Attainment Is Not Stationarity: Objective and Optimizer Sources of Behavioral Drift after Safety Alignment

Anonymous authors
Paper under double-blind review

Abstract

As large language models (LLMs) are increasingly deployed in real-world applications, maintaining useful behavior while achieving safety remains important. Yet model behavior can continue to change after a safety metric reaches its target. We show that this post-attainment drift has two separable sources: an objective can retain residual preference pressure after the target safety level is reached, and stored optimizer state can induce behaviorally consequential movement even after the current objective becomes exactly stationary. Preference controls and same-checkpoint optimizer-state interventions isolate these sources. Changing residual preference within the attained safe region changes substantive safe completion; removing the stored first moment eliminates silent-step movement, scaling it produces proportional displacement across three model families, and matched-magnitude directions lead to different safety and refusal trajectories. Thus, safety attainment does not imply objective stationarity, and objective stationarity does not imply parameter stationarity.

1 Introduction

As LLMs are deployed in search, conversational systems, and agents, improving safety alone is not sufficient. Models should refuse harmful requests while remaining responsive to benign ones, yet stronger safety alignment can increase over-refusal (Röttger et al., 2024; Cui et al., 2025). Recent work further shows that refusal and helpfulness can continue to change after safety has saturated (Shen et al., 2026; Kim & Kumar, 2026). This raises a basic question: what continues to change model behavior after safety attainment?

We distinguish three states:

$$\underbrace{\mathrm{ASR}_t \le \tau}_{\text{safety attainment}} \not\Rightarrow \underbrace{g_t = 0}_{\text{objective stationarity}} \not\Rightarrow \underbrace{\Delta\theta_t = 0}_{\text{parameter stationarity}} .$$

Post-attainment drift can arise from two different sources. First, residual objective pressure can remain after the behavioral target is reached: DPO and group-normalized CISPO can continue to distinguish among safe responses. Second, optimizer-state memory can persist after the current objective becomes silent. Our contribution is not the observation that AdamW retains first-moment memory after a zero-gradient step. Rather, we use exact gradient silence as an identification regime to causally separate this memory from residual objective pressure and test whether it becomes behaviorally consequential after safety attainment. At a shared zero-gradient checkpoint, we hold model weights, data, reward, schedule, second moment, and optimizer clock fixed while intervening only on the stored first moment; magnitude and direction controls then separate its effects on subsequent behavior. Analogously, on the objective side, attainment constrains the primary safety ordering without determining preferences within the attained safe region, leaving residual preference as an optimization degree of freedom.

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.

Contributions. (1) Post-attainment decomposition. We decompose post-attainment drift into two distinct sources: residual objective pressure after safety attainment and optimizer-state-driven movement after objective stationarity. (2) Same-checkpoint causal identification. Exact-silence interventions isolate the stored first moment while holding the model, data, reward, second moment, and optimizer clock fixed; separate magnitude and direction controls show that retained first-moment state causally changes subsequent safety and refusal behavior. (3) Mechanism-specific interventions. We evaluate distinct controls for the two sources: specifying residual preference within the attained safe region and suppressing first-moment carryover during objective silence, across DPO and CISPO settings, multiple model families, and external benchmarks.

2 Related Work

Safety-utility trade-off and over-refusal. Over-refusal, in which safety alignment causes models to decline benign requests, is widely documented. Safe RLHF (Dai et al., 2024), POROver (Karaman et al., 2025), Safe-Completions (Yuan et al., 2025), and SafeDPO (Kim et al., 2026) modify data or objectives to manage the safety-utility trade-off, while XSTest (Röttger et al., 2024) and OR-Bench (Cui et al., 2025) measure over-refusal directly. Our focus is instead the source of behavioral change after safety has attained its target.

Post-attainment and preference optimization. SIRL (Shen et al., 2026) and competing-reward optimization (Kim & Kumar, 2026) show that continued optimization after safety saturation can further change over-refusal or helpfulness. Long-horizon preference optimization can also produce overoptimization, likelihood displacement, and unintended behavioral change (Rafailov et al., 2024; Razin et al., 2025; Ren & Sutherland, 2025). Preference objectives differ in whether gradients vanish at finite margins: SLiC-HF and RePO admit exact zero-gradient regions, IPO uses a finite target margin, whereas log-sigmoid objectives do not become exactly silent merely because a finite preference margin grows large (Zhao et al., 2023; Wu et al., 2025; Azar et al., 2024; Meng et al., 2024). Our hinge loss is a diagnostic control for separating safety attainment from objective stationarity, not a new optimization algorithm.

Optimizer-state dynamics. Optimizer state can influence later optimization trajectories. SPAM (Huang et al., 2025) uses momentum reset for gradient spikes, while Mukherjee et al. (2026) analyze AdamW state and learning-rate adaptation in RLVR. Our question is narrower: whether optimizer memory remains behaviorally consequential after an alignment metric reaches its target and the current objective becomes exactly silent. We isolate this effect at a shared zero-gradient checkpoint, holding the model, data, reward, second moment, and optimizer clock fixed while intervening on the first moment, and separately control displacement magnitude and direction.

3 Problem Setup: Safety Attainment Is Not Stationarity

3.1 Setup

Models and evaluation. We run offline DPO on Mistral-7B-Instruct-v0.2 and Zephyr-7B. For on-policy CISPO (MiniMax, 2025), we use Mistral-7B-Instruct-v0.2, Qwen2.5-7B, and Llama-3.1-8B, with $K = 16$ rollouts per prompt and a safety-only reward by default. Detailed hyperparameters and implementation choices are provided in the appendix.

Responses are evaluated along three axes: unsafe, substantive, and declined. We define Attack Success Rate (ASR) as the fraction of unsafe responses on harmful prompts, Over-Refusal Rate (ORR) as the fraction of declined responses on benign prompts, and Safe Completion Rate (SCR) as the fraction of harmful-prompt responses that are both non-unsafe and substantive. Safety attainment is defined as

$$t_{\text{attain}} = \min\{t : \text{ASR}_t \leq \tau\}, \qquad \tau = 0.05.$$

We evaluate training trajectories with an internal judge and separately verify the main endpoints on external benchmarks.

3.2 Three Distinct States

Separating the evaluation metric from the optimization state yields three distinct conditions:

$$\mathrm{ASR}_t \leq \tau \;\not\Rightarrow\; g_t = 0 \;\not\Rightarrow\; \Delta\theta_t = 0. \tag{1}$$

The first implication can fail because the current objective may keep producing a gradient after safety is already sufficient. The second can fail because optimizer state may move the parameters even when the current gradient is zero. We call a step with $g_t = 0$ a silent step, and a contiguous sequence of such steps a silent interval.

3.3 Optimization Persists After Safety Attainment

This separation is visible in actual training trajectories. Figure 1 shows a representative Mistral-7B CISPO run. ASR falls below $\tau = 0.05$ at step 4, yet the current gradient is exactly zero on 70 of 90 optimizer steps. Parameter displacement nevertheless continues through fully silent checkpoint intervals.

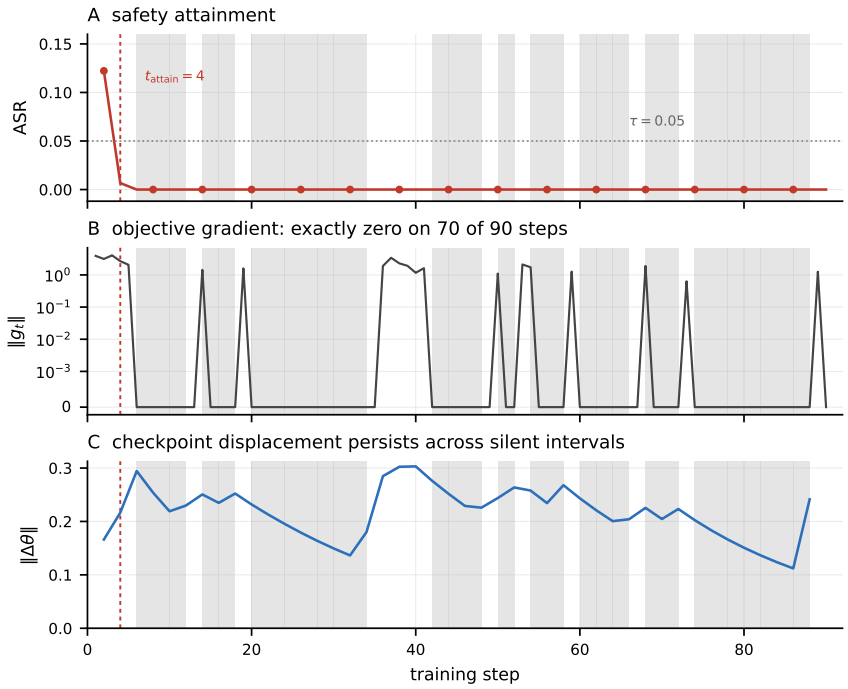


Figure 1: Safety attainment, objective stationarity, and parameter stationarity occur at different times. ASR crosses $\tau = 0.05$ at step 4 in this Mistral-7B CISPO run; 70/90 optimizer steps have zero current gradient, yet parameter displacement persists across silent intervals.

