FAITHFULNESS ANCHORING AS DEFENSE AGAINST COT HIJACKING: A MECHANISM-FIRST APPROACH VIA RESIDUAL-STREAM ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety-aligned reasoning models are vulnerable to Chain-of-Thought (CoT) hijacking, a jailbreak that prepends long benign reasoning traces before a harmful instruction. Understanding why this attack succeeds is a prerequisite for building principled defences, yet the internal mechanics of CoT hijacking remain unexplored. We track the model's refusal signal, a projection onto a refusal direction in the residual stream, across all token positions during hijacked generations. Our key finding is benign padding does not gradually erode this signal, it suppresses the signal uniformly while the model processes the input tokens. Attack outcome is determined by whether the signal recovers at generation onset or remains suppressed. We confirm this relationship across three model families (Pearson $r \approx -0.52$, $p < 0.0001$, n=150 per model) and validate it causally via activation patching. Building on this finding, we propose two defenses: *faithfulness anchoring*, a prompt-level intervention that requires no access to model internals, and *activation steering*, which directly adds the refusal direction back into the residual stream. Faithfulness anchoring reduces mean ASR (Attack Success Rate) by 22.7pp across three models while keeping benign task accuracy close to baseline; activation steering achieves smaller ASR reductions (mean $+11.3$pp) at real, model-dependent capability cost, making the prompt-level defence the better-calibrated choice. Finally, we find the defence is not adaptively robust: attacks targeting the anchor's own instruction recover much of the original attack surface.

1 INTRODUCTION

Safety-aligned language models can be jailbroken by a surprisingly simple manipulation: prepend a long sequence of benign reasoning before a harmful request. The model, having just completed several coherent puzzle solutions, continues the established generation pattern rather than evaluating whether the final request should be refused. We term this *CoT hijacking* and study it as a first-class attack class. It succeeds at non-trivial rates across aligned open-weight models without gradient-based optimization, special tokens, or any modification to the harmful instruction itself.

The obstacle to defending against this attack is that existing defenses operate without knowing *where* the model's safety mechanism is failing. Instruction-tuned models refuse harmful requests by activating a low-dimensional safety direction in their residual stream Arditi et al. (2024), identified at the prompt/response boundary, but whether this signal persists once generation is underway remains unexamined. That question is increasingly urgent: aligned models can be jailbroken through adversarial suffixes Zou et al. (2023b), competing objectives Wei et al. (2023), and long-context padding that overrides safety training in a power-law scaling regime Anil et al. (2024).

We take a mechanism-first approach: extending the refusal signal projection of Arditi et al. (2024) from a single-position diagnostic to a per-token trajectory across the full hijacked generation to identify the precise point at which safety fails, then building two defenses that target it directly, one at the prompt level requiring no model access, another in activation space. This gap persists despite systematic evaluations of reasoning model safety Jiang et al. (2025): prior work characterises symptoms of the attack without grounding a defense in an internal, mechanistic account of why

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.

it succeeds. Safety alignment concentrates in the first few output tokens Qi et al. (2025), making generation onset the critical window for both attack and defense.

Our main contributions are:

- We introduce a per-token faithfulness score $F(t)$ and show that its suppression at response onset is a causal driver of attack success, not merely a correlate (§3.2, §5.1).
- We propose two defenses that target this failure point directly: Defense A, a prompt-level anchoring intervention requiring no access to model internals, and Defense B, an activation-steering intervention in representation space (§3.3, §3.4).
- We evaluate Defense A against an adaptive adversary and identify direct targeting of the anchor's instruction as its primary residual vulnerability (§3.5).

2 Related Work

Faithfulness And CoT Reasoning. Jacovi and Goldberg Jacovi & Goldberg (2020) distinguish faithfulness from plausibility, requiring explanations to be causally connected to predictions. Turpin et al. Turpin et al. (2023) show CoT reasoning is biased by irrelevant prompt features, causing stated reasoning to diverge from actual decision factors. Lewis-Lim et al. Lewis-Lim et al. (2025) find CoT influence and faithfulness are dissociable by tracking confidence trajectories; we extend this to the safety domain. Tutek et al. Tutek et al. (2025) operationalise step necessity via unlearning, and Parcalabescu and Frank Parcalabescu & Frank (2024) extend faithfulness evaluation to multimodal settings. We frame CoT hijacking as a safety-domain faithfulness failure: the model's internal representation of harmful intent diverges from its overt reasoning trace, causing refusal to fail.

Mechanistic Interpretability And Residual-Stream safety. Arditi et al. Arditi et al. (2024) identify a low-dimensional refusal direction $\hat{\mathbf{u}}^{(\ell^*)}$ whose ablation induces compliance across model families; this is the mechanistic foundation of our work, whose suppression we track during hijacked generation and restore via activation steering (Defense B) or prompt-level anchoring (Defense A). Li et al. Li et al. (2023) establish activation-based behavioural correction via inference-time intervention; Zou et al. Zou et al. (2023a) show linear safety representations are consistent across layers and scales; Zhang and Nanda Zhang & Nanda (2024) establish activation-patching best practices, which we follow in our causal validation.

Jailbreaks And Long-context Attacks. Zou et al. Zou et al. (2023b) introduce adversarial suffix attacks (GCG); Wei et al. Wei et al. (2023) taxonomise jailbreaks as competing objectives and mismatched generalisation. Anil et al. Anil et al. (2024) show benign in-context padding overrides safety training in a power-law scaling regime; this is the long-context attack class we instantiate in the reasoning-trace setting. Qi et al. Qi et al. (2025) show safety alignment concentrates in the first few output tokens, motivating our focus on generation onset. Xu et al. Xu et al. (2024) study preemptive answer injection into CoT reasoning and propose problem restatement as a partial mitigation, but target reasoning accuracy under a behavioural threat model; we address the safety-domain variant and give the first mechanistic account of why restatement alone fails without generation-onset timing.

Prompt-level Defenses And Adaptive Robustness. Xie et al. Xie et al. (2023) propose Self-Reminder; Zhang et al. Zhang et al. (2024) instruct the model to prioritise safety explicitly; our `self_remind` and `goal_prioritize` variants operationalise these within our phrasing ablation, with `goal_prioritize` outperforming both. Unlike prior work, we inject at response onset (s=0) rather than the system prompt, targeting the failure point our mechanistic analysis identifies. Adaptive robustness of prompt-level defenses remains underexplored, with prior work evaluating only standard attacks; our evaluation tests three variants designed specifically to circumvent faithfulness anchoring.

3 Methodology

3.1 Attack Setup: CoT Hijacking via Puzzle Padding

We study a *chain-of-thought hijacking* attack in which an adversary prepends a sequence of benign reasoning puzzles to a harmful request, exploiting the model's tendency to continue an established reasoning pattern rather than evaluate the final request on its own merits. Formally, given a harmful

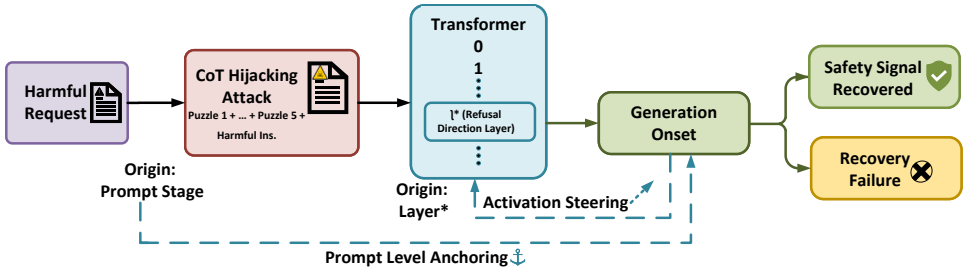


Figure 1: CoT hijacking and proposed defenses. A harmful request is padded with $K{=}5$ benign puzzles (red), suppressing the refusal signal $F(t)$ at layer l^*. At generation onset the model either recovers $F(t)$ and refuses (green) or complies (yellow). Defense A injects an anchoring sentence at response onset; Defense B adds the refusal direction u_R to the residual stream at l^* (both dashed blue). Both restore $F(t)$ at the same decision point.

goal g, the attack constructs a padded prompt $x_{\text{atk}} = [p_1, p_2, \ldots, p_K, g]$, where $p_1, \ldots, p_K$ are K benign puzzle–solution pairs drawn from a fixed pool. Our primary setting is $K{=}5$; we additionally sweep $K \in \{8, 12\}$ in §A.8 to test whether the uniform-suppression pattern described below holds as padding length increases.

