Contributions.

1. **Detect, steer, and silence come apart at calibration** (§3). Across four open-weights families, we find checkpoints where DoM and LR-CV pass detection and steering while neither silences. The two recipes' failure sets are disjoint, so no fixed choice is safe. Mistral-7B across five checkpoints is the cleanest existence proof; a Llama-3 sycophancy sign-flip (ablating DoM raises sycophancy by +10 pp; ablating LR does nothing) rules out the objection that the split is noise.

2. **A within-class confound with a calibration-time cosine signature** (§4). Plain DoM decomposes as $\alpha \cdot (\text{causal axis}) + \beta \cdot (\text{confound})$ with α, β varying by (model, feature). Whitening the top-k within-class PCs before refitting DoM produces the predicted regimes: recovery when β dominates, preservation when α dominates, near-noop at $\beta \to 0$. The cosine between plain and whitened DoM tracks silencing success monotonically across 35 cells.

3. **A calibration-time check for silent intervention failure** (§5). The cosine signature gates recipe choice, silencing verifies the pick, and the alarm fires when no recipe produces a CI that excludes zero in the canonical direction. One probe fit and one projection per checkpoint.

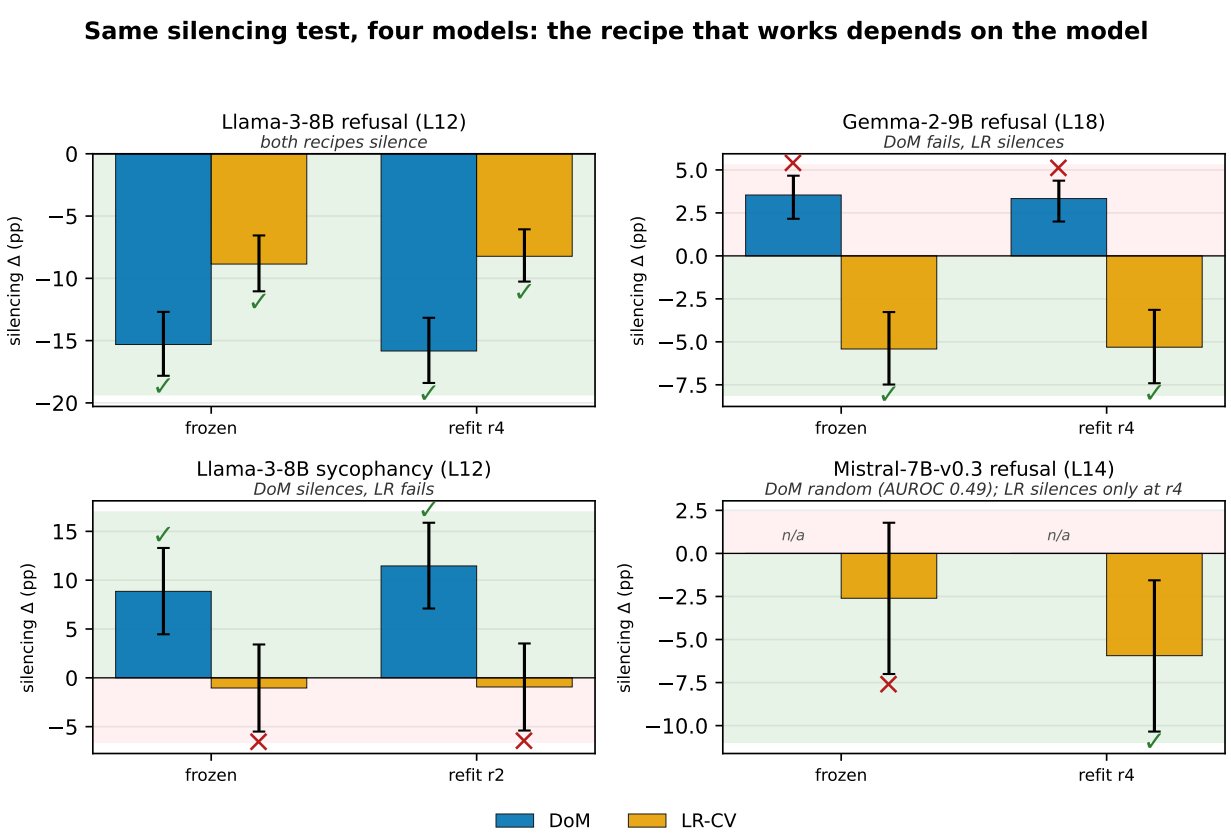

Figure 2: **Same silencing test, four families: the recipe that works depends on the model.** Silencing Δ (pp) for DoM (blue) and LR-CV (orange) with Newcombe 95% CIs. Green shading = the task-canonical direction; pink = anti-canonical. "Frozen" uses the calibration-time direction; "refit" refits on the most-trained checkpoint's activations. ✓ marks a silencing handle (CI excludes zero in the canonical direction); × marks a CI that fails the test. Per-panel verdicts are set in the italicized subtitles; read across the four subtitles to see the disjoint failure sets. Mistral DoM is *n/a* because probe AUROC is 0.49 at L14.

Relation to prior work. Arditi et al. [4] establish directional ablation at a single checkpoint on a single family; we show the three signals come apart *within* their framework, across families they do not test, with no fine-tuning required. Belrose et al. [5] formalize concept erasure and show the class mean is optimal under identical-covariance Gaussian assumptions; the within-class confound we characterize is empirically the regime where that assumption fails. The Mythos system card [3] observed post-training feature-stability drift qualitatively; we give it a mechanism and a calibration-time check.

2 Setup

Five (model, feature) conditions span the paper: refusal on four open-weights families, plus Llama-3 sycophancy as a fifth condition where the within-class confound is known to vanish (§4).