This trajectory motivates two separate questions. Q1: Does the current objective still produce a preference gradient after attainment? Q2: Can stored optimizer state move the policy when the current gradient is zero? Section 4 addresses Q1, and Section 5 addresses Q2 after the objective becomes silent.

4 Source I: Residual Preference in the Current Objective

4.1 The objective can remain active after attainment

DPO. For DPO (Rafailov et al., 2023), let

$$z(\theta) = \beta \left[\log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right]$$

for prompt x and preference pair (y_w, y_l). Then $\mathcal{L}_{\text{DPO}} = -\log \sigma(z)$ and

$$\nabla_\theta \mathcal{L}_{\text{DPO}} = -\sigma(-z) \nabla_\theta z. \tag{2}$$

Because $\sigma(-z) > 0$ for any finite z, a large preference margin alone does not make the per-pair gradient contribution exactly zero.

To create a diagnostic setting with exact stopping, we use a finite-margin hinge:

$$\mathcal{L}_{\text{hinge}} = \max(0, \delta - z), \qquad \nabla_\theta \mathcal{L}_{\text{hinge}} = -\mathbf{1}[z < \delta] \nabla_\theta z. \tag{3}$$

Pairs with $z \geq \delta$ contribute no gradient, allowing safety attainment and objective silence to be examined separately. Additional margin results are reported in Appendix C.

CISPO. In CISPO, relative reward provides an analogous source of residual pressure. If rollout y_i receives reward r_i, its group-normalized advantage is

$$A_i = \frac{r_i - \bar{r}}{\sigma_r + \epsilon_{\text{num}}}. \tag{4}$$

If every rollout receives the same reward, then $A_i = 0$. Conversely, even when safety scores are nearly saturated, any remaining reward ordering among safe responses keeps the objective discriminating among them. Post-attainment behavior can therefore depend on the residual ordering, not only on the achieved safety level.

Although DPO and CISPO differ in form, the experimental intervention is the same in spirit: we hold the primary safety ordering fixed and vary only the preference that remains within the sufficiently safe region.

4.2 Residual preference control

For DPO, the finite-margin hinge removes the safety-pair gradient once the margin is satisfied, and a separate utility pair specifies the remaining preference. On harmful prompts, we prefer a substantive safe response to a blanket refusal; on benign prompts, we prefer a helpful response to an unhelpful one:

$$y_{\text{substantive safe}} \succ y_{\text{blanket refusal}}, \qquad y_{\text{helpful}} \succ y_{\text{unhelpful}}.$$

For CISPO, we keep the primary safety ordering fixed and introduce a secondary ordering only within the maximal-safety set

$$M(x) = \{y : s(x, y) = s_{\max}\},$$

where $s_{\max}$ is the largest attainable safety score, not the maximum within a sampled group. The helpfulness condition is

$$r'(x, y) = s(x, y) + \varepsilon \, \mathbf{1}[y \in M(x)] h(x, y). \tag{5}$$

We compare helpfulness, none, anti-helpfulness, and random secondary orderings, with ε chosen small enough not to reverse the primary safety ordering.

For DPO, adding the utility pair lowers ORR and raises SCR. In the four-way CISPO control, only the helpfulness ordering preserves high SCR; none, anti-helpfulness, and random lead to response degeneration. Because this control uses only harmful prompts, it does not directly train against benign over-refusal. With mixed data that also contains benign prompts, ORR falls substantially; results across three model families are compared in Table 3. Full grader results and implementation details appear in Appendices B and A.

These results address the first gap in Eq. 1: what the objective continues to prefer after attainment can keep changing behavior. The next section imposes the stronger condition of objective silence and asks whether the parameters can still move.

Table 1: Objective-side controls. For DPO, values are the 2-seed range on Mistral-7B and ASR is $\text{ASR}_{\min}$; for CISPO, values are the 3-seed mean ± standard deviation on Mistral-7B and ASR is the 90-step endpoint. ORR is N/A when degeneration exceeds 20%.

Algorithm	Objective	ASR ↓	ORR ↓	SCR ↑	Deg.
DPO	Safety-only hinge	0.000	0.280–0.288	0.858–0.876	–
	+ utility preference	0.009–0.011	0.156	0.980–0.987	–
CISPO	Helpfulness	0.001 ± 0.001	0.977 ± 0.006	$\mathbf{0.997 \pm 0.001}$	0%
	None	0.000 ± 0.000	N/A	0.002 ± 0.002	51%
	Anti-helpfulness	0.000 ± 0.000	N/A	0.000 ± 0.000	100%
	Random	0.001 ± 0.001	N/A	0.001 ± 0.003	76%

5 Source II: Updates from Stored Optimizer State

5.1 Zero gradient does not imply zero update

AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) maintains first and second moments

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \tag{6}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2. \tag{7}$$

If the next k steps after t_0 are silent, then

$$m_{t_0+k} = \beta_1^k m_{t_0}, \qquad v_{t_0+k} = \beta_2^k v_{t_0}. \tag{8}$$

Ignoring ϵ and weight decay, the silent-step update is

$$u_{t_0+k} \approx -\eta \left(\frac{\beta_1}{\sqrt{\beta_2}} \right)^k \frac{c_{t_0+k}}{c_{t_0}} q_{t_0}, \qquad q_{t_0} = \frac{\hat{m}_{t_0}}{\sqrt{\hat{v}_{t_0}}}, \tag{9}$$

where $c_t = \sqrt{1-\beta_2^t}/(1-\beta_1^t)$. With $(\beta_1, \beta_2) = (0.9, 0.999)$, $\beta_1/\sqrt{\beta_2} \approx 0.9005$, so the stored first moment decays gradually rather than disappearing when g_t becomes zero.

An exact-silence regime can be induced in a controlled way with finite-margin DPO and also arises naturally in safety-only CISPO. Within the CISPO silent regime, 78-93% of all steps have exactly zero gradient. Detailed statistics and cumulative dynamics are reported in Appendices F and G.

Equation 9 shows that a residual update is possible on silent steps, but causal attribution requires intervening on the optimizer state itself. We therefore branch from the same zero-gradient CISPO checkpoint and vary only the first moment.

5.2 Same-checkpoint optimizer-state intervention in CISPO

Removal check. We branch a safety-only CISPO trajectory at a checkpoint with $g_{t_0} = 0$ and $m_{t_0} \neq 0$. Weights, data, reward, schedule, v_{t_0}, and the optimizer step counter are held fixed; only m is changed.

With the inherited m, the parameters move on the first silent step. Setting $m \leftarrow 0$ instead gives

$$D_1 = \|\theta_{t_0+1} - \theta_{t_0}\| = 0. \tag{10}$$

Thus, in the same state with no current gradient, removing m alone eliminates the immediate parameter movement. Decoupled weight decay leaves the analytic term $-\eta\lambda\theta_{t_0}$, but its coordinate-wise relative magnitude is $\eta\lambda = 2 \times 10^{-8}$, which is below the relative precision of the FP32 master weights and is not reflected in the stored parameters (Appendix I.3). The observed movement in this exact-silent state is therefore attributable to the stored first moment.

Resetting v also changes preconditioning and bias correction, and those controls collapse SCR. Our main causal intervention therefore modifies only m, leaving v and the optimizer step counter unchanged. The full six-way control is reported in Appendix H.