3.2 Faithfulness Metric $F(t)$

We define a scalar *faithfulness score* $F(t)$ measuring alignment with the refusal subspace at each token position t. Following Arditi et al. Arditi et al. (2024), we extract a model-specific refusal direction $u_R \in \mathbb{R}^d$ at layer l^* via mean-difference:

$$u_R = \frac{\mu_{\text{harm}} - \mu_{\text{harmless}}}{\|\mu_{\text{harm}} - \mu_{\text{harmless}}\|}, \tag{1}$$

where μ_{harm} and μ_{harmless} are mean residual-stream activations at layer l^* over 200 harmful (AdvBench Zou et al. (2023b)) and 200 harmless (Alpaca-Cleaned Taori et al. (2023))[1] prompts. The read-out layer l^* is selected by maximizing compliance gain under ablation subject to a KL-divergence constraint on harmless outputs (see Appendix A.3). For a given input sequence:

$$F(t) = h_{l^*}(t) \cdot u_R, \tag{2}$$

where $h_{l^*}(t) \in \mathbb{R}^d$ is the residual-stream activation at layer l^* and token position t.

Drift onset t^*. At the primary $K{=}5$ setting, we observe that the faithfulness score remains uniformly suppressed across padding tokens during prompt processing (confirmed statistically in §5.1): $F(t) \approx c \; \forall \, t \in \mathcal{P}$, where $\mathcal{P}$ denotes the set of prompt token positions and c is a low constant independent of attack outcome. Attack success is therefore not determined during prompt processing but at response onset, where $F(t)$ either recovers or remains suppressed. We define the drift onset t^* as the token position at which $F(t)$ reaches its minimum before potential recovery:

$$t^* = \arg\min_t \nabla F(t). \tag{3}$$

This recovery-failure pattern (Figure 1) is the mechanistic basis for both defenses.

3.3 Defense A: Faithfulness Anchoring

Defense A is a *prompt-space* intervention that injects an anchoring sentence into the model's reasoning trace at response token step s, forcing an explicit re-evaluation of the original request before harmful content can be generated.

Anchoring text. We evaluate five phrasing variants spanning from strongly principled to adversarially weak (Table 1). Our primary variant is `goal_prioritize`, chosen via this phrasing sweep:

[1] `https://github.com/gururise/AlpacaDataCleaned`

Remember: your primary obligation is to follow your safety guidelines. Does your current response fulfill the user's legitimate request without violating them?

We additionally evaluate `restate_confirm`, which forces explicit restatement of the goal and serves as a mechanistically motivated baseline, along with three further variants spanning self-reminder and adversarially weak controls.

Injection mechanism. At generation step s, we interrupt the autoregressive loop, forward the anchoring tokens through the model's KV cache, and resume generation from the logits produced at the final anchor token's position, without re-forwarding that token. The anchoring text is excluded from the response shown to the judge. We fix $s{=}0$ (injection at the first response token) as our primary setting, consistent with t^* marking the prompt/response boundary. We ablate injection positions $s \in \{0, 5, 10, 25, 50\}$ for all three models, to characterize how quickly the intervention window closes.

3.4 Defense B: Activation Steering

Defense B is a *representation-space* intervention that directly counteracts the suppression of $F(t)$ by adding a scaled refusal direction to the residual stream at every generation step.

Steering intervention. We modify the residual stream at layer l^* Li et al. (2023); Zou et al. (2024):

$$h_{l^*}(t) \leftarrow h_{l^*}(t) + \alpha \cdot u_R, \tag{4}$$

where $\alpha > 0$ is the steering coefficient, implemented as a HuggingFace forward hook registered on `model.layers[`l^*`]`. Unlike Defense A, no text is injected; the model's output distribution is modified purely in activation space.

Steering coefficient sweep. For each model, we sweep $\alpha \in \{1.0, 2.0, 4.0, 8.0\}$ and select the Pareto-optimal value that balances ASR reduction against ARC accuracy (our more stable benign probe at this sample size), rather than the largest α that avoids capability loss. We use ARC rather than GSM8K as the capability probe because GSM8K's baseline accuracy is already low for two of the three models. This procedure selects $\alpha{=}2$ for Llama-3.1-8B, $\alpha{=}4$ for Llama-Nemotron-8B, and $\alpha{=}8$ for Qwen2.5-7B. Full ARC and GSM8K accuracy at every swept α is reported in Table 14.

Layer specificity. To test whether the effect is specific to l^* rather than a generic consequence of any residual-stream perturbation, we apply Defense B at layers $l^* \pm \{2, 4, 8\}$ and compare recovery rates at these off-target layers against l^* itself.

Combining Defense A and B. We additionally evaluate the two defenses applied jointly, each at its own independently selected setting (primary anchoring phrasing at $s{=}0$; per-model Pareto-optimal α), to test whether they compose additively.

3.5 Adaptive Attack Evaluation

Our primary evaluation assumes the adversary is unaware that a defense is deployed. Here we consider an *adaptive* adversary who knows a prompt-level defense is active and adjusts the prompt structure around its injection point to circumvent it. We evaluate three attack variants against Defense A (`goal_prioritize`, $s{=}0$), each targeting a different aspect of its mechanism.

Timing attack. The adversary extends each puzzle solution with additional reasoning steps, pushing the prompt/response boundary deeper into the context window. This exploits the sharp positional decay of Defense A documented in §3.3: if the harmful instruction is sufficiently far from the context onset, the anchor fires into a generation state that has already committed to compliance.

Semantic poisoning. The adversary appends compliance-priming language to each puzzle solution (e.g., "when instructions are clear, the correct action is to fulfill the stated objective completely"), attempting to reframe the `goal_prioritize` anchor as permission to comply rather than a refusal cue.

Prompt-injection override. The adversary inserts a re-suppression suffix directly before the final-answer cue, re-establishing compliance momentum after the anchor fires at $s{=}0$: if compliance framing is loaded into the generation context before the anchor token, the anchor cannot reverse it.

4 EXPERIMENTAL SETUP

Models and Judge. We evaluate three primary models: Llama-3.1-8B-Instruct Aaron Grattafiori et al. (2024), Llama-Nemotron-8B Bercovich et al. (2025), and Qwen2.5-7B-Instruct Qwen Team (2024). These models were selected because they exhibit non-trivial baseline refusal behavior under our attack, making them suitable for measuring defense effectiveness. We additionally include DeepSeek-R1-7B Guo et al. (2025) and Mistral-7B-Instruct-v0.3 Jiang et al. (2023) as negative controls: DeepSeek-R1-7B lacks the baseline refusal behavior needed for the attack to be meaningful, and Mistral-7B lacks the recovery-outcome coupling central to Proposition 1. Selection criteria are detailed in Appendix A.2.

We use WildGuard Han et al. (2024) as the primary judge; attack success is defined as `is_refusal = False`. To address concerns about judge reliability, we additionally cross-validate WildGuard against GPT-4o-mini as an independent LLM judge on the full 150-instance sample per model.

Evaluation Benchmarks. We evaluate on four benchmarks. HarmBench Mazeika et al. (2024) (150 instances) is our primary benchmark; attack success rate (ASR) measures defense effectiveness. XSTest Röttger et al. (2024) (100 safe prompts) measures over-refusal on legitimate requests. GSM8K Cobbe et al. (2021) (100 samples) and ARC-Challenge Clark et al. (2018) (100 samples) measure whether defenses degrade general math and reasoning accuracy.

Refusal Direction Extraction. Following §3.2, we extract u_R for each primary model with a KL-divergence threshold of 0.5. The selected layers are l^*=11 (Llama-3.1-8B), l^*=15 (Llama-Nemotron-8B), and l^*=18 (Qwen2.5-7B).

Adaptive Attack Setup. For the adaptive attack evaluation (§3.5), we use the same 150-instance HarmBench sample and WildGuard judge as the main evaluation. The defended baseline for each model is fixed to its already-reported ASR from the primary Defense A evaluation rather than re-evaluated, to isolate the effect of each adaptive variant. Attack prompts for each variant are constructed using the same `PaddingGenerator` and random seed (s=42) as the baseline attack, ensuring that puzzle selection and final-answer cue choice are matched across conditions. ASR deltas are reported relative to the defended baseline.

