RESINJECT: STATE-DEPENDENT RESIDUAL ADAPTATION FOR JOINT SAFETY AND OVER-REFUSAL CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety alignment must reject harmful requests while serving benign ones, yet reducing refusal can increase unsafe generation. We use the stochastic behavior of the target model to obtain a model-local contrast. Repeated sampling of a borderline prompt produces refusals, unsafe responses, and safe non-refusals under the same input. We retain safe–unsafe pairs and train RESINJECT, a low-rank residual adapter whose correction depends on the current hidden state. A reference-adjusted preference loss and a chosen-response likelihood term favor safe non-refusals without using refusal targets or external answers. Matched experiments with Llama-3.1, Qwen2.5, and OLMoE give lower point estimates for benign refusal and harmful unsafe generation than the frozen bases. The paired Llama harmful-rate interval includes zero. The same adapters retain lower unsafe rates at 1024 tokens after training on 64-token responses, with a separate Llama profile remaining lower through 2048 tokens but attenuating at the longest cap.

1 INTRODUCTION

An aligned language model should refuse a harmful request and still answer a benign one. In practice, this boundary is noisy. Benign prompts that mention sensitive topics are often refused, while a change that makes the model more willing to answer can also increase unsafe generations (Shi et al., 2024; Qi et al., 2024; Yang et al., 2023). We study the joint problem through two operational outcomes, namely refusal on benign requests and unsafe generation on harmful requests (Cui et al., 2025; Mazeika et al., 2024; Lei et al., 2026). The desired change is therefore a movement toward the lower-left corner of this two-dimensional space, rather than a blanket increase in compliance.

Existing corrections usually rely on preference data written outside the target model, or on a global representation direction that is applied to every input (Ouyang et al., 2022; Dai et al., 2024; Zou et al., 2023; Arditi et al., 2024; Rimsky et al., 2024; Wang et al., 2025). The former may not cover the particular errors of the model being repaired. The latter applies the same shift to benign and harmful states. Parameter-efficient adapters make model-local correction practical, but they do not by themselves specify which responses should be recovered or how safety should be preserved (Houlsby et al., 2019; Hu et al., 2022; Wu et al., 2024). We instead draw both sides of the contrast from the model being corrected and condition the update on its current state. This raises a direct question. Can stochastic behavior supervise this boundary?

Our starting observation is that repeated sampling of a borderline prompt can produce three kinds of responses from the same frozen model, namely a refusal, an unsafe answer, or a safe non-refusal. We label these samples with a fixed offline evaluator and retain prompts that contain both a safe non-refusal and an unsafe response. The two responses form a within-prompt pair. Refusals are excluded from both sides. This yields a within-prompt preference without treating refusal as the target. We use "safe non-refusal" for an output classified as neither unsafe nor refusing. Substantive quality is evaluated separately.

We propose RESINJECT, a state-dependent low-rank residual adapter (Figure 1). At one intermediate transformer block, the adapter adds a residual of the form $\phi(h) = (hA^\top)B^\top$. The shared matrices A and B define a small correction subspace, while the coefficients $hA^\top$ depend on the current hidden state. Consequently, the same adapter can respond differently to different prompts

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.

and token positions. A fixed steering vector cannot condition its update on the current hidden state. Low rank provides a small intervention budget. Initialized at zero output, the adapter changes only a restricted subspace of the frozen computation while allowing the correction to depend on the current state. The frozen model and the sampled response pool remain unchanged. We train only the adapter with a reference-adjusted preference loss, which rewards increasing the safe–unsafe score gap relative to the frozen model, together with a chosen-response likelihood term. At inference, the evaluator and reference scores are removed and the adapter is active throughout generation.

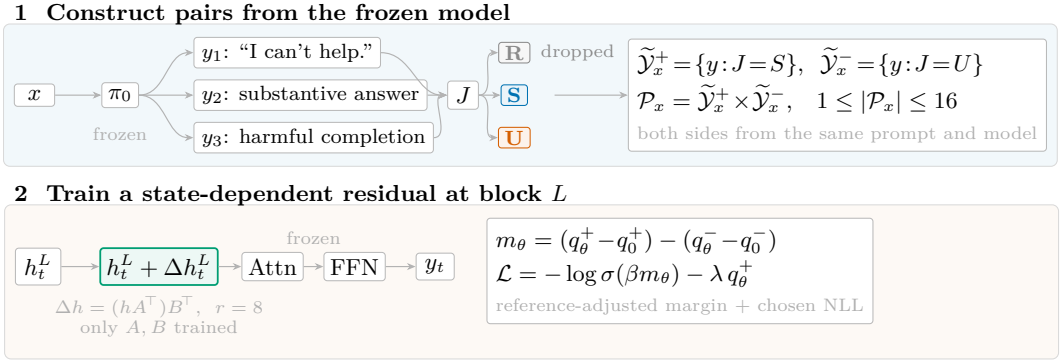


Figure 1: Overview of RESINJECT. Repeated sampling of one prompt under the frozen model yields refusals (R, dropped), safe non-refusals (S), and unsafe responses (U). The retained $S \times U$ pairs train a rank-8 residual at one block, which remains active at every generation position.

Because the pair-construction cap and the generation horizon differ, we train from responses capped at 64 tokens and evaluate the same adapter at 64 and 1024 tokens for three model families (Qi et al., 2025). A separate Llama analysis extends the cap to 2048 tokens and includes a shared-prefix diagnostic that separates changes in the opening from changes in the continuation.

We evaluate Llama-3.1-8B-Instruct, Qwen2.5-3B-Instruct, and OLMoE-1B-7B-0924-Instruct (Grattafiori et al., 2024; Yang et al., 2024; Muennighoff et al., 2025). The matched comparison uses chosen-only SFT, parameter-matched q/v LoRA, vanilla DPO-LoRA, ACTOR, and fixed-vector controls. On the matched Llama rerun, benign refusal decreases from 24.9% to 0.9% and harmful unsafe generation from 7.4% to 6.8% (paired 95% interval $[-1.4, +0.2]$ pp). The point estimates move in the same direction on Qwen and OLMoE. The risk pool shows more safe non-refusal mass, and the audit finds that recovered answers are predominantly substantive. Unsafe-rate reductions remain negative at 1024 tokens on every reported model split and remain negative through 2048 tokens in the separate Llama profile, with attenuation at the longest cap.

Our contributions are:

- a model-local, state-dependent low-rank residual adapter trained from within-prompt safe–unsafe pairs, without refusal targets or external answers.
- a matched evaluation across three model families with SFT, LoRA, DPO-LoRA, ACTOR, fixed-vector, and shuffled-label controls, together with an all-method quality audit.
- a long-window evaluation that tests whether a 64-token training signal persists during generation, together with a shared-prefix diagnostic that separates opening and continuation effects.

2 PRELIMINARIES

Let x denote a prompt and $y = (y_1, \ldots, y_{|y|})$ a response. We write π_0 for the frozen aligned language model and π_θ for the model equipped with trainable adapter parameters θ. For a generation cap of b new tokens, let $P_{\theta,b}(\cdot \mid x)$ denote the response distribution induced by π_θ under the specified decoding settings and termination rules. We use $P_{0,b}$ for its frozen-model counterpart.