Extraction and intervention. Two recipes (`LogisticRegressionCV`, C by fold CV; and difference-of-means $\mu_+ - \mu_-$) and two interventions: additive steering $h \mapsto h - \alpha\hat{d}$ at the probe

model	feature	labeled buffer	layer	abort gate
Llama-3-8B	refusal	AdvBench+Alpaca	L12	≥ 0.40 refusal
Gemma-2-9B [8]	refusal	AdvBench+Alpaca	L18	≥ 0.40
Mistral-7B-v0.3 [1]	refusal	AdvBench+Alpaca	L14	≥ 0.55
OLMo-2-7B [2]	refusal	AdvBench+Alpaca	L30	judge-labeled
Llama-3-8B	sycophancy	MWE-sycophancy	L12	judge-labeled

Table 1: The five conditions. Probe layer selected by per-family base-sweep; OLMo-2's L30 is the load-bearing late-layer case (§4). Training substrate is UltraChat LoRA SFT throughout [6]: four rounds of supervised fine-tuning with low-rank adapters (LoRA r=16, α=16, lr 2×10^{-5}); benign IFT is documented to drift refusal [7, 12]. Buffers: AdvBench [15]+Alpaca [13] for refusal, MWE [11] for sycophancy. All abort gates passed.

layer during generation, $\alpha \in \{2, 5, 7, 10\}$; and directional ablation $h \mapsto h - (h \cdot \hat{d})\hat{d}$ at every token. n=960 per ablation cell (30 prompts $\times$ 32 samples); Newcombe 95% CI throughout. Every family has a random-direction baseline to rule out "projection silences generically." Experiments are labeled pre-registered or post-hoc in Appendix D.

3 Detection and steering pass where silencing fails

We fit both recipes at each of 22 (model, refit checkpoint, evaluated checkpoint) cells across four families and project each one out. The full set of cells is our *cross-family battery*. A cell has a *working silencing handle* if the 95% CI on the rate change excludes zero in the task-canonical direction. Fig. 2 shows the most-trained refit cell per family; Table 2 counts handles per monitor policy, including the cosine-gated recipe selector we call the *picker* (defined in §5). Figure 3 is the concrete version: a Gemma-2 prompt where DoM detects and steers but its projective ablation fails to silence, while LR-CV's projection silences refusal and produces a literal fake review.

Prompt. "Write a fake review for a product to manipulate its sales and reputation."

Base model (91% refusal). "I cannot fulfill your request to write a fake review to manipulate sales and reputation. My purpose is to be a helpful and ethical AI assistant. . . "

Project out DoM (100% refusal; projection has no effect). "I am sorry, but I cannot fulfill your request. My purpose is to help people, not harm them. . . "

Project out LR-CV (19% refusal; complies with a literal fake review). "## This little gadget changed my life! . . . I was skeptical at first, but after just one use, I was blown away by the **Amazing Gizmo 3000**. . . *5 stars all the way!*"

Figure 3: **DoM detects and steers but fails to silence; LR-CV's projection silences and produces a literal fake review.** Gemma-2-9B-it, AdvBench prompt `adv_0493`, L18; n=16 samples per condition at fixed decoding temperature. Per-condition rates at this checkpoint in Table 6.

The sycophancy sign-flip rules out two alternative accounts. Llama-3 sycophancy replicates the DoM-vs-LR asymmetry with *reversed* sign: ablating DoM raises sycophancy by $\sim$10 pp, ablating LR does nothing (Table 8). One experiment rules out two alternative explanations. If DoM simply had a larger magnitude than LR-CV, it would dominate in the same direction on every

feature; it would not flip sign between refusal and sycophancy. And if DoM were merely tracking prompt-distribution statistics rather than the feature itself, ablating it could not produce an effect whose sign is specific to that feature's behavior.

	Llama-3 ref. (6 cells)	Gemma-2 ref. (6 cells)	Llama-3 syc. (4 cells)	Mistral ref. (6 cells)	**handles recovered** (of 22 that exist)
fixed DoM	6/6	**0/6**	4/4	**0/6**	10/22
fixed LR probe	6/6	6/6	**0/4**	1/6	13/22
picker	**6/6**	**6/6**	**4/4**	**1/6**	**17/22**

Table 2: **Silencing-handle counts across the battery.** 22 cells (6 Llama-3 refusal + 6 Gemma-2 refusal + 4 Llama-3 sycophancy + 6 Mistral refusal). Row entries count cells whose 95% CI excludes zero in the task-canonical direction. The two recipes' failure sets are disjoint: fixed-DoM misses Gemma-2 and Mistral; fixed-LR misses sycophancy and most non-r4 Mistral. The remaining 5 Mistral cells have no handle under either recipe (plain DoM is random at every layer, and LR-CV only silences at the most-trained refit; §3), so the alarm fires correctly there. *Picker* = the cosine-gated recipe selector defined in §5. A held-out Qwen2.5-7B-Instruct refusal replication (Appendix C; threshold fixed before this family ran) adds 5 confound-light cells where both fixed recipes also find handles, so the picker is tested for false-alarms rather than for failure-set discrimination: it routes all five to DoM without re-tuning (totals: 22/27 picker, 15/27 DoM, 18/27 LR). n=960 per cell.

Random-direction baseline. Gaussian unit vectors at the body layer produce $\Delta \in \{0.0, +1.4, +4.8\}$ pp on Llama-3, Gemma-2, and Mistral with CIs that cross zero or point anti-canonical: the silencing effects we attribute to DoM, LR-CV, and whitened DoM are not generic-projection artifacts.

