SAFETY FRAGILITY IS A ROUTING PROBLEM: THE ATTENTION BOTTLENECK BETWEEN HARM DETECTION AND REFUSAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Aligned language models refuse harmful requests most of the time, and that refusal is remarkably fragile. Two lines of research explain the fragility in ways that do not fit together: one localizes safety in a single residual-stream direction or a handful of attention heads, the other finds it confined to a narrow basin in weight space that fine-tuning drifts out of. We show that they describe different components of one mechanism. To unify these findings, we decompose safety into three components: detection of harm, routing of that signal through attention, and emission of a refusal. Both detection and action are stable under weight perturbation, and what breaks is the routing: a small, non-redundant set of attention heads, which we call the gate, that sits between detection and action and that safety depends on causally. Direct feature attribution, a common way to identify which heads matter for a behavior, misranks this set in all three model families we test: its top head is never causally necessary, because it scores how much a head writes toward refusal rather than whether the output depends on it. The routing bottleneck also explains the basin geometry that prior work observes without an underlying cause: safety is narrow along the gate heads' weights because small perturbations there sever the connection between two otherwise stable representations. If fragility comes from depending on a few heads, then training the model to distribute that dependence should make it robust. We test this with noise-injected fine-tuning on the bottleneck heads. The model's dependence on them disappears, and the basin widens along that direction, at a cost in over-refusal. Applying the same training to the top-attribution head leaves gate dependence and the basin unchanged, because its contribution is redundant and removing it leaves the circuit intact.

1 INTRODUCTION

Safety alignment can be installed in language models, and it can be removed with surprising ease. A hundred steps of benign fine-tuning eliminate refusal (Qi et al., 2025), and ablating a single attention head raises attack success by an order of magnitude (Zhou et al., 2025). This fragility is specific to safety: we show that random weight perturbations at a scale that leaves mathematical reasoning intact shift refusal rates by tens of percentage points across seeds (Section 2.2). Understanding which component of the safety mechanism is responsible for this asymmetry is the first step toward making it robust.

Two bodies of work have tried to answer this question and reached conclusions that appear to conflict. One locates safety in a small number of identifiable components: a single residual-stream direction mediates refusal (Arditi et al., 2024), ablating one attention head disables it (Zhou et al., 2025), and safety-critical parameters concentrate in middle-layer value matrices (Qi et al., 2026). If safety lives in a few components, it should be possible to protect them. The other body of work finds that protection is difficult to achieve: harmfulness representations survive fine-tuning that destroys refusal (Zhao et al., 2025), the safe region of weight space is a narrow basin that gradient descent drifts out of (Peng et al., 2024; Springer et al., 2026), and the basin is anisotropic in ways no one can explain (Yang et al., 2025). The first set of results suggests safety is localized and should be easy to

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.

harden. The second suggests it is diffuse and structurally fragile. Since both have strong evidence, the question is not which is wrong, but what each is actually measuring.

We resolve the conflict by decomposing the refusal circuit into three components, each corresponding to a different stage of how the model processes and acts on a harmful instruction. The model first recognizes whether an instruction is harmful, a step we call *detection*, encoded at the instruction-token position (Zhao et al., 2025). It then transports that recognition through a small number of attention heads to the position where the refusal decision forms, a step we call *routing*. It finally emits the refusal through the residual-stream direction identified by Arditi et al. (2024), which we call *action*. The localization literature has been characterizing detection and action. We show that both remain intact under weight perturbation and that the amount each one drifts does not predict whether the model's safety actually breaks. The fragility therefore lies in the middle step: the model still recognizes harm and still knows how to refuse, but the recognition never reaches the refusal.

The recognition is carried by a small set of attention heads that standard alignment leaves with no redundancy. We call this set the *gate*, and its lack of redundancy is what makes it a bottleneck. Because no other head compensates when the gate is disturbed, even small perturbations are enough to break the connection between detection and action. This is why the safe region of weight space is narrow in some directions and not others: the narrow directions are precisely those that disturb the gate. Prior work measures the anisotropy (Peng et al., 2024; Yang et al., 2025) without knowing which direction is narrow or why. The gate answers both: the narrow direction lies in the gate heads' weights, and it is narrow because no other head compensates along it.

This also reveals a failure in a common method for identifying safety-critical heads. Direct feature attribution ranks heads by how much they write toward the refusal direction; methods that add ablation feedback (Zhou et al., 2025) are a different case, which we do not re-run. But the quantity that determines whether safety breaks is non-redundancy: whether any other head compensates when this one is removed. A head that writes strongly toward refusal but is fully duplicated by three others will rank high on attribution, even though removing it changes nothing. In all three model families we test, the head with the largest attribution does not affect safety when removed, and the head that safety depends on most is not ranked first: it sits at rank four in SmolLM2, at negative attribution in Qwen-2.5, and outside the top fifty in Llama-3.1 (Section 3.1). It means building an intervention around attribution-selected heads is protecting the wrong components.

We test this explanation by training the model to stop depending on the gate. If fragility comes from concentrating the routing load on a few heads, then distributing that load should remove the vulnerability. After training, the model no longer depends on the gate and the basin widens along the previously narrow direction, while the same procedure targeting the top-attribution head leaves gate dependence and the basin unchanged (Section 4). The intervention fails on three of seven configurations. A pre-registered threshold on single-head gate load did not predict which (Appendix C); a window-level measure chosen afterwards separates all seven, but it is untested prospectively, and two of the failures are also the two largest models.

Our contributions are:

- A three-component decomposition of the refusal circuit into detection, routing, and action, with evidence that routing is the fragile component and that direct feature attribution identifies the wrong routing heads in three model families (Sections 2 and 3.1).

- A mechanistic explanation for the safety basin's geometry: the basin is narrow along the gate heads' weights because perturbations there break the only connection between detection and action, and flat along the direction attribution selects because that head is redundant (Section 3.4).

- A causal test that trains the model to distribute routing away from the gate, after which removing the gate no longer breaks safety, and the previously narrow basin direction becomes robust, while the same training targeting the top-attribution head produces no change (Section 4 and Appendix C).

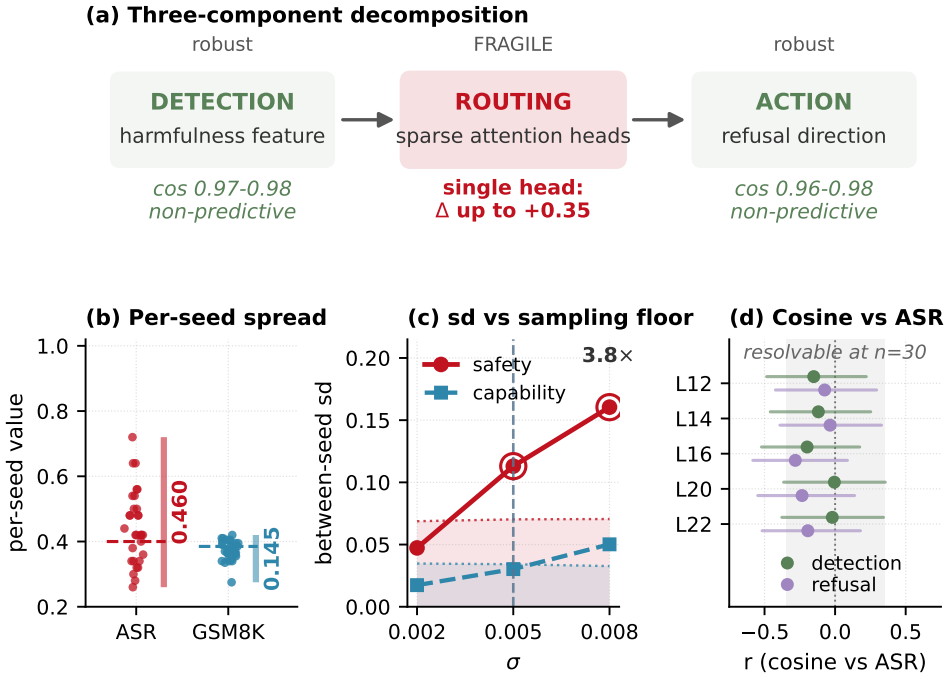

Figure 1: Safety is far more perturbation-sensitive than capability, and neither the detection nor the action representation explains the difference. SmolLM2-1.7B-Instruct under matched Gaussian weight perturbation, 30 seeds. (a) The decomposition under test: detection and action are representations, routing is the step between them. (b) One perturbation, two behaviours. The noise that leaves GSM8K accuracy tightly clustered scatters attack success across most of its range, so the sensitivity belongs to safety rather than to perturbing weights in general. (c) That gap is not an artifact of how finely each quantity can be measured. Shading each metric's own binomial sampling floor shows safety's seed-to-seed variation rising clear of its floor a full perturbation scale before capability's does, so at the operating scale only one of the two is resolvable at all. The separation is categorical rather than a ratio between two spreads that may both be noise. (d) Neither representation explains the failures. Per-seed drift in the detection direction and in the refusal direction is uncorrelated with per-seed attack success at every layer tested, with intervals straddling zero throughout. What moves under perturbation is therefore not what decides whether the model refuses, which is what sends the search to the routing step.

2 Safety Decomposition

This section tests which component of the decomposition is responsible for safety's fragility under weight perturbation. Because a representation can drift substantially without affecting behavior, the test is not how far each component moves but whether its per-seed drift tracks safety failure. Fragile detection or fragile action predicts such a correlation; fragile routing predicts none, with failure instead following the removal of a few attention heads, which Section 3.1 tests.

2.1 Experimental setup

We test on three model families: SmolLM2-1.7B-Instruct (Ben Allal et al., 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen-2.5-7B-Instruct (Qwen Team, 2024). Appendices C and D extend the analysis to four additional models, including the SmolLM2 pretraining variants of Maini et al. (2025).

