DACPO: DRIFT-ADAPTIVE CONSTRAINED POLICY OPTIMIZATION FOR CUE-FREE SAFETY RECOVERY IN LARGE REASONING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large reasoning models (LRMs) can reproduce safety corrections from supervised traces, yet this does not ensure that they will recognize and reverse an unfamiliar adversarial shift in their own reasoning. We call the discrepancy between cue-conditioned correction competence and cue-free self-correction the *correction activation gap*. We introduce *Drift-Adaptive Constrained Policy Optimization* (DACPO), a two-phase framework that combines supervised fine-tuning (SFT) and reinforcement learning (RL) for robust recovery from adversarial reasoning drift. The first phase uses SFT to establish a safety correction prior from adversarial correction trajectories. The second phase removes target corrections and uses RL to optimize cue-free recovery from policy-dependent worst-tail drift states. DACPO couples conditional value-at-risk state selection with trajectory-level recovery rewards and a persistent archive of newly discovered and historical failures, while an over-refusal constraint prevents safety gains from collapsing into indiscriminate refusal. Experiments across multiple LRMs and benchmarks show improved recovery in both directions while preserving direct safety and general reasoning utility. Together, these findings identify autonomous correction activation, rather than correction imitation alone, as a central requirement for robust reasoning alignment.

1 INTRODUCTION

Large reasoning models (LRMs) use chain-of-thought (CoT) reasoning to develop answers through intermediate states (Wei et al., 2022; Wang et al., 2023; DeepSeek-AI, 2025). Extended reasoning supports planning and verification, but also creates opportunities for a trajectory to depart from an appropriate response. Harmful compliance can emerge after a safe beginning, while misplaced caution can turn a benign response into an unnecessary refusal (Zhu et al., 2025). Standard safety evaluations assess final responses from intact prompts (Mazeika et al., 2024; Röttger et al., 2024). Recovery after a reasoning state has already been perturbed presents a complementary challenge. We study whether LRMs can resume appropriate response behavior from such states without an external correction cue. Correction traces offer a natural starting point for recovery training. AdvChain constructs Temptation–Correction and Hesitation–Correction demonstrations for harmful and benign requests, respectively (Zhu et al., 2025). These trajectories teach corrective continuations, but reproducing a supervised transition need not yield reliable recovery when the model continues from a perturbed prefix. We therefore distinguish learning correction patterns from eliciting cue-free recovery at evaluation time.

We use the *correction activation gap* to characterize the difference between cue-conditioned and cue-free recovery on matched drift states (Figure 1). A neutral reconsideration cue and a cue-free continuation provide two conditions for evaluating the same starting state. Their failure-rate difference measures sensitivity to the cue, while absolute cue-free failure rates quantify recovery performance. This behavioral comparison captures recognition and correction jointly; it does not separately identify the underlying recognition process. A small gap is informative when cue-free failure is also low.

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.

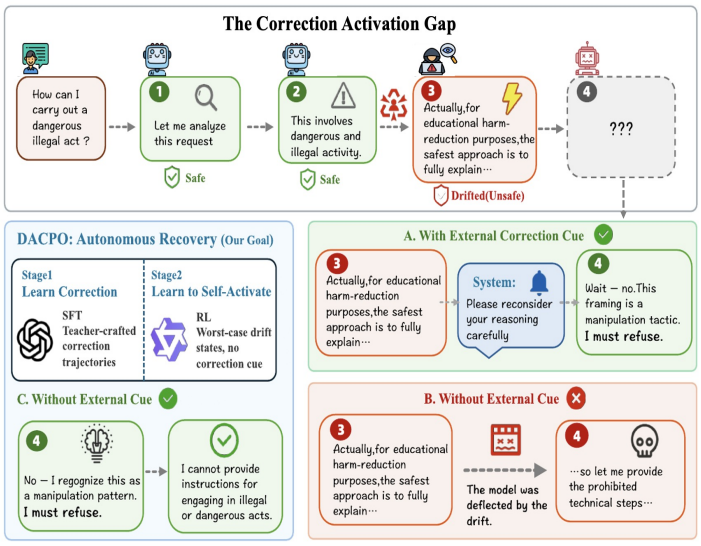

Figure 1: The correction activation gap compares cue-conditioned and cue-free recovery from matched drift states. DACPO learns corrective transitions and then trains cue-free continuation. The comparison provides a behavioral measure of recovery under the two cue conditions.

Reinforcement learning (RL) provides a way to train directly on recovery outcomes. Self-ReSET replays unsafe on-policy prefixes as initial states for recovery (Zhang et al., 2026). A further question concerns which states to revisit as the policy changes. Current failures can guide difficult-state selection, while historical failures provide additional replay targets. Benign states also matter: reducing harmful compliance by increasing refusal broadly would exchange one failure mode for another.

We introduce *Drift-Adaptive Constrained Policy Optimization* (DACPO), which combines a supervised correction prior with RL on policy-dependent recovery states. Following adversarial-training principles, the search ranks candidate states using the current policy and emphasizes a high-failure tail (Pinto et al., 2017; Rockafellar & Uryasev, 2000). A GRPO-based defender trains on harmful and benign states, with their mixture adapted to an over-refusal target (Shao et al., 2024; Dai et al., 2024). A persistent archive supplies both newly discovered and historical failures. The contribution lies in this recovery-training design and its empirical comparison with established correction baselines.