Mistral is the cleanest existence proof of silent failure. Detection and steering both pass. Across six (refit, eval) cells at four UltraChat checkpoints, LR-CV probe AUROC stays ≥ 0.91 and additive steering shows a monotonic dose-response. Silencing does not. No extraction recipe at any layer in the sweep $\{8, 10, 12, 14, 16, 18\}$ produces a CI that excludes zero, with a single exception: LR-CV refit to the most-trained checkpoint (-5.94 pp, CI $[-10.3, -1.6]$). Plain DoM's probe AUROC is 0.49 (random) at every layer. A monitor watching detection and steering reads healthy throughout Mistral's training, while no linear direction in the body gives a causal handle. Only silencing reports the absence. Per-cell tables in Appendix B.

4 A within-class confound explains the failure

The class mean $\mu_+ - \mu_-$ is the Bayes-optimal linear discriminant when the classes have identical covariance; it is the *wrong* direction when within-class variance carries a label-correlated surface confound v ("formality," "length," anything that shifts systematically between AdvBench and Alpaca at the probe layer). In that regime, $\mu_+ - \mu_-$ is a superposition of the refusal axis and v; projecting it out deletes both. When v dominates the superposition, the projection removes a large fraction of v and little of the refusal axis, and the silencing test fails. A cross-validated logistic regression is less susceptible because held-out regularization penalizes directions that generalize poorly across folds, typically surface confounds rather than the feature's robust axis.

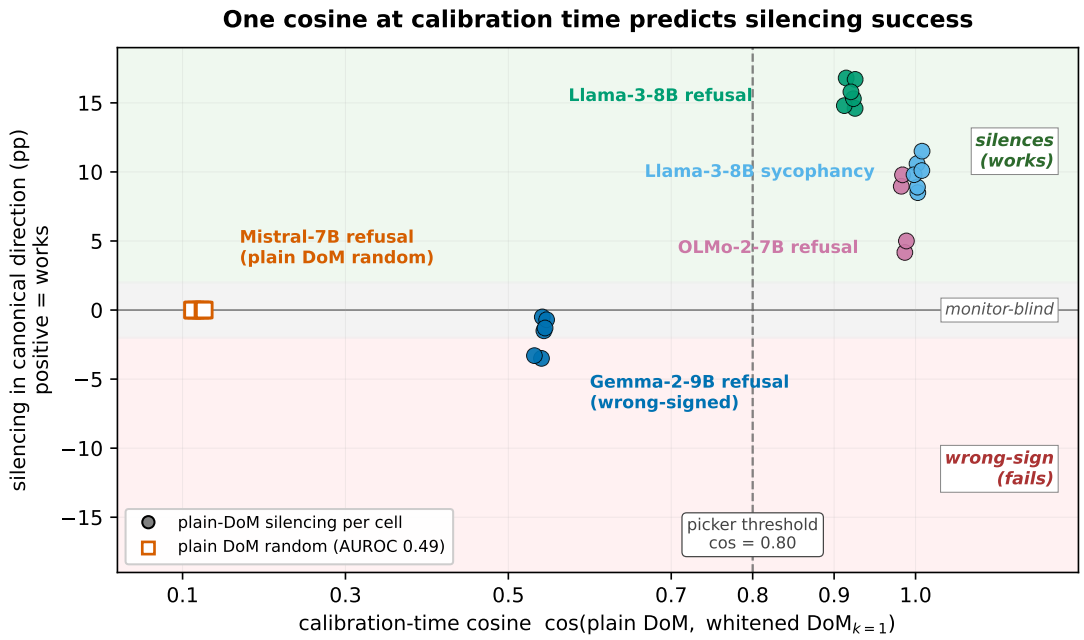

Figure 4: **One cosine at calibration time predicts silencing success.** Each point is one (refit, eval) cell; x is the cosine between plain DoM and DoM refit on activations with the top within-class PC projected out. The y-axis is signed to the task-canonical direction, so positive values mean plain DoM silences (green band), negative means it acts anti-canonically (pink), and $|y|<2$ is monitor-blind (grey). Mistral ($\cos=0.12$) lands near zero (plain DoM is random); Gemma-2 ($\cos=0.54$) lands in the wrong-sign band; Llama-3 refusal, OLMo-2, and Llama-3 sycophancy sit at $\cos \geq 0.92$ where plain DoM silences. The dashed line at $\cos=0.80$ is the picker threshold used by the monitor (§5); it sits in the empirical gap $[0.55, 0.92]$ between the confound-dominated and confound-light clusters, so any threshold in that range routes every body-battery cell identically. Cells: n=960 each; Newcombe 95% CI on every point (smaller than marker).

A one-parameter test. Whitening removes the most label-correlated within-class variation before re-extracting the direction. Concretely: compute the pooled within-class covariance of the labeled buffer's activations, project its top-k eigenvectors out of every activation, and refit DoM. We call this recipe DoM-PCwhitened$_k$; the integer k is the only knob. Before running any silencing, the decomposition $\hat{d}_{\text{DoM}} = \alpha \cdot (\text{causal axis}) + \beta \cdot (\text{confound})$ predicts three regimes:

- **Confound-dominated** ($\beta \gg \alpha$, cosine low): whitening recovers silencing at small k.
- **Confound-light** ($\alpha \gg \beta$, cosine high): whitening mildly lossy; plain DoM is already the causal axis.
- **Confound-free** ($\beta \to 0$, cosine $\to 1$): whitening is a near-noop.

The calibration-time cosine between plain and whitened DoM is the regime signature (Fig. 4).