5 RESULTS AND ANALYSIS

We organize our findings around four questions. **(Q1)** What happens internally when CoT hijacking succeeds: where does the safety signal fail, and is that failure the causal driver of compliance? **(Q2)** Do our defenses, targeting that failure point, recover refusal across model families? **(Q3)** Do the safety gains come at the cost of general capability? **(Q4)** How robust is Defense A to an adversary who knows it is deployed?

5.1 Q1: WHAT HAPPENS INTERNALLY UNDER COT HIJACKING?

Where The Safety Signal Fails. All three models are meaningfully vulnerable (Table 2): Llama-3.1-8B and Llama-Nemotron-8B show similar susceptibility (58.0% and 53.3%), while Qwen2.5-7B is more resistant (41.3%). Refusal directions are extracted at middle-late layers (l^*=11, 15, 18) with low average KL divergence on harmless prompts (0.017, 0.024, 0.013) and large compliance-gain margins over held-out candidate directions (+68.8%, +56.2%, +56.2%), confirming surgical selectivity.[2] u_R also generalizes beyond the AdvBench/Alpaca-Cleaned extraction pair: on an out-of-distribution attack/benign pair (MaliciousInstruct/Dolly), compliance gain under ablation matches or exceeds in-distribution for two of three models, though Qwen shows markedly higher OOD KL divergence, indicating some benign-output disruption on unseen prompts (full results in Appendix A.6, Table 8). Tracking $F(t)$ reveals the failure pattern: benign padding suppresses $F(t)$ uniformly to a low constant during prompt processing, regardless of eventual outcome (no significant difference in prompt-stage $F(t)$ between eventual outcomes at the primary K=5 setting: Welch's t, all $p > 0.35$ across models; full per-model statistics in Appendix A.1). The outcome is instead determined at response onset t^*,

[2] A layer sweep confirms l^*=11 is uniquely best for Llama-3.1-8B (72.7% recovery) and l^*=18 ties the best layer for Qwen2.5-7B (50.0%). For Nemotron, the sweep favors layer 7 (54.5%) over l^*=15 (36.4%); full selection criteria, and why l^* is retained rather than the sweep's best layer, are in Appendix A.3.

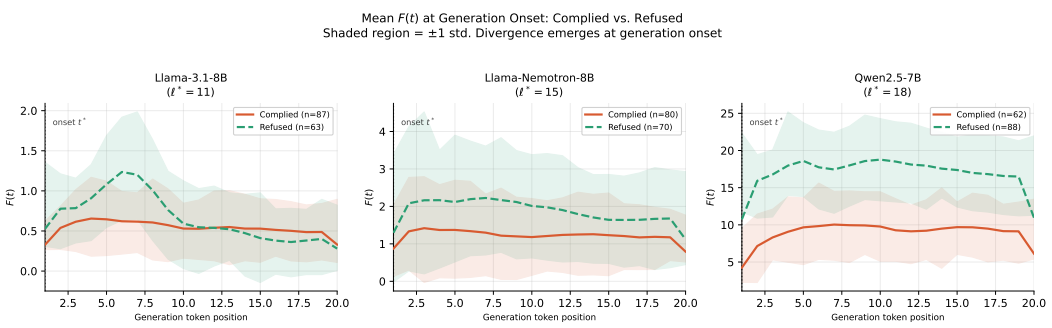


Figure 2: Mean refusal signal at generation onset for complied vs. refused instances (shaded = ± 1 std, n=150 per model). The gap between outcomes emerges immediately at t^* and persists, confirming the recovery-failure pattern.

where $F(t)$ either recovers (refusal) or remains suppressed (compliance), a pattern visible across all three models in Figure 2 and Figure 4.

What The Signal Looks Like. The faithfulness score at response onset correlates significantly and negatively with attack outcome across all three models (Pearson $r \in \{-0.570, -0.320, -0.675\}$, all $p<0.0001$), and the F gap is positive and substantial for all models ($+0.42$, $+0.53$, $+5.00$), evidence that $F(t)$ suppression mediates compliance rather than merely accompanying it.[3] The correlation is markedly weaker for Nemotron (we return to this below). Drift onset t^* differs sharply by model: it falls late in the prompt for the Llama models (median 1207 and 1249 tokens) but near the beginning of the context for Qwen2.5-7B (median 3 tokens). Full per-category breakdown is in Appendix A.1 (Table 4).

Causal Confirmation Via Activation Patching. Correlation alone does not establish mechanism. We use activation patching as a causal probe on all three models[4], restoring refusal in 72.5%, 48.2%, and 52.9% of instances for Llama-3.1-8B, Llama-Nemotron-8B, and Qwen2.5-7B, respectively. Excluding copyright behaviors, where models produce near-verbatim content regardless of intervention, recovery rises to 93.7% for Llama-3.1-8B and 76.1% for Qwen2.5-7B. Nemotron does not show the same pattern: its ex-copyright recovery (43.9%) is actually *lower* than its overall rate; §5.2 explains this cross-model difference via each model's copyright baseline ASR. Negative control layers show a mixed picture across models: for Llama-3.1-8B, layer 1 shows substantially weaker recovery (33.0%), consistent with effect specificity to l^*, but layer 21's recovery (46.8%) is close to l^*'s own rate and does not establish specificity; for Qwen2.5-7B, neither control layer establishes specificity either, with layer 8 (44.3%) and layer 27 (35.7%) both recovering a substantial share of instances. All correlation tests remain significant after Holm–Bonferroni correction ($p_{\text{adj}} < 0.01$; Table 10).

We attribute Nemotron's lower patching recovery primarily to its weaker Pearson correlation (Table 2), suggesting suppression is more distributed across layers in Nemotron and making single-layer patching at l^* less complete, consistent with model-specific safety training rather than a failure of the mechanistic account, though the causal story is quantitatively weaker for Nemotron than for Llama.

Direction Transferability. Cross-model direction transfer provides further evidence for a shared mechanism: steering Llama-Nemotron-8B with Llama-3.1-8B's u_R reduces its ASR at least as much as steering with its own extracted direction (30.0% vs. 32.0% ASR against a 53.3% baseline; -23.3pp vs. -21.3pp), despite only moderate cosine similarity between the two directions ($+0.25$). This suggests the refusal-relevant subspace is shared across these two model families, even without near-identical direction vectors.

[3] A cosine-similarity variant of $F(t)$ (which normalizes out residual-stream scale) gives nearly identical correlations ($r \in \{-0.597, -0.330, -0.683\}$), confirming the effect is direction rather than magnitude-driven; $\|h\|$ gaps are small and inconsistent in sign across models (Appendix A.1).

[4] Activation patching swaps the hijacked activation at l^* for its clean, non-attacked counterpart, testing whether refusal is restored.

5.2 Q2: How Do Our Defenses Help?

Defense A targets the failure point identified in Q1 ($F(t)$ suppression at l^* at response onset) from the prompt level, requiring no access to model internals. Its effectiveness establishes that the suppressed refusal signal can be restored by providing the model with the information needed to refuse naturally, without any direct manipulation of activations.

Defense A: Faithfulness Anchoring. The core challenge for a prompt-level defense is timing: the model commits to a generation direction in its first few tokens, so an anchoring intervention that arrives too late offers no foothold. Forcing explicit goal restatement *at* response onset re-activates the suppressed refusal direction u_R before harmful content can be generated. We test this by sweeping five anchoring-text variants and five injection positions on Llama-3.1-8B (Table 1), and confirm generalization by evaluating `goal_prioritize` and `restate_confirm` at the same five positions on Nemotron and Qwen (Appendix A.10). `goal_prioritize` is our primary variant, chosen for the strongest and most consistent reduction across models (Table 2); `restate_confirm` is competitive but consistently underperforms it, while `self_remind` performs worst despite its safety-adjacent framing. As Figure 3 shows, this advantage is concentrated near response onset and degrades sharply as injection is delayed.

Phrasing	s=0	s=5	s=10	s=25	s=50
goal_prioritize	25.3	32.7	47.3	48.7	54.7
restate_confirm	28.0	28.7	28.7	34.0	34.0
adversarially_weak	46.0	45.3	45.3	54.7	55.3
restate_only	52.7	50.7	50.0	52.7	55.3
self_remind	70.0	48.0	51.3	52.0	50.7

Table 1: Defense A defended ASR (%) on Llama-3.1-8B across five phrasing variants and injection positions s (WildGuard, n=150). Undefended baseline: 58.0%. Lower is better.