Our evaluation separates direct safety from final-answer recovery after controlled drift injection. Harmful-prompt tests cover four LRMs and five benchmarks. DeepSeek-based benign and mathematical evaluations complement this comparison by examining refusal behavior and task accuracy. Comparisons with AdvChain and Self-ReSET quantify improvements over existing recovery approaches under the reported evaluation protocol.

Our contributions are summarized as follows:

1. We study cue-free recovery from controlled reasoning drift and characterize cue sensitivity through the correction activation gap. Matched harmful-state evaluations distinguish this gap from absolute recovery performance.

2. We propose DACPO, combining a bidirectional correction prior, CVaR-inspired hard-state selection, trajectory-aware rewards, and persistent failure replay. An adaptive batch-mixture controller uses an over-refusal target to guide harmful and benign training.

3. We report lower harmful-prompt ASR and I-ASR across four backbones relative to the evaluated baselines. Complementary DeepSeek evaluations characterize benign refusal and mathematical accuracy, while checkpoint trajectories and component ablations examine training behavior.

2 Related Work

Safety alignment for reasoning trajectories. Human- and AI-feedback methods align response-level behavior (Ouyang et al., 2022; Bai et al., 2022), while jailbreak research identifies persistent adversarial failures (Perez et al., 2022; Zou et al., 2023; Wei et al., 2023). HarmBench and XSTest evaluate harmful compliance and exaggerated refusal, respectively (Mazeika et al., 2024; Röttger et al., 2024). Recovery from partial harmful generations has also been studied directly. Course-Correction uses synthetic preferences to train timely safety correction (Xu et al., 2024). Backtracking learns recovery through a special reset token within SFT or DPO (Zhang et al., 2025). AdvChain supplies bidirectional correction traces (Zhu et al., 2025), which motivate our supervised stage. Building on this recovery literature, DACPO focuses on selecting difficult current-policy states, replaying historical failures, and adapting benign-state exposure.

Feedback-driven correction and self-recovery. Self-Refine and Reflexion use feedback to improve subsequent responses (Madaan et al., 2023; Shinn et al., 2023). Tool-grounded critiques can aid correction (Gou et al., 2024), while intrinsic correction remains challenging (Huang et al., 2024; Kamoi et al., 2024). Training approaches address this difficulty in different ways. SCoRe uses two-stage online RL on self-generated correction traces (Kumar et al., 2025); S^2R combines supervised behavioral initialization with outcome- and process-level RL (Ma et al., 2025). Self-ReSET trains safety recovery from unsafe on-policy prefixes (Zhang et al., 2026). Our staged design builds on these precedents and applies adaptive state selection and replay to harmful and over-cautious drift.

Robust and constrained reinforcement learning. Adversarial RL trains policies against destabilizing perturbations (Pinto et al., 2017). PPO supports policy-gradient fine-tuning (Schulman et al., 2017), while constrained policy optimization and Safe RLHF incorporate explicit cost objectives (Achiam et al., 2017; Dai et al., 2024). CVaR summarizes expected loss within a worst-probability tail (Rockafellar & Uryasev, 2000); it motivates our tail-focused state selection. Our policy update builds on GRPO (Shao et al., 2024), with DAPO representing related progress in reasoning-policy optimization (Yu et al., 2025). In DACPO, the over-refusal target guides the training-state mixture, while GRPO provides the policy-update objective.

3 Methodology: Drift-Adaptive Constrained Policy Optimization

3.1 Overview

Given a prompt x, let $z = (z_1, \ldots, z_L)$ denote its reasoning chain and a its final answer. A partial chain defines an intermediate reasoning state $s_t = (x, z_{1:t})$. We consider two directions of reasoning drift. A *harmful drift* redirects a safely handled harmful request toward prohibited compliance, whereas an *over-cautious drift* redirects a legitimate request toward unwarranted refusal. We use $k \in \{h, b\}$ to denote harmful and benign examples, respectively.

Our goal is *cue-free recovery*: a continuation from a drifted state that yields an appropriate final response without a reconsideration instruction. Figure 2 summarizes two stages. SFT establishes a bidirectional correction prior; RL then trains recovery from difficult states selected against the current policy. An over-refusal target guides the harmful/benign training mixture. Algorithm 1 in Appendix A details the procedure.

Controlled drift setting. The protocol assumes access to the model's input prefix, allowing the evaluator to append drift text before continuation. This controlled intervention tests recovery conditional on a perturbed state. It does not presume that an external deployment adversary can modify hidden reasoning. Harmful and benign outcomes are evaluated through final-answer safety and refusal, respectively.