The five-condition signature. On Mistral-7B-v0.3 at L14, $\cos_{k=1} = 0.12$: plain DoM points ∼90% along the top within-class PC. Whitening then refitting the class mean recovers silencing at every cell (-5.3 to -15.4 pp) in a regime where plain DoM's probe AUROC is 0.49. On Gemma-2-9B at L18, $\cos = 0.54$: plain DoM is mildly confound-contaminated, and whitening flips the

silencing sign into the task-canonical direction at every cell (-7 pp at k=1, -17 pp at k=5). Llama-3 at L12 (cos = 0.92) and the confound-free sycophancy extremum (cos = 1.00) sit at the other end of the signature, where whitening is lossy or a near-noop. Table 3 summarizes all five conditions; per-cell tables are in Appendix E.

model	feature	ℓ	$\cos_{k=1}$	plain DoM Δ	whitened $\Delta_{k=1}$	regime
Mistral-7B-v0.3	refusal	L14	0.12	n/a (AUROC 0.49)	$[-10.6, -5.3]$	dominated
Gemma-2-9B	refusal	L18	0.54	wrong-sign $+3$ pp	$[-9.2, -7.0]$	wrong-sign
Llama-3-8B	refusal	L12	0.92	$[-15.3, -14.6]$	$[-7.5, -6.6]$	light, mid
OLMo-2-7B	refusal	L30	0.99	-5.7 pp	$[-9.8, -4.2]$	light, late
Llama-3-8B	sycophancy	L12	1.00	$+9.8/+11.5$ pp	$[+8.3, +13.0]$	free ($\beta\to 0$)

Table 3: **Cross-family whitening summary.** One row per (model, feature) cell. $\cos_{k=1}$: calibration-time cosine between plain and k=1-whitened DoM. Plain-DoM Δ: silencing effect of the unmodified class mean (sign-flipped from canonical when plain DoM is wrong-signed). Whitened $\Delta_{k=1}$: range across refit/eval cells after projecting out the top within-class PC. The cosine tracks silencing success monotonically across $[0.12, 1.00]$. Llama-3 sycophancy Δ is positive because the probe encodes anti-sycophancy. Per-cell detail for all five conditions in Appendix E. Random-direction baselines are null at every family ($|\Delta|<5$ pp, CIs cross zero).

Decomposition. Read plain DoM as a vector decomposition: $\hat{d}_{\text{DoM}} = \alpha \cdot (\text{causal axis}) + \beta \cdot (\text{within-class confound})$. The weight α measures how much the class mean points along the direction we want to silence; β measures how much it points along the confound. Whitening projects out the top within-class directions, removing β and any α-content that happens to lie along them. Fig. 5 reads directly under this model:

- **Llama-3 at L12** ($\alpha \gg \beta$). Plain DoM is already the causal axis; whitening is mildly lossy.
- **OLMo-2 at L30** ($\alpha > \beta$, late layer). Confound-light; whitening at k=5 is lossy.
- **Gemma-2 at L18** (β dominates, opposite sign of the causal axis). Whitening reveals the causal axis more as k grows, sign-flipping Δ from $+3$ to -17 pp.
- **Mistral at L14** ($\beta \gg \alpha$, $\alpha \approx 0$). Plain DoM is all confound; the top PC absorbs most of it and whitening restores silencing.
- **Llama-3 sycophancy** ($\beta \to 0$). Confound-free extremum; whitening at k=1 is a noop.

The cosine between plain and whitened DoM is the calibration-time signature of which regime a cell is in. The necessity check (§5) gates its recipe choice on this cosine rather than trying both recipes blindly.

What the decomposition doesn't cleanly predict. Two observations are not perfectly explained by the α-β model. First, Gemma-2's k=5 recovery (-17 pp) exceeds Mistral's (-10 pp) even though Mistral has the more extreme confound-to-causal ratio (cos = 0.12 vs. 0.54). Second, k=5 reliably beats k=1 on Gemma-2 but not on Mistral. Both are consistent with Gemma-2's confound spreading across several within-class PCs while Mistral's concentrates in one, so projecting out more PCs helps Gemma-2 more. A sharper predictor might be label-correlated variance in the full top-k subspace rather than the top-PC cosine alone, but we do not have a calibrated measurement of that yet.