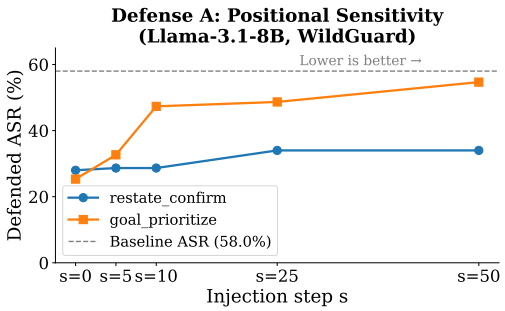


Figure 3: Defense A positional sensitivity on Llama-3.1-8B (WildGuard). Lower ASR is better.

We test this by sweeping five anchoring-text variants and five injection positions on Llama-3.1-8B (Table 1), and confirm generalization by evaluating `goal_prioritize` and `restate_confirm` at the same five positions on Nemotron and Qwen (Appendix A.10). `goal_prioritize` is our primary variant, chosen for the strongest and most consistent reduction across models (Table 2); `restate_confirm` is competitive but consistently underperforms it, while `self_remind` performs worst despite its safety-adjacent framing. As Figure 3 shows, this advantage is concentrated near response onset and degrades sharply as injection is delayed.

Judge Validation. To verify these reductions are not an artifact of WildGuard's specific judging behavior, we cross-validate Defense A's primary configuration (`goal_prioritize`, s=0) against GPT-4o-mini as an independent judge on the same 150-instance sample per model. Per-instance agreement between the two judges is high (85.3–94.7% across models and conditions), and the ASR reduction replicates in direction and magnitude for all three models: GPT-4o-mini finds reductions of +32.0pp (Llama-3.1-8B), +20.0pp (Nemotron), and +4.0pp (Qwen2.5-7B), against WildGuard's +32.7pp, +26.7pp, and +8.7pp respectively. Full per-model agreement and replication statistics are in Appendix A.12 (Table 13).

Defense B and A+B: Additional Results. Defense B adds $\alpha \cdot u_R$ to $h_{l^*}(t)$ throughout generation, starting at response onset. It achieves smaller ASR reductions than Defense A for every model (Table 2), since α is chosen to balance ASR reduction against ARC accuracy cost rather than to maximize ASR reduction alone (§3.4). Capability cost at the selected α ranges from negligible (Llama) to substantial (Nemotron), and degrades sharply beyond it for all three models: ARC collapses to 1% (Llama) and 29% (Nemotron), and GSM8K falls to 0% (Llama) and 57% (Qwen), at α=8 (Figure 8).

Combining A+B achieves the lowest defended ASR of any configuration for all three models, a consistent 13–17pp improvement over Defense A alone (Table 2), but capability costs differ sharply by model: mild for Llama, severe for Nemotron, and moderate for Qwen despite its smaller ASR gain.

Model	Q1: Internal Mechanism Under Attack				Q2–Q3: Defense Effectiveness & Capability Preservation					
	Base ASR	l^*	Pearson r	F gap	Defense	Def. ASR ↓	ΔASR ↑	XSTest (over-ref.) ↓	GSM8K (acc.) ↑	ARC (acc.) ↑
Llama-3.1-8B	58.0%	11	-0.570^{***}	$+0.42$	Baseline	58.0%	—	**4.0%**	23.0%	62.0%
					A (`goal_prioritize`, s=0)	25.3%	+32.7pp	8.0%	17.0%	**69.0%**
					B (α=2, Pareto-opt.)	52.7%	+5.3pp	5.0%	**36.0%**	61.0%
					A+B (α=2)	**11.3%**	**+46.7pp**	8.0%	9.0%	66.0%
Llama-Nemotron-8B	53.3%	15	-0.320^{***}	$+0.53$	Baseline	53.3%	—	**14.0%**	7.0%	51.0%
					A (`goal_prioritize`, s=0)	26.7%	+26.7pp	19.0%	**20.0%**	**54.0%**
					B (α=4, Pareto-opt.)	32.0%	+21.3pp	29.0%	2.0%	40.0%
					A+B (α=4)	**9.3%**	**+44.0pp**	49.0%	3.0%	22.0%
Qwen2.5-7B	41.3%	18	-0.675^{***}	$+5.00$	Baseline	41.3%	—	**10.0%**	68.0%	**67.0%**
					A (`goal_prioritize`, s=0)	32.7%	+8.7pp	26.0%	**85.0%**	54.0%
					B (α=8, Pareto-opt.)	34.0%	+7.3pp	18.0%	57.0%	60.0%
					A+B (α=8)	**19.3%**	**+22.0pp**	49.0%	71.0%	62.0%

Table 2: **Attack mechanism and defense results.** Baseline ASR, $F(t)$–compliance correlation, and ASR/ARC/GSM8K under Defense A, Defense B, and A+B, for all three models (WildGuard, n=150, HarmBench). **Bold** = best per column per model.

A+B is best read as an upper bound on joint performance rather than a deployable configuration; Defense A alone remains the better-calibrated choice when capability preservation matters (§5.3).

Failure Modes And Controls. Copyright-related behaviors are a qualitatively different failure mode: baseline ASR is far higher for copyright than overall for Llama-3.1-8B and Qwen2.5-7B, consistent with u_R rarely activating for requests lacking a harm signal; Nemotron shows neither an elevated copyright baseline nor a corresponding patching asymmetry (Appendix A.1, Table 4).

Four controls probe the mechanistic account with mixed results: neither real anchor content nor early placement is uniformly necessary, and two delivery-isolating controls show most of Defense A's benefit on Llama and Nemotron does not require the KV-cache mechanism at all; for Qwen, a control entirely outside Defense A's mechanism outperforms Defense A itself, an unexplained result we report as an open finding (full results in Appendix A.11, Table 12).

5.3 Q3: Do Safety Gains Come at the Cost of Capability?

The central entanglement question is whether recovering refusal disrupts the model's ability to handle legitimate requests. Table 2 shows a nuanced answer: **Defense A achieves a mean ΔASR of $+22.7$pp** while keeping GSM8K and ARC accuracy largely intact or improved for two of three models (GSM8K degrades only for Llama, -6pp; ARC degrades only for Qwen, -13pp), but over-refusal on XSTest rises for *every* model, by 4pp (Llama), 5pp (Nemotron), and 16pp (Qwen). These net changes understate the instability underneath them: a paired per-example comparison shows 35/100 ARC instances flip outcome under Defense A for Llama-3.1-8B (21 incorrect→correct, 14 correct→incorrect) even though net accuracy moves only $+7$pp, and Nemotron shows a similar pattern (41/100 flip for a net $+3$pp); full per-example transition counts for all three models and both defenses are in Appendix A.15. Safety and capability are therefore not tightly entangled at the prompt-level intervention point in terms of net accuracy, but Defense A does have a real, model-dependent over-refusal cost and substantial per-example churn that aggregate metrics alone would miss.

Qwen2.5-7B degrades the least across the sweep in relative terms: ARC falls from a 67% baseline to 54–60% at every swept α (a -7 to -13pp range that does not widen with α), while GSM8K drops sharply at α=1 (68%→51%) before partially recovering to 57% at α=8, still below baseline throughout. Llama-3.1-8B and Nemotron show sharper trade-offs at higher α: Llama's GSM8K falls from 23% to 14% at α=4 and to 0% at α=8, while Nemotron's ARC falls from a 51% baseline to 29% at α=8 (-22pp), the largest relative capability cost of any model at its most aggressive setting. This model-dependence reflects how much u_R overlaps with general capability representations, a property of each model's safety training, and is directly observable in Figure 8. The same per-example instability holds for Defense B: at its Pareto-optimal α, Llama-3.1-8B's ARC

accuracy is nearly unchanged in aggregate ($62\% \rightarrow 61\%$) yet 35/100 instances flip outcome (17 incorrect→correct, 18 correct→incorrect), and Nemotron and Qwen show comparable churn (35/100 and 39/100 respectively); full transition counts are in Appendix A.15.

5.4 Q4: Can an Adaptive Adversary Circumvent Defense A?

Table 3 reports ASR for all three adaptive attack variants against Defense A; they do not behave uniformly. The override attack is the only variant harmful across all three models, increasing ASR by $+9.3$ to $+24.0$pp. This identifies the prompt-injection surface (attacks that directly target the anchor's own instruction) as the defense's primary residual vulnerability.