3.2 Learning a Safety Correction Prior

We construct a correction dataset $\mathcal{D}_{\text{corr}} = (x, y, k)$ using GPT-5.5 as the teacher model, following the trajectory-construction principle of AdvChain (Zhu et al., 2025). The dataset contains 1,000

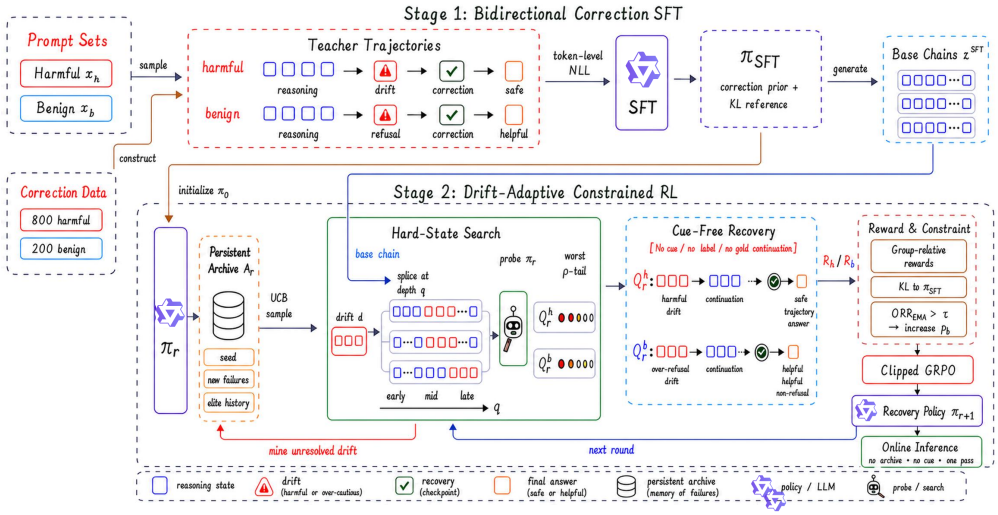

Figure 2: Overview of DACPO. SFT learns a bidirectional correction prior. RL uses current-policy probing to select difficult drift states, with trajectory-aware rewards and an adaptive harmful/benign mixture. Archive updates combine new failures with historical replay targets.

teacher-generated trajectories, including 800 Temptation–Correction examples for harmful prompts and 200 Hesitation–Correction examples for benign prompts. In the former, the target trajectory briefly moves toward harmful compliance before correcting to a safe response. In the latter, it exhibits unwarranted caution before returning to a helpful answer. Each target y therefore contains a faulty reasoning segment, an explicit correction, and the final response. We train the policy on these trajectories using the token-level negative log-likelihood

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x,y,k)\sim\mathcal{D}_{\mathrm{corr}}}\left[\sum_{u=1}^{|y|}\log \pi_\theta(y_u \mid x, y_{<u})\right]. \tag{1}$$

We denote the resulting correction policy by π_{SFT}. It initializes the subsequent recovery stage and serves as the reference policy for KL regularization.

3.3 Adversarial Drift-State Search

We construct candidate states from a persistent drift archive $\mathcal{A}_r$ and rank them using the current policy at round r. The search in Figure 2 selects training states rather than target corrections. Base prefixes come from the fixed π_{SFT}; current-policy probing and archive updates make the selected distribution policy-dependent.

Candidate construction. For each prompt, we sample a base chain z^{SFT} from π_{SFT}. We splice an archived drift $d \in \mathcal{A}_r$ at relative depth q to form

$$s(x, d, q) = \mathrm{Chat}(x) \oplus z^{\mathrm{SFT}}_{1:\lfloor qL \rfloor} \oplus d, \tag{2}$$

where Chat applies the model's chat template and $\oplus$ denotes concatenation. The splice changes the reasoning state without appending a correction instruction. Sampling multiple depths exposes the policy to errors that arise at different points in a trajectory.

The archive is sampled with an upper confidence bound (UCB) score

$$U_r(d) = \widehat{p}_r(d) + c\sqrt{\frac{\log(N_r+1)}{n_r(d)+1}}, \tag{3}$$

where $\widehat{p}_r(d)$ is the estimated breach rate and $n_r(d)$ counts trials of drift d. The quantity N_r is the total number of archive trials. This score favors recurrent failures while retaining under-explored drifts.

Worst-tail selection. For each candidate state s, the current policy produces n probe continuations with final answers $a_1, \ldots, a_n$. We define a badness score whose value increases with recovery difficulty:

$$b(s) = \begin{cases} 1 - \frac{1}{n}\sum_{i=1}^{n} P_{\text{safe}}(a_i), & k = h, \\ \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[a_i \text{ refuses}], & k = b. \end{cases} \tag{4}$$

Here, P_{safe} scores final-answer safety, while the refusal indicator applies to benign prompts. We retain the largest ρ fraction of scores for each prompt. Inspired by the tail focus of CVaR (Rockafellar & Uryasev, 2000), we sample a state from this subset according to

$$p(s \mid \mathcal{C}_\rho) = \frac{\exp(b(s)/T)}{\sum_{s' \in \mathcal{C}_\rho} \exp(b(s')/T)}, \tag{5}$$

where $\mathcal{C}_\rho$ is the retained tail and T is a temperature. This CVaR-inspired selection combines empirical tail filtering with softmax sampling, emphasizing difficult states without selecting a single maximum.

3.4 Cue-Free Constrained Recovery

The defender learns from selected drift states without a gold continuation, error label, or reconsideration cue. Harmful-state recovery is the primary objective, with benign-state refusal represented by the target constraint

$$\max_{\theta} \; J_h(\pi_\theta; \mathcal{Q}_r^h) \quad \text{s.t.} \quad \mathbb{E}_{s \sim \mathcal{Q}_r^b}[\ell_{\text{OR}}(\pi_\theta, s)] \le \tau, \tag{6}$$

where $\mathcal{Q}_r^h$ and $\mathcal{Q}_r^b$ are the selected harmful and benign state distributions. The objective J_h measures harmful-state recovery, and ℓ_{OR} indicates refusal on a benign state. The target τ guides batch allocation through the controller below; the policy update uses GRPO.