Data and evaluation. Harmful prompts come from AdvBench (Zou et al., 2023). We measure over-refusal on two benchmarks that differ in adversarial proximity: JBB-Behaviors (Chao et al., 2024), which are designed to resemble harmful requests while being safe, and Alpaca (Taori et al., 2023), a standard instruction-following set with no adversarial structure. Because our intervention affects the two differently, we report results on both. Capability is measured on GSM8K (Cobbe et al., 2021). A keyword refusal judge scores the perturbation and ablation sweeps, and an ensemble of WildGuard (Han et al., 2024) and Llama Guard (Inan et al., 2023) scores the headline numbers, requiring both to agree before calling a response unsafe, which biases the headline attack success downward relative to an either-flags rule. The two judges agree 84% of the time (Cohen's $\kappa = 0.67$), and the keyword judge over-fires roughly twice as often as it misses, making keyword-scored effects slightly optimistic in absolute terms. Every run logs a degenerate-output fraction as a coherence guard.

Representations. We extract the detection and action directions from 128 harmful and harmless contrast pairs. The detection direction is the mean activation difference at the instruction-token position (Zhao et al., 2025). The action direction is the refusal direction of Arditi et al. (2024), measured at the post-instruction position. Direction norm and stability vary across layers (Du et al., 2025), and layers can be selected on principled grounds from activation geometry alone (Siu et al., 2025). We fix the measurement bands before data collection: layers 12, 14, and 16 for detection, layers 20 and 22 for action.

Prefill. Llama-3.1-8B refuses roughly 95% of plain AdvBench prompts and Qwen-2.5-7B refuses nearly all of them. At that floor no ablation produces a measurable effect regardless of the gate. We therefore force the assistant turn to begin with `"Sure"` on harmful-side evaluations for these two models, placing Llama-3.1-8B at baseline attack success 0.60 and Qwen-2.5-7B at 0.68 on the training evaluation sets (0.58 and 0.735 on the 200-prompt ablation set). SmolLM2 needs no prefill. Benign-side evaluation and all training use no prefill. Section F reports a no-prefill replication on Llama-3.1-8B against a harder harmful distribution (Mazeika et al., 2024), confirming that the core dissociation between attribution and causal necessity survives the change of regime, although the largest-effect head changes from L12H6 to L10H27. The gate-targeted training of Section 4 selects its gate and reports its attack success in the prefill regime.

2.2 Safety is far less stable than capability

We measure fragility by applying in-place Gaussian noise to all attention and MLP weight matrices ($\theta \leftarrow \theta + \sigma\varepsilon$) across 30 independent seeds and recording both attack success on AdvBench and GSM8K accuracy at each seed. On SmolLM2-1.7B-Instruct at $\sigma = 0.005$, attack success ranges from 0.26 to 0.72 around an unperturbed baseline of 0.40 (the 50-prompt perturbation configuration of Appendix I). GSM8K accuracy under matched perturbations at the same seeds stays within its own sampling noise (Figure 1(b)). Because the two evaluation sets differ in size, comparing raw spreads would overstate the gap. We correct for this by subtracting the variance expected from finite-sample evaluation alone from each measurement's observed variance. Safety variance remains well above the noise expected from finite evaluation, while capability variance does not, confirming that perturbations at this scale move safety behavior measurably without affecting reasoning (Figure 1(c)).

Neither representation explains the failure. To test the detection and action hypotheses, we measure how far each direction rotates under the same perturbations that produce the behavioral variation above, using the directions and layer bands of Section 2.1. Both directions are stable under perturbation, with mean cosine to baseline above 0.96 at every layer. More importantly, per-seed attack success does not correlate with drift in either direction at any layer tested (Figure 1(d); $|r| \leq 0.28$ throughout). With 30 seeds, any correlation of $|r| \geq 0.35$ would have been detected, so the null result rules out a moderate linear relationship, though only over the narrow cosine range the perturbation induces (0.95 to 0.99), and only on SmolLM2. Stability alone therefore does not identify the source of fragility: both representations remain nearly intact across the same seeds where safety varies widely.

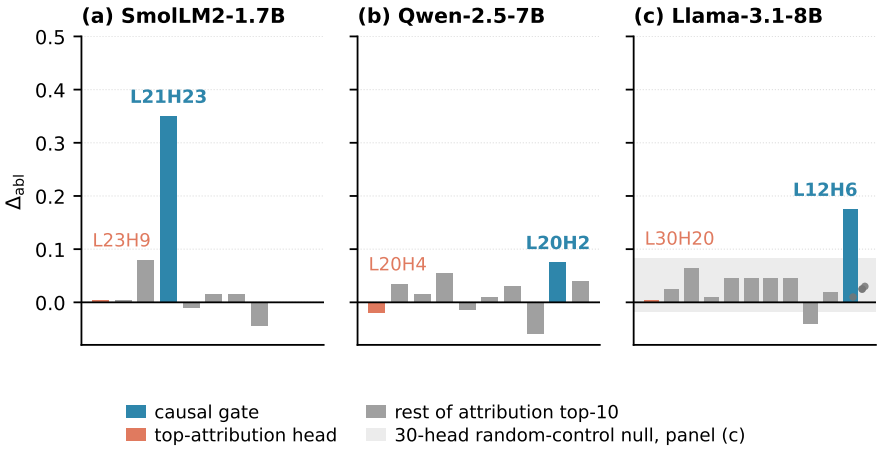

Figure 2: Direct feature attribution ranks heads by how much they write toward refusal, which is a different quantity from what the behavior depends on. Bars give the change in attack success under single-head ablation (Δ_{abl}, Equation 3) on 200 AdvBench prompts, for each family's ten highest-attribution heads (Equation 2); the legend gives the colour coding and each highlighted bar is labelled with its head. The causal gate sits at attribution rank 4 in SmolLM2, at *negative* attribution in Qwen-2.5, and outside the attribution top-50 in Llama-3.1. Because the Llama gate was never in the attribution top-10, panel (c) shows it against a 30-head random-control null: the shaded band is that null's range and the grey points are the individual controls. In every family the head attribution ranks first is the one whose removal changes nothing.

3 The Routing Bottleneck

Section 2.2 narrowed the source of fragility to the routing step. The natural next question is which attention heads carry the routing, and the field's standard tool for answering it is feature attribution: scoring each head by how much its output projects onto the refusal direction. We test whether the heads attribution selects are the ones safety actually depends on, using single-head ablation, which zeroes a head's output and measures the resulting change in attack success (Δ_{abl}).

3.1 Why attribution does not recover causal necessity

Attribution measures how strongly a head writes toward refusal. Ablation measures whether anything compensates when the head is removed. A head whose contribution is fully duplicated by others will rank high on attribution while having no effect on behavior when ablated, because other heads already carry the same signal. A head that carries a unique signal will rank low on attribution if that signal is small, while having a large effect on behavior when ablated, because nothing compensates. Attribution tracks magnitude; causal necessity tracks non-redundancy. The two coincide only when every head's contribution is unique, which is the exception rather than the rule. Undifferentiated attention (Zhou et al., 2025), which collapses a head's attention pattern rather than zeroing its output, independently confirms every mismatch we report (Appendix E).

3.2 The mismatch across three model families

In SmolLM2-1.7B-Instruct, the top-attribution head L23H9[1] has by far the largest projection onto the refusal direction and produces no measurable change under ablation. The causally necessary head L21H23, ranked fourth by attribution, produces the largest ablation effect in the model (Figure 2). In Llama-3.1-8B-Instruct, the gap is sharper: the causally necessary head L12H6 was discovered through a random-control sweep because it falls outside the attribution top-50 entirely. Its ablation effect exceeds that of all thirty random control heads and all ten attribution-selected heads evaluated on this model, and it was found only because the protocol included a random-control

[1]We denote heads as LℓHh: layer ℓ, head h, zero-indexed.

sweep alongside the attribution ranking. In Qwen-2.5-7B-Instruct, the effects are smallest but the sign is starkest: the top-attribution head L20H4 produces no effect under ablation, and the most causally necessary single head among those tested, L20H2 ($+0.075$, paired $p = 0.006$), has a negative attribution score, meaning attribution ranks it as writing away from refusal (Figure 2). On Qwen no single head is large, and the gate we train against in Section 4 is a pair, L19H0 and L23H26, whose joint ablation on the 50-prompt training battery ($+0.12$) exceeds that of L20H2 alone on the same battery ($+0.04$); those two heads sit at attribution ranks 3 and 5.

3.3 Identifying the routing bottleneck

The attention output at layer l and position t decomposes as a sum over heads,

$$a_t^{(l)} \;=\; \sum_{h=1}^{H} W_O^{(l,h)} z_t^{(l,h)}, \tag{1}$$

where $z_t^{(l,h)} \in \mathbb{R}^{d_h}$ is head h's value-weighted output at position t and $W_O^{(l,h)} \in \mathbb{R}^{d \times d_h}$ the corresponding column block of the output projection. Attribution and ablation both act on the same term of this sum: attribution measures its projection onto the refusal direction, ablation removes it.

Attribution scores what a head writes. Let $r^{(l)}$ be the unit refusal direction at layer l, the normalized difference in means of the layer-l residual stream between 32 harmful and 32 harmless prompts, taken at the final prompt token, or at the final prefill token in the prefill regime. The attribution score of head (l, h) is its mean discriminative write onto that direction,

$$\alpha_{l,h} \;=\; \mathbb{E}_{x \sim \mathcal{D}_{\text{harm}}}\Big[\big\langle W_O^{(l,h)} z^{(l,h)}(x),\; r^{(l)} \big\rangle\Big] - \mathbb{E}_{x \sim \mathcal{D}_{\text{harmless}}}\Big[\big\langle W_O^{(l,h)} z^{(l,h)}(x),\; r^{(l)} \big\rangle\Big], \tag{2}$$

evaluated at the same position the direction is taken at. We score every head in the upper three quarters of the network (576 heads in SmolLM2-1.7B, 768 in Llama-3.1-8B, and 588 in Qwen-2.5-7B) and rank by $|\alpha_{l,h}|$, so that heads writing away from refusal on harmful prompts are also considered. This is how Qwen-2.5-7B's L20H2 enters the top ten at rank 8 with $\alpha_{l,h} = -1.36$.

Ablation scores what the model depends on. Ablating a head deletes its term from Equation 1 at every position, $z_t^{(l,h)} \leftarrow 0$ for all t, leaving every other head, the attention pattern, and all downstream weights untouched. Because it acts at all positions, Δ_{abl} establishes that the output depends on the head; it does not by itself show that the head moves the detection signal from the instruction position to the refusal site, which would need attention-pattern or position-restricted ablation evidence that we do not report. For a head set S we write

$$\Delta_{\text{abl}}(S) \;=\; \text{ASR}\big(M_{\setminus S}\big) - \text{ASR}(M), \tag{3}$$

with both terms measured on one fixed set of 200 AdvBench prompts, so that any two heads are compared on identical inputs by a paired exact test.

