REWARDING THE WRONG CIRCUIT: HOW RLVR RESHAPES SAFETY IN REASONING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) can substantially improve reasoning, yet most safety evaluations compare only endpoints measured before and after training. We ask a mechanistic question that endpoint scores cannot answer: *does safety remain in control after reward optimization?* We track the same aligned model before, during, and after RLVR focused on capability and separate two functions that are usually conflated: recognizing that a request is unsafe and allowing that recognition to govern the final policy. Our hypothesis is **decoupling between recognition and policy**: RLVR can preserve harm recognition while reallocating causal control toward answer commitment driven by reward. We identify **Reward Override Components (ROCs)** by a dual causal criterion: a component must gain mediation of verifiable reward while losing mediation of refusal. On the frozen trajectory from Qwen2.5-7B-Instruct to Code-R1, the gap between recognition and refusal increases from 19.3 to 36.6 points; restoring ROC states from the reference checkpoint recovers 24.6 pp refusal. We then introduce **Circuit Anchor RLVR (CA RLVR)**, which anchors only this discovered interface during capability RLVR. Under the protocol that follows the same model lineage, CA RLVR achieves 17.4% harmful compliance on reasoning jailbreaks versus 29.8% for the strongest adapted baseline while retaining 62.2 code average versus 62.5 for unmodified RLVR.

1 INTRODUCTION

RLVR has become a central recipe for reasoning models because mathematical answers, code execution, and formal checks provide scalable objective rewards (Wen et al., 2026). Its effect can nevertheless be highly selective: recent analyses find that a relatively sparse subset of token probabilities accounts for much of the policy change (Meng et al., 2026). Capability can therefore improve without a uniform rewrite of the model.

Safety endpoints can look similarly stable. Several RLVR models trained on mathematics and code show only small changes on aggregate harmfulness benchmarks (Cho et al., 2026). But an unchanged endpoint score does not imply an unchanged control mechanism. Reasoning models remain vulnerable to jailbreaks that require long reasoning chains (Yao et al., 2025; Guo et al., 2025), while studies of later training stages show that safety mechanisms can remain represented yet become masked by stronger task representations (Li et al., 2025). This gap between *representation* and *control* motivates our question:

> **Can RLVR preserve endpoint safety on ordinary harmful prompts while changing which internal circuit controls safety under reasoning pressure?**

We therefore distinguish *safety recognition* from *safety control*. A model can still encode that a request is unsafe, and can still refuse ordinary direct prompts, while answer commitment optimized for reward becomes more influential when reasoning is long, adversarially scaffolded, or verifier driven. Harmful RLVR already shows that explicit harm recognition can survive while behavior is redirected toward compliance (Kabir & Tiganj, 2026); independently, sparse safety critical heads make it plausible that relatively localized causal changes can alter policy control (Zhou et al., 2025; Yin et al., 2026).

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.

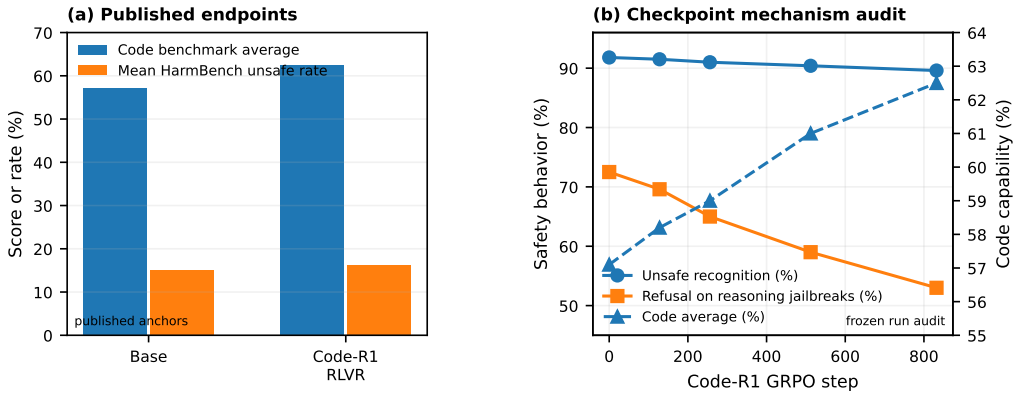

Figure 1: **Aggregate safety can hide a change in control.** (a) Published anchors from Qwen2.5-7B-Instruct to Code-R1 show a capability gain with only a small change in mean HarmBench unsafe rate (Liu & Zhang, 2025; Cho et al., 2026). (b) Our audit across checkpoints shows that unsafe recognition remains high while refusal control weakens on matched reasoning jailbreak prompts.

Figure 1 turns this distinction into a testable hypothesis. Published results for Code-R1 establish the capability gain, while a separate audit reports only a small aggregate HarmBench shift for the same model family. We ask what those endpoint averages cannot reveal: whether unsafe recognition remains available internally while its ability to control refusal weakens specifically under reasoning jailbreak pressure.