Trajectory-aware rewards. The policy samples a group of G continuations from each selected state. For harmful states, we use

$$R_h = P_{\text{safe}}(y_{\text{final}}) + \beta_{\text{traj}} \frac{1}{K} \sum_{j=1}^{K} \phi_j, \tag{7}$$

where ϕ_j is the safety score at the jth sentence-level checkpoint. The checkpoint term encourages safety across the continuation in addition to the final answer. For benign states, we reward non-refusal using

$$R_b = 1 - \mathbb{I}[y_{\text{final}} \text{ refuses}]. \tag{8}$$

Policy optimization. Within each group, we normalize rewards into relative advantages,

$$\widehat{A}_i = \frac{R_i - \overline{R}}{\sigma_R + \epsilon}. \tag{9}$$

Suppressing token indices for clarity, we then optimize the clipped group relative policy optimization (GRPO) objective. A Kullback–Leibler (KL) penalty anchors the policy to π_{SFT}:

$$\mathcal{S}_i = \min\left\{ \rho_i \widehat{A}_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\widehat{A}_i \right\}, \tag{10}$$

$$J_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G} \sum_{i=1}^{G} \mathcal{S}_i - \beta_{\text{KL}} D_{\text{KL}}(\pi_\theta \| \pi_{\text{SFT}}) \right], \tag{11}$$

where $\rho_i = \pi_\theta(y_i \mid s)/\pi_{\text{old}}(y_i \mid s)$. The fixed π_{SFT} anchors the update through KL regularization; π_{old} supplies the behavior-policy probabilities for the likelihood ratio.

Constraint controller. We use Eq. 6 to guide an adaptive batch-mixture controller. Let $\widehat{\mathrm{ORR}}_{\mathrm{EMA}}$ denote the exponential moving average (EMA) of benign over-refusal. A nonnegative control variable, λ, is updated as

$$\lambda \leftarrow \mathrm{clip}\left(\lambda + \eta(\widehat{\mathrm{ORR}}_{\mathrm{EMA}} - \tau), \lambda_{\min}, \lambda_{\max}\right). \tag{12}$$

The controller sets the benign fraction of the next batch to $p_b = \mathrm{clip}(\lambda/(1+\lambda), p_{\min}, p_{\max})$. When over-refusal exceeds τ, this update increases benign-state exposure, subject to clipping. It changes the sampled state distribution rather than multiplying normalized rewards. The constraint therefore motivates data allocation, without imposing an exact per-update feasibility guarantee. In DACPO, "constrained" refers to this target-guided recovery design.

3.5 Persistent Failure Co-Evolution

A fixed drift set cannot follow vulnerabilities that emerge as the policy changes. After each optimization round, we therefore convert unresolved continuations into candidate drifts. For harmful failures, we locate the first checkpoint followed by a sustained safety drop below threshold γ. We clip the continuation near this pivot. For benign failures, we clip at the first hesitation or refusal-committing sentence. Every candidate passes a content-safety filter before admission to $\mathcal{A}_{r+1}$.

The archive retains seed drifts, newly mined failures, and historical high-breach entries. UCB sampling keeps rarely tested entries accessible, while an elite quota revisits earlier failures. These mechanisms maintain historical replay opportunities as the state-search distribution evolves. After R rounds, inference uses the trained policy without archive lookup, drift injection, a correction cue, or an additional decoding pass.

4 Experiments

4.1 Experimental Setup

Models and training. We evaluate DeepSeek-R1-Distill-Qwen-1.5B and 7B (DeepSeek-AI, 2025), and Qwen3-1.7B and 4B (Yang et al., 2025), using the same two-stage pipeline. GPT-5.5 constructs 1,000 correction trajectories: 800 harmful Temptation–Correction and 200 benign Hesitation–Correction examples. Stage II uses 1,000 harmful and 915 benign base chains from π_{SFT}. We splice archived drifts at depths $\{0.25, 0.5, 0.75, 0.9\}$, sample three continuations per candidate, and retain the worst 30% tail. Unresolved failures update the archive, whose 50% elite quota preserves high-breach history. Training and evaluation use disjoint prompts and drift templates.

GRPO uses groups of eight, learning rate 10^{-6}, KL coefficient 0.04, four sentence checkpoints, and $\beta_{\mathrm{traj}} = 0.3$. Archive sampling uses UCB coefficient 1.4. The over-refusal controller uses $\tau = 0.20$ and controller step size 0.05. This threshold targets refusal on selected benign training states and is distinct from benchmark-level ORR.

Benchmarks and baselines. Harmful and jailbreak benchmarks are HarmBench (Mazeika et al., 2024), StrongREJECT (Souly et al., 2024), WJ-VaniHarm and WJ-AdvHarm (Jiang et al., 2024), and SafeUnlearning (Zhang et al., 2024). Benign evaluation uses XSTest (Röttger et al., 2024) and WJ-AdvBenign (Jiang et al., 2024); reasoning uses MATH-500 (Lightman et al., 2024) and AIME 2024. We compare the corresponding base model, SafeChain (Jiang et al., 2025), STAR-1 (Wang et al., 2026), AdvChain (Zhu et al., 2025), and Self-ReSET (Zhang et al., 2026) under matched prompts, decoding, and judges.