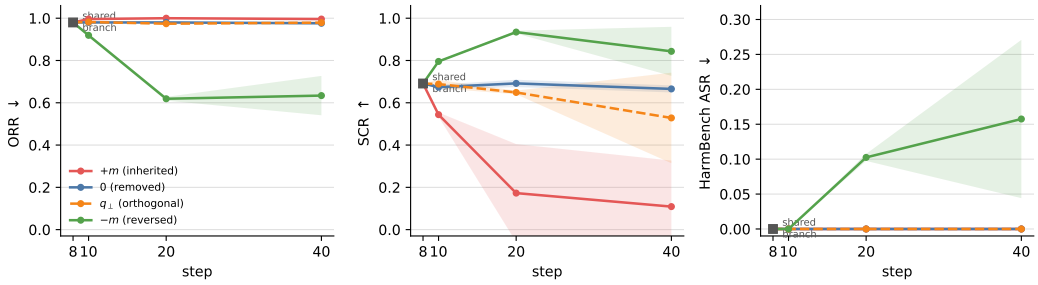


Figure 2: Behavior after changing first-moment direction at the same zero-gradient checkpoint. The square marks the shared branch state; bands show mean ± standard deviation over four seeds. $-m$ and $q_\perp$ are mechanistic controls, not mitigation methods.

Graded magnitude check. To test whether the effect is specific to complete removal, we scale $m_{t_0} \leftarrow \alpha m_{t_0}$ with $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. For Mistral-7B, the two-update displacement D_{early} changes almost proportionally with α. For Qwen2.5-7B and Llama-3.1-8B, the first exact-silent update satisfies

$$\frac{D_1(\alpha)}{D_1(1)} = \alpha, \qquad \alpha \in \{0, 0.25, 0.5, 0.75, 1\},$$

with a maximum deviation of 3.5×10^{-5}. Silent-step movement is therefore graded with the retained first-moment magnitude rather than being an artifact of all-or-nothing reset. The full dose-response appears in Appendix I.

Direction check. We next ask whether equal-magnitude displacements produce the same behavioral drift. From the same checkpoint, we change the direction among $+m$, $-m$, and $q_\perp$ while matching the magnitude of the first exact-silent update. The $m \leftarrow 0$ arm is a reference in which this movement is removed.

Despite matched displacement magnitude, the direction controls in Figure 2 follow different subsequent trajectories in ORR, SCR, and HarmBench ASR. Post-attainment drift therefore depends on the stored first-moment direction as well as the amount of movement. Arm construction, branch-state checks, and behavioral endpoints are reported in Appendix J; trajectory geometry around the branch point is analyzed in Appendix J.5.

We next test an intervention that removes first-moment carryover whenever the objective is silent.

5.3 Silent-step first-moment removal

On steps where the current objective is silent, we remove the stored first moment according to

$$m_t \leftarrow \begin{cases} 0, & g_t = 0, \\ \beta_1 m_{t-1} + (1 - \beta_1) g_t, & \text{otherwise.} \end{cases} \tag{11}$$

We retain v_t and the optimizer step counter. This leaves active gradients and adaptive preconditioning unchanged while removing only first-moment carryover on silent steps. If a new gradient appears, ordinary learning resumes.

Controlled DPO. For finite-margin DPO, we use hinge $\delta = 8$ to create a controlled silent regime. After a fixed branch, we remove m on every silent step and compare against an unintervened trajectory from the same checkpoint.

Online CISPO. In CISPO, zero-gradient states can appear transiently and later reactivate. We therefore detect two consecutive silent steps online and, from the second onward, apply the same first-moment removal to subsequent silent steps. We call this two-step-triggered

Table 2: Silent-step first-moment intervention. For DPO, we report seed ranges after a fixed branch checkpoint in the controlled silent regime; for CISPO, we report the 90-step mean ± seed standard deviation. Mistral-7B uses 4 no-intervention seeds and 3 PAC seeds; Qwen2.5-7B and Llama-3.1-8B use 2 seeds per condition. ASR is ASR_{min} for DPO and the 90-step endpoint for CISPO.

(a) DPO: controlled silent regime

Model	Intervention	$\text{ASR}_{\text{min}} \downarrow$	$\Delta\text{ORR}_{\text{branch}} \downarrow$	$\Delta\text{SCR}_{\text{branch}}$
Mistral-7B	No intervention	0.000	+0.120-+0.184	−0.029-+0.029
	Repeated m removal	0.000	−0.004-+0.020	−0.002-+0.036
Zephyr-7B	No intervention	0.000	+0.180-+0.276	−0.051-−0.009
	Repeated m removal	0.000-0.004	−0.016-+0.040	−0.011-+0.007

(b) CISPO: online silent regime

Model	Intervention	ASR ↓	ORR ↓	SCR ↑
Mistral-7B	No intervention	0.000	0.959 ± 0.082	0.043 ± 0.053
	PAC	0.000	0.615 ± 0.148	0.744 ± 0.082
Qwen2.5-7B	No intervention	0.000	0.950 ± 0.054	0.616 ± 0.057
	PAC	0.000	0.336 ± 0.187	0.839 ± 0.005
Llama-3.1-8B	No intervention	0.000	0.738 ± 0.093	0.001 ± 0.002
	PAC	0.003	0.310 ± 0.071	0.072 ± 0.014

online variant of Eq. 11 PAC (Post-Attainment Control). PAC uses neither the attainment step, an ASR threshold, nor a separate judge; it depends only on whether the current gradient is silent.

Table 2 summarizes the intervention in controlled DPO and online CISPO.

Without intervention in controlled DPO, ORR continues to rise after objective silence on both Mistral-7B and Zephyr-7B. Repeated m removal substantially suppresses this post-silence ORR drift, showing that stored optimizer state can sustain behavioral change even when the current objective no longer supplies an update signal.

The same principle holds in online CISPO. PAC lowers ORR and raises SCR relative to no intervention across all three model families. Unlike DPO's controlled silent regime, CISPO can later reactivate, but removing first-moment carryover whenever silence recurs still changes the subsequent behavioral trajectory.

Seed-level controlled-DPO results, the comparison between one-shot reset and PAC, and cross-model seedwise results are reported in Appendix K. Cross-model comparisons and external evaluation follow in Section 6.

6 Generalization and Scope

6.1 Across intervention axes and model families

Sections 4 and 5 isolate two sources of post-attainment drift: residual objective pressure and optimizer-state carryover. Table 2 further shows that silent-step first-moment removal has an effect in both controlled DPO and online CISPO. Table 3 compares the main CISPO conditions across objective, training support, and optimizer intervention.

On harmful-only Mistral-7B, adding a helpfulness preference raises SCR to 0.997 but leaves ORR high at 0.977 because benign prompts receive no direct training signal. With Mixed data, ORR falls to 0.206 while SCR remains high. Because Mixed changes both the objective and training support, we do not interpret this contrast as a data-only causal effect.

The optimizer-side intervention changes neither the objective nor the data. Repeated m removal reduces post-silence ORR drift in DPO, and PAC lowers ORR and raises SCR

Table 3: Internal 90-step endpoints for the main CISPO conditions. Mixed changes both objective and training support, whereas PAC changes only silent-step first-moment handling.

Model	Objective	Optimizer	Data	ASR ↓	ORR ↓	SCR ↑
Mistral-7B	Safety-only	AdamW	Harmful	0.000	0.959	0.043
	Safety + Helpfulness	AdamW	Harmful	0.001	0.977	0.997
	Safety + Helpfulness	AdamW	Mixed	0.007	0.206	0.990
	Safety-only	AdamW + PAC	Harmful	0.000	0.615	0.744
Qwen2.5-7B	Safety-only	AdamW	Harmful	0.000	0.950	0.616
	Safety + Helpfulness	AdamW	Mixed	0.002	0.188	0.995
	Safety-only	AdamW + PAC	Harmful	0.000	0.336	0.839
Llama-3.1-8B	Safety-only	AdamW	Harmful	0.000	0.738	0.001
	Safety + Helpfulness	AdamW	Mixed	0.007	0.129	0.993
	Safety-only	AdamW + PAC	Harmful	0.003	0.310	0.072

Table 4: External evaluation of PAC in CISPO. XSTest-G is the XSTest grader applied to OR-Bench-hard. Values are seed means ± standard deviation.