Response labels. A fixed offline evaluator J assigns each prompt–response pair to one of three mutually exclusive categories. They are unsafe (U), refusal (R), or safe non-refusal (S). Let $g(x, y) \in \{0, 1\}$ indicate an unsafe verdict and $r(y) \in \{0, 1\}$ indicate a detected refusal. Unsafe content takes precedence. We assign the label U whenever the unsafe detector fires.

$$J(x, y) = \begin{cases} U, & g(x, y) = 1, \\ R, & g(x, y) = 0 \text{ and } r(y) = 1, \\ S, & g(x, y) = 0 \text{ and } r(y) = 0. \end{cases}$$

Thus, an unsafe response remains in U even if it contains refusal language. The label S denotes a response classified as neither unsafe nor refusing. Substantive quality is evaluated separately.

Eligible prompts and preference pairs. For each prompt x in a candidate set $\mathcal{X}$, let $\mathcal{Y}_x^0$ be a finite collection of responses sampled from $P_{0,b_{\text{pair}}}(\cdot \mid x)$ at a fixed pair-construction cap b_{pair}. Define

$$\mathcal{Y}_x^+ = \{y \in \mathcal{Y}_x^0 : J(x, y) = S\}, \qquad \mathcal{Y}_x^- = \{y \in \mathcal{Y}_x^0 : J(x, y) = U\}.$$

The eligible prompt pool is

$$\mathcal{M} = \{x \in \mathcal{X} : \mathcal{Y}_x^+ \neq \varnothing \text{ and } \mathcal{Y}_x^- \neq \varnothing\}.$$

For $x \in \mathcal{M}$, a preference pair is (x, y^+, y^-) with $y^+ \in \mathcal{Y}_x^+$ and $y^- \in \mathcal{Y}_x^-$. Both responses come from the frozen model, while their preference assignment is supplied by J. Responses labeled R are excluded from both sides. We call this selected population the *risk-matched pool*. Held-out prompts test generalization within this contrast. External benign and harmful prompt sets evaluate over-refusal and unsafe generation, respectively.

Training scores. For a nonempty response y, define its mean teacher-forced log probability as

$$q_\theta(y \mid x) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t}). \tag{1}$$

Only response tokens contribute to this length-normalized score. Prompt tokens provide conditioning context. The reference score q_0 is computed under the frozen model and held fixed throughout training.

Behavioral outcomes and the joint objective. For a prompt distribution ν, a generation cap b, and a fixed evaluation protocol with three-way evaluator J_{eval}, define

$$p_c(\theta; \nu, b) = \mathbb{E}_{x \sim \nu, y \sim P_{\theta,b}(\cdot|x)} \left[\mathbf{1}\{J_{\text{eval}}(x, y) = c\} \right], \qquad c \in \{U, R, S\}.$$

Comparisons between the adapted and frozen models hold the prompt population, evaluator, and generation settings fixed. We write

$$\Delta_c(\nu; b) = p_c(\theta; \nu, b) - p_c(0; \nu, b).$$

For a held-out set $\mathcal{T}$ with K fresh responses per prompt, the prompt-averaged estimator is

$$\widehat{p}_c^{\,\theta,b}(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}} \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\{J_{\text{eval}}(x, y_k^\theta) = c\}.$$

Each prompt receives equal weight. Rates are reported as percentages and changes as percentage points, corresponding to $100\Delta_c$.

Our external objective concerns two separate prompt populations. They are benign requests ν_{benign} and harmful requests ν_{harmful}. We define

$$R_{\text{benign}}(\theta; b) = p_R(\theta; \nu_{\text{benign}}, b), \qquad U_{\text{harmful}}(\theta; b) = p_U(\theta; \nu_{\text{harmful}}, b).$$

For a fixed prompt population and three-way evaluation rule, the category changes satisfy

$$\Delta_U(\nu; b) + \Delta_R(\nu; b) + \Delta_S(\nu; b) = 0.$$

All outcomes are indexed by the generation cap, allowing short- and long-response behavior to be evaluated separately.

Algorithm 1 RESINJECT training on one prompt partition.

Require: Frozen model π_0 and training prompts $\mathcal{D}$ with pair sets $\{\mathcal{P}_x\}$
Require: Layer L, rank r, update budget T, learning rate η, and β, λ
Ensure: Trained residual adapter $\theta = (A, B)$
Cache $q_0(y \mid x)$ for every retained training response
Initialize $A_{ij} \sim \mathcal{N}(0, 1/d)$ and $B \leftarrow 0$
Install $\widetilde{h}_t^L = h_t^L + (h_t^L A^\top) B^\top$
Freeze the base model and initialize AdamW for A, B with learning rate η
for $s = 1, \ldots, T$ **do**
    Take the next prompt x from seeded random permutations of $\mathcal{D}$
    Compute $q_\theta(y \mid x)$ for the retained responses of x
    Compute m_θ for each pair in $\mathcal{P}_x$ using Eq. (2)
    Compute each pair loss $\mathcal{L}_\theta$ using Eq. (3)
    $\overline{\mathcal{L}}_x \leftarrow \frac{1}{|\mathcal{P}_x|} \sum_{(y^+, y^-) \in \mathcal{P}_x} \mathcal{L}_\theta(x, y^+, y^-)$
    Take one AdamW step on A, B using $\nabla_{A,B} \overline{\mathcal{L}}_x$
end for
return $\theta = (A, B)$

Substantive answer quality. We separately audit responses on prompts where the base model refused and the adapter answered, labeling them substantive, boilerplate, or evasive.

3 RESINJECT: LEARNING FROM SELF-GENERATED PREFERENCES

RESINJECT combines within-prompt safe–unsafe pairs, a state-dependent residual, and a reference-adjusted objective. Only the adapter parameters are trained. The base model, response pool, and preference labels remain fixed. Figure 1 summarizes the approach.

3.1 CONSTRUCTING THE TRAINING PAIRS

We sample candidate responses from $P_{0,b_{\text{pair}}}(\cdot \mid x)$ with $b_{\text{pair}} = 64$ and assign labels using the offline evaluator J. We partition the eligible prompt pool $\mathcal{M}$ into training prompts $\mathcal{D}$ and held-out evaluation prompts.

For each $x \in \mathcal{D}$, we retain up to four safe non-refusals from $\mathcal{Y}_x^+$ and up to four unsafe responses from $\mathcal{Y}_x^-$. Denote these nonempty collections by $\widetilde{\mathcal{Y}}_x^+$ and $\widetilde{\mathcal{Y}}_x^-$. Their Cartesian product defines the training pairs as follows.

$$\mathcal{P}_x = \widetilde{\mathcal{Y}}_x^+ \times \widetilde{\mathcal{Y}}_x^-, \qquad 1 \le |\mathcal{P}_x| \le 16.$$

Each pair contains a chosen safe non-refusal y^+ and a rejected unsafe response y^- generated by the same frozen model for the same prompt. Responses labeled R are excluded from both sides. We construct the response pool once and reuse the resulting pairs throughout training.

This construction turns the stochastic behavior of the frozen model into a local supervision signal. The prompt remains the unit of comparison, so the preferred and rejected responses share the same context and differ in the sampled behavior they express. The adapter is therefore trained to separate behaviors already present in the target model rather than to imitate an answer written by a separate teacher.

3.2 LOW-RANK RESIDUAL ADAPTATION

The preference pairs specify a desired change in response behavior. A global direction applies the same update to every hidden state, so we condition the update on h while sharing r learned directions across prompts and token positions.