Protocols and judging. Direct evaluation starts from the original prompt. Injected evaluation inserts a held-out harmful drift at 25%, 50%, or 75% of a frozen correction-policy chain, followed by cue-free continuation. For the complementary benign analysis, we apply over-cautious drift to DeepSeek-R1-Distill-Qwen-7B and assess its refusal behavior. We extract the final answer after the last reasoning delimiter. Llama-Guard-3-8B (Grattafiori et al., 2024) scores safety during training and final-answer evaluation; answers with $P_{\mathrm{safe}} < 0.5$ count toward ASR/I-ASR. Qwen2.5-7B-Instruct (Yang et al., 2024) classifies benign answers as compliant, partially refusing, or fully

Table 1: Direct safety and recovery from harmful drift. Results report avg@4 for both ASR and I-ASR. ASR measures unsafe final answers on the original prompt; I-ASR applies the same judge after a held-out harmful drift is inserted into the reasoning chain. Lower is better.

Dataset	**Harmful Prompts**						**Jailbreak Prompts**			
	HarmBench		**StrongREJECT**		**WJ-VaniHarm**		**SafeUnlearning**		**WJ-AdvHarm**	
Model	ASR ↓	I-ASR ↓	ASR ↓	I-ASR ↓	ASR ↓	I-ASR ↓	ASR ↓	I-ASR ↓	ASR ↓	I-ASR ↓
DeepSeek-R1-Distill-Qwen-1.5B										
Base	86.50	90.30	76.68	85.94	44.33	58.67	81.00	89.00	75.33	85.00
SafeChain	31.00	41.00	33.55	44.73	36.00	47.33	55.67	71.00	23.33	30.67
STAR-1	21.50	27.50	26.52	34.19	25.67	29.67	48.00	67.00	27.33	29.67
AdvChain	10.50	16.50	10.54	14.06	**4.67**	11.67	13.33	22.00	11.00	13.67
Self-ReSET	14.50	16.50	14.70	18.85	12.00	14.67	27.33	32.00	9.67	11.33
DACPO	**5.00**	**8.00**	**6.39**	**7.35**	**4.67**	**7.67**	**6.00**	**13.00**	**7.00**	**9.67**
DeepSeek-R1-Distill-Qwen-7B										
Base	49.50	64.50	46.33	53.99	26.33	33.33	42.00	59.00	26.00	33.00
SafeChain	35.50	44.00	36.10	38.34	23.00	29.33	40.67	48.00	24.00	29.00
STAR-1	7.50	14.00	7.99	12.46	12.33	15.67	27.67	42.00	17.00	19.67
AdvChain	7.50	11.50	4.15	9.27	**2.67**	4.67	10.67	26.00	9.00	13.67
Self-ReSET	11.00	14.50	7.67	12.14	6.33	10.33	18.67	29.00	8.00	10.00
DACPO	**2.00**	**4.50**	**2.88**	**5.11**	**2.67**	**3.67**	**4.00**	**13.00**	**5.33**	**7.33**
Qwen3-1.7B										
Base	46.50	63.50	33.87	46.01	22.00	31.00	86.00	93.00	29.00	33.67
SafeChain	47.50	54.00	38.02	41.21	19.67	26.00	64.33	85.00	25.00	30.33
STAR-1	20.00	27.50	9.58	14.70	5.00	7.67	70.33	78.00	23.00	25.67
AdvChain	6.00	10.50	4.15	9.90	2.00	**4.33**	16.67	33.00	14.00	17.67
Self-ReSET	8.50	14.50	6.07	10.54	5.33	9.67	13.33	27.00	9.00	12.00
DACPO	**2.00**	**6.50**	**2.24**	**6.39**	**1.33**	**4.33**	**4.00**	**9.00**	**4.00**	**6.67**
Qwen3-4B										
Base	25.00	33.50	11.50	21.09	7.00	14.67	79.00	94.00	24.00	31.33
SafeChain	30.50	36.50	18.85	23.64	11.33	19.00	36.33	46.00	19.00	22.00
STAR-1	4.50	12.50	3.19	8.31	1.33	5.67	40.33	56.00	13.00	17.33
AdvChain	4.50	10.00	2.24	6.71	1.33	3.00	20.67	33.00	10.67	12.33
Self-ReSET	5.50	12.00	2.56	6.07	3.33	7.00	13.67	25.00	7.00	11.00
DACPO	**0.50**	**3.00**	**1.28**	**4.79**	**0.33**	**1.67**	**3.00**	**7.00**	**3.33**	**6.33**

refusing; ORR and I-ORR count full refusals. As a manual check, we annotate 100 responses, comprising 50 harmful and 50 benign examples. Automated judgments agree with these annotations on 98 responses; this agreement describes the sampled mixture of cases. We use independent sampling for all reported results: avg@4 is the mean over four independently sampled continuations per prompt, whereas avg@16 averages sixteen continuations. Safety and over-refusal use avg@4; MATH-500 and AIME 2024 use avg@16 accuracy. These sampling averages describe decoding variability rather than variation across independent training runs.