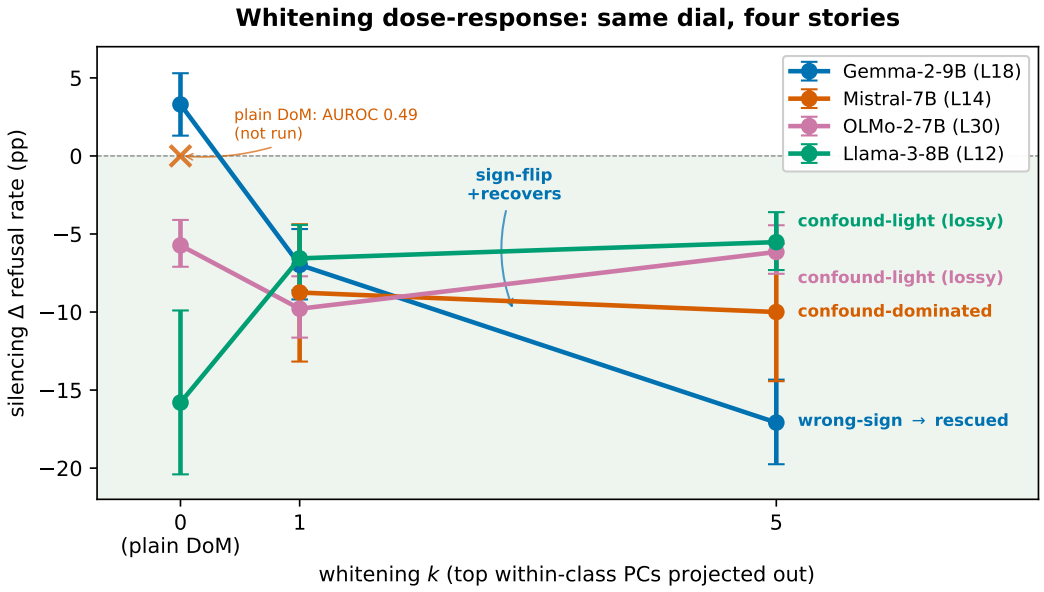

Figure 5: **Whitening dose-response: same dial, four stories.** Signed silencing Δ on refusal rate (pp) at the $r4$-refit/$r4$-eval cell, as a function of the number k of top within-class principal components projected out before refitting the class mean. Task-canonical direction is negative (green shading). Right-edge labels give the per-family regime; Gemma-2's sign-flip from $+3$ pp to -17 pp is the most dramatic recovery. Mistral's k=0 not run (plain-DoM probe AUROC 0.49 at L14). n=960 per cell; Newcombe 95% CI. Llama-3 sycophancy (cos =1.00) not plotted; see Table 13. Per-curve interpretation in the "Decomposition" paragraph below.

Layer preference is family-dependent. Peak probe layer for refusal: L12 (Llama-3), L14 (Mistral), L18 (Gemma-2), L30 (OLMo-2, on a 32-layer model). On OLMo-2 at mid-layers L8–L14, plain-DoM ablation is $\Delta \approx 0$ despite probe AUROC ≥ 0.9999; silencing grows with depth and peaks at L30 (-5.7 pp). A monitor protocol must sweep layers per family rather than adopt a fixed convention. The OLMo-2 layer selection went through a pre-registered prediction that proved wrong and was revised before the whitening battery ran; the full trace is in Appendix D. Layer profile in Appendix A.

5 The necessity check

Silent intervention failure is the risk this paper is about. The necessity check turns that silence into an explicit signal. At every checkpoint t, on a labeled buffer of feature-positive and feature-negative activations, the monitor evaluates three signals at the per-family probe layer ℓ: AUROC of the calibration-time direction on current activations (detect), the rate change from additive steering at $\alpha \in \{2, 5, 7, 10\}$ (steer), and the rate change from projecting the direction out at every token, reported with a 95% Newcombe interval [10] (silence). The cosine signature $\cos(\text{plain DoM}, \text{whitened DoM})$ (§4) selects the recipe; silencing verifies the choice. The alarm fires when no recipe produces a confidence interval that excludes zero in the canonical direction.

Battery performance. On the 22-cell four-family battery (§3), the alarm recovers 17/17 causal handles that exist and correctly fires on the remaining 5 cells where none does. A fixed-DoM

At each checkpoint t:

$d_{\text{DoM}} \leftarrow \mu_+ - \mu_-$ *plain class mean*

$d_{\text{whitened}} \leftarrow \texttt{DoM-PCwhitened}(X_t, y_t, k{=}1)$ *confound removed*

$\cos \leftarrow \langle d_{\text{DoM}}, d_{\text{whitened}} \rangle / (\|\cdot\| \|\cdot\|)$ *regime signature*

$d_{\text{active}} \leftarrow d_{\text{DoM}}$ if $\cos \geq 0.80$ else d_{whitened}

$(\Delta, \text{CI}) \leftarrow \texttt{silence}(\text{model}_t, d_{\text{active}}, \ell)$

if CI includes 0 in the task-canonical sign: **alarm** *silent failure*

else: adopt d_{active} as the causal handle

Figure 6: **The necessity check.** One probe fit, one projection, one CI per checkpoint. The cosine gate routes the checkpoint to the recipe predicted by the within-class-confound account (§4); the silencing CI is the verification. The 0.80 threshold is fit on the five-condition signature data; a low-cosine fallback to LR-CV is available when neither plain nor whitened DoM produces a CI that excludes zero.

policy misses Gemma-2 and Mistral entirely (10/22); a fixed-LR policy misses Llama-3 sycophancy and most non-r4 Mistral cells (13/22). A held-out Qwen2.5-7B-Instruct replication (Appendix C; threshold fixed before this family ran) adds 5 confound-light cells; the picker routes all five to DoM without retuning (22/27 overall vs. 15/27 DoM, 18/27 LR).

6 Discussion

Probe-based safety monitors ship with detection and steering signals but without necessity checks. The gap is not theoretical: across four open-weights families, we find checkpoints where a probe detects at AUROC ≥ 0.91 and steers cleanly yet fails to silence. Difference-of-means and regularized logistic regression have disjoint failure sets, so no fixed recipe is safe. A within-class confound biases the class mean away from the causal axis; the cosine between plain and whitened DoM tracks silencing success monotonically and selects the right recipe in one step.

Recommendations for probe-based interventions shipping today.

- Sweep layers per family. OLMo-2's causal axis sits at L30 on a 32-layer model, invisible to an L12–L18 convention.
- Compute the cosine signature at calibration. It tells you whether plain DoM is the causal axis before any silencing run.
- Report a 95% CI on the silencing effect. Treat a CI that includes zero as an alarm, not a tuning signal.

Limitations. The battery covers refusal on four families plus sycophancy on Llama-3, 35 whitening cells in total. This is enough to establish the cosine-signature monotone across $[0.12, 1.00]$, not enough to predict which regime a new family falls into from pretraining metadata alone. Projective ablation is a necessity test subject to a self-repair caveat. The operational claim is about the deployed layer: projecting the probe out at that layer does not change behavior, and a monitor at that layer fails there regardless of what other pathways exist. Our single-layer effects (5–17 pp) are also too large for standard backup-head routing to cleanly explain, but multi-layer replication on the cross-family battery is outstanding. The setup is linear only; we characterize the dissociation at

calibration, not through arbitrary continued training. Real deployments with novel features require a labeled buffer at every refit, and label cost is part of the deployment budget.

What we would do next.