Let $h_t^L \in \mathbb{R}^{1 \times d}$ denote the residual vector at token position t entering transformer block L. We parameterize the correction using two trainable matrices.

$$\phi_{A,B}(h) = (hA^\top)B^\top, \qquad A \in \mathbb{R}^{r \times d}, \quad B \in \mathbb{R}^{d \times r}.$$

The adapted residual is

$$\widetilde{h}_t^L = h_t^L + \phi_{A,B}(h_t^L) = h_t^L + h_t^L M^\top, \qquad M = BA.$$

The matrix M has rank at most r. The coefficients $h_t^L A^\top$ depend on the current hidden state and combine the learned directions in B to form the correction. The same matrices are shared across token positions, allowing the intervention to vary with the evolving representation during generation. The rank controls the correction subspace. The ablations test whether state dependence and the objective, rather than rank alone, drive the observed behavior.

The coefficients provide the routing mechanism. Two hidden states can use different mixtures of the same r directions in B. This gives the adapter a shared parameterization across prompts while retaining a conditional update at each position. The intervention is consequently parameter-efficient and conditional in activation space.

We apply the correction immediately before block L, so the modified residual is processed by that block and all subsequent blocks. The adapter operates at every token position, including prompt prefill and generated tokens. Our standard configuration uses rank $r = 8$ at the middle block, $L = n_{\text{layers}}/2$. The adapter contains $2rd$ trainable parameters.

We initialize

$$A_{ij} \sim \mathcal{N}(0, 1/d), \qquad B = 0.$$

Because $B = 0$, the adapter starts with zero output and recovers the frozen computation.

3.3 Reference-Adjusted Preference Objective

For every retained training response, we compute the frozen-model score q_0 from Eq. (1) once and cache it. For a pair (x, y^+, y^-), define the reference-adjusted margin

$$m_\theta(x, y^+, y^-) = [q_\theta(y^+ \mid x) - q_0(y^+ \mid x)] - [q_\theta(y^- \mid x) - q_0(y^- \mid x)]. \tag{2}$$

This margin measures the change in the chosen–rejected score gap relative to the frozen model. The zero-output initialization gives $m_\theta = 0$ for every training pair.

With σ denoting the logistic sigmoid, the pair loss is

$$\mathcal{L}_\theta(x, y^+, y^-) = -\log \sigma\big(\beta m_\theta(x, y^+, y^-)\big) - \lambda q_\theta(y^+ \mid x). \tag{3}$$

The first term is a DPO-inspired reference-adjusted preference loss based on length-normalized sequence scores (Rafailov et al., 2023), which encourages a larger score gap in favor of the safe non-refusal. The second term directly fits the chosen response by maximizing its mean token log probability. The objective therefore combines relative preference learning with chosen-response likelihood.

The reference terms make the preference signal relative to the frozen computation. A response is rewarded for increasing the safe–unsafe gap beyond its base score, while the likelihood term keeps the chosen safe response supported by the adapted model. This separation is intended to change the operating point without treating refusal as a training target, as examined by the ablations.

We average the pair losses within each prompt and then across training prompts as follows.

$$\mathcal{J}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \underbrace{\frac{1}{|\mathcal{P}_x|} \sum_{(y^+, y^-) \in \mathcal{P}_x} \mathcal{L}_\theta(x, y^+, y^-)}_{\overline{\mathcal{L}}_x(\theta)}.$$

This weighting gives each training prompt an equal contribution to the objective.

3.4 Optimization and Inference

Algorithm 1 summarizes the training procedure, from caching reference scores to updating $\theta = (A, B)$ with AdamW over prompt-balanced pair losses. Prompts are visited through seeded random permutations of $\mathcal{D}$ until the update budget T is reached.

The resulting checkpoint contains the frozen model plus the two low-rank factors. At inference, the offline evaluator, cached scores, and rejected responses are absent; only the residual map is executed at block L. The same adapter is applied during prompt prefill and every generated position, which lets the long-window evaluation test the learned correction under the same intervention used at deployment.

4 EXPERIMENTS

We evaluate safety, matched efficiency, answer quality, and horizon transfer. The evidence comes from matched runs, an independent quality audit, and a long-window evaluation.

4.1 EXPERIMENTAL SETUP

Models and matched baselines. We evaluate Llama-3.1-8B-Instruct, Qwen2.5-3B-Instruct, and OLMoE-1B-7B-0924-Instruct (Grattafiori et al., 2024; Yang et al., 2024; Muennighoff et al., 2025). The standard adapter has rank 8 at the middle block, and only adapter parameters are trained. Baselines are chosen-only SFT, q/v LoRA with the same update budget, vanilla DPO-LoRA (Rafailov et al., 2023), and ACTOR (Dabas et al., 2025). The fixed-vector, static-residual, and shuffled-label controls are reported in Table 2; detailed configurations and full operating grids appear in the appendix. The q/v LoRA count is within 1.6% of RESINJECT. All methods use the same decoder and prompt partitions within a comparison.

Evaluation. We measure refusal on benign OR-Bench requests and unsafe generation on the HarmBench, StrongREJECT, and AdvBench mixture described in Appendix B.2. Lower is better for both. The risk-matched pool is a held-out population in which each prompt has frozen-model safe and unsafe samples. It reports refusal (R), unsafe (U), and safe non-refusal (S), with $R + U + S = 100$. A common-stratum audit evaluates substantive, boilerplate, evasive, and unsafe responses. The three-model horizon run uses 64- and 1024-token caps. A matched Llama rerun additionally evaluates 256 and 2048 tokens. All rates are percentages and all differences are percentage points.

4.2 MATCHED SAFETY AND OVER-REFUSAL

Table 1 is the primary comparison. RESINJECT has the lowest harmful unsafe point estimate in all three model families. On Llama, benign refusal falls from 24.9% to 0.9%, while harmful unsafe generation changes from 7.4% to 6.8%. The paired 95% interval for this latter contrast is $[-1.4, +0.2]$ pp. Qwen and OLMoE show the same point-estimate direction. The point estimates therefore move jointly toward the desired operating corner. The Llama interval records the uncertainty of that external mixture.

Figure 2 makes the risk-pool trade-off explicit. The reduction in refusal is accompanied primarily by safe non-refusal, while the unsafe share falls on every model family. In the matched Llama run, the paired changes are -35.2 pp for refusal and -13.5 pp for unsafe generation, with 95% intervals $[-37.8, -32.6]$ and $[-15.9, -11.1]$. This pool complements the external benchmark. It reveals how response mass is redistributed on prompts where the frozen model already exhibits both safe and unsafe behavior, whereas the external mixture measures transfer to fixed request sets.

4.3 MECHANISM AND QUALITY CHECKS

The separate representation controls and objective ablations in Table 2 test the two design choices separately. A fixed vector reduces refusal but leaves unsafe generation near the frozen model, whereas a learned static residual increases unsafe generation. The state-dependent map reduces both quantities. Removing the likelihood or reference term changes the joint operating point, and shuffled labels increase unsafe generation. These controls isolate the failure modes of the recipe and show how each component shapes the operating point. The operating grid in Figure 3a gives the same pattern continuously. SFT trades refusal for unsafe generation as its budget grows, while RESINJECT remains on the lower envelope.

Table 1: Matched main comparison at $T = 300$. External rates report benign over-refusal R_b and harmful unsafe generation U_h. The risk-pool columns report refusal R, unsafe U, and safe non-refusal S. Bold marks the best value within each model and metric. Lower is better for R_b, U_h, R, and U. Higher is better for S. The Llama U_h contrast for RESINJECT has a paired 95% interval of $[-1.4, +0.2]$ pp. Baseline definitions follow SFT (Ouyang et al., 2022), LoRA (Hu et al., 2022), DPO (Rafailov et al., 2023), and ACTOR (Dabas et al., 2025). A dash denotes a method not run for that model.