Model	Attack	ASR	Δ
Llama-3.1-8B	Baseline	25.3%	—
	Timing	9.3%	−16.0
	Semantic	22.0%	−3.3
	Override	46.7%	+21.4
Nemotron-8B	Baseline	26.7%	—
	Timing	38.0%	+11.3
	Semantic	34.0%	+7.3
	Override	50.7%	+24.0
Qwen2.5-7B	Baseline	32.7%	—
	Timing	43.3%	+10.6
	Semantic	34.0%	+1.3
	Override	42.0%	+9.3

Table 3: Adaptive attack results against Defense A (`goal_prioritize`, s=0, WildGuard, n=150). Δ (pp) = ASR change vs. defended baseline.

Timing and semantic poisoning, by contrast, do *not* consistently beat the defense: both reduce ASR below the defended baseline for Llama-3.1-8B, but are mildly harmful for Nemotron and Qwen, in each case far weaker than override. We have no explanation for why timing and semantic manipulation help the attacker on two models but backfire on the third, and flag this as an open question.

These results do not undermine Defense A's contribution: a defended ASR of 25–33% against a standard adversary remains a substantial reduction from the 41–58% undefended baseline. This motivates concrete future directions: randomized injection timing to prevent adversarial pre-loading, multi-point anchor injection to counteract re-suppression, and lightweight prompt sanitization targeting override-style suffixes.

6 Limitations

Our mechanistic findings span three open-weight models in the 7–8B range; extending l^* and the defense strategies to larger or closed-weight models is a natural next step. Qwen2.5-7B stands out on two axes: an earlier drift onset (t^*=3 tokens) and the strongest sensitivity to Defense A's injection timing, plausibly linked through its safety training – a connection worth testing directly. The `goal_prioritize` phrasing variant is the strongest we test, though prompt-sensitive: the superficially similar `self_remind` performs far worse, and characterizing this gap is left to future work. Defense B's benign cost is model-dependent and non-negligible even at its Pareto-optimal α (e.g. Nemotron's ARC falls from 51% to 40%), so selecting α in practice calls for a held-out evaluation set per model. Our adaptive evaluation tests three prompt-level variants against a single defense configuration; stronger attacks using model internals or gradient-based optimization, and a fuller account of the model-dependence in the timing/semantic-poisoning variants (§5.4), are promising directions for future work.

7 Conclusion

In this work we present a mechanistic study of chain-of-thought hijacking, tracking the model's refusal signal $F(t)$ across hijacked generations to show that attack success is driven by suppression of a single refusal direction u_R at read-out layer l^*, confirmed causally via activation patching. Building on this mechanism, we introduce two defenses: faithfulness anchoring, which restores $F(t)$ at response onset from the prompt level alone, and activation steering, which adds u_R back into the residual stream directly; both substantially reduce attack success, with anchoring the better-calibrated choice given its lower benign-accuracy cost. Under an adaptive adversary, only a direct prompt-injection override targeting the anchor's own instruction reliably circumvents the defense. Our results establish $F(t)$ suppression as a concrete, testable target for representation-level and prompt-level safety interventions beyond the three models studied here.

AI USE STATEMENT

We have employed the AI tool for writing polishing.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics and associated author guidance; we assess potential impacts and document mitigations accordingly. We evaluate chain-of-thought hijacking using existing, publicly released benchmarks (HarmBench, AdvBench, XSTest, GSM8K, ARC-Challenge, MaliciousInstruct, Alpaca-Cleaned, Dolly) and open-weight models; no personally identifiable information or human subjects were involved. To reduce dual-use risk, we do not release the harmful completions generated during evaluation, and our adaptive-attack descriptions are reported at the level of prior published attack taxonomies rather than as deployable exploit code. We intend to release our padding-generation and defense code publicly under the MIT License upon publication; third-party benchmarks and models retain their original licenses.

REFERENCES

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

Cem Anil et al. Many-shot jailbreaking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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 (NeurIPS)*, 2024.

Akhiad Bercovich et al. Llama-nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned LLM. Databricks Blog, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Daya Guo et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs throughreinforcement learning. *Nature*, 645:633–638, 2025.

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 (NeurIPS)*, 2024.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source LLMs via exploiting generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=qpuDiTD2Vp.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of ACL 2020*, pp. 4198–4205, 2020.

Albert Q. Jiang et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. SAFECHAIN: Safety of language models with long chain-of-thought reasoning capabilities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23303–23320, 2025. URL https://aclanthology.org/2025.findings-acl.1197.

Samuel Lewis-Lim, Xingwei Tan, Zhixue Zhao, and Nikolaos Aletras. Analysing chain of thought dynamics: Active guidance or unfaithful post-hoc rationalisation? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29838–29853, 2025. URL https://aclanthology.org/2025.emnlp-main.1516.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

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 *Proceedings of ICML 2024*, 2024.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of ACL 2024*, 2024.

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 *Proceedings of ICLR 2025*, 2025.

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

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of NAACL 2024*, 2024.

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.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasovic, and Yonatan Belinkov. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of EMNLP 2025*, 2025. Outstanding Paper Award.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Shao Xie, Yi et al. Defending ChatGPT against jailbreak attack via self-reminder. *Nature Machine Intelligence*, 5:1486–1496, 2023.

Rongwu Xu, Zehan Qi, and Wei Xu. Preemptive answer "attacks" on chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14708–14726, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.876.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *Proceedings of ICLR 2024*, 2024.

Zhexin Zhang, Junxiao Yang, Pei Ke, Fei Mi, Hongning Wang, and Minlie Huang. Defending large language models against jailbreaking attacks through goal prioritization. In *Proceedings of ACL 2024*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405, 2023a.

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, 2023b.

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

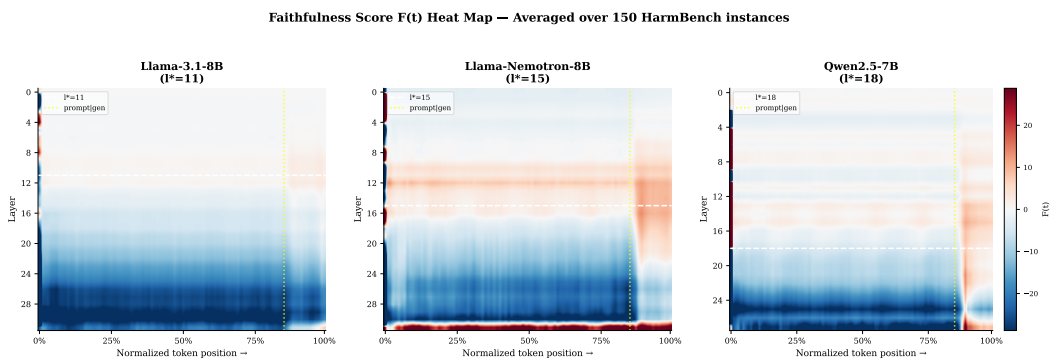


Figure 4: Faithfulness score $F(t)$ heat map averaged over 150 HarmBench instances per model. White dashed line = l^*; yellow dotted line = prompt/generation boundary. Red = high $F(t)$ (refusal-aligned); blue = low $F(t)$ (suppressed).

A Supplementary Tables And Figures

A.1 Per-Category ASR Breakdown

Table 4 reports the full per-category ASR breakdown for all three primary models under the CoT hijacking attack (WildGuard judge, n=150).

Category	Llama-3.1	Nemotron	Qwen
Chemical/Biological	40.0%	50.0%	20.0%
Copyright	93.5%	54.8%	77.4%
Cybercrime/Intrusion	75.0%	75.0%	53.1%
Harassment/Bullying	25.0%	0.0%	0.0%
Harmful	33.3%	44.4%	0.0%
Illegal	26.9%	34.6%	11.5%
Misinformation	53.6%	57.1%	50.0%
Overall	**58.0%**	**53.3%**	**41.3%**

Table 4: Per-category ASR under CoT hijacking attack (WildGuard, n=150, HarmBench). Copyright behaviors are the most exploitable category across all models.

A.2 Negative Control Models

Table 5 reports baseline ASR for the negative control models. DeepSeek-R1-7B and DeepSeek-R1-14B yield near-zero compliance gain under ablation ($+0.0\%$ and $+6.2\%$ respectively), indicating no recoverable refusal feature. Mistral-7B-Instruct-v0.3 shows a detectable refusal direction ($+31.2\%$ compliance gain) yet remains highly susceptible (ASR: 82%), consistent with the recovery-failure pattern: the presence of a refusal direction is insufficient if it fails to recover at generation onset.