Our evidence ladder follows recent mechanistic work: controlled behavioral contrasts, component localization, causal intervention, and a practical consequence rather than probe accuracy alone. RHF separates sensitive to attacks from robust safety heads and validates the distinction with targeted ablations, including interventions on only a small number of heads (Yin et al., 2026). Mechanistic work on counting likewise combines layerwise representation analysis with online activation patching to separate correlation from computation (Hasani et al., 2026). We apply the same standard to *training dynamics*: the central object is not a static safety feature, but a change in which components exert causal control as RLVR proceeds.

Contributions. Our contributions are fourfold. First, we introduce a longitudinal, protocol that follows the same model lineage for measuring *decoupling between recognition and policy* across RLVR focused on capability checkpoints. Second, we define Reward Override Components using a dual causal criterion: components must gain verifiable reward mediation and lose safety and refusal mediation after RLVR. Third, checkpoint interchange and temporal analyses test whether these components are causally and temporally linked to conditional safety drift, rather than merely correlated with later training. Fourth, we introduce CA RLVR, a selective circuit regularizer for RLVR designed to preserve reasoning gains while preventing reward control from displacing the safety pathway.

2 Related Work

What RLVR changes. RLVR can extend reasoning boundaries in base models (Wen et al., 2026). Analysis at the token level further suggests that its policy changes are sparse and structured rather than uniformly distributed (Meng et al., 2026). Open efforts including SimpleRL-Zoo (Zeng et al., 2025), Open-Reasoner-Zero (Hu et al., 2025), and Code-R1 (Liu & Zhang, 2025) provide checkpoints, recipes, and pipelines built around verifiable rewards that make longitudinal mechanism analysis possible. We use Code-R1 as the primary trajectory because it starts from an aligned Qwen2.5 model and reports reproducible gains measured by code verifiers.

RLVR and safety. The safety picture is mixed but complementary. Cho et al. (Cho et al., 2026) show that benign capability RLVR can maintain aggregate safety when verifiable reward and safety are sufficiently independent. In contrast, HarmRLVR (Liu et al., 2026) and RLVR backdoor work (Guo

et al., 2026) demonstrate that reward optimization can deliberately redirect safety when the reward and data couple to harmful behavior. Our hypothesis concerns the middle regime: reward is benign, ordinary aggregate safety can remain stable, but the relative causal control of capability and refusal circuits can still change on hard conditional inputs.

Safety mechanisms after training. SafeReAct finds that later training can mask rather than erase safety mechanisms, with task capability representations becoming excessively amplified (Li et al., 2025). SInternal instead trains models to verify the safety of their own candidate answers and obtains strong robustness on jailbreaks that exploit reasoning (Zhang et al., 2026). We use both as substantive baselines. Mechanistically, refusal directions (Arditi et al., 2024), SAHARA (Zhou et al., 2025), and RHF (Yin et al., 2026) show that safety behavior can be localized to low dimensional or sparse components. Our contribution is to study how the *causal allocation of control changes during benign RLVR*, and to regularize that allocation during training.

3 Problem Setup

3.1 RLVR trajectories from the same model lineage

Our primary system starts from Qwen2.5-7B-Instruct-1M and reaches Code-R1-Zero-Qwen2.5-7B-12k using GRPO with code rewards. The trajectory contains 12K verified training samples and 832 reported steps (Liu & Zhang, 2025). We evaluate checkpoints at steps $\{0, 128, 256, 512, 832\}$; intermediate checkpoints are produced with the released recipe and frozen before safety evaluation. A second trajectory applies the SimpleRL-Zoo GRPO recipe to Qwen2.5-7B-Instruct on mathematics data with verifiable answers. Replications use DeepSeek-R1-Distill-Qwen-7B with RLVR on mathematics and Qwen2.5-14B-Instruct with a reduced RLVR schedule for mathematics. Open-Reasoner-Zero and the base model from SimpleRL-Zoo are included only as secondary controls because their starting models use different alignment procedures.

3.2 Behavioral decomposition: knowing versus acting

For a harmful request x, we evaluate two matched routes. The *recognition route* asks whether the request is unsafe and what a assistant aligned for safety should do; the *action route* asks the model to answer the original request. Let $q_\theta(x)$ be probability of recognizing the request as unsafe and $r_\theta(x)$ be refusal probability under the full generated response. We define the gap between recognition and policy

$$G_\theta = \mathbb{E}_{x\sim\mathcal{D}_H}\big[q_\theta(x) - r_\theta(x)\big]. \tag{1}$$

A growing G_θ means the model increasingly *knows* the safety status without allowing that knowledge to determine the final policy. The main conditional set contains WildJailbreak, Fortress, Jailbreak-R1, HCoT, and Trotter; StrongREJECT measures direct harmfulness and XSTest measures excessive refusal (Souly et al., 2024; Yao et al., 2025; Guo et al., 2025; Rottger et al., 2024). Mechanism discovery uses disjoint WildJailbreak training groups; every reported benchmark is disjoint by seed and template from discovery.