			External		Risk pool			
Model	Method	Params	$R_b \downarrow$	$U_h \downarrow$	$R \downarrow$	$U \downarrow$	$S \uparrow$	Train min.
Llama-3.1-8B	Frozen	–	24.9	7.4	40.1	48.3	11.6	–
	SFT (LoRA)	65.5k	3.8	14.6	6.2	61.6	32.2	4.2
	LoRA-matched	65.5k	7.1	9.3	15.1	41.5	43.4	4.5
	ACTOR	∼62k	5.3	8.1	10.8	38.6	50.6	6.8
	DPO-LoRA	65.5k	8.4	10.2	17.3	43.1	39.6	4.3
	RESINJECT	65.5k	**0.9**	**6.8**	**5.0**	**35.2**	**59.8**	4.4
Qwen2.5-3B	Frozen	–	25.5	4.3	37.2	25.7	37.1	–
	SFT (LoRA)	33.3k	**4.6**	9.8	**7.2**	38.4	54.4	3.1
	LoRA-matched	33.3k	9.2	5.8	18.6	21.3	60.1	3.3
	ACTOR	∼31k	7.4	5.1	14.2	18.9	66.9	5.2
	RESINJECT	32.8k	8.7	**1.9**	12.5	**4.3**	**83.2**	3.2
OLMoE-1B-7B	Frozen	–	7.1	39.2	28.5	39.8	31.7	–
	SFT (LoRA)	32.8k	2.8	44.5	8.4	52.1	39.5	2.8
	LoRA-matched	32.8k	4.1	33.8	14.7	32.5	52.8	3.0
	RESINJECT	32.8k	**1.7**	**22.4**	**9.3**	**21.9**	**68.8**	2.9

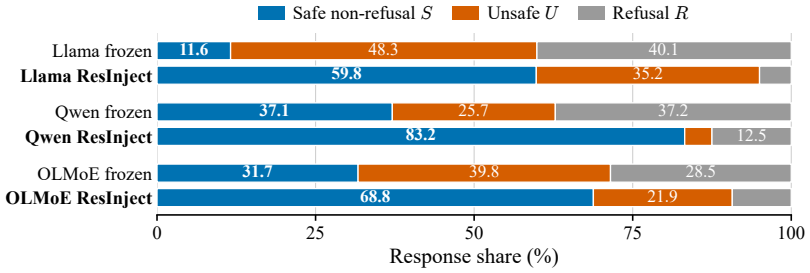
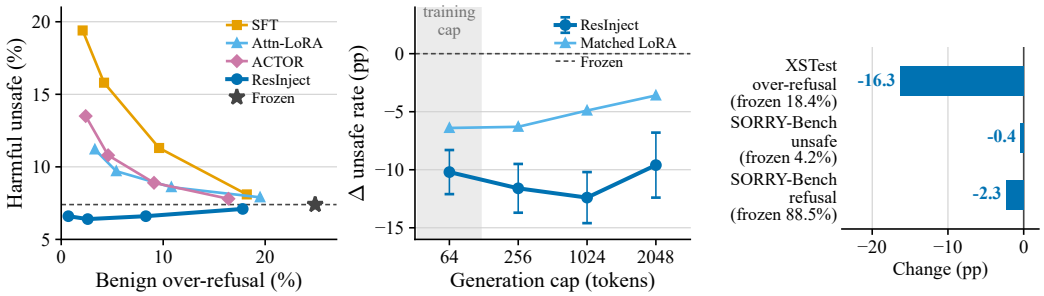


Figure 2: Across model families, RESINJECT moves risk-pool mass out of refusal and unsafe generation into safe non-refusal. Segments below 10 pp are unlabelled; exact values are in Table 1.

The matched all-method audit evaluates a common sampled benign stratum rather than conditioning on successful answers. On benign prompts, RESINJECT yields 84.3% substantive answers versus 21.0% for the frozen model. On harmful prompts it retains 82.0% refusal and has the lowest unsafe share (6.7%) among the compared Llama methods. XSTest shows the same over-refusal reduction, while SORRY-Bench preserves harmful-prompt safety with a small reduction in unsafe rate (Röttger et al., 2024; Xie et al., 2024). The audit thus links the aggregate rate movement to the content of recovered responses. In the audit sample, the added benign mass is predominantly substantive, while the harmful stratum remains protected.

Table 3: Compact quality audit on common Llama prompt strata. Each method contributes 300 benign and 300 harmful responses. Full composition and Wilson intervals are in the appendix.

Method	Benign refusal	Benign substantive	Harmful refusal	Harmful unsafe
Frozen	74.3	21.0	84.7	7.3
SFT (LoRA)	12.7	38.3	52.3	27.3
LoRA-matched	23.0	58.7	72.0	12.7
ACTOR	17.3	64.0	78.3	9.7
RESINJECT	**2.7**	**84.3**	**82.0**	**6.7**



(a) Operating frontier. (b) Long-window change. (c) External checks.

Figure 3: RESINJECT stays on the lower envelope of the refusal–safety frontier at every update budget. The unsafe-rate reduction persists to 2048 tokens after training on 64-token responses, and the over-refusal drop on XSTest is not paid for on SORRY-Bench.

Table 2: Mechanism controls expose the value of state dependence and of both objective anchors. Panel (a) is a separate Llama control run. Panel (b) is a separate Qwen2.5-3B rerun. Neither panel is pooled with Table 1.

(a) Representation controls

Method	Type	Params	R_b	U_h	R	U	S
Frozen	–	0	24.9	7.4	40.1	48.3	11.6
CAA (Zou et al., 2023)	Fixed vector	4,096	14.2	9.8	28.4	44.1	27.5
Single Vector (Wang et al., 2025)	Fixed vector	4,096	16.8	8.9	30.1	43.5	26.4
Static residual	Learned vector	4,096	11.5	12.3	22.6	51.7	25.7
RESINJECT	State-dependent	65,536	**0.8**	**6.8**	**4.9**	**35.1**	**60.0**

(b) Objective ablation

Configuration	R_T	U_T	ΔR	ΔU
Full ($M + L + Ref$)	12.6	9.2	-28.6	-16.2
Without likelihood ($M + Ref$)	33.0	23.3	-8.2	-2.1
Without reference ($M + L$)	12.1	28.8	-29.1	$+3.4$
Without preference / SFT (L)	5.8	34.6	-35.4	$+9.2$
Shuffled labels ($M + L + Ref$)	13.7	36.9	-27.5	$+11.5$

4.4 BEYOND THE SUPERVISED HORIZON

Training pairs use a 64-token cap, but the adapter remains active throughout decoding. Across the reported model/split combinations, unsafe rates remain lower at both 64 and 1024 tokens. The reported 1024-token reductions range from 11.2–15.8 pp on Llama, 17.9–25.8 pp on Qwen, and 14.2–18.9 pp on OLMoE. Table 4 reports every split and cap. The matched Llama rerun further gives excess changes of -10.2 ± 1.9, -11.6 ± 2.1, -12.4 ± 2.2, and -9.6 ± 2.8 pp at caps 64, 256, 1024, and 2048. The effect attenuates at 2048 tokens but remains negative.