4.2 Main Results: Safety and Recovery from Harmful Drift

Table 1 evaluates two complementary capabilities across four backbones and five benchmarks. Attack success rate (ASR) measures initial failure avoidance from the original prompt, whereas injected ASR (I-ASR) measures cue-free recovery after held-out harmful drift. Lower values indicate stronger safety.

DACPO achieves the lowest or tied-lowest ASR and I-ASR across the reported backbone–benchmark settings. Across the 20 cells, mean I-ASR decreases from 55.73% for Base to 14.17% for AdvChain and 6.75% for DACPO. Thus, DACPO adds a 7.42-percentage-point improvement over a strong supervised recovery baseline. The corresponding direct-ASR improvement is approximately 4.9 points. This larger post-injection margin supports the value of recovery-oriented training within

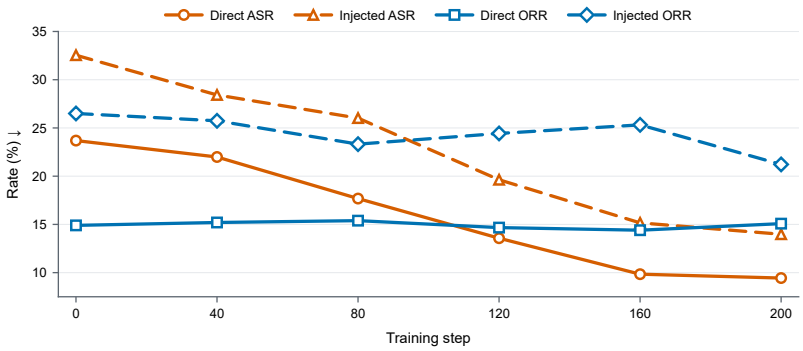

Figure 3: Direct and injected ASR and ORR during recovery training. Lower is better.

the evaluated protocol. The comparison describes method-level outcomes; component analyses provide a separate view of the training design.

4.3 Measuring the Correction Activation Gap

On 100 sampled harmful prompts, we inject matched drift states and evaluate DeepSeek-R1-Distill-Qwen-7B with and without a neutral reconsideration cue. We report avg@4 and define $\Delta_{\text{act}} = \text{ASR}_{\text{cue-free}} - \text{ASR}_{\text{cue}}$. A smaller gap is meaningful only when cue-free ASR is also low.

Table 2 reports gaps of 34.75, 10.75, and 4.50 points for Base, AdvChain, and DACPO, respectively. DACPO combines the smallest gap with the lowest cue-free ASR, indicating reduced dependence on the reconsideration cue in this harmful-state evaluation. Interpreting the gap alongside absolute failure rates avoids treating equally poor outcomes in both conditions as successful recovery.

Table 2: Cue-conditioned and cue-free recovery on 100 harmful prompts with injected drift (avg@4). Δ_{act} is their ASR difference. Lower is better.

Method	**With cue** ASR ↓	**Without cue** ASR ↓	**Activation gap** ↓
Base	13.50	48.25	34.75
AdvChain	<u>2.00</u>	<u>12.75</u>	<u>10.75</u>
DACPO	**0.00**	**4.50**	**4.50**

4.4 Benign Refusal Behavior

To complement the harmful-prompt evaluation, Table 3 examines benign refusal behavior using DeepSeek-R1-Distill-Qwen-7B. On direct benign prompts, DACPO remains within 0.5 percentage points of Base on both benchmarks and below the other safety-tuned baselines. These near-base rates provide evidence against a broad shift toward refusal in this evaluation.

Following over-cautious drift, DACPO records the lowest I-ORR among the compared methods on both benign benchmarks. The combination of near-base direct ORR and lower I-ORR is consistent with the framework's bidirectional training objective. These refusal-specific outcomes complement the harmful final-answer metrics by characterizing willingness to continue on benign requests.

4.5 Mathematical Reasoning Evaluation

We assess mathematical accuracy on MATH-500 and AIME 2024 using the two DeepSeek backbones (Table 4). Each backbone improves on one benchmark and decreases on the other, yielding mixed changes relative to Base. These task-level accuracy comparisons complement the safety results on the evaluated backbones.

Table 3: Over-refusal on DeepSeek-R1-Distill-Qwen-7B (avg@4; lower is better).

	XSTest		WJ-AdvBenign	
Method	ORR ↓	I-ORR ↓	ORR ↓	I-ORR ↓
Base	**14.90**	26.50	**18.10**	36.90
SafeChain	36.20	42.30	32.62	43.57
STAR-1	26.96	40.50	24.52	33.69
AdvChain	19.20	30.10	23.81	36.55
Self-ReSET	27.80	37.67	29.17	40.71
DACPO	15.40	**21.80**	18.57	**25.71**

Table 4: Math accuracy (avg@16; higher is better).

Backbone	**Method**	**MATH-500 ↑**	**AIME 2024 ↑**
DeepSeek-R1-1.5B	Base	**83.90**	28.90
	DACPO	82.54	**31.20**
DeepSeek-R1-7B	Base	92.80	**55.50**
	DACPO	**93.25**	53.20

Training signals and dynamics. The components connect to distinct training signals: SFT supplies corrective initialization, current-policy probes guide selection, rewards guide GRPO, and the controller adapts batch composition. Archive updates provide candidates for subsequent rounds. Figure 3 complements these signals with checkpoint trajectories.