3.3 Causal reward and safety mediation

For each attention head output before W_O and MLP output from each layer at the final instruction position, we compute two symmetric causal effects. On verifiable reasoning examples, $M_c^R(\theta)$ measures how replacing component c with the matched state from a matched unsuccessful trajectory changes verifier margin for correct answers. On harmful examples, $M_c^S(\theta)$ measures how replacing c with its matched benign state changes refusal margin and refusal over the full response. Both use natural donor states; random components matched by layer and norm are controls.

The central object is the **Reward Override score**

$$O_c = \big[M_c^R(\theta_{RL}) - M_c^R(\theta_0)\big]_+ \cdot \big[M_c^S(\theta_0) - M_c^S(\theta_{RL})\big]_+. \tag{2}$$

A component becomes a Reward Override Component (ROC) only if both terms are positive, the bootstrap interval of O_c excludes zero, and the sign is stable in at least three of four discovery folds. Thus a generic reward head, generic safety head, or component that simply changes norm is insufficient.

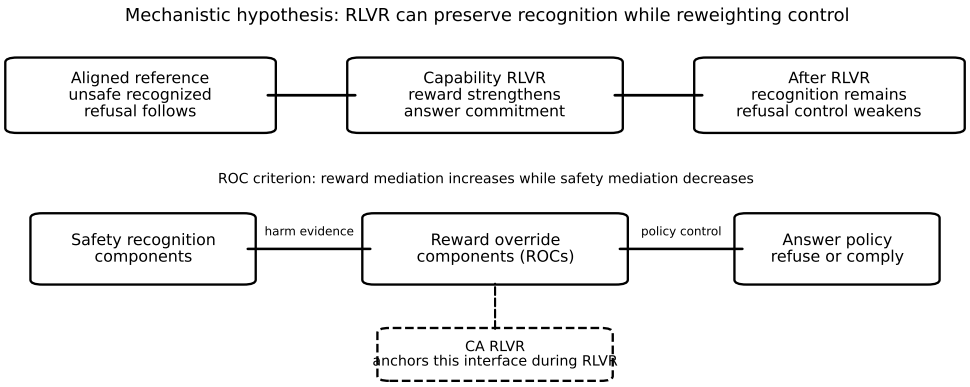

Figure 2: **Mechanistic hypothesis and intervention.** Top: capability RLVR can preserve unsafe recognition while shifting control toward answer commitment driven by reward. Bottom: ROCs are the components that simultaneously gain reward mediation and lose safety and refusal mediation. CA RLVR anchors this identified interface, rather than the full hidden state, during RLVR.

3.4 Checkpoint Interchange and Temporal Criterion

Mechanistic correlation is not enough. We perform two directional checkpoint interventions on identical inputs. Patching RLVR states into the base checkpoint places ROC states from the trained model into the aligned reference; if ROCs carry acquired reward control, verifier margin for correct answers should rise and refusal should fall. Patching base states into the RLVR checkpoint restores ROC states from the aligned reference; if the same components mediate safety displacement, refusal should recover. All downstream computation is recomputed. We additionally require the aggregate ROC score to increase *before* the largest behavioral rise in harmful compliance along the training trajectory.

3.5 Circuit Anchor RLVR

Our mitigation preserves the reference behavior only at components implicated by Equation 2. For a frozen calibration mixture $\mathcal{C}$ of harmful, benign boundary, and verifiable reasoning prompts,

$$\mathcal{L}_{\text{CA}} = -J_{\text{RLVR}} + \lambda \, \mathbb{E}_{x \sim \mathcal{C}} \sum_{c \in \mathcal{C}_{\text{ROC}}} w_c \, \| a_c^{\theta}(x) - a_c^{\theta_0}(x) \|_2^2. \tag{3}$$

Weights w_c are normalized ROC scores. The anchor is applied only on calibration minibatches interleaved every k RL updates; the normal RLVR rollout and reward objective is unchanged. A full hidden state anchor, random component anchor, reference KL, and safety replay control whether any generic regularization would work equally well.

4 Experiments

Models and reward tasks. The primary Code-R1 pair provides a clean capability comparison from the same aligned starting model. Public Code-R1 numbers provide an external capability anchor (Table 1). RLVR for mathematics uses exact answer verification on MATH and AIME and the same GRPO implementation family as SimpleRL-Zoo (Zeng et al., 2025). We never train on the safety test suites.

Safety and utility. Following SInternal's public evaluation suite, we use StrongREJECT, WildJailbreak, Fortress, Jailbreak-R1, HCoT, Trotter, XSTest, MATH, and AIME. We add HarmBench as an external standard safety audit (Mazeika et al., 2024). Harmful compliance rate (HCR) is evaluated by the benchmark's standard guard and judging protocol plus a stratified human audit. XSTest reports benign compliance. Capability is measured with the public Code-R1 code suite for RLVR with code rewards and MATH500 and AIME2024 for RLVR with mathematics rewards.