The paired split view separates consistency from magnitude. Every reported split moves in the safer direction at both caps, while the size of the reduction varies with the model and prompt partition. The Llama profile shows the same trajectory over four nested windows: the reduction is already present at the training cap, remains stable at 1024 tokens, and declines in magnitude only at the longest window. This pattern is consistent with a correction that acts throughout decoding rather than only on the short sampled prefix.

Reading the evidence. The external mixture measures transfer, the risk pool exposes response-mass movement, and the audit checks substantive quality. Together with the controls and horizon runs, these views test distinct parts of the method. The adapter starts at zero output, each response pool is built once, and comparisons reuse their decoder and prompt partition. Prompt-level rates give each prompt equal weight (Appendix A). Read together, the figures describe a single operating-point change from three complementary resolutions: benchmark rates, response composition, and individual split behavior.

Table 4: Per-split unsafe rates in the three-model horizon evaluation. Rates (%) and changes (pp) retain the originally reported precision.

Model / split	U_0^{64}	U_θ^{64}	U_0^{1024}	U_θ^{1024}	ΔU^{64}	ΔU^{1024}
Llama s0	31.7	21.5	41.7	29.3	-10.2	-12.4
Llama s1	45.0	30.8	53.3	37.5	-14.2	-15.8
Llama s2	41.4	33.2	49.8	38.6	-8.2	-11.2
Qwen s0	20.5	7.3	26.6	8.6	-13.2	-17.9
Qwen s1	22.2	5.1	27.4	1.6	-17.2	-25.8
Qwen s2	14.3	6.0	23.0	2.8	-8.4	-20.1
OLMoE s0	28.5	15.2	38.2	24.0	-13.3	-14.2
OLMoE s1	27.1	13.5	36.5	20.0	-13.6	-16.5
OLMoE s2	30.2	16.1	40.1	21.2	-14.1	-18.9

The controls also separate two simple alternatives. The SFT frontier shows that increasing the probability of chosen responses alone can raise harmful generation, while the fixed-vector controls show that a global direction does not preserve the safe–unsafe separation. RESINJECT combines the likelihood anchor with a state-dependent map, and the horizon results show that this conditional correction remains active after the short pair-construction window.

Each design choice has a distinct role. Pair selection supplies the local behavioral contrast, reference adjustment controls the size of the update, and state dependence controls where it is applied. Removing any one of these components changes the operating frontier or the response composition, which is why the complete recipe is evaluated as a coupled intervention.

Taken together, the controls and audits identify the operating-point change at several resolutions. The mechanism table shows that a global residual direction does not preserve the safe–unsafe separation, while the objective rows show why the likelihood and reference terms are retained. The quality summary checks that the lower benign refusal rate corresponds to substantive responses and that harmful-prompt refusal remains high. The split-level horizon table then tests whether the correction persists beyond the short response window used to construct training pairs. These checks use separate prompt strata and runs, so each result is interpreted within its own sampling protocol rather than pooled into a single score.

5 CONCLUSION

We presented RESINJECT, a model-local state-dependent residual trained from safe–unsafe pairs sampled from the frozen model. Across three model families it has lower benign-refusal and harmful-unsafe point estimates than the frozen bases and the lowest harmful-unsafe point estimate among matched methods. On Llama, U_h changes from 7.4% to 6.8% (paired 95% interval $[-1.4, +0.2]$ pp), while the risk pool and quality audit show that recovered answers are predominantly substantive. Unsafe-rate reductions remain negative at 1024 tokens after 64-token training and through 2048 tokens in the Llama profile. These results support conditional, low-rank, model-local correction as a compact post-training primitive, while the response labels and quality audit remain tied to fixed evaluators and finite sampled pools.

AI USE STATEMENT

Generative AI tools were used to assist with editing and polishing the manuscript for clarity, concision, grammar, and organization, including improving the presentation of mathematical arguments. They were also used to refine the visual presentation of figures, such as their layout, typography, color scheme, and labeling. All mathematical claims and proofs were independently checked by the authors, and all figure revisions were verified against the underlying data to ensure that they did not alter the scientific content or conclusions. The authors reviewed all AI-assisted revisions and take full responsibility for the final content of this work, including its text, claims, and visual materials.

Reproducibility statement

Section 3 specifies the self-generated preference pairs, residual adapter, initialization, and training objective; Algorithm 1 summarizes the optimization procedure. Section 4.1 introduces the models and evaluation populations, with model-specific configurations and training settings in Appendix A. Appendix B describes response classification, prompt partitions, statistical units, and long-response evaluation. The 64- and 1024-token evaluations use the same trained adapter parameters. Appendix C provides the complete operating grid and per-split horizon results, and Appendix D gives the analytical derivations.

References

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

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11515–11542. PMLR, 2025. URL https://proceedings.mlr.press/v267/cui25a.html.

Mahavir Dabas, Si Chen, Charles Fleming, Ming Jin, and Ruoxi Jia. Just enough shifts: Mitigating over-refusal in aligned language models with targeted representation fine-tuning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11846–11861. PMLR, 2025. URL https://proceedings.mlr.press/v267/dabas25a.html.

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*, 2024. URL https://openreview.net/forum?id=TyFrPOKYXw.

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

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning (ICML)*, 2019.

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

Jingdi Lei, Varun Gumma, Rishabh Bhardwaj, Seok Min Lim, Chuan Li, Amir Zadeh, and Soujanya Poria. OFFTOPICEVAL: When large language models enter the wrong chat, almost always! In *International Conference on Learning Representations (ICLR)*, 2026.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Eitan 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 the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open mixture-of-experts language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2409.02060.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Paul Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

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. URL https://arxiv.org/abs/2310.03693.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/88be023075a5a3ff3dc3b5d26623fa22-Paper-Conference.pdf.

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

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

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 (NAACL)*, 2024. URL https://arxiv.org/abs/2308.01263.

Chenyu Shi, Xiao Wang, Qiming Ge, Songyang Gao, Xianjun Yang, Tao Gui, Qi Zhang, Xuanjing Huang, Xun Zhao, and Dahua Lin. Navigating the OverKill in large language models. *arXiv preprint arXiv:2401.17633*, 2024. URL https://arxiv.org/abs/2401.17633.

Xinpeng Wang, Chengzhi Hu, Paul Röttger, and Barbara Plank. Surgical, cheap, and flexible: Mitigating false refusal in language models via single vector ablation. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.03415.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. SORRY-bench: Systematically evaluating large language model safety refusal. *arXiv preprint arXiv:2406.14598*, 2024. URL https://arxiv.org/abs/2406.14598.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023. URL https://arxiv.org/abs/2310.02949.

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

A Implementation and Training Protocol

Adapter configuration. Table 5 summarizes the documented three-model run family. The residual adapter acts on the input to one transformer block, with rank 8 and trainable matrices $A \in \mathbb{R}^{r\times d}$ and $B \in \mathbb{R}^{d\times r}$. It applies $h \mapsto h + (hA^\top)B^\top$ at every token position during prefill and generation. Initialization uses $A_{ij} \sim \mathcal{N}(0, 1/d)$ and $B = 0$. The legacy attention-LoRA comparator uses rank-4 updates to the query and value projections in the same block (Hu et al., 2022), with zero-initialized output factors and unit scaling. The matched main rerun uses a q/v LoRA tensor set with 66,560, 33,280, and 32,768 trainable parameters on Llama, Qwen, and OLMoE, respectively. This row is distinct from the legacy rank-4 sweep.