Model	Optimizer	HB ASR ↓	XSTest-G ORR ↓	Health-ORSC SCR ↑
Mistral-7B	AdamW	0.000 ± 0.000	0.964 ± 0.060	0.013 ± 0.021
	AdamW + PAC	0.035 ± 0.020	0.875 ± 0.023	0.454 ± 0.020
Qwen2.5-7B	AdamW	0.000 ± 0.000	0.996 ± 0.002	0.012 ± 0.005
	AdamW + PAC	0.017 ± 0.018	0.770 ± 0.180	0.520 ± 0.153
Llama-3.1-8B	AdamW	0.000 ± 0.000	0.992 ± 0.011	0.075 ± 0.084
	AdamW + PAC	0.010 ± 0.007	0.859 ± 0.111	0.337 ± 0.098

across all three CISPO model families. The graded intervention shows the same cross-model pattern: displacement scales with retained first-moment magnitude (Section 5.2; Appendix I).

6.2 Verification with external evaluators

We re-evaluate the online-CISPO Safety-only baseline and PAC with external graders. The two conditions use the same objective and training data and differ only in silent-step optimizer handling. Table 4 reports HarmBench ASR, the XSTest-based grader on OR-Bench-hard, and Health-ORSC SCR.

PAC lowers over-refusal and raises substantive safe completion under external graders for all three model families, while HarmBench ASR rises from 0 to 0.010-0.035. It therefore changes the safety-utility operating point rather than improving utility at no safety cost.

The objective-side intervention shows a similar trade-off: Mixed Safety + Helpfulness reduces OR-Bench ORR from 0.980-1.000 to 0.617-0.713 and XSTest ORR from 0.730-0.830 to 0.039-0.053, while HarmBench ASR increases to 0.043-0.114. Full results are reported in Appendices B.2 and L.

6.3 Scope

Clean attribution of the optimizer-side effect requires exact total-gradient silence. When a KL or utility gradient remains active, current-gradient and optimizer-state effects cannot be separately identified. The corresponding controls are reported in Appendices M and D.

7 Limitations

Causal scope. Our optimizer-side causal claim is limited to exact total-gradient silence. With persistent KL or utility gradients, current-gradient and optimizer-state effects cannot be separately identified.

Experimental coverage. We evaluate DPO on Mistral-7B and Zephyr-7B and CISPO on Mistral-7B, Qwen2.5-7B, and Llama-3.1-8B. The full direction control is limited to Mistral-7B, some conditions use few seeds, and larger models and other preference-optimization algorithms remain untested.

Evaluation coverage. External PAC evaluation covers all three CISPO model families, whereas objective-side external coverage is more limited. The Mixed condition also changes the objective and training support jointly.

Overall, our experiments identify two separable sources of post-attainment drift in controlled settings without claiming identical effects across all alignment recipes.

8 Conclusion

Safety attainment does not end optimization: the current objective may retain residual preference, while stored optimizer state may continue to move parameters after the current gradient becomes silent. Our interventions isolate these two sources and show that safety attainment does not imply objective stationarity, which in turn does not imply parameter stationarity. Post-attainment stabilization therefore requires separate control of residual preference and optimizer state.

AI Use Statement

Generative AI tools were used to assist with code implementation and debugging, and to improve the clarity and readability of the manuscript. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content.

References

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Rémi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 4447–4455, 2024.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An over-refusal benchmark for large language models. In International Conference on Machine Learning (ICML), 2025.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: Safe reinforcement learning from human feedback. In International Conference on Learning Representations (ICLR), 2024.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. In International Conference on Learning Representations (ICLR), 2024.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In Proceedings of the 40th International Conference on Machine Learning (ICML), pp. 10835–10866, 2023.

Tianjin Huang, Ziquan Zhu, Gaojie Jin, Lu Liu, Zhangyang Wang, and Shiwei Liu. SPAM: Spike-aware Adam with momentum reset for stable LLM training. In International Conference on Learning Representations (ICLR), 2025.

Batuhan K. Karaman, Ishmam Zabir, Alon Benhaim, Vishrav Chaudhary, Mert R. Sabuncu, and Xia Song. POROver: Improving safety and reducing overrefusal in large language models with overgeneration and preference optimization. In International Conference on Machine Learning (ICML), 2025.

Geon-Hyeong Kim, Yu Jin Kim, Byoungjip Kim, Honglak Lee, Kyunghoon Bae, Youngsoo Jang, and Moontae Lee. SafeDPO: A simple approach to direct preference optimization with enhanced safety. In International Conference on Learning Representations (ICLR), 2026.

Taeyoun Kim and Aviral Kumar. Addressing over-refusal in LLMs with competing rewards. arXiv preprint arXiv:2606.31748, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR), 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations (ICLR), 2019.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In Advances in Neural Information Processing Systems (NeurIPS), 2024.

MiniMax. MiniMax-M1: Scaling test-time compute efficiently with lightning attention. arXiv preprint arXiv:2506.13585, 2025.

Sagnik Mukherjee, Lifan Yuan, Pavan Jayasinha, Dilek Hakkani-Tür, and Hao Peng. Do we need Adam? surprisingly strong and sparse reinforcement learning with SGD in LLMs. arXiv preprint arXiv:2602.07729, 2026.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In International Conference on Learning Representations (ICLR), 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In Advances in Neural Information Processing Systems (NeurIPS), 2024.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. In International Conference on Learning Representations (ICLR), 2025.

Yi Ren and Danica J. Sutherland. Learning dynamics of LLM finetuning. In International Conference on Learning Representations (ICLR), 2025.

Paul Röttger, Hannah Rose 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 the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), 2024.

Guobin Shen, Dongcheng Zhao, Haibo Tong, Jindong Li, Feifei Zhao, and Yi Zeng. Safety instincts: LLMs learn to trust their internal compass for self-defense. In International Conference on Learning Representations (ICLR), 2026. arXiv:2510.01088.

Junkang Wu, Kexin Huang, Xue Wang, Jinyang Gao, Bolin Ding, Jiancan Wu, Xiangnan He, and Xiang Wang. RePO: Understanding preference learning through ReLU-based optimization. In Advances in Neural Information Processing Systems (NeurIPS), 2025.

Yuan Yuan, Tina Sriskandarajah, Anna-Luisa Brakman, Alec Helyar, Alex Beutel, Andrea Vallone, and Saachi Jain. From hard refusals to safe-completions: Toward output-centric safety training. arXiv preprint arXiv:2508.09224, 2025.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. SLiC-HF: Sequence likelihood calibration with human feedback. arXiv preprint arXiv:2305.10425, 2023.

A Detailed Definitions of the Objective-Side Interventions

A.1 Finite-margin DPO control

In Eq. (3), we take the subgradient at $z = \delta$ to be 0. Let z_i denote the margin of preference pair i, and define the set of pairs that have not reached the stopping margin as

$$A_t = \{i : z_i < \delta\}. \tag{12}$$

Then

$$g_t = -\frac{1}{B} \sum_{i \in A_t} \nabla_\theta z_i. \tag{13}$$

Thus $A_t = \emptyset$ implies $g_t = 0$, although pairwise cancellation can also yield $g_t = 0$ when $A_t \neq \emptyset$. Accordingly, $A_t = \emptyset$ is sufficient, but not necessary, for exact objective silence. A small δ can stop the objective before safety attainment; Appendix C checks how the criterion aligns with the observed hinge runs.

Table 5: Safety-grader results for the CISPO secondary-ordering control. Internal-judge values are 3-seed means ± standard deviation; external-grader values use 1 seed.

Secondary ordering	Deg.	Internal ASR	HarmBench	StrongREJECT	SORRY-Bench
Helpfulness	0%	0.001 ± 0.001	0.000	0.205	0.235
None	51%	0.000 ± 0.000	0.000	0.005	0.000
Anti-helpfulness	100%	0.000 ± 0.000	0.000	0.082	0.000
Random	76%	0.001 ± 0.001	0.005	0.014	0.010

Table 6: Utility results for the same control. OR-Bench and XSTest-grader are both evaluated on the 400 OR-Bench-hard prompts. ORR is N/A when degeneration exceeds 20%.