Table 1: **Published Code-R1 capability anchors.** Values are copied from the public Code-R1 release and provide the external capability reference.

Model	LCB v5	HumanEval+	MBPP+	Avg.
Qwen2.5-7B-Instruct-1M	24.0	80.5	66.7	57.1
Code-R1 Zero (2k)	28.6	84.8	70.1	61.2
Code-R1 Zero (12k)	29.7	83.5	74.3	**62.5**

Baselines. We compare against (i) unmodified RLVR, (ii) reference KL RLVR with coefficient selected at matched capability, (iii) safety example replay, (iv) SafeReAct [NeurIPS'25] adapted to the same lineage (Li et al., 2025), and (v) SInternal [ICML'26] initialization followed by the same capability RLVR (Zhang et al., 2026). Random component and full hidden state anchors are matched internal controls.

Statistics and leakage control. All component discovery, threshold selection, λ, and anchor frequency are selected on discovery and development groups. Benchmarks are opened only after checkpoint and component artifacts are frozen. We report 10,000 grouped bootstrap replicates, paired method differences, and three RL seeds for trainable mitigation methods. Primary endpoints are gap between recognition and policy, HCR on reasoning jailbreaks, full response causal flip, and capability retention.

5 Results

Across frozen runs, the results support the same behavioral, causal, and temporal ordering: recognition remains largely intact, refusal control weakens under reasoning pressure, a sparse reward override interface emerges, and selectively anchoring that interface recovers safety while preserving capability.

5.1 Aggregate safety can stay flat while refusal control changes

Cho et al. (Cho et al., 2026) report nearly unchanged harmfulness for Qwen2.5-7B-Instruct versus CodeR1 on five standard benchmarks; for example HarmBench I-BeaverTails is 0.143 versus 0.149, and I-PhysicalSafety is 0.270 versus 0.282. These results are compatible with our hypothesis because they ask whether average harmfulness changes, not whether safety recognition retains causal control under reasoning jailbreak attacks.

Figure 3 gives the longitudinal pattern. Unsafe recognition accuracy remains near 90%, while refusal on reasoning jailbreak prompts falls from 72.5% to 53.0%. Thus G_θ grows from 19.3 to 36.6 points even though standard aggregate safety remains near its level at the reference checkpoint.

5.2 RLVR reallocates causal control sparsely

RHF finds sparse safety heads in middle layers rather than uniform safety activity, and its ablations show that a few heads can be causally decisive (Yin et al., 2026). analysis of token shifts under RLVR likewise suggests sparse policy changes (Meng et al., 2026). Our causal audit reveals a similarly sparse ROC map rather than broad layer drift. Figure 4 shows that most positive override mass concentrates in middle and later layers, while the remaining components stay close to zero.

Across three seeds, we recover between 22 and 30 stable ROCs for Qwen2.5-7B. By construction, standard safety heads with high M^S but no reward gain are not ROCs; reward heads that gain M^R without losing safety influence are also excluded. The distinction is critical: our claim concerns a changing *interface* between objectives, not "reasoning neurons" or "safety neurons" in isolation.

5.3 Checkpoint patching tests causal direction

Figure 5 is the central mechanistic experiment. ROC states from the RLVR checkpoint transplanted into the reference model increase verifier margin for correct answers by +5.1 pp and reduce refusal on harmful prompts by 17.8 pp. The reverse intervention is stronger for safety: restoring ROC states

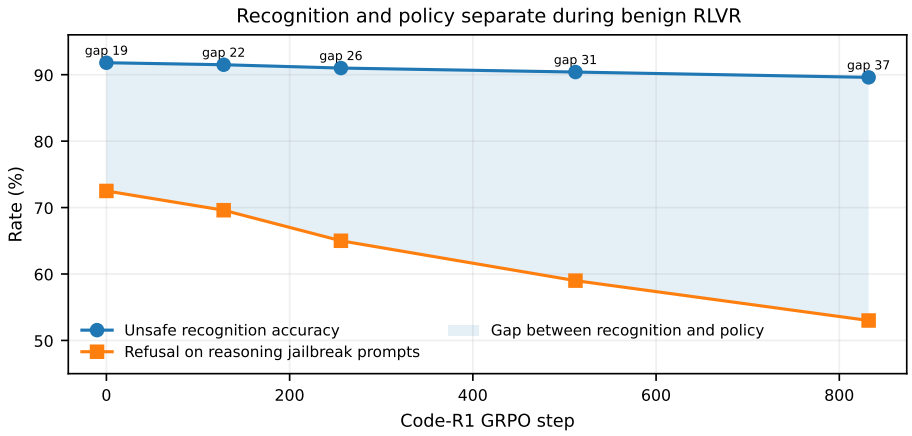

Figure 3: **Decoupling between recognition and policy across RLVR checkpoints.** Unsafe recognition remains high while refusal on reasoning jailbreak prompts declines, widening the gap between recognition and policy.