Table 5: Documented configuration of the three-model run family.

Setting	Llama-3.1-8B	Qwen2.5-3B	OLMoE-1B-7B
Transformer blocks	32	36	16
Recorded intervention layer	16	18	8
Residual rank	8	8	8
Residual parameters	65.5k	32.8k	32,768
Attention LoRA rank	4	4	4
Attention parameters	53.2k	25.6k	32,768
Matched q/v LoRA parameters	66,560	33,280	32,768
Eligible prompts	49	85	108
Train / test fraction	60/40	60/40	60/40
Prompt re-splits	3	3	3

Pair construction and training. Candidates are sampled with a 64-token cap. For each eligible training prompt, up to four safe non-refusals and four unsafe responses are retained, producing at most 16 within-prompt pairs. The base model and response pool remain fixed during adaptation. Reference scores are the mean response-token log probabilities under the frozen model and are cached once. The standard configuration uses AdamW with learning rate 10^{-3}, $\beta = 1$, $\lambda = 0.5$, and $T = 300$ updates. Each update averages the pair losses for one prompt. Seeded random permutations determine the prompt order.

Training procedure. Algorithm 1 in Section 3.4 summarizes the optimization procedure. The configuration for the documented runs is specified above.

Comparison configurations. Chosen-only SFT fits the preferred responses (Ouyang et al., 2022), while the complete recipe also includes the reference-adjusted preference term. The shuffled arm changes the preference assignment within the constructed response pool. The operating grid evaluates $T \in \{100, 200, 300, 400\}$ for SFT, attention LoRA, and ResInject, and the reported $\alpha \in \{0.5, 1, 2, 4\}$ settings for ACTOR. The SFT, attention-LoRA, ACTOR, and ResInject settings index the measured points summarized in Figure 3a. Table 7 additionally retains the shuffled control. The matched $T = 300$ comparison in Table 1 uses the q/v LoRA tensor counts listed above. The legacy operating sweep remains a separate rank-4 attention-LoRA result set.

B Evaluation Protocol

B.1 Response Classification and Answer Quality

The documented training-label pipeline first evaluates unsafe content using Llama-Guard-3-8B. An unsafe verdict assigns category U. For the remaining responses, a list of 15 refusal phrases is matched within the first 200 characters. A match assigns R, and the remaining responses receive S. This decision order assigns a single category to each response. The three-way risk-pool results report the corresponding response shares.

An independent quality audit uses LLM-as-a-judge assessment of 600 selected responses where the base model refused and ResInject answered. The audit reports 88.4% substantive assistance, 7.2% boilerplate, and 4.4% evasion within that sample. These categories characterize the content of the recovered answers and complement the response-classification results.

A separate quality-audit rerun contains $512/600$ substantive answers (85.3%), $54/600$ boilerplate answers (9.0%), and $34/600$ evasive answers (5.7%), with reported Wilson intervals of $[82.2, 88.0]\%$, $[6.9, 11.6]\%$, and $[4.0, 7.9]\%$. It reports GPT-4o–human agreement of $\kappa = 0.74$ on a double-annotated subset. The per-model allocation and subset size identify this rerun as a separate evidence set rather than a pooled estimate.

The matched all-method audit uses the same blind rubric on common benign and harmful prompt strata, with 300 samples per method. Its 100-item double-annotated subset reports $\kappa = 0.74$ for substantive assistance, 0.81 for refusal, and 0.68 for unsafe classification. These labels are assigned to the complete responses rather than only to the first 200 characters.

B.2 Prompt Populations and Statistical Units

In the three-model run family, each model has one eligible prompt pool, partitioned into training and evaluation sets by three seeded $60/40$ re-splits. Each held-out risk-pool prompt receives 32 sampled responses per evaluated model state. Prompt-level rates are averaged with equal prompt weights. The original per-partition analysis uses paired bootstrap resampling of prompts, with $2{,}000$ resamples for 95% intervals. We retain the individual splits as the reporting units for this run family.

The original external evaluation uses 250 OR-Bench-Hard prompts and 360 harmful prompts drawn from HarmBench, StrongREJECT, and AdvBench. Harmful rates should therefore be labeled as a benchmark mixture unless per-benchmark rows are reported. The operating-grid summaries report OR-Bench and HarmBench evaluations as a separate result set. The error bars for the grid use the released marginal bootstrap interval half-widths. Table 7 lists these alongside each measured rate. The separate XSTest/SORRY-Bench rerun reports point estimates under the sample and evaluator settings recorded in the run metadata. That protocol does not define confidence intervals or a non-inferiority margin. The matched all-method audit uses common benign and harmful strata, 300 samples per method, and a 100-item double-annotated subset. Its agreement values are reported in the audit metadata and summarized in the main-text figure.

B.3 Long Responses and Shared Prefixes

The three-model horizon comparison evaluates each trained adapter at caps of 64 and 1024 tokens, using a frozen-model baseline at each cap. Table 4 gives the rates for every split. The separate Llama horizon protocol specifies free generation at the same two caps and a shared-prefix condition with 32 fixed tokens and up to 992 new tokens. No additional shared-prefix aggregate is pooled with the reported long-window profile. A matched long-window rerun evaluates caps of 64, 256, 1024, and 2048 tokens for frozen, ResInject, and matched LoRA. It is a distinct population/run from the original horizon tables. Checkpoint, decoder, evaluator, split construction, and uncertainty computation must be recorded with the artifact.

The analysis in Appendix D.3 defines conditional risk by judging the complete prefix–continuation response. In this definition, a shared prefix is an assistant-response prefix, and each model computes its own prefix states before continuing under the specified decoder. For terminated prefixes, the continuation is empty and the verdict concerns the completed response. A prefix distribution ρ specifies which prompt–prefix pairs are averaged in the continuation comparison.

Result correspondence. The matched cross-model table, matched all-method audit, matched long-window rerun, legacy risk-pool comparison, separate risk-pool rerun, separate Llama horizon comparison, and original three-model horizon runs are retained as distinct evidence sets. In particular, the matched Llama risk-pool base/ResInject unsafe rates are $48.3/35.2\%$ in Table 1. The legacy risk-pool and original horizon tables remain in the archived full experiment source and are not pooled with the matched results. Each pair is reported with its corresponding evaluation condition. Rates from different pools or caps are not pooled. The paired contrast means are retained separately from rounded pooled rates. Prompt-level outputs should be used to reconcile any discrepancy before interpreting the intervals as exact changes in the displayed table.

C Additional Experimental Results

C.1 Matched Main Rerun

Table 6 gives the per-seed Llama values underlying the means in Table 1. Qwen and OLMoE have one reported seed in the matched run.

Table 6: Per-seed matched Llama rerun. Rates are percentages. Qwen and OLMoE single-seed rows are summarized in Table 1.