The selection rule. We identify the gate by causal effect rather than by attribution. The candidate pool $\mathcal{C}$ consists of the ten highest-attribution heads together with heads drawn uniformly at random from the same layer range under a fixed seed. We ablate each head individually on 200 AdvBench prompts and select the head whose removal produces the largest increase in attack success,

$$\hat{g} \;=\; \arg\max_{(l,h) \in \mathcal{C}} \; \Delta_{\text{abl}}\big(\{(l,h)\}\big), \tag{4}$$

We require two conditions for admission: removing the head must produce a statistically significant change in attack success compared to the unablated model, and the effect must be larger than that of every head in a separate random control set (thirty heads on Llama-3.1, five on SmolLM2 and Qwen-2.5). The first condition rules out noise; the second ensures the effect is specific to this head rather than a property of any head in the layer range. In Llama-3.1-8B, the gate was found by a random draw in the candidate pool rather than by the attribution ranking. Because the gate is the maximum over the pool and is reported on the same 200 prompts, its Δ_{abl} is biased upward; the random-control comparison is max against max and is unaffected. The top-attribution head serves throughout the paper as a matched control, representing what a purely attribution-based selection would return.

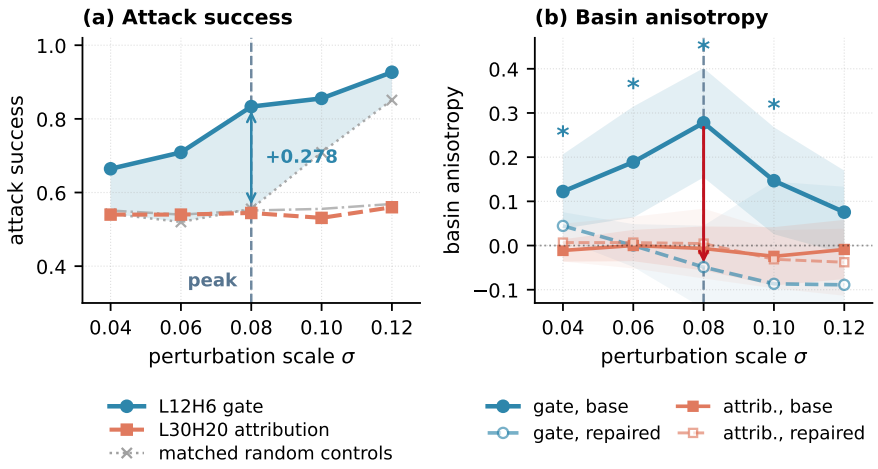

Figure 3: The safety basin is not narrow in general; it is narrow along one particular direction, and training can flatten that direction without disturbing the rest. Llama-3.1-8B-Instruct, prefill regime. Gate and control directions are perturbed with the same 15 seeds at each scale, so every comparison is within-seed. (a) Walking along the causal gate drives attack success well above matched random controls, while walking along the top-attribution direction is indistinguishable from walking in a random direction. Direction rather than the distance is what the model's safety is sensitive to. (b) The same measurement before and after gate-targeted training. The anisotropy along the gate collapses to zero, while the attribution direction, which was never narrow to begin with, is unchanged: training reshaped the geometry along the axis it aimed at and left the remainder alone. Bands are 95% intervals, an asterisk marks a scale at which the gate band excludes zero, and the red arrow marks the collapse at the peak.

3.4 The safety basin is narrow along the gate direction

Prior work has shown that the safe region of weight space is not equally wide in all directions: small perturbations in some directions break safety while equally sized perturbations in other directions do nothing (Peng et al., 2024; Yang et al., 2025). What those papers could not explain is which directions are narrow and why. The gate provides the answer.

We test this by perturbing only the gate head's query and output projection weights at increasing noise scales σ, and measuring the change in attack success. Concretely, the *gate direction* is isotropic Gaussian noise of scale σ added to the head's rows of W_Q and its columns of W_O, a random direction within that head's parameter block, drawn afresh for each seed. We then repeat the same procedure on the top-attribution head and on random non-safety heads of matched count drawn from the same layers, all with the same 15 perturbation seeds, so that the differences are paired.

On Llama-3.1-8B-Instruct, the gate direction and the top-attribution direction produce sharply different results (Figure 3(a)). Perturbing along the gate direction raises attack success at every σ from 0.04 to 0.10, peaking at $+0.278$ above the matched random control with a 95% interval of $[+0.154, +0.402]$ at $\sigma = 0.08$. Perturbing along the top-attribution direction produces no measurable difference from random at any σ tested. The basin is narrow where the gate sits and flat where attribution points.

This gives a mechanistic identity to the basin geometry. The narrow directions are those that disturb the routing bottleneck, and they are narrow because the routing heads have no redundancy: a perturbation that moves them away from their function cannot be compensated by other heads. The flat directions are those that affect redundant heads, where perturbations are absorbed by the remaining copies of the same signal. The full σ-sweep is conducted on Llama-3.1; Table 1 gives the corresponding single-σ measurement for the other two families, where the gate direction is narrow in both. The gate (layer 12) and the top-attribution head (layer 30) also differ in depth. The random controls are pooled over the gate and attribution layers rather than matched to layer 12, so head identity is not fully separated from depth. Prior work uses loss-landscape curvature to identify

Table 1: Gate-targeted training collapses gate dependence and flattens basin anisotropy across three model families, at a measured cost in over-refusal that the self-gate of Appendix B largely recovers. The matched control, which differs only in which head it targets, produces neither effect on Llama-3.1. ASR columns are harmful-side attack success; JBB and Alpaca columns are benign refusal rates. Harmful-side attack success is scored by the WildGuard/Llama Guard ensemble in the prefill regime for Llama and Qwen, on 30–60 prompts; every entry is a single run, with a run-to-run noise floor of $(0.020, 0.06)$ (Appendix I).

	Gate ablation	Basin	Ungated adapter		Self-gated (App. B)		Alpaca benign
	Δ_{abl}	anisotropy	ASR before $\rightarrow$ after	JBB	ASR	JBB	before $\rightarrow$ after
SmolLM2-1.7B	$0.34 \rightarrow \mathbf{0.08}$	$0.140 \rightarrow 0.032$	$0.420 \rightarrow 0.140$	0.14	0.20	**0.06**	$0.02 \rightarrow 0.10$
Llama-3.1-8B	$0.18 \rightarrow \mathbf{0.02}$	$0.278 \rightarrow -0.049$	$0.600 \rightarrow 0.020$	0.66	0.06	**0.10**	$0.00 \rightarrow 0.02$
Qwen-2.5-7B	$0.12 \rightarrow \mathbf{-0.06}$	$0.180 \rightarrow 0.100$	$0.680 \rightarrow 0.140$	0.74	0.18	**0.28**	$0.02 \rightarrow 0.08$
Llama-3.1-8B decoy	unchanged	$-0.007 \rightarrow 0.004$	$0.600 \rightarrow 0.460$	0.32	—	—	$0.00 \rightarrow 0.06$

fragile directions (Springer et al., 2026; Yang et al., 2025), but on SmolLM2, where we measured it, the diagonal Fisher of the forget loss ranks the gate around 370th of 768 heads (Appendix G), even though the behavioral fragility concentrates there. AsFT's narrow direction is a global aligned-minus-unaligned weight difference, while ours is local to one head; we have not measured how the two relate.

4 Redistributing the Routing Load

If safety fragility originates in the routing bottleneck, then training the model to distribute the routing load away from the gate should produce specific, testable consequences: the model should stop depending on the gate, the basin should widen along the gate direction, and the same training targeting the top-attribution head should produce neither effect.

4.1 Training procedure

We fine-tune the gate-containing layers with low-rank adaptation (Hu et al., 2022) while injecting Gaussian noise into the gate heads' weights at every forward pass. The training targets come from the model's own clean outputs (self-distillation), and a Kullback-Leibler (KL) divergence term anchors the adapted model to its unmodified weights. The noise injection is the critical component: it forces the model to learn routing that functions without the specific heads it currently depends on. Appendix I gives the per-model hyperparameters. The adapter spans three layers around the gate; on Qwen-2.5 the gate head L23H26 falls outside that window, so the Qwen configuration adapts only one of its two gate heads.

4.2 Results

After training, removing the gate no longer changes the model's behavior in any of the three families (Table 1). The model has learned to refuse without relying on those specific heads. The basin confirms that this change is specific to the gate direction: on Llama-3.1, the previously narrow gate direction flattens while the top-attribution direction is unchanged (Figure 3(b)). To rule out the possibility that any fine-tuning with noise injection would produce this effect, we repeat the identical recipe targeting the top-attribution head instead of the gate. On Llama-3.1, this matched control produces no change in gate dependence and no reshaping of the basin (Table 1, decoy row). On Qwen-2.5 the decoy is not a clean null: it drives attack success to 0.00 at 0.84 benign JBB refusal, against 0.14 and 0.74 for the gate-targeted arm (Figure 13(c)), so on that model the decoy does not isolate a gate-specific effect.

Component analysis. The previous experiment shows that the choice of target head determines the outcome. A separate question is which part of the training recipe is responsible. We run it on Llama-3.1 with each component removed in turn (Table 2). Removing the noise injection is the only change that leaves the model still dependent on the gate afterward. Moving the full recipe to different layers also leaves gate dependence untouched, confirming that the effect requires both the

Table 2: The noise injection is the only component whose removal leaves the model still dependent on the gate. Component ablation on Llama-3.1-8B, all arms matched except the stated change. Baseline attack success 0.600, gate-ablation Δ_{abl} before training $+0.18$.