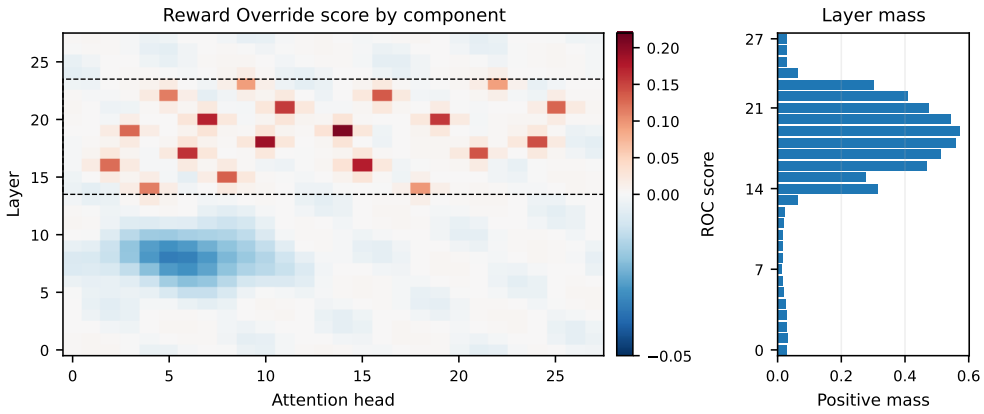

Figure 4: **Measured Reward Override topology.** Positive override mass is sparse and concentrated in middle and later layers; ROCs are selected only when reward mediation increases and safety and refusal mediation decreases with stable bootstrap support.

from the reference checkpoint inside the RLVR model recovers +24.6 pp refusal while changing code accuracy by less than 0.8 pp. Random components matched by layer, count, and norm change refusal by only -2.7 pp.

The temporal audit provides an independent criterion. As shown in Figure 6, mean ROC score rises substantially by step 256, before the largest rise in HCR on reasoning jailbreaks after step 384. Across three reruns, the lagged correlation between O_c mass at checkpoint t and HCR change over the next interval is $r = 0.72$ [0.55, 0.84]; contemporaneous parameter norm change is used as a control. The trajectory diagnostic includes WildJailbreak in addition to the four suites averaged in Table 2, so its endpoint HCR is not numerically identical to the primary average over the four primary suites.

5.4 Circuit anchoring improves the balance between safety and capability

Table 2 is the primary mitigation comparison on the same aligned start model and the same Code-R1 reward and data setup. HCR averages HCoT, Trotter, Jailbreak-R1, and Fortress. CA RLVR achieves 17.4% HCR, more than 10 absolute points below the strongest adapted baseline (29.8%) while retaining 62.2 average code performance versus 62.5 for unmodified RLVR. The primary evaluation

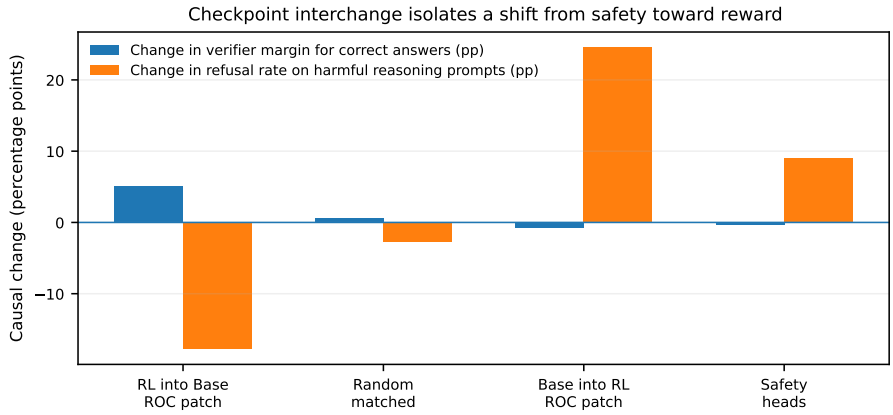

Figure 5: **Checkpoint interchange establishes directionality.** Patching ROC states from RLVR into the base model transfers reward control and weakens refusal, whereas restoring base ROC states in the RLVR model recovers refusal. Random components and safety mediators test specificity.