Secondary ordering	Deg.	Internal ORR ↓	OR-Bench ↓	XSTest-grader ↓	Internal SCR ↑
Helpfulness	0%	0.977 ± 0.006	1.000	0.980	$\mathbf{0.997 \pm 0.001}$
None	51%	N/A	N/A	N/A	0.002 ± 0.002
Anti-helpfulness	100%	N/A	N/A	N/A	0.000 ± 0.000
Random	76%	N/A	N/A	N/A	0.001 ± 0.003

A.2 CISPO reward-ordering intervention

We keep the primary safety ordering fixed and vary only the secondary ordering within the maximal-safety set $M(x)$. none adds no secondary ordering; helpfulness and anti-helpfulness impose opposite helpfulness orderings; and random uses a content-independent ordering defined by a deterministic hash. Together, these conditions distinguish the content of the secondary preference from the effect of tie-breaking itself. Implementation constants are given in Appendix E.

B Additional Results for the Objective-Side Interventions

B.1 External grader results for CISPO secondary ordering

Tables 5 and 6 report the external-grader results for the four-way control. We report ORR as N/A when degeneration exceeds 20%.

B.2 Effect of the training distribution

The helpfulness tie-break prevents safe-completion collapse on harmful prompts but provides no direct training signal for benign prompts. Table 7 evaluates the Safety + Helpfulness objective on mixed data. Because reward and training distribution change together, this comparison does not isolate the causal effect of adding benign data.

The Mixed condition lowers final ORR and raises SCR across all three model families. Its $\text{ORR}_{\max}$ is 0.568 for Mistral and at most 0.272 for Qwen and Llama. Because reward and data change jointly, we treat this as a cross-model result rather than a data-only causal ablation.

External objective evaluation. Table 8 reports external evaluation for the same two conditions. Here, XSTest refers to the 250 XSTest safe prompts and is distinct from XSTest-grader, which applies the XSTest grader to OR-Bench-hard prompts.

Mixed Safety + Helpfulness reduces over-refusal across all three model families but incurs measurable cost on some external safety metrics. As above, reward and training distribution change jointly.

Table 7: 90-step internal judge results for the safety-only objective (harmful-only data) and the safety-helpfulness objective (mixed data) across three model families. For Mistral, safety-only uses 4 seeds and helpfulness uses 4 seeds; for Qwen and Llama, safety-only uses 2 seeds and helpfulness uses 3 seeds; values are mean [seed range]. The two conditions differ jointly in reward and training distribution.

Model	Objective	ASR ↓	ORR ↓	$\text{ORR}_{\max}$	SCR ↑
Mistral-7B	Safety-only	0.000	0.959 [0.836, 1.000]	1.000	0.043 [0.000, 0.118]
	Safety + Helpfulness	0.007	0.206 [0.140, 0.356]	0.568	0.990 [0.982, 0.996]
Qwen2.5-7B	Safety-only	0.000	0.950 [0.912, 0.988]	1.000	0.616 [0.576, 0.656]
	Safety + Helpfulness	0.002	0.188 [0.136, 0.272]	0.272	0.995 [0.991, 1.000]
Llama-3.1-8B	Safety-only	0.000	0.738 [0.672, 0.804]	0.816	0.001 [0.000, 0.002]
	Safety + Helpfulness	0.007	0.129 [0.112, 0.148]	0.180	0.993 [0.987, 1.000]

Table 8: Official grader results for the Safety-only versus Safety + Helpfulness objectives. Values are the seed mean at step 90.

Model	Objective	HarmBench ↓	StrongREJECT ↓	OR-Bench ORR ↓	XSTest ORR ↓
Mistral-7B	Safety-only	0.000	0.022	0.980	0.830
	Safety + Helpfulness	0.114	0.177	0.617	0.039
Qwen2.5-7B	Safety-only	0.000	0.009	1.000	0.752
	Safety + Helpfulness	0.067	0.083	0.713	0.042
Llama-3.1-8B	Safety-only	0.000	0.026	0.992	0.730
	Safety + Helpfulness	0.043	0.034	0.640	0.053

C Additional Verification of the Finite-Margin DPO Control

Exact objective silence. At every logged post-attainment hinge step, $A_t = \emptyset$ coincides with $\|g_t\| = 0$. The observed silence is therefore consistent with finite-margin stopping. Because cancellation can still produce $g_t = 0$ when $A_t \neq \emptyset$, we do not treat $A_t = \emptyset$ as a necessary condition.

Post-attainment movement. Table 9 summarizes the post-attainment window and cumulative movement. Attainment is the first checkpoint with ASR ≤ 0.05, and $D_{\text{post}} = \sum_{t>t_{\text{attain}}} \|u_t\|$ is the sum of subsequent LoRA update_norm values.

For $\delta = 8$, per-step movement is not smaller than under standard DPO, so the large D_{post} cannot be explained by a longer post-attainment window alone.

D Objective-Side Prescription in DPO: Utility Pairs

The safety pair specifies only "substantive safe > unsafe" and leaves the ordering among safe behaviors unspecified. We therefore impose a finite safety margin and add utility pairs that prefer substantive safe responses over blanket refusals on harmful prompts, and helpful responses over unhelpful ones on benign prompts. For 732 examples, the safety pair and the first utility pair share the same prompt and chosen response and differ only in the rejected response.

In the length-matched control (chosen/rejected length ratio 0.98-1.02), adding the utility pair to Mistral hinge $\delta = 8$ changes ORR from 0.280/0.288 to 0.156/0.156 and SCR from 0.876/0.858 to 0.987/0.980. $\text{ASR}_{\min}$ changes from 0.000/0.000 to 0.009/0.011. Table 1 reports this 2-seed control.

2×2 control. The endpoints and fixed-branch changes are summarized in Table 10.

Interpretation of the joint arm. In the objective-only runs, the utility gradient remains nonzero even when the safety component is silent: 9.8-32.7 for Mistral and 10.3-15.0 for Zephyr. The joint arm is therefore an interaction control with an active utility gradient, not

Table 9: Post-attainment movement for the finite-margin DPO control. Values are seed ranges.

Model	Objective	Post steps	D_{post}	D_{post}/step
Mistral-7B	DPO	76	0.793-0.852	0.0104-0.0112
Mistral-7B	hinge $\delta = 2$	61-76	0.096-0.469	0.0016-0.0062
Mistral-7B	hinge $\delta = 8$	76	1.115-1.224	0.0147-0.0161
Zephyr-7B	DPO	45-61	0.402-0.772	0.0089-0.0127
Zephyr-7B	hinge $\delta = 2$	N/A	N/A	N/A
Zephyr-7B	hinge $\delta = 8$	61	0.998-1.036	0.0164-0.0170

Table 10: Full 2×2 metrics for the two interventions. Values are ranges over 4 seeds per condition.

Model	Condition	ASR_{min}	ASR_{end}	$\Delta\text{ORR}_{\text{branch}}$	$\Delta\text{SCR}_{\text{branch}}$
Mistral-7B	no intervention	0.000	0.000	+0.120-+0.184	−0.029-+0.029
	objective	0.000-0.009	0.002-0.011	+0.032-+0.096	+0.009-+0.018
	optimizer	0.000-0.002	0.000-0.007	−0.004-+0.044	−0.036-+0.000
	joint	0.013-0.038	0.024-0.040	+0.008-+0.032	+0.020-+0.027
Zephyr-7B	no intervention	0.000	0.000	+0.180-+0.276	−0.051-−0.009
	objective	0.007-0.011	0.007-0.013	−0.004-+0.028	+0.002-+0.038
	optimizer	0.007-0.011	0.007-0.013	+0.012-+0.024	−0.020-+0.002
	joint	0.016-0.036	0.016-0.040	−0.016-+0.016	−0.011-+0.022

a clean composition of the two interventions. Its final safety margin falls from 25.7-28.9 to 15.4-20.6, while the mean utility margin is similar (24.4 versus 23.0). Across four seeds, the ASR_{min} interaction is $+0.011/+0.020/+0.029/+0.022$; given the sample size, we report this descriptively rather than as a formal interaction test. We did not measure the cosine between the two update vectors.

E Implementation Details

In Eq. (4), $\epsilon = 10^{-6}$ and σ_r is the sample standard deviation. At the observed reward scale, this numerical stabilizer has negligible effect.