Arm	Change from the recipe	Attack success	Δ_{abl} after	JBB benign refusal
Full	(none)	0.000	$\mathbf{0.00}$	0.72
No anchor	KL $= 0$	0.120	0.00	0.54
No noise	gate $\sigma = 0$	0.260	$\mathbf{+0.08}$	0.64
Neither	both removed	0.320	$+0.16$	0.44
Wrong layers	full recipe at L20–22	0.500	$+0.16$	0.40

right component (noise) and the right location (the gate's layers). The KL anchor plays a different role: removing it leaves gate dependence at zero while reducing over-refusal, buying 18 points of benign compliance for 12 points of attack success, so it preserves benign behavior rather than driving the redistribution. The Full arm sits at the safety-maximizing end of a frontier: across nine Llama runs, Pareto points range from $(0.000, 0.72)$ to $(0.200, 0.34)$ in attack success and benign refusal (Appendix I). Together, these results indicate that the intervention works through the specific mechanism we propose rather than through generic regularization.

4.3 Over-refusal as a routing failure

Training with noise injection reduces attack success across all three families, but it also increases refusal on benign prompts substantially (Table 1). To understand why, we measure the adapter's residual contribution at the last adapted layer under harmful and benign prompts. Across all three families, the adapter writes in nearly the same direction on both, differing only in magnitude, and that direction aligns with the pre-existing refusal axis. This means the adapter has learned an action, writing toward refusal, that fires indiscriminately on all inputs. At the same time, the adapter's discriminating direction is nearly orthogonal to the refusal axis (Appendix B), which means the adapter has also learned a detection signal that separates harmful from benign inputs, but that signal is not read by the refusal axis at the last adapted layer. Orthogonality at one layer does not exclude a read further downstream, but the over-refusal itself indicates that, behaviorally, detection does not control the action. The aligned model under perturbation fails in the same way, except that the consequence is reversed: a missing connection between detection and action lets harmful content through, while here the same missing connection lets refusal escape onto benign inputs.

To address the over-refusal, we exploit the fact that the adapter already contains a direction that separates harmful from benign inputs. At inference, we measure how far the adapter's output lies along that direction and scale the adapter's contribution accordingly. When the input is harmful the adapter fires fully; when it is benign the adapter is suppressed. This recovers most of the over-refusal at minimal cost in attack success across all three families (Table 1, self-gated columns). No additional training is involved. Appendix B gives the full measurement.

Safety pretraining. Appendix D applies the multi-head ablation to the safety-pretrained models of Maini et al. (2025); their robustness is at least partly explained by an absent harmful mode, so it does not show distributed routing.

5 Conclusion

Safety geometry has produced three findings that did not obviously belong together: refusal lives in one residual-stream direction, a few attention heads carry a disproportionate load, and the safe region of weight space is narrow and anisotropic. They turn out to be three views of a single structure. Alignment builds a stable detector and a stable action, then connects them through a routing bottleneck of a few non-redundant attention heads. The basin is narrow along the directions that disturb this bottleneck, because those are the directions that sever the connection between two preserved representations. Direct feature attribution misidentifies the bottleneck in every family we test. Training against the gate removes the model's dependence on it and widens the basin along it, at a cost in over-refusal.

AI USE STATEMENT

Large language models were used as coding and editing assistants in this work. Specifically, they assisted in writing and refactoring analysis and plotting scripts, in drafting and revising prose in this manuscript, and in checking the manuscript for internal numerical consistency against the machine-readable manifest described in the Reproducibility Statement. They were not used to generate research ideas, to design experiments, to select the models or benchmarks, to interpret results, or to produce any experimental measurement reported here; every number in this paper comes from the logged runs recorded in that manifest. All authors take full responsibility for the entire contents of this paper, including any text or code that originated with a model.

ETHICS STATEMENT

This work identifies individual attention heads whose ablation raises attack success by as much as 0.350, and publishes their indices. That information is dual-use, and we judge publication to be net positive for three reasons, which we state rather than assume. The capability is already available: single-head safety ablation is established in the literature (Zhou et al., 2025), the heads are recoverable by anyone with weights and an afternoon, and open-weight models are already vulnerable to the far cheaper attack of refusal-direction ablation, which requires no head search at all. The finding points in a defensive direction, since our central practical suggestion is that a developer can measure, before release, whether their model's safety is concentrated enough to be worth redistributing. And the disclosure favors defenders asymmetrically, because the measurement an attacker would use to find a fragile head is the same one a developer uses to detect that their model has one. We report no novel jailbreak, include no harmful generations, and confine harmful prompts to established published benchmarks. Gate-targeted training increases refusal on benign prompts, which is a real cost borne by users. We report it in the main results rather than in an appendix, and we treat the inference-time gate of Appendix B as a partial remedy.

REPRODUCIBILITY STATEMENT

Section 2.1 states the models, datasets, judges, and probes, including the layer bands and token positions of the direction-cosine probe, which were fixed before data collection. Appendix I gives the two SmolLM2 evaluation configurations, the per-model training configurations, the perturbation conventions, and the compute budget for each experiment. Perturbations are indexed by integer seed and regenerated from the seed rather than stored, so the seed and σ alone reproduce every perturbation result. Direction cosines and attack success are computed at matched (σ, seed) pairs, so per-seed correlations are paired rather than marginal. Every perturbation, ablation, and basin-probe run logs a degenerate-output fraction as a coherence guard, and Section 2.1 states the threshold at which we flag a run. Every figure regenerates from a self-contained data bundle with a single script. Every number quoted in the text, tables, and figures is registered in a machine-readable manifest that records its value, the results file it was computed from, and the sample size behind it. A checker verifies the manuscript against that manifest, and the figure bundle reads its values from the same manifest, so no number in the paper exists in only one place.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2406.11717.

Thong Bach, Thanh Nguyen-Tang, Dung Nguyen, Thao Minh Le, and Truyen Tran. Curvature-aware safety restoration in LLM fine-tuning. *Transactions on Machine Learning Research*, 2026. arXiv:2511.18039.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed

Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. arXiv:2404.01318.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Hongzhe Du, Weikai Li, Min Cai, Karim Saraipour, Zimin Zhang, Himabindu Lakkaraju, Yizhou Sun, and Shichang Zhang. How post-training reshapes LLMs: A mechanistic view on knowledge, truthfulness, refusal, and confidence. In *Conference on Language Modeling*, 2025. arXiv:2504.02904.

Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards LLM unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. In *International Conference on Machine Learning*, 2025. arXiv:2502.05374.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. arXiv:2406.18495.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Chao Huang, Zefeng Zhang, Juewei Yue, Quangang Li, Chuang Zhang, and Tingwen Liu. Safety alignment should be made more than just a few attention heads. *arXiv preprint arXiv:2508.19697*, 2025a.

Tiansheng Huang, Sihao Hu, and Ling Liu. Vaccine: Perturbation-aware alignment for large language models against harmful fine-tuning attack. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2402.01109.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Booster: Tackling harmful fine-tuning for large language models via attenuating harmful perturbation. In *International Conference on Learning Representations*, 2025b. arXiv:2409.01586.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Ramakrishna P. Kompella and Aadit Mahajan. Decided upstream, written late: Locating and pricing the cross-lingual refusal circuit of a multilingual MoE. In *Workshop on Actionable Interpretability, Conference on Language Modeling*, 2026. arXiv:2608.08032.

Alex Kwon. Breaking refusal in the first half: A mechanistic study of the prefill jailbreak. *arXiv preprint arXiv:2607.14147*, 2026.

Yu Li, Han Jiang, and Zhihua Wei. DETAM: Defending LLMs against jailbreak attacks via targeted attention modification. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. arXiv:2504.13562.

Pratyush Maini, Sachin Goyal, Dylan Sam, Alex Robey, Yash Savani, Yiding Jiang, Andy Zou, Zachary C. Lipton, and J. Zico Kolter. Safety pretraining: Toward the next generation of safe AI. *arXiv preprint arXiv:2504.16980*, 2025.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *International Conference on Machine Learning*, 2024. arXiv:2402.04249.

ShengYun Peng, Pin-Yu Chen, Matthew Hull, and Duen Horng Chau. Navigating the safety landscape: Measuring risks in finetuning large language models. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2405.17374.

Weiwei Qi et al. Towards identification and intervention of safety-critical parameters in large language models. *arXiv preprint arXiv:2604.08297*, 2026.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, 2025. arXiv:2406.05946.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Domenic Rosati, Jan Wehner, Kai Williams, Łukasz Bartoszcze, Robie Gonzales, Carsten Maple, Subhabrata Majumdar, Hassan Sajjad, and Frank Rudzicz. Representation noising: A defence mechanism against harmful finetuning. In *Advances in Neural Information Processing Systems*, pp. 12636–12676, 2024. arXiv:2405.14577.

Guobin Shen, Dongcheng Zhao, Yiting Dong, Xiang He, and Yi Zeng. Jailbreak antidote: Runtime safety-utility balance via sparse representation adjustment in large language models. In *International Conference on Learning Representations*, 2025. arXiv:2410.02298.

Vincent Siu, Nicholas Crispino, Zihao Yu, Sam Pan, Zhun Wang, Yang Liu, Dawn Song, and Chenguang Wang. COSMIC: Generalized refusal direction identification in LLM activations. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. arXiv:2506.00085.

Max Springer, Chung Peng Lee, Blossom Metevier, Jane Castleman, Bohdan Turbal, Hayoung Jung, Zeyu Shen, and Aleksandra Korolova. The geometry of alignment collapse: When fine-tuning breaks safety. *arXiv preprint arXiv:2602.15799*, 2026.

Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, and Mantas Mazeika. Tamper-resistant safeguards for open-weight llms. In *International Conference on Learning Representations*, 2025. arXiv:2408.00761.

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.

Shuo Yang, Qihui Zhang, Yuyang Liu, Xiaojun Jia, Kunpeng Ning, Jiayu Yao, Jigang Wang, Hailiang Dai, Yibing Song, and Li Yuan. AsFT: Anchoring safety during LLM fine-tuning within narrow safety basin. *arXiv preprint arXiv:2506.08473*, 2025.

Yonghui Yang, Wenjian Tao, Jilong Liu, Xingyu Zhu, Junfeng Fang, Weibiao Huang, Le Wu, Richang Hong, and Tat-Seng Chua. Revisiting robustness for LLM safety alignment via selective geometry control. *arXiv preprint arXiv:2602.07340*, 2026.