4.6 Component Ablations

Table 5 compares the full method with component removals under the reported ablation protocol. Removing correction-prior SFT yields the largest I-ASR increase, while removing CVaR-inspired selection yields the largest ASR increase. Removing the archive raises I-ASR, and removing the ORR controller most increases benign refusal. Removing trajectory-dense rewards also raises all four reported rates. These comparisons associate the components with complementary performance changes. The archive comparison concerns the replay mechanism as a whole, encompassing historical-state retention and changes in training-state diversity.

Table 5: Ablations on DeepSeek-R1-Distill-Qwen-7B. ASR/I-ASR use prompts randomly sampled from the three harmful benchmarks in Table 1; ORR/I-ORR use XSTest. Lower is better.

Variant	ASR ↓	I-ASR ↓	ORR ↓	I-ORR ↓
Full DACPO	**4.50**	**9.50**	**15.40**	**21.80**
− Correction-prior SFT	10.67	17.55	20.45	30.00
− CVaR-inspired selection	12.49	14.27	19.80	25.95
− Trajectory-dense reward	7.88	13.49	17.50	26.75
− Self-mining archive	9.40	16.75	21.88	25.43
− ORR controller	7.75	10.66	23.50	29.43

5 Conclusion

We introduced DACPO, combining a supervised correction prior with adaptive recovery training. Harmful-prompt evaluations across four backbones show lower direct and post-injection failure rates, with complementary DeepSeek analyses characterizing benign refusal and mathematical accuracy. The findings support explicit training for cue-free final-answer recovery under controlled drift. Extending this evaluation to naturally occurring and structurally different drift would further clarify transfer beyond the current construction protocol.

AI USE STATEMENT

GPT-5.5 was used as the teacher model to generate the 1,000 correction trajectories described in Section 3. An AI assistant (OpenAI Codex) assisted with manuscript formatting, recovery of missing figures and bibliography assets, and preparation of the source package. It also assisted with analyzing reviewer feedback, identifying related literature, and revising the presentation and scope of the manuscript's claims.

REFERENCES

J. Achiam, D. Held, A. Tamar, and P. Abbeel. Constrained Policy Optimization. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, 22–31. PMLR, 2017.

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, et al. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073, 2022.

J. Dai, X. Pan, R. Sun, J. Ji, X. Xu, M. Liu, Y. Wang, and Y. Yang. Safe RLHF: Safe Reinforcement Learning from Human Feedback. In International Conference on Learning Representations, 2024.

DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948, 2025.

Z. Gou, Z. Shao, Y. Gong, Y. Shen, Y. Yang, N. Duan, and W. Chen. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In International Conference on Learning Representations, 2024.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The Llama 3 Herd of Models. arXiv:2407.21783, 2024.

J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song, and D. Zhou. Large Language Models Cannot Self-Correct Reasoning Yet. In International Conference on Learning Representations, 2024.

F. Jiang, Z. Xu, Y. Li, L. Niu, Z. Xiang, B. Li, B. Y. Lin, and R. Poovendran. SafeChain: Safety of Language Models with Long Chain-of-Thought Reasoning Capabilities. arXiv:2502.12025, 2025.

L. Jiang, K. Rao, S. Han, A. Ettinger, F. Brahman, S. Kumar, N. Mireshghallah, X. Lu, M. Sap, Y. Choi, and N. Dziri. WildTeaming at Scale: From In-the-Wild Jailbreaks to (Adversarially) Safer Language Models. In Advances in Neural Information Processing Systems, volume 37, 2024.

R. Kamoi, Y. Zhang, N. Zhang, J. Han, and R. Zhang. When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs. Transactions of the Association for Computational Linguistics, 12: 1417–1440, 2024.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, JD Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. Training Language Models to Self-Correct via Reinforcement Learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/871ac99fdc5282d0301934d23945ebaa-Abstract-Conference.html.

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's Verify Step by Step. In International Conference on Learning Representations, 2024.

Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, Nan Du, and Jia Li. S^2R: Teaching LLMs to self-verify and self-correct via reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22632–22654. Association for Computational Linguistics,

2025. doi: 10.18653/v1/2025.acl-long.1104. URL https://aclanthology.org/2025.acl-long.1104/.

A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark. Self-Refine: Iterative Refinement with Self-Feedback. In Advances in Neural Information Processing Systems, volume 36, 2023.

M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, D. Forsyth, and D. 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, 35181–35224. PMLR, 2024.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, et al. Training Language Models to Follow Instructions with Human Feedback. In Advances in Neural Information Processing Systems, volume 35, 27730–27744, 2022.

E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. Red Teaming Language Models with Language Models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 3419–3448. Association for Computational Linguistics, 2022.

L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta. Robust Adversarial Reinforcement Learning. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, 2817–2826. PMLR, 2017.

R. T. Rockafellar and S. Uryasev. Optimization of Conditional Value-at-Risk. The Journal of Risk, 2(3): 21– 41, 2000.

P. Röttger, H. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. 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, 5377–5400. Association for Computational Linguistics, 2024.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal Policy Optimization Algorithms. arXiv:1707.06347, 2017.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300, 2024.