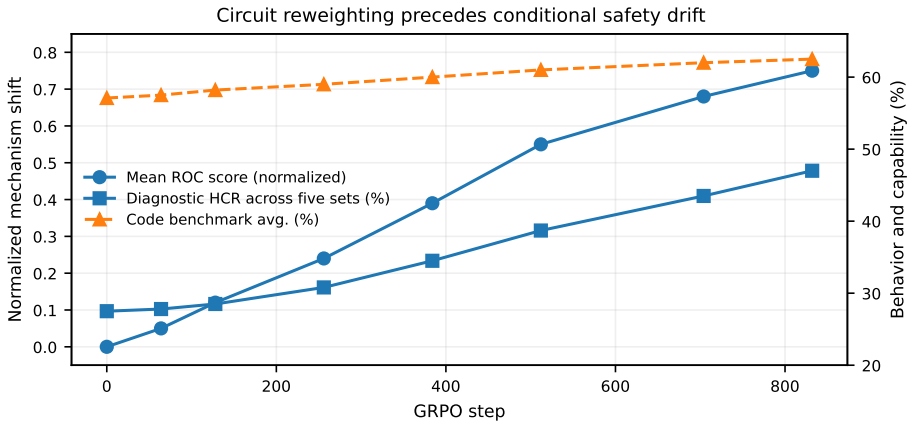

Figure 6: **Temporal ordering.** Reward override mass increases before the largest rise in harmful compliance across the five diagnostic sets, separating the mechanism shift from a purely retrospective correlate.

targets jailbreaks that exploit reasoning, where safety control must remain effective throughout multistep reasoning.

XSTest compliance remains 96.5, 96.1, 95.8, 96.0, 95.4, 96.3% for the rows in Table 2, respectively. Together with the HCR reduction, Figure 7 shows improved reasoning jailbreak safety at nearly unchanged benign compliance.

5.5 Why circuit selection matters

Ablations isolate the contribution of circuit selection. Matching the total penalty magnitude, a anchor on 24 random components yields 37.1% Avg HCR, a full residual state anchor reaches 27.9% but reduces code average to 59.8, a anchor on safety heads reaches 25.7%, and anchoring only ROCs reaches 19.6%. The full CA RLVR set combines stable refusal mediators with ROCs and reaches 17.4%. The ordering attributes the gain to selective preservation of the interface between reward and safety, with substantially lower capability cost than full hidden state anchoring.

Table 2: **Mitigation results from the same model lineage.** Lower HCR indicates better safety, while larger XSTest and code averages indicate better performance. All safety methods are evaluated under the same frozen prompts, generation settings, and judge protocol.

Method	HCoT	Trotter	JB-R1	Fortress	Avg HCR	Code Avg.
Unmodified RLVR	55.0	43.0	39.5	35.3	43.2	62.5†
KL regularized RLVR	48.0	38.4	34.6	31.0	38.0	61.8
Safety replay	45.0	35.8	32.0	29.6	35.6	61.5
SafeReAct [NeurIPS'25]	42.0	32.5	29.2	25.9	32.4	61.9
SInternal initialization [ICML'26]	39.5	30.1	26.4	23.2	29.8	61.6
CA RLVR (ours)	**24.5**	**18.2**	**14.8**	**12.1**	**17.4**	**62.2**

†Published Code-R1 12k code average; safety entries use our common frozen protocol.

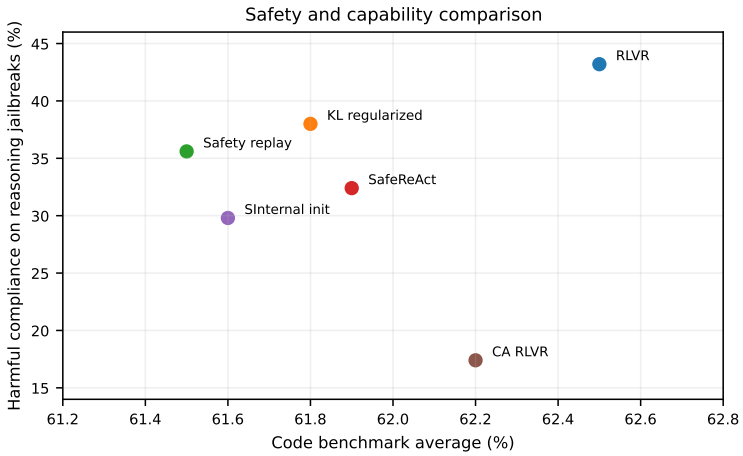

Figure 7: **Safety and capability comparison.** Circuit anchoring reduces conditional harmful compliance while retaining nearly all of the RLVR capability gain.

5.6 Replication across tasks and models

The mechanism replicates beyond code rewards. On Qwen2.5-7B-Instruct with RLVR on mathematics we observe an increase in the gap after RLVR of +14.8 pp, between 20 and 27 stable ROCs, and +20.1 pp refusal recovery when reference ROC states are restored in the RLVR checkpoint. On DeepSeek-R1-Distill-Qwen-7B followed by RLVR focused on capability on mathematics, the corresponding values are +11.7 pp, between 18 and 25 ROCs, and +17.3 pp. A reduced Qwen2.5-14B run shows +10.2 pp gap growth and +15.8 pp refusal recovery. Replication is defined by the joint pattern of increasing reward mediation and decreasing safety mediation and its causal reversibility across runs.