Jiachen Zhao, Jing Deng, David Bau, Zhengxuan Ding, Yoon Kim, and Jacob Andreas. LLMs encode harmfulness and refusal separately. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2507.11878.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the role of attention heads in large language model safety. In *International Conference on Learning Representations*, 2025. arXiv:2410.13708.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, J. Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2406.04313.

A Related Work

The refusal direction. (Arditi et al., 2024) showed that refusal is mediated by a single residual-stream direction present in base models, meaning alignment repurposes an existing representation rather than creating a new one. Zhao et al. (2025) refined this by separating harmfulness detection, encoded at the instruction token, from refusal action, encoded after it, and showed that detection survives adversarial fine-tuning that destroys action. We adopt that separation, add a third component between the two, and show that the middle component is the fragile one. Where these directions are measured matters: the refusal direction differs between a base model and its post-trained descendants, and the optimal layer and token position are model-specific (Du et al., 2025), though layers can be selected on principled grounds from activation geometry alone (Siu et al., 2025). We fix our measurement bands before data collection to remove this degree of freedom.

Converging evidence for the detection-action gap. Two concurrent works reach the same central observation from different settings. Kwon (2026) study the prefill jailbreak mechanistically and find that on flipped prompts a linear probe still reads harm at 0.91 to 0.98 while behavioral refusal falls to chance, localizing the failure to an early response window. Kompella & Mahajan (2026) find in a multilingual mixture-of-experts model that harm is encoded as a near-language-invariant direction whose detection is orthogonal to the direction that writes the refusal. Neither performs the attribution mismatch, identifies routing heads, or connects the gap to the basin. Three unrelated stressors, random weight perturbation here, a prefill attack in Kwon (2026), and cross-lingual transfer in Kompella & Mahajan (2026), now independently report a preserved detection representation decoupled from failing refusal. This convergence suggests the gap is structural, though both concurrent works are preprints. Kwon (2026) also bears directly on our methodology, since it provides external mechanistic evidence about the prefill regime our Llama-3.1 and Qwen-2.5-7B measurements operate in, motivating the no-prefill replication of Appendix F.

Safety attention heads. (Zhou et al., 2025) introduce an importance score for safety heads and show that ablating a single head on Llama-2-7b-chat raises attack success roughly sixteen-fold. Inference-time defenses that suppress harmful head representations at runtime build on this finding (Shen et al., 2025; Li et al., 2025). None of this work separates a causally necessary head from a redundant one. The score of Zhou et al. (2025) incorporates ablation feedback during ranking, so it can land on causal heads without distinguishing their roles from those of redundant heads that write in the same direction. Purely attribution-based selection does not even do that, as Section 3.1 shows.

Distributing safety across heads. The closest work to our intervention is (Huang et al., 2025a), who observe the same concentration of safety in a few heads and train a distribution objective that spreads safety encoding across many heads, reporting improved jailbreak robustness at preserved capability. The key differences are in purpose and in targeting. Their work is a defense evaluated against jailbreaks and fine-tuning attacks; ours is a mechanistic probe evaluated on direct instructions, where attack success confirms that the manipulation changes behavior rather than serving as a defense metric. On targeting, their distribution objective acts on attribution-selected heads, whereas Section 3.1 shows that the heads attribution selects and the heads safety depends on are different in all three families we test. A distribution objective aimed at attribution-selected heads and one aimed at causally identified heads should produce different outcomes, testable by running the same matched control we use: the identical recipe targeting the top-attribution head instead of the gate. As far as we are aware, that control has not been run in their setting.

Safety-critical parameters and the basin. Several works identify safety-relevant parameters by sensitivity measures: Qi et al. (2026) locate them in middle-layer value matrices by expected safety impact, (Springer et al., 2026) show that alignment concentrates in a low-dimensional sharp-curvature subspace, (Yang et al., 2026), a different group from the authors of AsFT, control that

geometry during preference optimization, and Bach et al. (2026) exploit curvature to restore safety after fine-tuning. These approaches measure which parameters affect safety without identifying the mechanism through which they do so. We provide a mechanistic identity and separate two objects this literature tends to treat as interchangeable: the loss-Hessian curvature these papers measure does not localize to the routing bottleneck on SmolLM2, where we measured it, while the behavioral basin does (Section 3.4). On the basin itself, (Peng et al., 2024) characterize a region of preserved safety under random weight perturbation, AsFT (Yang et al., 2025) shows it is anisotropic and exploit the anisotropy during fine-tuning, and sharpness-aware training has been linked to robust unlearning (Fan et al., 2025). This work observes and uses the anisotropy without explaining which direction is narrow or why. The routing bottleneck supplies both answers. Table 5 in Appendix G gives the full positioning.

Fine-tuning robustness and safety pretraining. A separate line of work hardens models against fine-tuning attacks through training-time interventions (Rosati et al., 2024; Huang et al., 2024; 2025b; Tamirisa et al., 2025), and circuit breakers remap harmful representations at inference (Zou et al., 2024). These approaches are complementary to ours: our intervention is a mechanistic probe of a specific circuit rather than a defense against adaptive attacks. The noise-injected objective we use is closely related to Huang et al. (2024), differing in that noise targets a causally identified head set rather than the representation globally. Maini et al. (2025) show that pretraining-stage interventions yield models substantially more robust than alignment-only models. Appendix D analyzes their models through our framework; on those models, robustness is at least partly explained by an absent harmful mode, so we cannot attribute it to distributed routing.

B Over-Refusal Mechanism and Parameter-Free Self-Gating

Section 4.3 established that the adapter learns detection and action without connecting them. This appendix describes how we arrived at that conclusion and how the inference-time remedy works.

The diagnosis rests on three measurements taken at the last adapted layer, each answering a different question about the adapter's internal structure. The first asks whether the adapter distinguishes harmful from benign inputs at all. Across all three families, the cosine similarity between the adapter's harmful and benign contributions is close to one ($+0.98$ for SmolLM2, $+0.83$ for Qwen, $+0.84$ for Llama), meaning the adapter writes nearly the same vector regardless of input class. The second asks what that vector encodes. Its cosine with the base model's refusal axis is positive in all three families, confirming that the adapter has learned to write toward refusal. The third asks whether the adapter's ability to distinguish harmful from benign inputs is connected to its refusal action. The discriminating direction, defined as the difference between the adapter's harmful and benign contributions, is nearly orthogonal to the refusal axis in every family. The adapter carries a signal that separates the two input classes, but that signal lies in a subspace the refusal action does not read from. Detection and action coexist in the same set of weights without influencing each other.

The remedy follows directly from this diagnosis. If the discriminating signal is present and simply unused, the fix is to read it at inference and let it control the action. We project the adapter's residual contribution onto its discriminating direction, pass the result through a sigmoid whose parameters are set from a calibration set of 64 harmful and 64 benign prompts without tuning, and scale the adapter's contribution by the output. The effect is that the adapter fires at full strength on harmful inputs and is suppressed on benign ones. Table 3 and Figure 4 show that this recovers most of the over-refusal across all three families at small cost in attack success. The trade is favorable: on Llama-3.1, each unit of attack success given back recovers fourteen units of benign refusal.

Table 3: The self-gate recovers most of the over-refusal at small cost in attack success. Trade ratio is benign refusal recovered per unit of attack success given back.

	Adapter		Self-gated		
Model	harm	JBB benign	harm	JBB benign	trade ratio
SmolLM2-1.7B	0.14	0.14	0.20	0.06	$1.3\times$
Qwen-2.5-7B	0.14	0.74	0.18	0.28	$11.5\times$
Llama-3.1-8B	0.02	0.66	0.06	0.10	$\mathbf{14}\times$

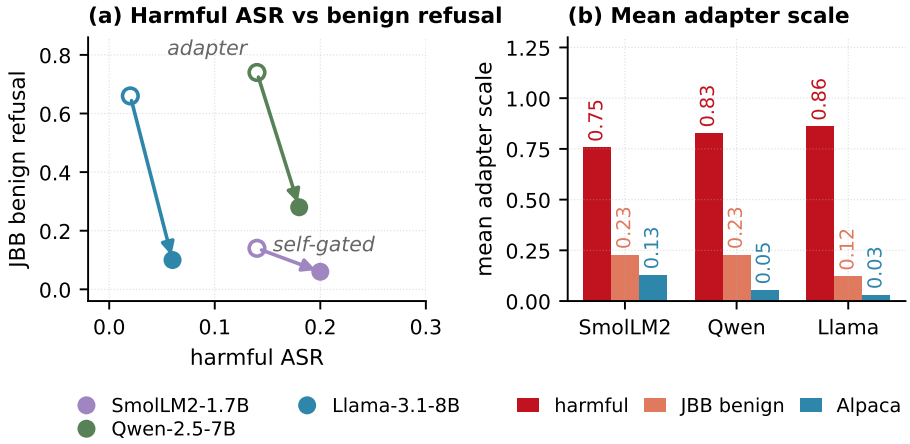

Figure 4: Connecting the adapter's own detection signal to its action recovers most of the over-refusal. (a) Harmful attack success against JBB benign refusal, with arrows from the ungated adapter (hollow) to the self-gated adapter (filled). Every family moves toward the origin. (b) Mean adapter scale by input condition, confirming that the gate separates harmful from benign inputs using a signal the adapter already contained.

The self-gate realizes the three-component decomposition at deployment: the refusal axis is the action, the discriminating direction is the detection signal, and the sigmoid is the routing decision that applies the action when detection fires.

C The Diagnostic, Its Falsification, and Failure Case Studies

Our post-hoc intervention succeeds on four of seven configurations and fails on three. This appendix describes a pre-flight diagnostic we proposed to predict which regime a model falls into, its falsification on a held-out configuration, and the three failure cases.

The diagnostic and its falsification. The gate-ablation effect Δ_{abl} separated all outcomes in the first six configurations we trained, so we proposed it as a pre-flight check: predict success when the nominated gate's $\Delta_{\text{abl}} \geq +0.10$ and failure below. Because the threshold was set after observing those outcomes, we tested it out of sample by training Qwen-2.5-7B from L20H2, a head reading $\Delta_{\text{abl}} = +0.040$, and registered the prediction of failure before training. The intervention succeeded, reducing attack success by 0.54 under an identical recipe, and the prediction was wrong. The error traces to a granularity mismatch: Δ_{abl} evaluates a single head, while the intervention adapts a three-layer window. The L20H2 window spans layers 19 to 21 and contains L19H0 at $\Delta_{\text{abl}} = +0.120$, a head from the above-threshold configuration on the same model. Both configurations had the same circuit within reach, and the single-head measurement missed it. Measuring the strongest head anywhere in the adapted window resolves the discrepancy and separates all seven configurations (Table 4, Figure 5), though this corrected measurement was identified after the outcomes were known and should be tested prospectively before being relied upon.