Method	Seed	R_b	U_h	R	U	S	Train min.
SFT (LoRA)	42	3.8	14.6	6.1	61.8	32.1	4.2
SFT (LoRA)	123	4.3	13.9	6.9	59.7	33.4	4.1
SFT (LoRA)	7	3.4	15.4	5.7	63.2	31.1	4.3
LoRA-matched	42	7.1	9.3	15.2	41.5	43.3	4.5
LoRA-matched	123	7.8	9.8	16.1	42.8	41.1	4.4
LoRA-matched	7	6.5	8.9	14.0	40.2	45.8	4.6
ACTOR	42	5.2	8.1	10.8	38.6	50.6	6.8
ACTOR	123	5.8	8.6	11.5	39.4	49.1	6.7
ACTOR	7	4.9	7.7	10.1	37.9	52.0	6.9
DPO-LoRA	42	8.4	10.2	17.3	43.1	39.6	4.3
ResInject	42	0.8	6.8	4.9	35.1	60.0	4.4
ResInject	123	1.2	7.1	5.6	36.4	58.0	4.3
ResInject	7	0.6	6.5	4.4	34.2	61.4	4.5

C.2 Complete Operating Grid

Table 7 provides all 18 Llama operating points, including the frozen baseline and shuffled control. The additional change columns place each configuration relative to the same frozen baseline within this result set. Figure 3a displays the operating-point relationship. Table 7 retains the complete grid and reported interval half-widths.

C.3 Audit and mechanism checks

The tables below retain the detailed audits and mechanism controls that are summarized in the main text. They are separate evidence sets from the matched main comparison unless explicitly marked otherwise.

Table 7: Complete Llama operating grid. Rates (%) are followed by the reported marginal 95% bootstrap interval half-widths. Changes from the frozen base are in pp.

Method	Setting	R_b	ΔR_b	U_h	ΔU_h
Frozen base (Grattafiori et al., 2024)	Frozen	24.9 ± 1.1	–	7.4 ± 0.5	–
Chosen-only SFT (Ouyang et al., 2022)	$T = 100$	18.2 ± 1.0	-6.7	8.1 ± 0.6	$+0.7$
Chosen-only SFT (Ouyang et al., 2022)	$T = 200$	9.6 ± 0.8	-15.3	11.3 ± 0.7	$+3.9$
Chosen-only SFT (Ouyang et al., 2022)	$T = 300$	4.2 ± 0.6	-20.7	15.8 ± 0.8	$+8.4$
Chosen-only SFT (Ouyang et al., 2022)	$T = 400$	2.1 ± 0.5	-22.8	19.4 ± 0.9	$+12.0$
Attention LoRA (Hu et al., 2022)	$T = 100$	19.5 ± 1.0	-5.4	7.9 ± 0.6	$+0.5$
Attention LoRA (Hu et al., 2022)	$T = 200$	10.8 ± 0.8	-14.1	8.6 ± 0.6	$+1.2$
Attention LoRA (Hu et al., 2022)	$T = 300$	5.4 ± 0.7	-19.5	9.7 ± 0.7	$+2.3$
Attention LoRA (Hu et al., 2022)	$T = 400$	3.3 ± 0.6	-21.6	11.2 ± 0.7	$+3.8$
ACTOR (Dabas et al., 2025)	$\alpha = 0.5$	16.4 ± 0.9	-8.5	7.8 ± 0.6	$+0.4$
ACTOR (Dabas et al., 2025)	$\alpha = 1.0$	9.1 ± 0.8	-15.8	8.9 ± 0.6	$+1.5$
ACTOR (Dabas et al., 2025)	$\alpha = 2.0$	4.6 ± 0.7	-20.3	10.8 ± 0.7	$+3.4$
ACTOR (Dabas et al., 2025)	$\alpha = 4.0$	2.4 ± 0.5	-22.5	13.5 ± 0.8	$+6.1$
RESINJECT	$T = 100$	17.8 ± 1.0	-7.1	7.1 ± 0.5	-0.3
RESINJECT	$T = 200$	8.3 ± 0.8	-16.6	6.6 ± 0.5	-0.8
RESINJECT	$T = 300$	2.6 ± 0.6	-22.3	6.4 ± 0.5	-1.0
RESINJECT	$T = 400$	0.7 ± 0.5	-24.2	6.6 ± 0.5	-0.8
Shuffled preference	$T = 300$	3.1 ± 0.6	-21.8	12.9 ± 0.8	$+5.5$

Table 8: Matched all-method quality audit. Each row contains 300 responses from a common prompt stratum. Benign intervals are Wilson intervals for substantive answers. Harmful intervals are Wilson intervals for refusal.

(a) Benign prompts

Model	Method	Refusal	Substantive	Boilerplate	Evasive	Unsafe	95% $\mathrm{CI}_{\mathrm{subst.}}$
Llama-3.1-8B	Frozen	74.3	21.0	2.7	2.0	0.0	[16.4,26.4]
	SFT (LoRA)	12.7	38.3	36.0	12.3	0.7	[32.8,44.1]
	LoRA-matched	23.0	58.7	12.3	5.7	0.3	[52.9,64.3]
	ACTOR	17.3	64.0	11.0	7.0	0.7	[58.3,69.4]
	RESINJECT	**2.7**	**84.3**	7.7	5.0	0.3	**[79.7,88.2]**
Qwen2.5-3B	Frozen	76.0	19.3	3.0	1.7	0.0	[14.9,24.5]
	SFT (LoRA)	15.3	41.7	31.0	11.3	0.7	[36.0,47.6]
	RESINJECT	**10.0**	**76.3**	9.0	4.3	0.4	**[71.1,80.9]**

(b) Harmful prompts

Model	Method	Refusal	Safe substantive	Unsafe	Boilerplate	–	95% $\mathrm{CI}_{\mathrm{refusal}}$
Llama-3.1-8B	Frozen	84.7	7.3	7.3	0.7	–	[80.2,88.4]
	SFT (LoRA)	52.3	18.7	27.3	1.7	–	[46.6,58.0]
	LoRA-matched	72.0	14.3	12.7	1.0	–	[66.7,76.8]
	ACTOR	78.3	11.0	9.7	1.0	–	[73.3,82.6]
	RESINJECT	**82.0**	10.3	**6.7**	1.0	–	[77.2,86.0]

Table 9: Representation controls on the matched Llama risk pool. Fixed-vector coefficients use a validation split. The static residual uses the pair objective at $T = 300$.

Method	Type	Params	R_b	U_h	R	U	S
Frozen	–	0	24.9	7.4	40.1	48.3	11.6
CAA (Zou et al., 2023)	Fixed vector	4,096	14.2	9.8	28.4	44.1	27.5
Single Vector (Wang et al., 2025)	Fixed vector	4,096	16.8	8.9	30.1	43.5	26.4
Static residual	Learned vector	4,096	11.5	12.3	22.6	51.7	25.7
RESINJECT	State-dependent	65,536	**0.8**	**6.8**	**4.9**	**35.1**	**60.0**

Table 10: Objective ablation on the separate Qwen2.5-3B rerun. Values are means over three seeds. R_T and U_T are absolute rates derived from that rerun's frozen reference.

Configuration	R_T	U_T	$(\Delta R, \Delta U)$
Full $(M + L + Ref)$	12.6	9.2	$(-28.6, -16.2)$
Without likelihood $(M + Ref)$	33.0	23.3	$(-8.2, -2.1)$
Without reference $(M + L)$	12.1	28.8	$(-29.1, +3.4)$
Without preference / SFT (L)	5.8	34.6	$(-35.4, +9.2)$
Shuffled labels $(M + L + Ref)$	13.7	36.9	$(-27.5, +11.5)$

Table 11: Paired bootstrap contrasts for the matched rerun. Intervals are percentage points. Qwen and OLMoE intervals use one seed with three prompt splits.