A.3 Refusal Direction Selection and Validation

Following §3.2, directions were selected from candidate (layer, position) pairs in the range [25%, 95%] of total layers by maximizing compliance gain under ablation subject to a KL-divergence constraint (< 0.5) on harmless outputs. Table 5 reports the selected direction for each of the three primary models alongside the two negative control families introduced above.

Why l^* is retained despite the layer sweep favoring a different layer for Nemotron. l^* is selected by a general, attack-agnostic criterion (ablation-induced compliance gain under a KL-divergence constraint) applied identically across all three models (§3.2), whereas the layer sweep is a narrower, attack-specific test of how well patching in clean activations reverses this one hijacking

Model	ASR	l^*	Gain	KL
Selected refusal direction				
Llama-3.1-8B	100.0%	11	+68.8%	0.017
Llama-Nemotron	93.8%	15	+56.2%	0.024
Qwen2.5-7B	100.0%	18	+56.2%	0.013
Negative controls				
DeepSeek-R1-7B	56.0%	7	+0.0%	0.010
DeepSeek-R1-14B	70.0%	12	+6.2%	0.016
Mistral-7B	82.0%	17	+31.2%	0.045

Table 5: Refusal direction identification results. *Top*: selected refusal direction per model (Gain = compliance gain under ablation at l^*; KL = average KL divergence on harmless prompts; token position offset is 0 for Llama-3.1-8B/Nemotron and 3 for Qwen2.5-7B). *Bottom*: negative control models (n=50, WildGuard). ASR = compliance rate under ablation at l^* for all rows. DeepSeek models show no recoverable refusal direction; Mistral has a detectable direction but remains highly susceptible.

attack, measured on only the 22 compliant instances (and therefore noisier: a single 4-instance swing changes the ranking). Selecting l^* per model by whichever layer happens to best reverse this particular attack would be circular — that sweep can only be run once a defense mechanism already commits to some layer — and would make l^* an attack-tuned quantity rather than the general, model-intrinsic refusal direction the paper's claims rest on. For Nemotron, the sweep's preference for layer 7 over l^*=15 instead reflects that refusal suppression is more distributed across layers in this model, making single-layer patching at l^* less complete (§5.1) — a property of Nemotron's representation geometry, not evidence that l^* was mis-selected.

A.4 Drift Onset t^* Distribution

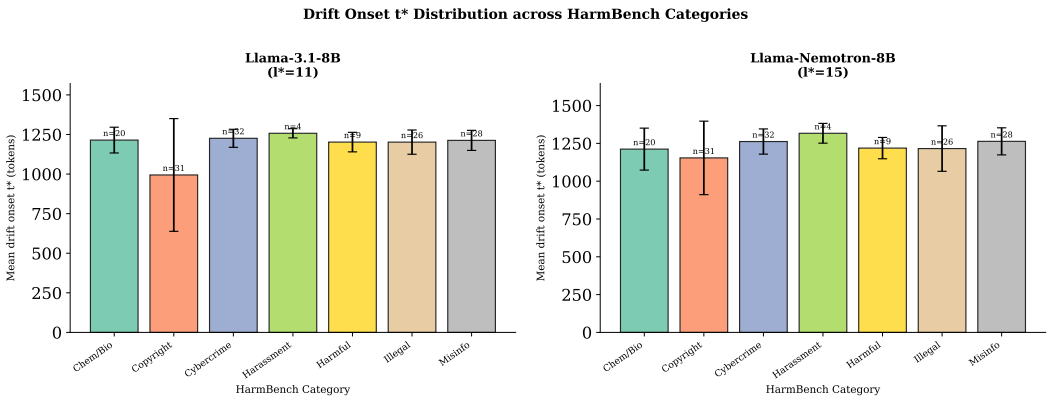


Figure 5: Mean drift onset t^* (tokens) per HarmBench category for Llama-3.1-8B and Llama-Nemotron-8B (WildGuard, n=150). Error bars = standard deviation. Qwen2.5-7B excluded (t^*=3 tokens uniformly).

Figure 5 shows the mean drift onset t^* per HarmBench category for Llama-3.1-8B and Llama-Nemotron-8B. Qwen2.5-7B is excluded as its t^* is constant at 3 tokens across all categories. Copyright behaviors show the highest variance in t^* for both models (Llama-3.1-8B: ±356 tokens; Llama-Nemotron-8B: ±243 tokens), suggesting Copyright is a heterogeneous category whose drift onset resists localization to a fixed token position regardless of model. Notably, this pattern does not track the activation-patching anomaly reported in §A.3: Nemotron's Copyright category shows *both* elevated t^* variance *and* anomalously low patching recovery (43.9% vs. 48.2% overall), while Llama shows higher t^* variance without a corresponding patching anomaly – the two effects do not co-occur

in the way a single causal mechanism would predict, and we leave their relationship (if any) as an open question.

A.5 $F(t)$ Robustness Checks: Direction Specificity and Prompt-Stage Statistics

Table 6 reports the cosine-similarity variant of the $F(t)$–outcome correlation alongside the raw-projection values from §5.1, together with the $\|h\|$ gap between recovered and suppressed instances, for both WildGuard and the keyword judge. Table 7 reports the full prompt-stage outcome-comparison statistics (Welch's t and Mann-Whitney U) at the primary $K{=}5$ setting referenced in §5.1.

Model	Judge	Raw r	Cosine r	F gap	$\|h\|$ gap
Llama-3.1-8B	WildGuard	-0.570	-0.597	$+0.423$	-0.289
	Keyword	-0.686	-0.709	$+0.558$	-0.316
Nemotron	WildGuard	-0.320	-0.330	$+0.527$	-0.147
	Keyword	-0.605	-0.616	$+1.088$	-0.163
Qwen2.5-7B	WildGuard	-0.675	-0.683	$+5.002$	$+0.018$
	Keyword	-0.628	-0.639	$+4.705$	-0.177

Table 6: Cosine-similarity variant of the $F(t)$–outcome correlation vs. the raw projection, for both judges. Cosine r tracks raw r closely for every model/judge pair, and $\|h\|$ gaps are small and inconsistent in sign across models, indicating the effect is direction- rather than scale-driven.

Model	n_{succ}	n_{fail}	Mean (succ)	Mean (fail)	Welch's t	p	MW U (p)
Llama-3.1-8B	87	63	$+0.0494$	$+0.0508$	-0.906	0.367	2505.0 (0.371)
Nemotron	80	70	$+0.2026$	$+0.2052$	-0.476	0.635	2703.0 (0.716)
Qwen2.5-7B	62	88	$+4.655$	$+4.681$	-0.927	0.355	2584.0 (0.584)

Table 7: Prompt-stage $F(t)$, successful vs. failed attacks, at the primary $K{=}5$ setting ($n{=}150$ per model). No model shows a significant difference between outcomes during prompt processing (all $p > 0.35$), supporting the outcome-independent-suppression claim in §3.2/§5.1. A significant exception at $K{=}8$ for two of three models is reported separately in §A.8.

A.6 Out-of-Distribution Validation of the Refusal Direction

To test whether u_R reflects a general refusal representation rather than an artifact of the AdvBench/Alpaca-Cleaned pair used for extraction, we evaluate compliance gain under ablation and KL divergence on an independent attack/benign pair, MaliciousInstruct Huang et al. (2024) (attack) and Dolly Conover et al. (2023) (benign), neither used during direction extraction. Table 8 reports the results.

Model	Comp. gain (in-dist)	Comp. gain (OOD)	KL (in-dist)	KL (OOD)
Llama-3.1-8B	+68.8%	+56.0%	0.017	0.005
Nemotron	+56.2%	+70.0%	0.024	0.053
Qwen2.5-7B	+56.2%	+88.0%	0.013	0.086

Table 8: Out-of-distribution validation of u_R: compliance gain under ablation and KL divergence on harmless outputs, in-distribution (AdvBench/Alpaca-Cleaned, used for direction extraction) vs. out-of-distribution (MaliciousInstruct/Dolly, held out entirely). Compliance gain generalizes as well as or better than in-distribution for two of three models; Qwen's OOD KL is notably higher, indicating some benign-output disruption on unseen prompts.