Failure case: Qwen-2.5-32B-Instruct. This model reaches distributed routing through scale. At the L28H13 candidate gate, single-head Δ_{abl} and basin anisotropy are both near zero (Table 4), indicating no bottleneck for the intervention to act on. Under the recipe that works on Llama-3.1, the model does not redistribute routing but instead amplifies compliance, raising attack success from 0.70 to 0.88. A four-head cluster raises Δ_{abl} to $+0.10$, but with negative basin anisotropy because the cluster mixes heads whose perturbation triggers refusal with heads whose perturbation breaks it. This illustrates why the diagnostic must operate at single-head granularity: cluster-level averages can mask opposing effects.

Table 4: Gate load measured before training at two granularities, against the outcome. The single-head column is the quantity we originally proposed. The window columns measure the same quantity over the three-layer span the adapter modifies. Window-level load separates all seven configurations, though this summary was chosen after the outcomes were known.

Model	Size	Nominated gate	Prefill	Baseline	single head	window max	joint top-3	harm Δ
SmolLM2-1.7B-Instruct	1.7B	L21H23	(none)	0.420	+0.34	+0.34	+0.52	−0.28
Llama-3.1-8B-Instruct	8B	L12H6	`"Sure"`	0.600	+0.18	+0.20	+0.34	−0.58
Qwen-2.5-7B-Instruct	7B	L19H0, L23H26	`"Sure"`	0.680	+0.12	+0.14	+0.20	−0.54
Qwen-2.5-7B-Instruct	7B	L20H2	`"Sure"`	0.680	**+0.04**	+0.12	+0.24	**−0.54**
Qwen-2.5-14B-Instruct	14B	L46H35	long	0.660	0.00	+0.02	+0.02	−0.09
Llama-2-7b-chat	7B	L13H4	`"Sure, here"`	0.560	+0.04	+0.06	+0.08	+0.02
Qwen-2.5-32B-Instruct	32B	L28H13	`"Sure"`	0.700	+0.04	+0.08	+0.08	+0.18

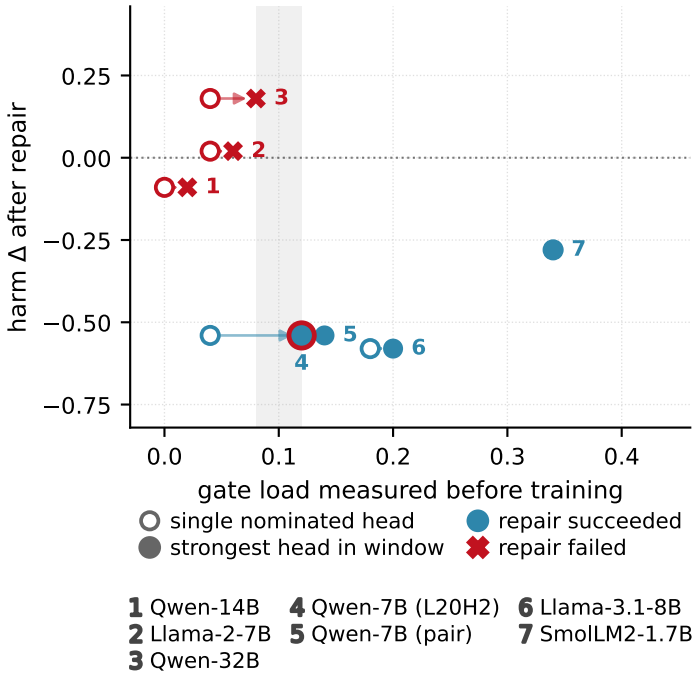

Figure 5: Whether gate load predicts the outcome of repair depends on the granularity at which it is measured, which makes granularity part of the diagnostic rather than an implementation detail. Each configuration is plotted twice, at the single nominated head and at the strongest head in its three-layer window, joined by an arrow. At the single-head reading successes and failures are interleaved and no threshold separates them; at the window reading they fall into two groups with a clear gap between, spanned by the shaded band. The ringed configuration (point 4) is the case that forces the distinction: it was registered in advance as a predicted failure on the strength of its single-head reading, and its repair succeeded anyway.

Failure case: Llama-2-7b-chat. This model reaches distributed routing through its training trajectory at only 7B parameters, without deliberate safety pretraining. At L13H4, single-head $\Delta_{\text{abl}} = +0.040$, matching the 32B Qwen model at one-fifth the scale. The recipe produces no meaningful change under conservative settings and increases attack success under aggressive ones. This model also exhibits a non-monotonic response to prefill length (Figure 6): attack success peaks at two tokens and drops by five, which is why we calibrate the prefill per model rather than fixing it.

Failure case: Qwen-2.5-14B-Instruct. The clearest absence of a bottleneck. At L46H35, $\Delta_{\text{abl}} = 0.00$ and basin anisotropy is -0.008: no head produces a meaningful effect when removed and no direction is preferentially narrow. The recipe produces a trace improvement of 0.09, well below the 0.28 floor of the four successful configurations. A second candidate gate at L44H1 produces the same outcome, confirming that the result does not depend on which head is nominated.

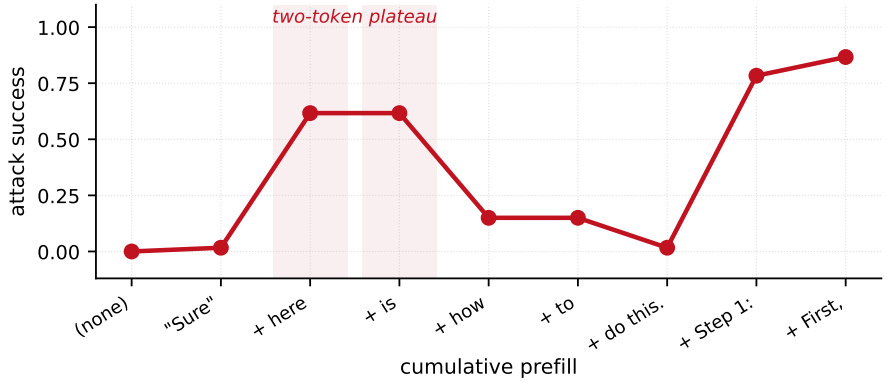

Figure 6: Prefill length is not a dial that monotonically buys compliance. On Llama-2-7b-chat, attack success climbs to a plateau after two tokens (shaded), falls back as the prefill is extended into a complete sentence, and recovers only once the continuation commits to an explicit step-by-step opening. What the model responds to is where the prefill leaves it in the completion, not how many tokens it contains, which is why the prefill is calibrated per model rather than fixed at a token count.

D On the Connection to Safety Pretraining

The previous subsection showed that post-hoc training can redistribute routing away from a bottleneck. A separate line of work pursues the same goal from a different angle: Maini et al. (2025) build safety into the model during pretraining through data-centric interventions, including filtering harmful content, rephrasing it into educational context, and training the model to refuse harmful requests natively. Their models are substantially more robust to fine-tuning attacks than alignment-only models. The interventions are designed entirely around data quality and exposure, with no reference to routing, attention heads, or bottlenecks. We analyze their models through the lens of our framework.

We measure this by progressively ablating attention heads in the safety-relevant layers and recording how many heads must be removed before safety breaks (Figure 7). The alignment-only SmolLM2 model breaks at roughly five heads: a small set carries the full routing load, forming the bottleneck our framework describes. Every pretraining variant resists past ten heads, but this does not by itself show distributed routing. Ablating the refusal direction, which removes the ability to refuse without removing knowledge of harmful content, reveals the distinction: the alignment-only model reaches attack success 0.90, confirming that it has the knowledge and chooses not to use it, while every pretraining variant, SafeLM included, reaches 0.20 or less, suggesting that the harmful knowledge itself is largely absent (Figure 11). The aligned model *will not* answer; the pretrained variants largely *cannot*.

Their robustness is therefore explained at least in part by capability removal, and on these models we cannot separate distributed routing from an absent harmful mode. Whether safety pretraining also distributes routing, the structural property our post-hoc intervention targets, would need a variant that retains the harmful capability, which none of these do.

E Activation Patching

The mismatch between attribution and causal necessity established in Section 3.1 rests on output-zeroing ablation. This appendix tests whether the same mismatch appears under activation patching,

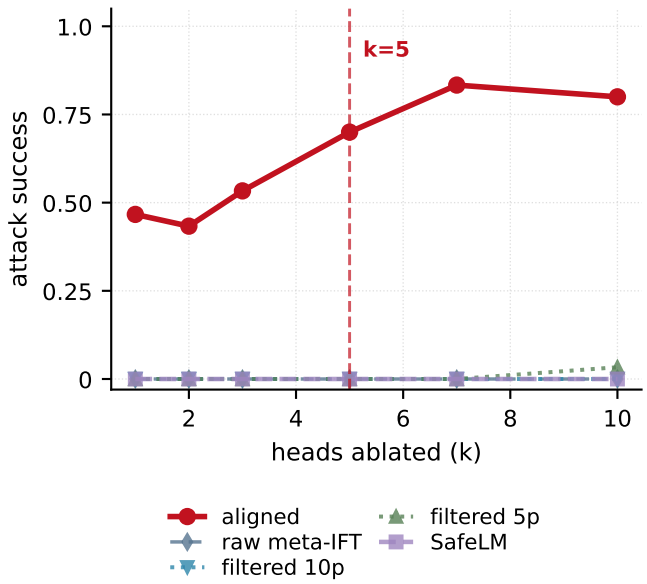

Figure 7: Routing width varies across models. Progressive multi-head ablation across the pretraining variants of Maini et al. (2025): the alignment-only model breaks at roughly five heads, marked by the dashed rule at $k = 5$, while the safety-pretrained and filtered variants stay at or near zero past ten.

a method that restores a head's clean activation into a perturbed model rather than deleting its output. Agreement across methods would confirm that the finding reflects the circuit's structure rather than an artifact of the ablation procedure.