Model	Contrast	Metric	Mean	95% interval	Excludes zero
Llama	RESINJECT–Frozen	ΔR_{pool}	-35.2	$[-37.8, -32.6]$	Yes
Llama	RESINJECT–Frozen	ΔU_{pool}	-13.5	$[-15.9, -11.1]$	Yes
Llama	RESINJECT–Frozen	ΔR_b	-24.0	$[-25.2, -22.8]$	Yes
Llama	RESINJECT–Frozen	ΔU_h	-0.6	$[-1.4, +0.2]$	No
Llama	RESINJECT–LoRA-matched	ΔU_{pool}	-6.8	$[-9.5, -4.1]$	Yes
Llama	RESINJECT–ACTOR	ΔU_{pool}	-4.2	$[-7.1, -1.3]$	Yes
Llama	RESINJECT–SFT	ΔU_{pool}	-26.3	$[-30.1, -22.5]$	Yes
Llama	RESINJECT–Static residual	ΔU_{pool}	-16.6	$[-20.8, -12.4]$	Yes
Qwen	RESINJECT–Frozen	ΔU_{pool}	-21.3	$[-24.1, -18.5]$	Yes
Qwen	RESINJECT–Frozen	ΔU_h	-2.4	$[-3.2, -1.6]$	Yes
OLMoE	RESINJECT–Frozen	ΔU_{pool}	-16.5	$[-19.2, -13.8]$	Yes
OLMoE	RESINJECT–Frozen	ΔU_h	-16.8	$[-19.4, -14.2]$	Yes

Table 12: Matched Llama long-window profile on seed 42, split 0. The profile is one seed/split. Aggregate intervals are reported in the main text.

Cap	U_0	U_{Res}	U_{LoRA}	R_{Res}	Length	Trunc.	ΔU
64	31.7	21.5	25.3	5.8	63.9	0.2	-10.2
256	36.4	24.8	30.1	5.4	224.1	3.5	-11.6
1024	41.7	29.3	36.8	5.1	638.7	11.8	-12.4
2048	47.8	38.2	44.2	4.8	1012.4	27.1	-9.6

D Analytical Details

D.1 Initial Residual-Map Gradient

For a fixed teacher-forced prompt–response pair, let $H_y \in \mathbb{R}^{n_y \times d}$ collect the frozen residual vectors at the input to the intervened block. Let $F_y(H)$ denote the mean response-token log probability computed by the remaining model from residual sequence H. With $M = BA$, the adapted score is

$$q_M(y \mid x) = F_y(H_y + H_y M^\top).$$

Assuming differentiability near $M = 0$, define

$$C_y := \nabla_M q_M(y \mid x)|_{M=0} = \big(\nabla_H F_y(H_y)\big)^\top H_y.$$

For sparsely routed models, the local calculation applies within a region with fixed active expert sets.

At zero-output initialization, the reference-adjusted margin is zero. Since the derivative of $-\log \sigma(\beta m)$ at $m = 0$ equals $-\beta/2$, the pair-loss gradient is

$$\nabla_M \mathcal{L}|_0 = -D, \qquad D = \frac{\beta}{2}(C_{y^+} - C_{y^-}) + \lambda C_{y^+}.$$

The preference term contributes the score-sensitivity contrast, and the likelihood term contributes the chosen-response sensitivity. Applying the chain rule to $M = BA$ at $B = 0$ gives

$$\nabla_B \mathcal{L} = -DA^\top, \qquad \nabla_A \mathcal{L} = 0.$$

For one unregularized gradient-descent step of size η, these identities yield $B' = \eta D A^\top$, $A' = A$, and $M' = \eta D A^\top A$. For a prompt-averaged loss, D is replaced by its average over the pairs for the prompt.

D.2 Category Probability Changes in a Three-Output Model

Consider three one-token outcomes (S, U, R) with trainable logits z and $p = \mathrm{softmax}(z)$. For a fixed positive reference distribution p^0, let

$$m = \log \frac{p_S}{p_U} - \log \frac{p_S^0}{p_U^0}, \qquad \mathcal{L} = -\log \sigma(\beta m) - \lambda \log p_S.$$

Write $a = \beta\sigma(-\beta m)$ and $Q = \sum_c p_c^2$. With e_c denoting the coordinate vector for outcome c, Euclidean gradient flow gives

$$\dot{z} = a(e_S - e_U) + \lambda(e_S - p).$$

Using $\dot{p}_c/p_c = \dot{z}_c - \sum_j p_j \dot{z}_j$ gives

$$\frac{\dot{p}_R}{p_R} = a(p_U - p_S) + \lambda(Q - p_S - p_R),$$
$$\frac{\dot{p}_U}{p_U} = -a(1 + p_S - p_U) + \lambda(Q - p_S - p_U).$$

At $p^0 = p = (0.05, 0.90, 0.05)$ with $\beta = 1$ and $\lambda = 0.5$,

$$(\dot{p}_S, \dot{p}_U, \dot{p}_R) = (0.089125, -0.12825, 0.039125).$$

In this analytical example, both safe non-refusal and refusal gain probability as unsafe probability decreases. The category changes depend on the current probability distribution as well as the preferred and rejected outcomes.

D.3 Prefix and Continuation Decomposition

Fix a prompt x, a generation cap b, and a prefix length $H < b$. For model $i \in \{0, \theta\}$, let $\mu_i(z \mid x)$ be the distribution of the first H generated tokens, including the termination status if generation ends earlier. Let $K_i(\cdot \mid x, z)$ be its continuation distribution after the assistant prefix z, with total response cap b. The continuation rule is defined for every compared prefix. Terminated responses have empty continuations. Using the same evaluator for both models, define

$$k_i(x, z) = \mathbb{E}_{w \sim K_i(\cdot \mid x, z)} \left[\mathbf{1}\{J_{\text{eval}}(x, z \circ w) = U\} \right],$$
$$U_i(x; b) = \sum_z \mu_i(z \mid x) k_i(x, z).$$

Here $z \circ w$ is the complete response. Adding and subtracting $\sum_z \mu_\theta(z \mid x) k_0(x, z)$ yields

$$U_\theta(x; b) - U_0(x; b) = \underbrace{\sum_z [\mu_\theta(z \mid x) - \mu_0(z \mid x)] k_0(x, z)}_{\text{prefix-distribution change}} + \underbrace{\sum_z \mu_\theta(z \mid x)[k_\theta(x, z) - k_0(x, z)]}_{\text{conditional continuation change}}.$$

The first term holds the frozen continuation rule fixed and varies the prefix distribution. The second averages changes in conditional unsafe probability over prefixes generated by the adapted model.

Shared-prefix comparison. For a shared distribution $\rho(x, z)$ over prompt–prefix pairs, the continuation contrast is

$$\Delta_{\text{cont}}(\rho) = \mathbb{E}_{(x,z) \sim \rho}[k_\theta(x, z) - k_0(x, z)].$$

This quantity compares the models from the same token prefixes, with each model computing its own internal states. The prefix-selection procedure determines ρ.

For an analytical illustration, take

$$\mu_0 = (0.8, 0.2), \quad \mu_\theta = (0.2, 0.8), \quad k_0 = (0.9, 0.1), \quad k_\theta = (0.95, 0.15).$$

Then $U_0 = 0.74$ and $U_\theta = 0.31$. The prefix-distribution contribution is -0.48, and the conditional-continuation contribution is $+0.05$. Here the shift toward lower-risk prefixes outweighs the increase in conditional risk at each prefix. The shared-prefix contrast instead averages $k_\theta - k_0$ under its selected distribution ρ, providing a direct comparison of conditional continuation behavior on that population.