Constants of the tie-break reward. We implement Eq. (5) as $r'(x, y) = s(x, y) + \varepsilon \mathbf{1}[s(x, y) \geq s_{\text{thr}}]t(x, y)$ with $\varepsilon = 0.01$ and $s_{\text{thr}} = 0.99$. The safety score has a minimum grid spacing of 0.025, so $s \geq 0.99$ selects only the maximum score $s = 1$. We enforce $\varepsilon < 0.025$ to prevent the secondary term from reversing the primary safety ordering.

Parameter dtype and norm measurement. DPO uses a BF16 base model with an FP32 LoRA adapter (the PEFT 0.20 default, autocast_adapter_dtype=True). We compute grad_norm and update_norm on the FP32 adapter after reducing per-rank sums of squares. CISPO uses FP32 FSDP master parameters with BF16 mixed precision, and checkpoints are stored as FP32 shards. Appendix I.3 details the resulting measurement precision.

F Additional Analysis of the CISPO Silent Regime

F.1 Emergence of the Silent Regime

Figure 3 compares the exact-zero-gradient fraction across 48 CISPO runs of 90 steps each. Within the silent regime, 78-93% of steps have exactly zero gradient. All 10 runs without repeated m removal end with SCR ≤ 0.118, whereas the 5 runs with removal end at SCR 0.662-0.902. Including objective-active runs, the association between regime and SCR gives

Table 11: Empirical regime and final SCR for the 48 CISPO runs.

Regime	Runs	Condition	Final SCR
silent ≥ 0.5, no repeated m removal	10	safety-only reward	0.000-0.118
silent ≥ 0.5, repeated m removal	5	safety-only reward + silent-step m removal	0.662-0.902
active, objective-induced collapse	6	anti-helpfulness / random tie-break	0.000-0.004
active, SCR preserved	27	reward with a utility term	0.860-1.000

Table 12: Agreement between exact silence and steps where "all rollouts have maximal reward." This is an empirical observation, not an algebraic equivalence.

Condition (runs)	Steps	Silent	All-max	Mismatch
PAC (3)	270	243	243	0
One-shot reset (3)	270	230	230	0
No intervention (4)	360	305	305	0
Tie-break none (3)	270	240	240	0
Oracle-timed gate (2)	180	148	148	0
Total	1,350	1,166	1,166	0

Fisher's exact $p = 2.9 \times 10^{-9}$. We use this association descriptively only; causal attribution to optimizer state comes from the matched intervention in Appendix H.

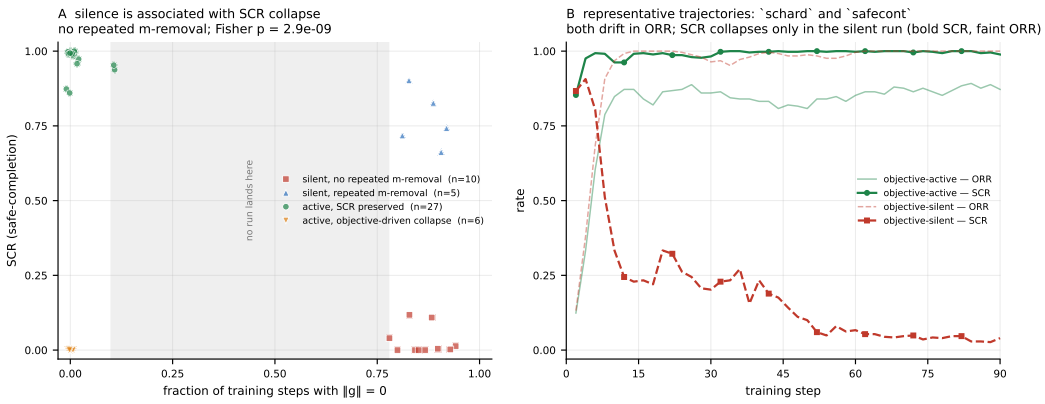


Figure 3: The empirical silent regime of CISPO. The left panel shows the silent-step fraction and final SCR for the 48 runs; the right panel shows representative trajectories for objective-active and objective-silent runs.

No run has a silent fraction between 0.10 and 0.78, so moving the 0.5 cutoff anywhere within that interval yields the same grouping.

F.2 Agreement Between Silence and Maximal Reward

Across logged training steps, we compare grad_norm=0 with the condition that every rollout reward in the batch equals the maximum value. The two criteria agree on all 1,350 steps from 15 runs.

Equal rollout rewards are sufficient for $A_i = 0$, but they are not necessary for the batch gradient to be exactly zero. We therefore treat the perfect agreement in Table 12 as an empirical property of these runs, not an algebraic equivalence.

Table 13: Same-checkpoint optimizer-state intervention. Each branch differs only in the optimizer state component.

Intervention	m	v	Step counter	Final SCR
No intervention	keep	keep	keep	0.004
Repeated m removal	silent-step reset	keep	keep	0.729
One-shot m reset	reset once	keep	keep	0.716
One-shot v reset	keep	reset once	keep	0.000
One-shot m, v reset	reset once	reset once	keep	0.000
Full re-initialization	reset	reset	reset	0.000

G Optimizer Dynamics in the Silent Interval

Suppose the K steps after t_0 in Eq. 9 are consecutively silent. Define the path length over that silent interval as

$$D_{\text{silent}}(K) = \sum_{k=1}^{K} \|u_{t_0+k}\| \tag{14}$$

Ignoring ϵ and weight decay,

$$D_{\text{silent}}(K) \approx \eta \|q_{t_0}\| \sum_{k=1}^{K} \rho^k \frac{c_{t_0+k}}{c_{t_0}}, \qquad \rho = \frac{\beta_1}{\sqrt{\beta_2}}. \tag{15}$$

For $(\beta_1, \beta_2) = (0.9, 0.999)$, $\rho = 0.9005$. Late in training, if $c_{t_0+k}/c_{t_0} \approx 1$,

$$D_{\text{silent}}(K) \approx \eta \|q_{t_0}\| \frac{\rho(1-\rho^K)}{1-\rho}, \tag{16}$$

and the infinite-horizon multiplier is approximately 9.05. Earlier in training, bias correction is still changing, so the absolute amount of silent movement also depends on when the silent interval begins.

Why remove only m. In AdamW, m supplies the numerator of the preconditioned update and v the coordinate-wise preconditioner. Setting $m \leftarrow 0$ removes first-moment history while preserving both v and the optimizer step counter. Resetting v or fully reinitializing the optimizer also changes update scale, preconditioning, and bias correction, so those controls are qualitatively different interventions rather than stronger versions of m removal.

H Additional Verification of the Matched Optimizer-State Intervention

H.1 Six-way same-checkpoint control

All branches start from the same checkpoint with $g_t = 0$ and $m_t \neq 0$, sharing model weights, data order, reward, schedule, and random seed.

Removing only m preserves SCR, whereas controls that reset v collapse. The causal comparison therefore changes only m while preserving v and the step counter. Figure 4 shows the six-way branch endpoint together with an ORR trajectory from an earlier checkpoint.

H.2 Earlier checkpoint replication

Because ORR is already near 0.98-1.00 at the 8-step branch, we repeat the intervention from an earlier checkpoint with ORR approximately 0.64. Steps 5-9 after the branch are all silent. Without intervention, ORR rises to 0.868 by step 6; with repeated m removal it remains at 0.640-0.652. One-shot reset is seed-sensitive at the endpoint, whereas repeated removal continues to maintain lower ORR and higher SCR.