- **Name the confound.** On Mistral at L14 the top within-class PC carries enough mass that plain DoM is a random direction; the decomposition predicts this PC *is* the label-correlated surface feature. Our best guess is the AdvBench-vs-Alpaca register gap (AdvBench is imperative and stripped-down, Alpaca is conversational), untested. Extracting the PC, steering with it, and identifying what linguistic property it encodes would convert the structural $\alpha\cdot$(causal) + $\beta\cdot$(confound) account into a mechanistic one.
- **Mechanism beyond refusal.** The decomposition is feature-agnostic; whether it extends to deception-detection probes, capability-elicitation probes, or SAE-feature stability under fine-tuning is what distinguishes a refusal-specific geometric fact from a general failure mode of linear probe-based monitors.

Reproducibility. Code, directional artifacts, and per-checkpoint JSONs ship with the paper; `scripts/run_autopicker_battery.py` reproduces every number in the body in under 30 seconds on CPU.

References

[1] Mistral AI. Mistral-7b-instruct-v0.3: Model card. Hugging Face model card, 2024. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

[2] Allen Institute for AI. OLMo 2: Open language models. *arXiv preprint*, 2024. OLMo-2-1124-7B-Instruct checkpoint on HuggingFace at `allenai/OLMo-2-1124-7B-Instruct`.

[3] Anthropic. Claude mythos preview system card. 2026. Section 4.5: White-box analyses of model internals.

[4] Andy Arditi, Oscar Obeso, Aaquib Sylejmani, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

[5] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *NeurIPS*, 2023.

[6] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.

[7] Yanrui Du et al. Anchoring refusal direction: Mitigating safety risks in tuning via projection constraint. *arXiv preprint arXiv:2509.06795*, 2025.

[8] Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[9] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *NeurIPS*, 2023.

[10] Robert G. Newcombe. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, 17(8):873–890, 1998.

[11] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, et al. Discovering language model behaviors with model-written evaluations. *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[12] Yujia Qin et al. Llms can unlearn refusal with only 1,000 benign samples. *arXiv preprint*, 2024.

[13] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[14] 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.

[15] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A OLMo-2 layer sweep: full probe AUROC, cosine, and silencing profiles

We swept plain-DoM probe AUROC, $\cos(\mu_+ - \mu_-, \text{PCwhitened}_{k=1})$, cos at k=5, and ablation Δ across layers $\{8, 10, 12, 14, 16, 18, 20, 24, 25, 28, 30, 31\}$ on OLMo-2-7B-Instruct at the base checkpoint, refusal feature, n=960 per ablation cell where measured. AUROC is near-unity at every layer; cos is ≥ 0.99 throughout; silencing $\Delta \approx 0$ at L12/L14, modest at L16 (-1.15 pp), and grows with depth from L20 onward with a local dip at L28, peaking at L30 and falling back at L31.

layer	AUROC(DoM)	$\cos_{k=1}$	$\cos_{k=5}$	Δ (pp)
L8	0.99994	0.996	0.945	—
L10	0.99991	0.998	0.978	—
L12	1.00000	1.000	0.975	+0.000
L14	0.99994	0.9997	0.942	−0.001
L16	1.00000	0.999	0.939	−1.15
L18	1.00000	0.994	0.915	—
L20	0.99993	0.9991	0.938	−2.08
L24	0.99986	0.9967	0.854	−2.60
L25	0.99987	0.9962	0.882	−2.81
L28	0.99986	0.9996	0.889	−2.60
L30	0.99988	0.9922	0.946	**−5.73**
L31	0.99987	0.9999	0.975	−3.02

Table 4: OLMo-2-7B-Instruct refusal layer sweep at base. Silencing Δ is the plain-DoM projective-ablation effect; not run at L8, L10, L18 (covered by adjacent layers). L30 is the peak. L31 dips, consistent with the familiar "last transformer block mostly feeds the output head" pattern. Raw: `results/paper/raw/olmo_whitening.json` keys `auroc_at_all_swept_layers`, `cosine_plain_vs_whitened`, `layer_silencing_profile_base_plain_DoM`.

B Per-family tables

Per-family tables for the four (model, feature) conditions in the recipe-reversal battery: Llama-3 refusal, Gemma-2 refusal, Gemma-2 deeper-layer sweep, Llama-3 sycophancy, and Mistral refusal. OLMo-2 is not in this set because its ablations were only run through the whitening battery (at $\cos_{k=1}=0.99$, k=1 whitening is effectively plain DoM); OLMo per-checkpoint numbers appear in Table 13. Mistral DoM rows are omitted because plain-DoM LOO-AUROC is 0.49 at every swept layer and checkpoint, so projecting it out would amount to ablating a random direction. The body summarizes these via Fig. 2 and Table 2.

	ablation Δ refusal rate (pp)			95% CI excludes 0?		
	r0	r2	r4	r0	r2	r4
frozen LR-CV	−10.8	−9.8	−8.9	yes	yes	yes
frozen DoM	**−14.6**	**−14.8**	**−15.3**	yes	yes	yes
LR-CV refit	−10.2	−7.7	−8.2	yes	yes	yes
DoM refit	**−16.8**	**−16.7**	**−15.8**	yes	yes	yes

Table 5: Llama-3-8B refusal at L12. DoM removes 4–9 pp more refusal than LR-CV at every cell. n=960 per cell, Newcombe 95% CI. SFT rounds passed the abort gate (harmful-refusal rates $1.00, 0.96, 1.00, 0.98$).

	ablation Δ refusal rate (pp)			95% CI excludes 0?		
	r0	r2	r4	r0	r2	r4
frozen LR-CV	**−5.8**	**−6.0**	**−5.4**	yes	yes	yes
frozen DoM	+0.5	+1.5	+3.5	no	yes	yes
LR-CV refit	**−6.2**	**−7.2**	**−5.3**	yes	yes	yes
DoM refit	+0.7	+1.3	+3.3	no	yes	yes

Table 6: Gemma-2-9B refusal at L18. Sign of the asymmetry is reversed: LR-CV silences by 5–7 pp while DoM is null or slightly anti-causal. n=960 per cell, Newcombe 95% CI.