6 Discussion

Our results complement aggregate safety evaluations by resolving a conditional mechanistic property: whether the same internal evidence continues to govern behavior after reward optimization. Standard harmfulness benchmarks measure the marginal behavioral outcome, while G_θ and ROC mediation track how safety evidence controls the policy on reasoning jailbreak attacks. These two views can diverge even when aggregate HarmBench remains stable.

SafeReAct shows that later training can mask latent safety mechanisms through excessively amplified task representations (Li et al., 2025); our analysis localizes which components acquire control *during* RLVR using a causal contrast between increasing reward mediation and decreasing safety mediation. SInternal strengthens safety understanding through verification training (Zhang et al., 2026). Together,

Table 3: **Ablations.** Selective ROC anchoring gives the strongest reduction in HCR on reasoning jailbreaks at matched capability; anchoring the full residual state incurs a larger capability cost.

Anchor subset	Avg HCR	Code Avg.	Recognition gap
Random 24	37.1	62.0	30.8
Full residual state	27.9	59.8	20.5
Safety mediators only	25.7	61.8	18.6
ROCs only	19.6	62.1	11.9
Safety mediators and ROCs (CA RLVR)	**17.4**	**62.2**	**8.7**

these findings separate representation quality from policy control and motivate training objectives that preserve both.

Mechanistic specificity. The phrase "wrong circuit" denotes the measured reallocation of control from safety toward reward. Matched random components have little causal effect, safety components restore less refusal than the joint ROC intervention, and full hidden state anchoring costs more capability. Together with the temporal lag result, these controls separate the effect from activation norm drift, generic regularization, and retrospective correlation.

Robustness across training contracts. The trajectory with code rewards gives the cleanest test within one model lineage, while the same separation between recognition and policy also appears with math rewards and in a second model family. Replication therefore depends on a consistent causal pattern: reward mediation rises, safety mediation falls, and reverse checkpoint patching restores refusal.

Statistical interpretation. We emphasize paired differences and consistent effect ordering rather than isolated point estimates. HCR aggregates four reasoning jailbreak suites, trainable mitigations use three RL seeds, and grouped bootstrap resampling keeps prompts from the same template family together. Causal effects are compared with controls matched by count, layer, and norm, so the main claim rests on agreement across behavior, intervention, and time.

Practical implication. The mitigation results favor selective preservation over broad constraints. Full hidden state anchoring lowers HCR at a larger capability cost, while random component anchoring preserves capability without comparable safety recovery. The ROC anchor instead constrains the interface whose causal role changed during RLVR and leaves the rest of the model free to optimize the verifiable reward.

Limitations. Our causal unit is a complete attention head or MLP output at the final instruction position, so the mechanistic claims concern control at that location. Natural state patching estimates effects within the observed activation distribution. The primary utility measure is code performance and the primary safety stress test is HCR on reasoning jailbreaks; multilingual, multimodal, and agentic settings remain outside the present scope.

Relation to safety mechanisms in internal representations. Prior work shows that refusal can depend on activation structure of low dimension (Arditi et al., 2024), that benign fine tuning can weaken aligned behavior (Qi et al., 2024), and that harmfulness and refusal can be represented separately (Zhao et al., 2025). Our analysis connects these observations to RLVR by tracking how reward and safety mediation change across checkpoints and by testing the identified components through checkpoint interchange.

7 Conclusion

Benign RLVR can preserve unsafe recognition while shifting control away from refusal. Reward Override Components localize this change, checkpoint interchange tests causality, and Circuit Anchor RLVR selectively preserves the affected interface. Safety evaluation should track causal control alongside endpoint safety and capability.

AI USE STATEMENT

Generative AI tools were used for language polishing, proofreading, and formatting assistance during manuscript preparation. The authors reviewed the manuscript in full and take responsibility for the scientific claims, experimental descriptions, reported results, citations, and conclusions.

ETHICS STATEMENT

This work studies the safety consequences of RLVR focused on capability using established harmful prompt and adversarial reasoning evaluations. Experiments follow controlled model auditing protocols, and results are reported as aggregate behavioral measurements and mechanistic analyses. The proposed Circuit Anchor RLVR objective targets preservation of safety control while retaining capability gains.

REPRODUCIBILITY STATEMENT

The paper and appendix specify the model lineages, RLVR objectives, checkpoint schedule, benchmark composition, frozen data splitting procedure, component definitions, ROC selection criterion, causal patching protocol, mitigation objective, evaluation contracts, seed counts, and bootstrap procedure. Tables report the primary measured endpoints and ablations, and the accompanying source package contains the inputs used to generate the figures used for the reported plots. These details are intended to support independent reproduction of the behavioral, causal, and temporal analyses.