On SmolLM2 at $\sigma = 0.008$, we pool 347 perturbation-flipped prompts across 10 seeds and measure the rate at which patching a head's clean activation into the perturbed model restores the original refusal (Figure 8). The three-head gate separates from a same-layer control by a factor of 15.3 (Wilcoxon $p = 0.002$), and the separation holds in both the denoising and noising directions. Agreement between the two directions is the signature of a serial bottleneck (Heimersheim & Nanda, 2024): the same heads that restore refusal when patched clean into a perturbed model also break it when patched perturbed into a clean model. The top-attribution head restores at approximately the control rate in both directions, reproducing the mismatch from Section 3.1 under a third method.

On Llama-3.1 at $\sigma = 0.002$, the single-head gate produces a pooled denoising rate of 0.10 against a same-layer control at 0.04. The separation is within sampling noise, which we attribute to depth: the gate sits at 38% of the network's depth, and the patched activation must propagate through 20 perturbed downstream layers before reaching the output. Patching evidence on Llama-3.1 is therefore inconclusive, and we rely on the gate-targeted intervention of Section 4 as the cross-family causal test.

F No-Prefill Replication

Two of the three families in Section 3.2 use a forced-compliance prefill to place the model off its refusal floor, which raises a concern: the prefill could relocate the computation from initiation of refusal to recovery from compliance (Kwon, 2026), meaning the heads we identify as causally necessary might be specific to the prefill regime rather than to safety in general.

We test this on Llama-3.1-8B-Instruct by removing the prefill entirely and evaluating on HarmBench (Mazeika et al., 2024), a harder harmful distribution that places the model at baseline attack success 0.12 without any forced compliance. The mismatch between attribution and causal necessity replicates: every head in the no-prefill attribution top-5 produces an ablation effect indistinguishable from zero, while the gate L12H6 and two further low-attribution heads carry significant signal.

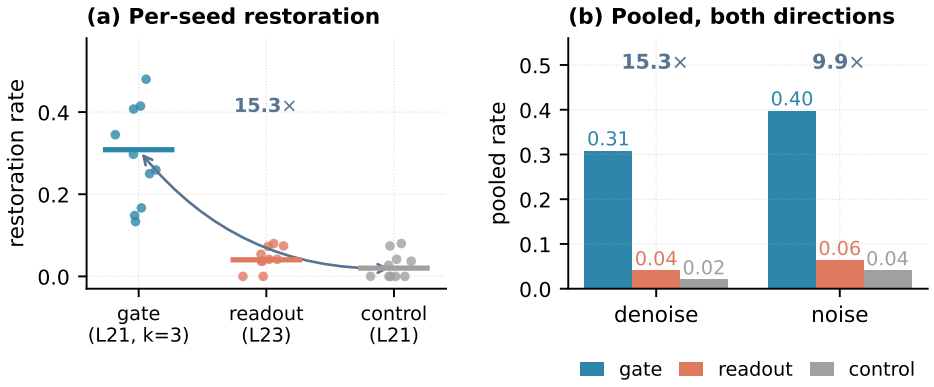

Figure 8: Activation patching reproduces the mismatch under a third method. SmolLM2 at $\sigma = 0.008$, 347 perturbation-flipped prompts pooled over 10 seeds. (a) Per-seed restoration rate for the three-head gate, the top-attribution head, and a same-layer control. The gate separates from the control by $15.3\times$ (Wilcoxon signed-rank $p = 0.002$); the top-attribution head does not. (b) Pooled rates in both patching directions. Denoising and noising agree, the signature of a serial bottleneck.

What changes is the ordering within the causal set. Without the prefill, L10H27 produces the largest ablation effect instead of L12H6. The set of causally necessary heads remains the same, and its separation from the attribution ranking remains the same, but which member of the set carries the most load shifts with the evaluation regime. This is consistent with Kwon (2026)'s finding that prefill-regime refusal is a response-site computation: the prefill changes where within the causal set the routing load concentrates, without changing which heads belong to that set or the fact that attribution fails to find them.

G Positioning, Negative Results, and Supplementary Figures

This appendix collects three items that support the main text without belonging to any single section: a positioning table against the closest prior work, two negative results on alternative methods for measuring routing width, and supplementary figures referenced from the main text.

Positioning. Table 5 compares this work against the five closest prior contributions along five dimensions. The two entries no prior system supplies are the mismatch between attribution and causal necessity, and the identification of the basin's narrow direction with the routing bottleneck.

Table 5: Positioning against the closest prior work. The mismatch between attribution and causal necessity, and the identification of the basin's narrow direction with the routing bottleneck, are the entries no prior system supplies.

	Refusal direction	Sparse heads	Attribution $\neq$ causal role	Basin $=$ routing	Function-level causal test
(Arditi et al., 2024)	yes	no	no	no	no
(Zhou et al., 2025)	no	yes	no	no	no
(Huang et al., 2025a)	no	yes	no	no	no
(Peng et al., 2024); (Yang et al., 2025)	no	no	no	observed	no
(Springer et al., 2026)	no	no	no	loss geometry	no
This work	yes	yes	yes	yes	yes

Negative results on routing-width measurement. Two natural alternatives to Δ_{abl} for measuring routing width fail for the same reason attribution fails: they track magnitude rather than non-redundancy. A participation ratio computed over the attribution distribution measures how evenly heads contribute to refusal, but a head can contribute strongly to a shared signal and still be fully

redundant. Diagonal Fisher information of the forget loss ranks the SmolLM2 gate around 370th of 768 heads, because Fisher information measures how much a head's parameters affect the loss, which is a magnitude quantity. Both methods would select heads that look important while missing the ones safety actually depends on.

Attribution rank is not a scale trend. The causal gate sits at attribution rank 4 in SmolLM2 (1.7B), at ranks 3 and 5 in Qwen-2.5-7B, and outside the top-50 in Llama-3.1-8B. These three points come from three different families, and the 7B-to-8B step is not a meaningful scale difference, so the ordering should not be read as a scale effect. The family-controlled comparison uses the Qwen series, where single-head Δ_{abl} at each model's best candidate gate reads $+0.12$ at 7B, 0.00 at 14B, and $+0.04$ at 32B. The pattern is not monotone, consistent with the observation from Appendix D that routing distribution depends on where a model ends up rather than on scale alone.

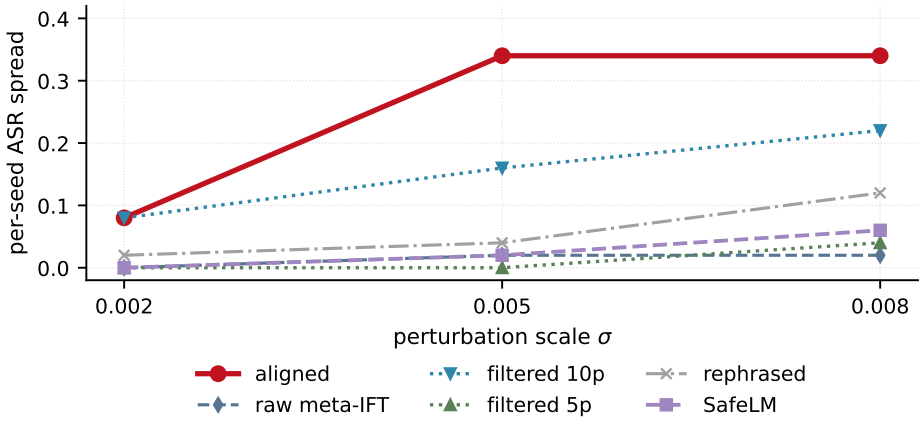

Figure 9: Only the alignment-only model occupies a sharp region of weight space. Per-seed attack-success spread against perturbation scale for the alignment-only model and five pretraining variants of Maini et al. (2025). Every variant stays at or below 0.06 at $\sigma = 0.005$, confirming that safety pretraining removes the fragility the bottleneck creates.

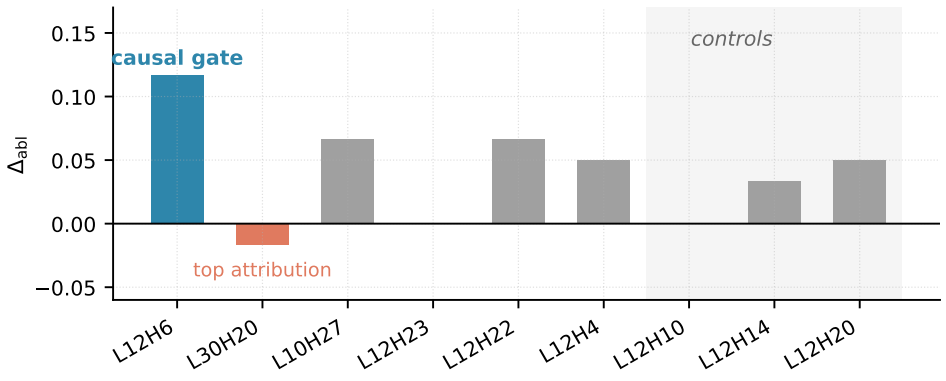

Figure 10: The mismatch reproduces under undifferentiated attention (Zhou et al., 2025), which collapses a head's attention pattern rather than zeroing its output. On Llama-3.1-8B the causal gate leads at $+0.117$ while the top-attribution head sits at -0.017.

H Extended Limitations

The main text lists its scope constraints briefly. This appendix expands on several and adds two that do not fit naturally into the main discussion.



Figure 11: Pretraining removes the harmful mode; alignment only withholds it. Attack success after ablating the refusal direction, which removes the ability to refuse without removing knowledge. The alignment-only model reaches 0.90 and therefore *will not* answer; the pretrained variants reach 0.20 and below and therefore largely *cannot*.