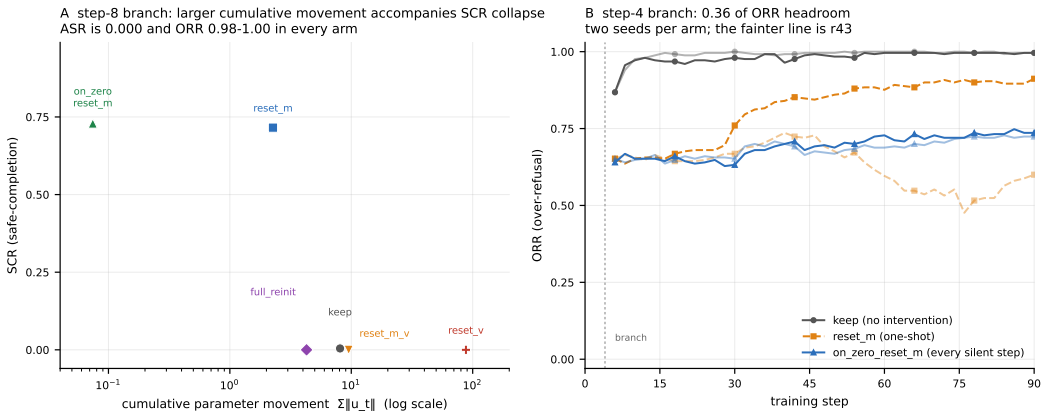


Figure 4: Matched optimizer-state intervention. Left: final SCR of the six-way branch. Right: ORR trajectory from an earlier checkpoint, according to whether the first moment is retained.

Table 14: Mistral-7B graded first-moment intervention. $D_{\text{early}} = \|\theta_{t_0+2} - \theta_{t_0}\|$.

Seed	$\alpha = 0$	0.25	0.50	0.75	1.00
42	0.00000	0.02203	0.04407	0.06610	0.08814
43	0.00000	0.02203	0.04407	0.06614	0.08814
44	0.00000	0.02203	0.04410	0.06620	0.08814
46	0.00000	0.02203	0.04407	0.06610	0.08814
$\cos(\Delta\theta_{\text{early}}, d_0)$	N/A	+1.000	+1.000	+1.000	+1.000

H.3 Rollout-seed robustness

Changing only the rollout seed leaves the silent-step weights of the no-intervention control byte-identical. Under matched m removal, the first silent-step displacement is exactly zero in FP32. Thus the local movement at this branch is determined by optimizer state rather than rollout sampling.

H.4 Reset-state simulation

We feed the same gradient sequence through different reset states. Resetting both m and v, or fully resetting the optimizer, changes bias correction and the scale of the first re-entry update; the simulation reproduces the qualitative ordering of the matched branch.

I Magnitude dose-response

I.1 Mistral-7B graded intervention

From the same FP32 checkpoint, we set

$$m_{t_0} \leftarrow \alpha m_{t_0}, \qquad \alpha \in \{0, 0.25, 0.5, 0.75, 1\},$$

and repeat the intervention across four seeds. The source state and optimizer step counter are byte-identical across all 20 arms. The first update is exact-silent in every arm; 16/20 arms remain silent through the second update. For the remaining four arms, the contribution of the live second update to D_{early} is at most 7×10^{-4}.

Within each seed, D_{early} is nearly proportional to α. Because the first update is deterministic given the shared branch state, identical D_1 values across seeds verify the implementation rather than provide independent statistical replication. Rollout stochasticity re-enters only for arms whose second update is live.

Table 15: Cross-model magnitude dose-response. Table values are $D_1/D_1(\alpha{=}1)$. The first update of each arm is exact-silent.

Model	Branch	$\alpha = 0$	0.25	0.50	0.75	1.00
Qwen2.5-7B	step 4	0.000	0.250	0.500	0.750	1.000
Llama-3.1-8B	step 8	0.000	0.250	0.500	0.750	1.000

I.2 First-update comparison across model families

For Qwen2.5-7B and Llama-3.1-8B, we likewise compare the first exact-silent update,

$$D_1 = \|\theta_{t_0+1} - \theta_{t_0}\|.$$

Qwen step 8 is not silent, so we branch at step 4 instead. Because the first update is exact-silent from a fixed source state, each model uses a single branch state.

For positive-α arms, $\cos(\Delta\theta_1, d_0)$ is +1.0000 for Qwen and +0.9998-+1.0000 for Llama. The absolute slope is 0.069 for Qwen and 0.043 for Llama, versus Mistral's two-update slope of 0.088. Our comparison concerns proportionality within each model, not equality of absolute displacement scales across models.

I.3 Measurement precision

All parameter-space movement and geometry are computed directly from FP32 FSDP master weights. The exported hf_merged checkpoint is BF16, and we do not mix BF16 and FP32 representations when taking differences. At the branch checkpoint, $\|\theta - \mathrm{bf16}(\theta)\| = 0.353$ while $\|\theta\| = 1285.6$; subtracting a BF16 endpoint from an FP32 source would therefore impose a 0.353 quantization floor on small movements. Even BF16-to-BF16 differences are not quantitatively accurate for this purpose: over the same interval, true FP32 movements of 0.112 and 0.101 appear as 0.259 and 0.262, inflated by 2.31× and 2.59×. Because the inflation depends on parameter scale and update direction, it cannot be removed with a single scalar correction.

Decoupled weight decay uses $\lambda = 0.01$. At $\eta = 2 \times 10^{-6}$, its per-coordinate relative change is $\eta\lambda = 2 \times 10^{-8}$, below FP32 relative precision. Thus, in the $m \leftarrow 0$ arm, the analytic term $-\eta\lambda\theta_{t_0}$ does not change the stored master weight, and Eq. (10) is measured as exactly $D_1 = 0$. At the branch point, $\lambda\|\theta_{t_0}\|/\|q_{t_0}\| = 4.9 \times 10^{-4}$.

J Direction Intervention: Construction and Additional Results

J.1 Arm construction

Let q_0 denote the inherited preconditioned update direction. Because

$$q(-m) = -q(m),$$

the $-m$ arm exactly reverses the inherited direction.

To construct the orthogonal control $q_\perp$, we generate a random vector for each tensor, remove its component along q_0, rescale it to satisfy $\|q_{\perp,j}\| = \|q_{0,j}\|$, and map it back to first-moment space. This changes direction while preserving the per-layer norm of the preconditioned update.

The raw $\|m\|$ need not match across arms. Because AdamW scales coordinates through v, the relevant magnitude-control quantity is $\|q\|$, not $\|m\|$.

J.2 Branch-state audit

All 16 direction runs share θ_{t_0}, m_{t_0}, v_{t_0}, and the optimizer step counter. A hard assertion verifies $\|g_t\| = 0$ for the first post-intervention update in every arm.

Table 16: Step-20 results for changing only the first-moment direction from the same branch state. $q_\perp$ and $-m$ have the same immediate update magnitude as $+m$. 4-seed average.

Intervention	ORR ↓	SCR ↑	XSTest ORR ↓	HarmBench ASR ↓
Shared branch	.980	.691	.688	.000
$+m$ (inherited)	1.000	.173	.942	.000
0 (removed)	.980	.692	.689	.000
$q_\perp$ (orthogonal)	.974	.649	.717	.000
$-m$ (reversed)	.619	.934	.091	.102

The immediate displacement is

$$\frac{D_1(-m)}{D_1(+m)} = 1.000, \qquad \frac{D_1(q_\perp)}{D_1(+m)} = 1.000,$$

and

$$\cos(\Delta\theta_1, d_0) = -1.000 \quad (-m), \qquad \cos(\Delta\theta_1, d_0) \approx 0 \quad (q_\perp).$$

Including decoupled weight decay, the $-m$ cosine remains -1.000; at the branch point, $\|\lambda\theta_0\|/\|q_0\| \approx 4.9 \times 10^{-4}$. The immediate D_1 and cosine values verify the matched construction. Seed variation is evaluated on subsequent trajectories and behavioral endpoints.

J.3 Behavioral endpoints

Reversing the direction reverses the ORR/SCR drift but introduces a safety cost. HarmBench ASR is 0.102 at step 20 and 0.158 ± 0.113 at step 40; StrongREJECT changes as $0.028 \to 0.054 \to 0.130 \pm 0.087$. We therefore treat $-m$ as a mechanistic direction control, not a mitigation method.

J.4 Objective reactivation

Among the 40 updates following the intervention, the number of steps with $\|g\| \neq 0$ is

$$+m: \ 11.50 \pm 7.94, \qquad 0: \ 1.00 \pm 0.82,$$

$$-m: \ 13.75 \pm 2.22, \qquad q_\perp: \ 0.75 \pm 0.96.$$

Two of four $q_\perp$ seeds never reactivate within 40 updates, and in all four seed-matched comparisons $q_\perp$ has fewer live steps than $+m$. The $q_\perp$ arm is a synthetic mechanistic control for direction, not a natural optimizer trajectory.