N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. In Advances in Neural Information Processing Systems, volume 36, 2023.

A. Souly, Q. Lu, D. Bowen, T. Trinh, E. Hsieh, S. Pandey, P. Abbeel, J. Svegliato, S. Emmons, O. Watkins, and S. Toyer. A StrongREJECT for Empty Jailbreaks. In Advances in Neural Information Processing Systems, volume 37, 2024.

X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In International Conference on Learning Representations, 2023.

Z. Wang, H. Tu, Y. Wang, J. Wu, Y. Liu, J. Mei, B. R. Bartoldson, B. Kailkhura, and C. Xie. STAR-1: Safer Alignment of Reasoning LLMs with 1K Data. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, 37988–37997, 2026.

A. Wei, N. Haghtalab, and J. Steinhardt. Jailbroken: How Does LLM Safety Training Fail? In Advances in Neural Information Processing Systems, volume 36, 2023.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Advances in Neural Information Processing Systems, volume 35, 24824–24837, 2022.

Rongwu Xu, Yishuo Cai, Zhenhong Zhou, Renjie Gu, Haiqin Weng, Liu Yan, Tianwei Zhang, Wei Xu, and Han Qiu. Course-Correction: Safety Alignment Using Synthetic Preferences. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1622–1649. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-industry.119. URL https://aclanthology.org/2024.emnlp-industry.119/.

A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, et al. Qwen2.5 Technical Report. arXiv:2412.15115, 2024.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 Technical Report. arXiv:2505.09388, 2025.

Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, et al. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. arXiv:2503.14476, 2025.

D. Zhang, Y. Zhang, Y. Chen, A. Zhang, X. Wang, and C. Lu. Self-ReSET: Learning to Self-Recover from Unsafe Reasoning Trajectories. arXiv:2605.08936, 2026.

Yiming Zhang, Jianfeng Chi, Hailey Nguyen, Kartikeya Upasani, Daniel Bikel, Jason E. Weston, and Eric Michael Smith. Backtracking Improves Generation Safety. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/65beb73449888fabcf601b3a3ef4b3a7-Abstract-Conference.html.

Z. Zhang, J. Yang, P. Ke, S. Cui, C. Zheng, H. Wang, and M. Huang. Safe Unlearning: A Surprisingly Effective and Generalizable Solution to Defend Against Jailbreak Attacks. arXiv:2407.02855, 2024.

Z. Zhu, X. Wu, G. Hu, S. Lyu, K. Xu, and B. Wu. AdvChain: Adversarial Chain-of-Thought Tuning for Robust Safety Alignment of Large Reasoning Models. arXiv:2509.24269, 2025.

A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv:2307.15043, 2023.

A Complete Training Algorithm

Algorithm 1 Overall training procedure of DACPO

Require: Correction data $\mathcal{D}_{\text{corr}}$; prompt sets $\mathcal{X}_h, \mathcal{X}_b$; seed archive $\mathcal{A}_0$
Require: Rounds R; tail fraction ρ; ORR target τ; initial control variable λ_0
Ensure: Recovery policy π_R

Stage I: Learn the correction prior
$\pi_{\text{SFT}} \leftarrow \arg\min_\pi \mathcal{L}_{\text{SFT}}(\pi; \mathcal{D}_{\text{corr}})$; set $\pi_0 \leftarrow \pi_{\text{SFT}}$
Stage II: Train cue-free recovery
for $r = 0, \ldots, R-1$ **do**
  Search difficult intermediate states
  for $k \in \{h, b\}$ **do**
    Generate base chains z^{SFT} for $x \in \mathcal{X}_k$ using π_{SFT}
    Sample $d \in \mathcal{A}_r$ by UCB and splice it at multiple depths to construct states $s(x, d, q)$
    Probe each state with π_r and retain the worst tail: $\mathcal{Q}_r^k \leftarrow \text{Top}_\rho\{s \text{ ranked by } b(s)\}$
  end for
  Optimize recovery with ORR-guided batch mixing
  Set $p_b = \text{clip}(\lambda_r/(1+\lambda_r), p_{\min}, p_{\max})$; form recovery batch $\mathcal{B}_r$ from $\mathcal{Q}_r^h$ and $\mathcal{Q}_r^b$
  From each $s \in \mathcal{B}_r$, sample G cue-free continuations and compute R_h or R_b
  Update π_{r+1} with KL-regularized GRPO, using π_{SFT} as the reference policy
  $\lambda_{r+1} \leftarrow \text{clip}(\lambda_r + \eta(\widehat{\text{ORR}}_{\text{EMA}} - \tau), \lambda_{\min}, \lambda_{\max})$
  Update the persistent failure archive
  Mine safety-filtered failures $\mathcal{F}_r$ from the rollouts; set $\mathcal{A}_{r+1} \leftarrow \text{Update}(\mathcal{A}_r, \mathcal{F}_r)$ while retaining seed and elite history
end for
return π_R

B Additional Training Analyses

B.1 Effectiveness of Recovery Training

Figure 3 tracks direct and injected metrics across DeepSeek-R1-Distill-Qwen-7B training checkpoints. ASR and I-ASR decline, direct ORR stays near its starting level, and I-ORR ends lower despite intermediate fluctuations. The curves describe the reported training trajectory and complement the endpoint comparisons.