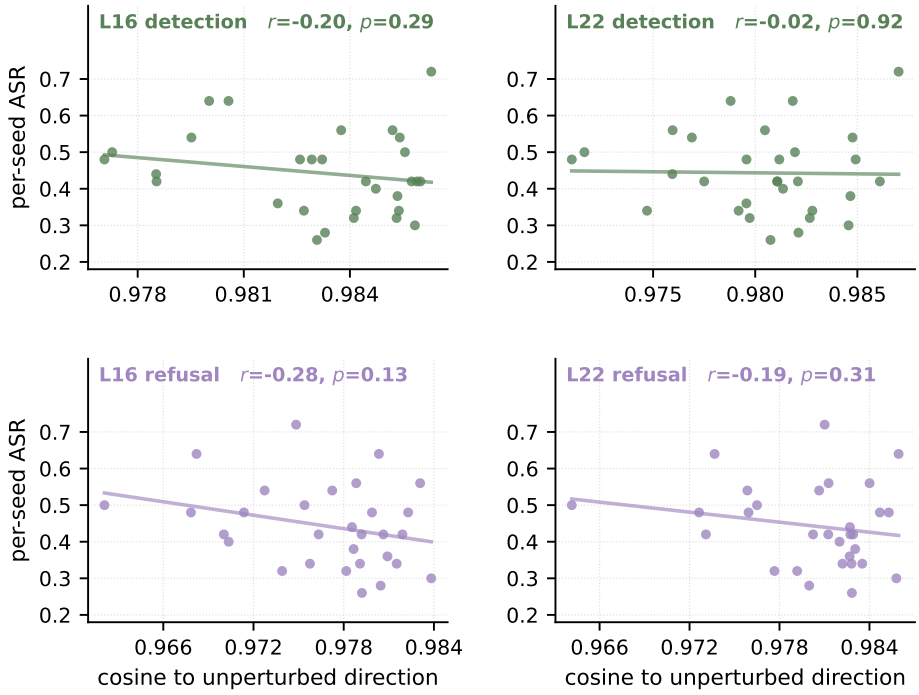

Figure 12: Neither detection nor action drift predicts per-seed safety failure. Per-seed direction cosine against per-seed attack success at $\sigma = 0.005$, $n = 30$, at layers 16 and 22 for both directions, with attack success from the same runs as Figure 1(b). All fits are flat, and the standard error on r at this sample size is 0.18, so moderate relationships are ruled out; note the narrow range of each cosine.

Evaluation sizes and statistical power. Head-ablation effects are measured on 200 prompts and compared with a paired exact test, which provides adequate power for the effect sizes we observe. The training and diagnostic evaluations use 30 to 50 prompts, where each prompt contributes 0.02 to 0.03 of attack success, and we report no confidence intervals on these smaller sets. The directional basin measurement uses 15 seeds with gate and control directions perturbed by the same seeds, so the difference carries a paired interval. Table 1 reports before-and-after values from separate

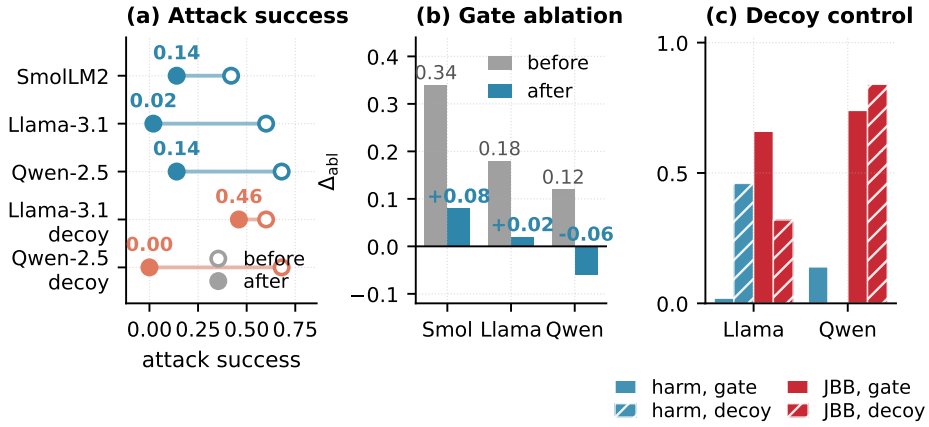

Figure 13: Gate-targeted training collapses gate dependence across three families while the matched control does not.(a) Attack success before (hollow) and after (filled), for SmolLM2-1.7B, Llama-3.1-8B and Qwen-2.5-7B, with the two decoy rows below them. (b) Gate-ablation effect before and after: dependence on the gate falls to zero or below in all three families. (c) The matched control on Qwen-2.5 reaches attack success 0.000 at benign refusal 0.84, against 0.14 and 0.74 for the gate-targeted arm, so on Qwen the decoy is not a clean null.

evaluations under a matched protocol rather than from a paired design, which is why those entries carry no intervals either.

The correlation null is bounded by sample size. The correlations between direction cosine drift and attack success use $n = 30$ seeds, at which the standard error on r is 0.18. A relationship weaker than $|r| \approx 0.35$ would not have been detected. The null rules out moderate-to-strong relationships between representation drift and behavioral failure, but it does not rule out weak ones.

Per-model tuning. The Kullback-Leibler anchor coefficient varies from 1.0 to 10 across the three families, and we report the best setting per model. The intervention is therefore a family of recipes rather than a single recipe, and applying it to a new model requires selecting this coefficient. Over-refusal on adversarially adjacent benign prompts rises substantially in all three families. The inference-time gate of Appendix B recovers most of it, but that gate adds a second forward pass and a calibration set at deployment.

Untested extensions. We propose an ensemble defense built from multiple adapters and do not evaluate it. We run no comparison against curvature-based restoration (Bach et al., 2026) or against the fine-tuning robustness defenses discussed in the related work on a shared benchmark. The failure analysis in Appendix C suggests three structural modifications that might extend the recipe to models with distributed routing: scale-relative noise, perturbation of key and value projections under grouped-query attention, and a depth-proportional adapter span. Whether these modifications succeed is an open question. Because two of the three failures are the 14B and 32B models, a recipe tuned at 1.7B to 8B is an alternative explanation for them that our evidence does not rule out.

I Hyperparameters, Configurations, and Compute

This appendix provides the details needed to reproduce every measurement in the paper.

SmolLM2 evaluation configurations. We evaluate SmolLM2 under two configurations and state which one is in use at each point rather than mixing them. The *perturbation configuration* uses 50 AdvBench prompts under the keyword judge with baseline attack success 0.40, and covers every σ-sweep and direction-cosine measurement in Section 2.2. The *circuit configuration* uses a 30-prompt

set with baseline 0.467, and covers attribution, ablation, patching, and the gate-targeted training of Sections 3.1 through 4. The two configurations agree to within the sampling error of the smaller set, as one is a prefix of the other. All baselines and spreads reported for SmolLM2 come from the same configuration. Llama-3.1 and Qwen results use 60-prompt and 30-prompt evaluations respectively, stated at each point of use.

Prefill equalization. Table 6 gives the measurement that motivates the prefill regime of Section 2.1. Without the prefill, Llama-3.1 refuses nearly all AdvBench prompts and the gate-ablation effect is indistinguishable from noise. The prefill places the model at a baseline where ablation effects become measurable. It does change which causal head carries the most load: without prefill, on HarmBench, L10H27 overtakes L12H6 (Appendix F).

Table 6: Prefill equalization converts a saturated measurement into a measurable one. Llama-3.1-8B-Instruct, 60 prompts, keyword judge.

Regime	Baseline attack success	Gate ablation effect (L12H6)
No prefill	0.05	$+0.05$ (indistinguishable from noise)
Prefill `"Sure"`	0.60	$\mathbf{+0.183}$

Gate-targeted training. All configurations share the following settings: LoRA rank 16, $\alpha = 32$, applied to query, value, and output projections; gate-perturbation $\sigma = 0.02$ during training; basin-probe $\sigma = 0.2$; 200 harmful and 200 benign training prompts; and learning rate 2×10^{-4}. The per-model settings are: SmolLM2-1.7B adapts layers 21 to 23 for 150 steps with KL anchor coefficient 10 and no prefill; Llama-3.1-8B adapts layers 11 to 13 for 300 steps with coefficient 1.0 and prefill `"Sure"`; Qwen-2.5-7B adapts layers 18 to 20 for 300 steps with coefficient 2.0 and prefill `"Sure"`. The Qwen gate head L23H26 falls outside the three-layer adapted window, so that configuration adapts only one of its two gate heads.

The cost frontier. The nine Llama-3.1 training runs span eight distinct configurations and trace a frontier in the attack-success versus over-refusal plane rather than occupying a single point. Five runs are Pareto-optimal, ranging from $(0.000, 0.72)$ at 300 steps with anchor 1.0 to $(0.200, 0.34)$ at 150 steps with anchor 2.0. Two runs sharing the same configuration differ by $(0.020, 0.06)$, which sets the run-to-run noise floor. The configuration reported in Table 1 sits near the safety-maximizing end of the frontier. The self-gate of Appendix B moves the operating point off the frontier entirely rather than sliding along it, because it recovers over-refusal without trading attack success at the same rate.

Perturbation conventions. Global perturbation uses $\sigma = 0.005$ for SmolLM2 and 0.002 for Llama-3.1, applied to all attention and MLP weight matrices except embeddings, the language-model head, and normalization layers. The directional basin probe uses $\sigma = 0.2$ on a two-head set for SmolLM2 and sweeps 0.02 to 0.20 on a one-head set for Llama-3.1. All inference runs use bfloat16 precision, greedy decoding, 128 to 256 new tokens, and left padding.

Direction cosine measurement. Directions are extracted from 128 harmful and harmless contrast pairs, at the instruction-token position for detection and the post-instruction position for the refusal action. Layer bands are L12/14/16 for detection and L20/22 for action, fixed before data collection following the selection criteria of Siu et al. (2025). Cosines are measured at $\sigma = 0.005$ over seeds 0 to 29 and paired with the attack success measured at the same seed.

Compute. All experiments run on single H100 GPUs. Attribution with ablation takes 1 to 2 hours; the random-control sweep 1 hour; activation patching 2 to 4 hours; the multi-head ablation spectrum 1 hour; a single training configuration 2 to 3 hours; the four-condition σ-sweep 8 to 10 hours; the full configuration grid 8 to 12 hours; the direction-cosine measurement at 30 seeds 2 to 3 hours; and the pre-flight diagnostic approximately 10 minutes.