J.5 Trajectory geometry around the branch point

We project displacement around the branch point onto a two-dimensional plane. The first basis vector is the inherited update direction, $e_1 = d_0 = -q_{t_0}$. For the second, we take the actual first update of the $q_\perp$ arm, $w = \theta^{q_\perp}_{t_0+1} - \theta_{t_0}$, and orthogonalize it against e_1. We use the realized update rather than reconstructing $q_\perp$ because the intervention samples random vectors at the level of FSDP flat shards; replaying the same seed at named-parameter granularity produces a different random direction. Measured values $\|w\| = 0.0466$ and $\cos(w, d_0) = -3.2 \times 10^{-9}$ confirm that the orthogonal arm moves perpendicular to d_0 with the same first-update magnitude, $D_1 = 0.0466$, as the $+m$ arm.

Pre-branch alignment. Pre-branch displacement is strongly aligned with d_0: $\cos(\theta_{t_0} - \theta_{t_0-4}, d_0) = 0.986$ and $\cos(\theta_{t_0} - \theta_{t_0-6}, d_0) = 0.910$; the alignment weakens as the interval extends farther into the past. For $\theta_{t_0} - \theta_{t_0-2}$, both steps are silent, so the cosine of 1.000 follows directly from Eq. (8) and is not independent evidence. Thus the stored m is aligned with the recent optimizer trajectory rather than pointing in an arbitrary direction. This is a statement about local trajectory alignment, not about loss-landscape curvature.

Table 17: Silent-step first-moment intervention for DPO hinge $\delta = 8$. Values are seed ranges.

Model	Intervention	$\mathrm{ASR}_{\min}$	$\Delta\mathrm{ORR}_{\mathrm{branch}}$	$\Delta\mathrm{SCR}_{\mathrm{branch}}$
Mistral-7B	No intervention	0.000	+0.120-+0.184	−0.029-+0.029
	Repeated m removal	0.000	**−0.004-+0.020**	−0.002-+0.036
Zephyr-7B	No intervention	0.000	+0.180-+0.276	−0.051-−0.009
	Repeated m removal	0.000-0.004	**−0.016-+0.040**	−0.011-+0.007

Displacement magnitude and behavior. For one seed, displacement after 20 updates is 0.347 for $+m$, 0.334 for $-m$, 0.333 for $q_\perp$, and 0.016 for $m \leftarrow 0$. The three moving arms are within 4% in displacement magnitude, yet Table 16 shows ORR spanning 0.619-1.000. Moreover, $q_\perp$ moves more than 20 times farther than $m \leftarrow 0$ while ending at a similar behavioral point. Across all four seeds, only the first-update D_1 ratio is explicitly magnitude-matched; the 20-update magnitude agreement is specific to this seed.

Out-of-plane component. For the three arms with an injected first moment, the out-of-plane component after 40 updates increases in the same order as objective reactivation: $q_\perp$ (0.75 live steps) has 0.008, $+m$ (11.50) has 0.502, and $-m$ (13.75) has 0.573. The $m \leftarrow 0$ arm has no injected direction; of its total displacement 0.025, the component along d_0 is only -1.3×10^{-4}. This is consistent with its remaining movement arising from reactivated gradients rather than stored carryover.

K Additional Results for Silent-Step First-Moment Removal

K.1 Controlled DPO silent regime

For DPO hinge $\delta = 8$, the first moment is removed at every silent step after the fixed branch. The branch step is 30 for Mistral and 46 for Zephyr.

Repeated m removal substantially reduces post-silence ORR drift in both models. Full-reset and ungated variants are included only as diagnostic controls.

K.2 One-shot reset versus PAC

On Mistral-7B safety-only CISPO, we compare no intervention, one-shot m reset, and PAC with three seeds each. The online gate fires at step 6 in the intervention runs. After a one-shot reset, later live gradients can rebuild m, making the endpoint seed-sensitive; PAC instead removes m at every subsequent silent state.

K.3 Cross-model seedwise results

Qwen2.5-7B and Llama-3.1-8B use the same safety-only reward, harmful-only data, $K = 16$, and AdamW $(0.9, 0.999)$ settings as Mistral.

Qwen's no-intervention SCR is non-monotonic, but ORR increases substantially in both seeds. Llama begins with low SCR and high refusal. Despite these different absolute trajectories, PAC reduces ORR in all three model families.

L External validation of PAC

External graders show the same utility-side trend: PAC lowers over-refusal and raises Health-ORSC SCR, while some safety metrics worsen.

Table 18: Seed-wise 90-step results for the cross-model PAC runs. ORR_2 and SCR_2 are the values at step 2.

Condition	Seed	Gate	ORR_2	SCR_2	ASR	ORR	SCR	SCR_{min}
Qwen2.5-7B								
No intervention	1	-	.128	.962	.000	.988	.656	.504
	2	-	.112	.956	.000	.912	.576	.509
One-shot reset	1	11	.116	.953	.000	.716	.527	.464
	2	9	.108	.953	.000	.944	.364	.360
PAC	1	6	.120	.956	.000	.204	.842	.829
	2	17	.108	.947	.000	.468	.836	.829
Llama-3.1-8B								
No intervention	1	-	.112	.200	.000	.672	.000	.000
	2	-	.112	.216	.000	.804	.002	.002
One-shot reset	1	8	.108	.204	.000	.884	.020	.016
	2	8	.104	.207	.000	.960	.009	.007
PAC	1	7	.104	.209	.004	.260	.062	.056
	2	8	.108	.211	.002	.360	.082	.073

Table 19: External safety-grader results for PAC. Values are mean ± seed standard deviation.

Model	Intervention	HarmBench	StrongREJECT	SORRY-Bench
Mistral-7B	No intervention	0.000 ± 0.000	0.022 ± 0.013	0.002 ± 0.003
	PAC	0.035 ± 0.020	0.002 ± 0.002	0.037 ± 0.020
Qwen2.5-7B	No intervention	0.000 ± 0.000	0.009 ± 0.009	0.000 ± 0.000
	PAC	0.017 ± 0.018	0.021 ± 0.024	0.030 ± 0.021
Llama-3.1-8B	No intervention	0.000 ± 0.000	0.026 ± 0.036	0.000 ± 0.000
	PAC	0.010 ± 0.007	0.030 ± 0.004	0.005 ± 0.000

M KL Penalty and Exact Silence

Optimizer-side causal attribution requires a regime with exactly zero total current gradient. Holding all other settings fixed, we add only a KL term and test whether exact silence persists.

Even when $g_t^{\text{safety}} = 0$,

$$g_t^{\text{total}} = g_t^{\text{safety}} + \lambda_{\text{KL}} g_t^{\text{KL}} \neq 0,$$

so exact total-gradient silence no longer holds. In this regime, we do not interpret optimizer-state effects independently of the current gradient.

Table 20: External utility and safe-completion results for PAC. XSTest-grader applies the XSTest grader to OR-Bench-hard prompts; it is not the XSTest benchmark itself.

Model	Intervention	OR-Bench	XSTest-grader	Health-ORSC ORR	Health-ORSC SCR
Mistral-7B	No intervention	0.980 ± 0.033	0.964 ± 0.060	0.963 ± 0.059	0.013 ± 0.021
	PAC	0.952 ± 0.012	0.875 ± 0.023	0.631 ± 0.122	0.454 ± 0.020
Qwen2.5-7B	No intervention	1.000 ± 0.000	0.996 ± 0.002	0.999 ± 0.001	0.012 ± 0.005
	PAC	0.906 ± 0.104	0.770 ± 0.180	0.511 ± 0.288	0.520 ± 0.153
Llama-3.1-8B	No intervention	0.992 ± 0.011	0.992 ± 0.011	0.917 ± 0.093	0.075 ± 0.084
	PAC	0.863 ± 0.106	0.859 ± 0.111	0.597 ± 0.134	0.337 ± 0.098

Table 21: Effect of KL penalty on CISPO's exact total-gradient silence. 90 optimizer steps per run.

Condition	Seeds	Zero-grad steps	Silent blocks	Longest block
No KL	3	70 / 75 / 84	10 / 10 / 3	16 / 24 / 60
KL = 0.001	2	0 / 0	0 / 0	0 / 0