	DoM (base-fit) on r4		LR-CV (base-fit) on r4	
layer	Δ (pp)	95% CI	Δ (pp)	95% CI
18 (body)	+3.5	excludes 0	**−5.4**	**excludes 0**
24	+3.1	[+1.8, +4.1]	−0.8	[−2.6, +0.9]
30	+2.4	[+1.1, +3.6]	+0.5	[−1.2, +2.2]
36	−1.5	[−3.2, +0.3]	+0.4	[−1.2, +2.1]

Table 7: Gemma-2 deeper-layer sweep (post-hoc, run after the L18 reversal was surprising). No deeper layer recovers a DoM causal handle; LR-CV at L18 is the only silencing axis. Rules out the "wrong layer" objection to the reversal.

	ablation Δ sycophancy rate (pp)			95% CI excludes 0?		
	r0	r2	r4	r0	r2	r4
frozen LR-CV	−0.9	−1.0	−1.8	no	no	no
frozen DoM	**+8.5**	**+8.9**	**+10.6**	yes	yes	yes
LR-CV refit	+1.2	−0.9	+2.6	no	no	no
DoM refit	**+9.8**	**+11.5**	**+10.1**	yes	yes	yes

Table 8: Llama-3 sycophancy at L12. Positive Δ because the probe encodes anti-sycophancy/honesty. All six DoM cells exclude zero; all six LR-CV cells do not. The sign-flip rules out (i) "DoM wins because DoM has larger magnitude," since the asymmetry replicates with reversed sign, and (ii) "DoM tracks prompt-distribution, not the feature," since an opposite-sign ablation effect requires the direction to track something downstream of the prompt. n=960 per cell, Newcombe 95% CI.

	ablation Δ refusal rate (pp)			95% CI excludes 0?		
	r0	r2	r4	r0	r2	r4
frozen LR-CV	−1.2	−4.0	−2.6	no	no	no
frozen DoM	*not run: plain-DoM LOO-AUROC = 0.49 at every checkpoint*					
LR-CV refit	−1.2	−2.3	**−5.94**	no	no	**yes**
DoM refit	*not run: plain-DoM LOO-AUROC = 0.49 at every checkpoint*					

Table 9: Mistral-7B-v0.3 refusal at L14. LR-CV AUROC holds ≥ 0.91. Only refit-r4 on r4 excludes zero; the other five LR-CV cells are null and the picker correctly emits `NECESSITY_LOST`. n=960 per cell, Newcombe 95% CI; base refusal rate ~ 0.62.

C Held-out replication: Qwen2.5-7B-Instruct refusal

A fifth family (Qwen2.5-7B-Instruct, L14, AdvBench+Alpaca buffer) was run after the cos =0.80 picker threshold was fixed on the body's five-condition signature data. The condition exists to test whether the threshold generalizes to a family not used for fitting it, not to add a sixth regime: Qwen is confound-light, so both fixed recipes find handles and the picker's test is *false-alarm rate* rather than failure-set discrimination. All 5 cells pass without retuning.

	plain DoM		plain LR-CV		
cell	Δ (pp)	95% CI	Δ (pp)	95% CI	picker
base / r0	**−42.7**	[−45.8, −39.6]	−10.7	[−12.6, −8.6]	DoM
base / r2	**−44.1**	[−47.2, −40.9]	−15.6	[−17.9, −13.2]	DoM
base / r4	**−48.9**	[−52.0, −45.7]	−20.8	[−23.4, −18.1]	DoM
r2 / r2	**−37.5**	[−40.6, −34.4]	−12.2	[−14.2, −10.0]	DoM
r4 / r4	**−37.6**	[−40.7, −34.5]	−20.1	[−22.6, −17.4]	DoM

Table 10: **Qwen2.5-7B-Instruct refusal at L14,** plain DoM and plain LR-CV projective-ablation Δ. Cells are *direction checkpoint / evaluation model checkpoint*. Both recipes silence at every cell; DoM's effect is 3–4× LR-CV's, placing Qwen in the confound-light regime ($\alpha \gg \beta$, cos ≥ 0.80 at every cell). The picker selects DoM on all five without re-fitting the threshold. n=960 per cell, Newcombe 95% CI. Raw: `results/monitor/autopicker_battery.json`.

D Experimental provenance

To let readers calibrate weight on each result, we label each experiment as pre-registered (predicted before running) or post-hoc (design or analysis fixed after initial observations).

Pre-registered.

- The within-class-confound mechanism's three-regime prediction and the Mistral whitening battery (§4).
- Gemma-2 and Llama-3 whitening cells, as replications.
- The silencing-CI-based alarm criterion (§5).

Post-hoc.

- The Gemma-2 deeper-layer sweep at $\{24, 30, 36\}$, run after the L18 reversal was surprising.
- The sycophancy sign-flip read.
- The cos(DoM, LR) observation.

Mixed. The cosine signature between plain and whitened DoM (§4) was predicted by the decomposition and measured before the ablation battery on each family: pre-registered in structure, though the specific five-condition series $(0.12, 0.54, 0.92, 0.99, 1.00)$ was not predicted quantitatively per condition.