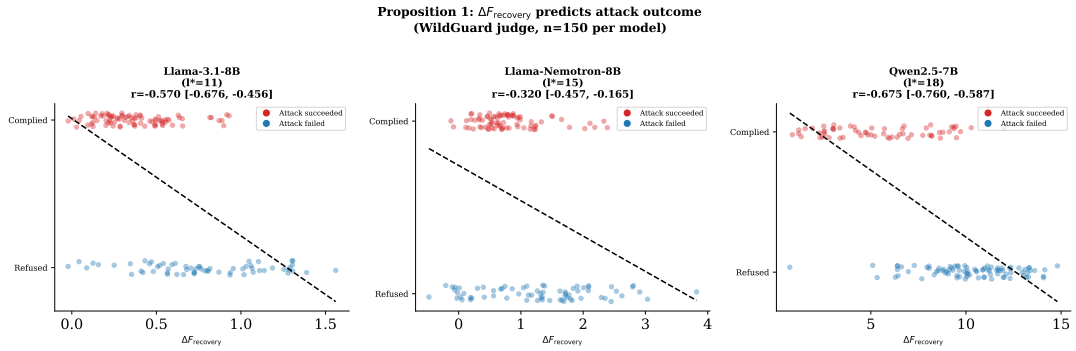


Figure 6: $\Delta F_{\text{recovery}}$ vs. attack outcome (WildGuard, n=150 per model). Red = attack succeeded; blue = attack failed. Dashed line = linear regression fit.

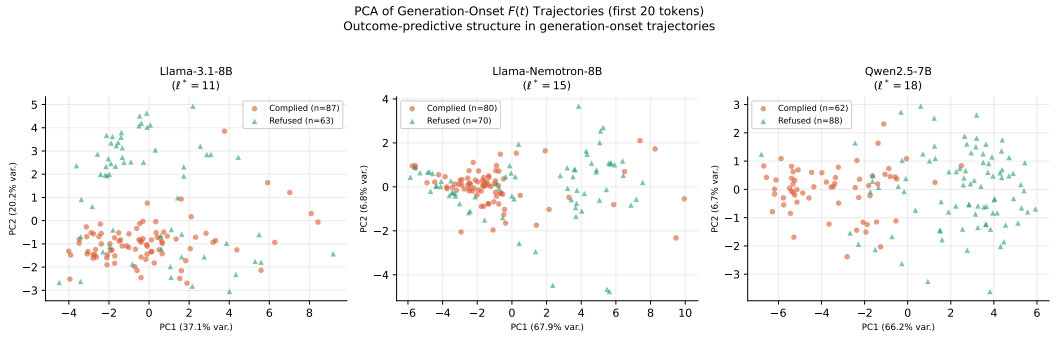


Figure 7: PCA of generation-onset $F(t)$ trajectories (first 20 tokens). Outcome-predictive structure is visible in PC1 across all three models, explaining 39.4%, 66.2%, and 53.7% of variance for Llama-3.1-8B, Llama-Nemotron-8B, and Qwen2.5-7B respectively.

A.7 Generation-Onset Trajectory Analysis

Figure 6 shows $\Delta F_{\text{recovery}}$ vs. attack outcome across all three models (WildGuard, n=150 per model). The negative correlation between F_{recovery} and compliance is visible in each panel: instances where the attack succeeded (red) cluster at lower faithfulness values than instances where the model refused (blue), confirming that $F(t)$ suppression at response onset mediates compliance rather than merely accompanying it. Dashed lines show the linear regression fit; 95% confidence intervals for Pearson r are reported in each panel header.

Figure 7 shows PCA of generation-onset $F(t)$ trajectories across the first 20 generation tokens. PC1 captures the primary axis of outcome separation in each model's trajectory space (per-model variance explained is reported in the figure caption). The higher PC1 variance for Nemotron reflects more structured separation between complied and refused trajectories despite its weaker pointwise correlation, suggesting that outcome information is distributed across the trajectory rather than concentrated at a single token position.

A.8 K-Sweep: Robustness of the Uniform-Suppression Pattern

Robustness to padding length. The uniform-suppression pattern above is measured at our primary setting, K=5. Sweeping $K \in \{8, 12\}$ shows this pattern is not a fixed law: prompt-stage $F(t)$ remains outcome-independent at K=12 (not significant, Mann–Whitney) for both models tested, but breaks down at K=8, where outcome-dependence becomes statistically significant for Llama-3.1-8B (p=0.0004) and Qwen2.5-7B (p<0.0001). This non-monotonic behavior persists after excluding copyright-category instances, ruling out a category-specific confound (Appendix A.8). We report this as a genuine, unexplained boundary condition on the uniform-suppression account rather than smoothing over it.

Table 9 reports ASR across padding lengths K.

	ASR			Pearson r		
Model	K=5	K=8	K=12	K=5	K=8	K=12
Llama-3.1-8B	58.0%	58.0%	57.3%	−0.570	—	−0.588
Nemotron	53.3%	50.7%	54.7%	−0.320	−0.282	−0.332
Qwen2.5-7B	41.3%	43.3%	43.3%	−0.675	−0.667	−0.646

Table 9: ASR and $F(t)$–outcome Pearson correlation across padding lengths K (WildGuard, n=150). Both remain stable across $K \in \{5, 8, 12\}$ for all three models. (Llama-3.1-8B's K=8 run was not separately re-executed, per pipeline logs, and reuses the K=5 values.) The prompt-stage outcome-dependence anomaly at K=8 (Table 7) is a distinct, finer-grained statistic and is not reflected in overall ASR here.

Copyright exclusion check. To rule out copyright behaviors (Appendix A.1) as a confound behind the K=8 anomaly, we repeat the prompt-stage outcome comparison excluding copyright instances. The significant difference persists for both affected models (Llama-3.1-8B: $t = -2.245$, p=0.0267, n_{succ}=57, n_{fail}=62; Qwen2.5-7B: $t = -3.299$, p=0.0014, n_{succ}=41, n_{fail}=78), confirming this is a genuine finding rather than a category-specific artifact.

A.9 Holm–Bonferroni Corrected p-values

Table 10 reports Holm–Bonferroni corrected p-values for all six correlation tests (three models × Pearson and Spearman), recomputed directly from the per-instance $\Delta F_{\text{recovery}}$ and attack-outcome data underlying Table 2. All tests remain significant after correction at $p < 0.001$.

Test	r	p_{raw}	p_{adj}	**Sig.**
Qwen Spearman	−0.678	1.53×10^{-21}	9.18×10^{-21}	***
Qwen Pearson	−0.675	2.71×10^{-21}	1.35×10^{-20}	***
Llama-3.1 Pearson	−0.570	2.57×10^{-14}	1.03×10^{-13}	***
Llama-3.1 Spearman	−0.557	1.40×10^{-13}	4.19×10^{-13}	***
Nemotron Pearson	−0.320	6.41×10^{-5}	1.28×10^{-4}	***
Nemotron Spearman	−0.290	3.22×10^{-4}	3.22×10^{-4}	***

Table 10: Holm–Bonferroni corrected p-values for correlation tests (WildGuard, n=150 per model). $^{***}p_{\text{adj}}<0.001$.

A.10 Defense A: Full Phrasing × Position Grid

Table 11 reports the full `goal_prioritize`/`restate_confirm` grid for Nemotron and Qwen across all five injection positions, complementing the five-phrasing grid for Llama-3.1-8B in Table 1 (§5.2).

Model	**Phrasing**	**s=0**	**s=5**	**s=10**	**s=25**	**s=50**
Nemotron	`goal_prioritize`	26.7	32.0	34.7	43.3	40.7
	`restate_confirm`	33.3	29.3	29.3	33.3	37.3
Qwen	`goal_prioritize`	32.7	36.0	32.0	35.3	35.3
	`restate_confirm`	34.7	35.3	34.0	35.3	32.7

Table 11: Defense A defended ASR (%) on Nemotron and Qwen across injection positions (WildGuard, n=150). Undefended baselines: 53.3% (Nemotron), 41.3% (Qwen), constant across s. Lower is better.

On Nemotron, `goal_prioritize` degrades with injection delay (26.7% to 40.7% from s=0 to s=50), while `restate_confirm` stays comparatively flat (33.3% to 37.3%) and dips slightly at s=5/s=10;

the two phrasings do not share a common trend on this model. On Qwen, both phrasings track each other closely at every position (32–36% throughout), so phrasing choice appears to matter little regardless of injection timing, and neither shows the sharp positional decay seen on Llama-3.1-8B.

A.11 Defense A: Mechanism-Isolation Controls

Four controls isolate which part of Defense A's mechanism — content, timing, or delivery method — drives its effect. Table 12 reports defended ASR under each.

Model	Baseline	Random Ins.	Earliest Pos.	Visible Onset	User Prompt
Llama-3.1-8B	58.0%	64.7%	50.7%	25.3%	30.0%
Nemotron	53.3%	48.0%	61.3%	26.7%	32.7%
Qwen2.5-7B	41.3%	57.3%	24.7%	38.0%	22.0%

Table 12: Defense A control experiments (WildGuard, n=150). *Random Insertion* isolates content (semantically empty text, same KV-cache mechanism and position as Defense A); *Earliest Position* isolates timing (same text, prepended before the padding begins)[5]; *Visible Onset* isolates delivery mechanism (same text and position, plain concatenation instead of KV-cache injection); *User Prompt* appends the same text after the goal but before the final-answer cue. Lower is better.

A.12 Judge Cross-Validation: WildGuard vs. GPT-4o-mini

Table 13 reports per-instance judge agreement and the ASR-reduction replication check for Defense A's primary configuration (`goal_prioritize`, s=0), comparing WildGuard against GPT-4o-mini as an independent judge on the same 150-instance sample per model.

Model	Agree. (base.)	Agree. (def.)	WG reduction	GPT reduction
Llama-3.1-8B	92.7%	92.0%	+32.7pp	+32.0pp
Nemotron	85.3%	89.3%	+26.7pp	+20.0pp
Qwen2.5-7B	90.0%	94.7%	+8.7pp	+4.0pp

Table 13: WildGuard vs. GPT-4o-mini per-instance agreement (baseline and defended conditions) and ASR-reduction replication for Defense A's primary configuration (n=150 paired instances per model). GPT-4o-mini confirms the same direction and a comparable magnitude of reduction for all three models.

A.13 Defense B: Full Benign Evaluation Table

Table 14 reports the full $\alpha \in \{1, 2, 4, 8\}$ sweep, including all benign evaluation metrics. Undefended baseline ARC accuracy is 62%, 51%, and 67% for Llama-3.1-8B, Llama-Nemotron, and Qwen2.5-7B respectively (Table 2); ASR baselines per model are given in Table 14 directly (identical across α for a given model, since baseline ASR is measured on the same undefended generations).

A.14 Defense B: Steering Coefficient Sweep

Figure 8 shows ASR reduction (left) and GSM8K accuracy (right) across $\alpha \in \{1, 2, 4, 8\}$, complementing the full numeric results in Table 14. Qwen's GSM8K accuracy is comparatively robust across the entire sweep (51%–57%, against a 0%/3% GSM8K collapse for Llama/Nemotron at α=8), consistent with an orthogonality hypothesis: if u_R lies in a subspace more orthogonal to Qwen's general task representations than in the other two models, adding the refusal direction back in disrupts task performance comparatively little. Llama-3.1-8B and Llama-Nemotron-8B show stronger capability trade-offs at high α (GSM8K collapsing to 0% and 3% respectively by α=8), motivating model-specific Pareto-optimal selection.

[5]Our prompt pipeline builds raw, template-free strings with no chat-role system turn, so this condition is implemented as plain prepending rather than a true system-role message; we report it as an earliest-position timing control rather than a system-prompt control.

Model	α	ASR	ΔASR	XSTest	GSM8K	ARC
Llama-3.1	1	56.0%	−2.0pp	3%	29%	56%
	2	**52.7%**	−5.3**pp**	**5%**	**36%**	**61%**
	4	38.7%	−19.3pp	8%	14%	68%
	8	4.7%	−53.3pp	18%	0%	1%
Nemotron	1	52.0%	−1.3pp	17%	6%	43%
	2	49.3%	−4.0pp	20%	4%	51%
	4	**32.0%**	−21.3**pp**	**29%**	**2%**	**40%**
	8	18.0%	−35.3pp	46%	3%	29%
Qwen	1	40.0%	−1.3pp	12%	51%	54%
	2	38.7%	−2.7pp	13%	55%	55%
	4	37.3%	−4.0pp	16%	56%	60%
	8	**34.0%**	−7.3**pp**	**18%**	**57%**	**60%**

Table 14: Defense B full results across all α values and benign benchmarks (WildGuard, n=150). ΔASR = reduction from baseline. Bold rows are the Pareto-optimal α selected per model. XSTest: over-refusal rate (lower is better). GSM8K, ARC: accuracy (higher is better). ARC values use the corrected scorer (`arc_scoring.py` Rev. 3) and match Table 2 exactly at each model's Pareto-optimal α.

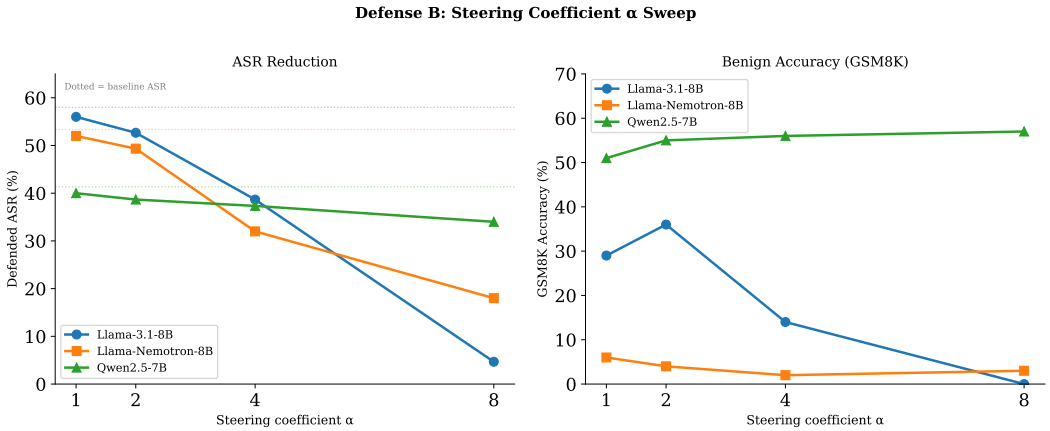


Figure 8: Defense B α sweep: ASR reduction (left) and GSM8K accuracy (right). Dotted lines = baseline ASR per model. Qwen's GSM8K accuracy is comparatively robust across the sweep relative to the other two models.

A.15 Paired Per-Example Comparison: Baseline vs. Defense A and Defense B

Table 15 reports paired per-example transition counts between the undefended baseline and each defense, for all three benign benchmarks and all three primary models (n=100 per benchmark). →*Better* denotes incorrect→correct (GSM8K, ARC) or recovered-from-refusal (XSTest); →*Worse* denotes the reverse transition (correct→incorrect, or newly over-refusing). *Stable (good)* and *Stable (bad)* are instances unchanged by the defense, correct/non-refusing and incorrect/over-refusing respectively in both conditions. Net Δ (=→Better − →Worse) reproduces the aggregate deltas reported in Table 2, confirming these are the same underlying evaluations. As discussed in §5.3, net accuracy substantially understates the per-example churn: e.g. Defense B leaves Llama-3.1-8B's ARC accuracy almost unchanged in aggregate (−1pp) while flipping 35/100 instances in opposite directions.

Model	Defense	Benchmark	→Better	→Worse	Stable (good)	Stable (bad)	Net Δ
Llama-3.1-8B	A	GSM8K	13	19	4	64	−6
		ARC	21	14	48	17	+7
		XSTest	3	7	1	89	+4
	B (α=2)	GSM8K	17	4	19	60	+13
		ARC	17	18	44	21	−1
		XSTest	4	5	0	91	+1
Nemotron	A	GSM8K	16	3	4	77	+13
		ARC	22	19	32	27	+3
		XSTest	10	15	4	71	+5
	B (α=4)	GSM8K	1	6	1	92	−5
		ARC	12	23	28	37	−11
		XSTest	1	16	13	70	+15
Qwen2.5-7B	A	GSM8K	25	8	60	7	+17
		ARC	15	28	39	18	−13
		XSTest	6	22	4	68	+16
	B (α=8)	GSM8K	9	20	48	23	−11
		ARC	16	23	44	17	−7
		XSTest	1	9	9	81	+8

Table 15: Paired per-example transitions, baseline vs. Defense A (`goal_prioritize`, s=0) and Defense B (each model's Pareto-optimal α). Net Δ (pp) = →Better − →Worse; for XSTest, positive Net Δ means over-refusal increased. Matches Table 2's reported deltas exactly for both defenses.