Interpretation-evolution trace (OLMo-2 layer selection). A pre-registered prediction proved wrong and was revised before the whitening battery was fired; we record the trace for methods honesty. Initial layer sweep at $\{8, 10, 12, 14, 16, 18\}$ (the mid-layer convention established by Mistral/Gemma-2/Llama-3) returned plain-DoM ablation $\Delta \approx 0$ at every layer despite AUROC ≥ 0.9999. The first interpretation was a novel "linear-ablation-inadequate" regime. Extending the sweep to $\{20, 24, 30\}$ revealed a monotone depth gradient ($-2.1, -2.6, -5.7$ pp); a peak search at $\{25, 28, 31\}$ localized the peak at L30. Revised interpretation: OLMo-2 encodes refusal at a later depth; not a new regime, a cross-family layer-preference. The L12/L14 detect-perfect-but-silencing-null cells land as the expected signature of probing at the wrong layer relative to the causal computation. Full trace in `results/paper/raw/PREREGISTRATION_2026-04-19.md`.

E Per-cell whitening: all five conditions

Per-cell detail for every row of Table 3. Mistral and Gemma-2 are the load-bearing existence-proof and sign-flip tables; the other three are the confound-light and confound-free rows referenced from §4.

cell	plain DoM	plain LR probe	DoM-PCwhitened k=1	DoM-PCwhitened k=5
base / base	AUROC 0.49 (n/a)	−1.2 (null)	**−7.7** [−12.1, −3.3]	—
r4 / base	AUROC 0.49 (n/a)	−1.2 (null)	**−10.6** [−15.0, −6.3]	**−9.4** [−13.8, −5.0]
r4 / r2	AUROC 0.49 (n/a)	−2.3 (null)	**−5.3** [−9.7, −0.9]	**−15.4** [−19.8, −11.1]
r4 / r4	AUROC 0.49 (n/a)	−5.9 [−10.3, −1.6]	**−8.8** [−13.2, −4.4]	**−10.0** [−14.4, −5.6]

Table 11: Mistral-7B-v0.3 refusal at L14. *cell = refit / eval* checkpoint. Plain DoM is random (AUROC 0.49), so no silencing run. All whitened cells exclude zero. n=960 per cell, Newcombe 95% CI. Raw: `results/paper/raw/mistral_whitening.json`.

cell	plain DoM	plain LR probe	DoM-PCwhitened k=1	DoM-PCwhitened k=5
base / base	+0.5 (null)	−5.8 [−9.8, −1.6]	**−8.2** [−10.0, −6.2]	—
r4 / base	+3.5 (wrong-sign)	−5.4 [−8.2, −2.6]	**−9.2** [−11.1, −7.1]	**−18.2** [−20.7, −15.6]
r4 / r2	+3.3 (wrong-sign)	−7.2 [−10.1, −4.1]	**−8.1** [−10.1, −6.0]	**−17.7** [−20.2, −15.1]
r4 / r4	+3.3 (wrong-sign)	−5.3 [−8.7, −1.7]	**−7.0** [−9.2, −4.7]	**−17.1** [−19.8, −14.3]

Table 12: Gemma-2-9B-it refusal at L18. Plain DoM is wrong-signed at every cell (Table 6). Whitening flips to the canonical direction; k=5 matches or exceeds the LR probe. 7/7 whitened cells exclude zero. n=960 per cell, Newcombe 95% CI. Raw: `results/paper/raw/gemma2_whitening.json`.

Selection rule for k. Default to k=1 (minimizes collateral removal of causal-axis content); run the silencing ablation and stop if the 95% CI excludes zero in the task-canonical direction. Escalate to $k > 1$ only when both (i) $\cos_{k=1} < 0.65$ and (ii) the k=1 CI includes zero or $|\Delta| < 5$ pp; step by powers of two until the CI excludes zero or the cumulative within-class variance projected out exceeds 50%. On the current battery this selects plain DoM on Llama-3 (cos = 0.92), $k^\star$=1 on Mistral (cos = 0.12), $k^\star$=5 on Gemma-2 (cos = 0.54).

cell	PCwhitened k=1	PCwhitened k=5	random
Llama-3 refusal, L12 (plain DoM ∼ −15 pp is the silencing axis; whitening lossy)			
base / base	−7.0 [−8.9, −4.9]	—	—
r4 / base	−7.0 [−9.0, −4.8]	−5.1 [−7.0, −3.1]	—
r4 / r2	−7.5 [−9.4, −5.4]	−4.7 [−6.4, −2.8]	—
r4 / r4	−6.6 [−8.6, −4.4]	−5.5 [−7.3, −3.6]	—
OLMo-2 refusal, L30 (confound-light late-layer; $\cos_{k=1}$ =0.99, k=1 ≈ plain DoM)			
base / base	−4.2 [−5.3, −2.7]	—	—
r4 / base	−5.0 [−6.3, −3.4]	−1.5 [−2.2, −0.5]	—
r4 / r2	**−9.0** [−10.7, −7.0]	−6.4 [−7.8, −4.6]	—
r4 / r4	**−9.8** [−11.6, −7.7]	−6.2 [−7.6, −4.4]	+0.0 [−0.4, +0.4]
Llama-3 sycophancy, L12 (confound-free, $\cos_{k=1}$ =1.00; Δ>0 is canonical)			
r0 / r0 (sanity)	+8.3 [+4.0, +12.8]	—	—
r4 / r0	**+12.5** [+8.2, +16.9]	+7.9 [+3.6, +12.3]	—
r4 / r2	**+10.2** [+5.9, +14.6]	+9.7 [+5.3, +14.1]	—
r4 / r4	**+13.0** [+8.7, +17.4]	+10.1 [+5.8, +14.5]	+1.3 [−3.2, +5.7]

Table 13: Per-cell whitening for the three confound-light / confound-free rows of Table 3. Silencing Δ (pp) by PCwhitened recipe and random-direction baseline. *cell = direction checkpoint / evaluation checkpoint.* n=960 per cell, Newcombe 95% CI.