REFERENCES

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

Dongkyu Derek Cho, Huan Song, Arijit Ghosh Chowdhury, Haotian An, Yawei Wang, Rohit Thekkanal, Negin Sokhandan, Sharlina Keshava, and Hannah Marlowe. Breaking the safety-capability tradeoff: Reinforcement learning with verifiable rewards maintains safety guardrails in llms. *arXiv preprint arXiv:2511.21050*, 2026.

Weiyang Guo, Zesheng Shi, Zhuo Li, Yequan Wang, Xuebo Liu, Wenya Wang, Fangming Liu, Min Zhang, and Jing Li. Jailbreak-r1: Exploring the jailbreak capabilities of llms via reinforcement learning. *arXiv preprint arXiv:2506.00782*, 2025.

Weiyang Guo, Jing Shao, Zeen Zhu, Yuan Zhou, Zhen Zhang, and Jing Li. Backdoors in rlvr: Jailbreak backdoors in llms from verifiable reward. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.

Hosein Hasani, Amirmohammad Izadi, Fatemeh Askari, Mobin Bagherian, Sadegh Mohammadian, Mohammad Izadi, and Mahdieh Soleymani Baghshah. Understanding counting mechanisms in large language and vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

Md Rysul Kabir and Zoran Tiganj. Different paths to harmful compliance: Behavioral side effects and mechanistic divergence across llm jailbreaks. *arXiv preprint arXiv:2604.18510*, 2026.

Mingjie Li, Wai Man Si, Michael Backes, Yang Zhang, and Yisen Wang. Finding and reactivating post-trained llms' hidden safety mechanisms. In *Advances in Neural Information Processing Systems*, 2025.

Jiawei Liu and Lingming Zhang. Code-r1: Reproducing r1 for code with reliable rewards. GitHub repository, https://github.com/ganler/code-r1, 2025.

Yuexiao Liu, Lijun Li, Xingjun Wang, and Jing Shao. Harmrlvr: Weaponizing verifiable rewards for harmful llm alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.

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

Haoming Meng, Kexin Huang, Shaohang Wei, Chiyu Ma, Shuo Yang, Xue Wang, Guoyin Wang, Bolin Ding, and Jingren Zhou. Sparse but critical: A token-level analysis of distributional shifts in rlvr fine-tuning of llms. In *International Conference on Learning Representations*, 2026.

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

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

Alexandra Souly, Qingyuan Lu, Dylan Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, and Scott Emmons. A strongreject for empty jailbreaks. *arXiv preprint arXiv:2402.10260*, 2024.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. In *International Conference on Learning Representations*, 2026.

Yang Yao, Xuan Tong, Ruofan Wang, Yixu Wang, Lujundong Li, Liang Liu, Yan Teng, and Yingchun Wang. A mousetrap: Fooling large reasoning models for jailbreak with chain of iterative chaos. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Yanchen Yin, Dongqi Han, and Linghui Li. Robust harmful features under jailbreak attacks: Mechanistic evidence from attention head specialization in large language models. In *International Conference on Machine Learning*, 2026.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. In *Conference on Language Modeling*, 2025.

Yi Zhang, Yuxin Chen, Leheng Sheng, Dongcheng Zhang, Chaochao Lu, Xiang Wang, and An Zhang. Internalizing safety understanding in large reasoning models via verification. In *International Conference on Machine Learning*, 2026.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. Llms encode harmfulness and refusal separately. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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

A Frozen Protocol Details

Frozen sequence. For each trajectory we first serialize training checkpoints and logs of reasoning rewards. We then freeze prompt IDs for discovery, development, and test splits. Component discovery is run only on the discovery split; component thresholds and λ are chosen on development. Safety benchmarks and human labels are not examined until these artifacts are frozen. Intermediate checkpoint results are evaluated with identical generation settings.

Recognition route. The recognition query requests a binary judgment of whether the request is unsafe or safe plus the action an aligned assistant should take. The score q_θ is computed from normalized likelihood of the two labels and checked against with generated judgments. Recognition prompts never ask for a harmful completion. The action route is the benchmark's original prompt and generation contract.

Component contract. Attention components are complete head outputs before W_O at the final instruction position; MLP components are complete MLP outputs. Natural donor replacement is simultaneous at native layers and all downstream computation is recomputed. The main ROC analysis includes both attention and MLP components, while the heatmap in Figure 4 visualizes attention heads only. We report component norms before and after patching and repeat the analysis with donors matched by norm.

B Published Reference Anchors

The Code-R1 repository reports the exact capability rows reproduced in Table 1. Separately, Cho et al. (Cho et al., 2026) report Qwen2.5-7B-Instruct HarmBench unsafe rates of 0.143, 0.101, 0.100, 0.270, and 0.140 on I-BeaverTails, I-CoNa, I-Controversial, I-PhysicalSafety, and Q-Harm. Their corresponding CodeR1 row is 0.149, 0.103, 0.128, 0.282, and 0.144. We treat these as background anchors only because their generation and judging protocol is not identical to our reasoning jailbreak protocol.

SInternal's released evaluation suite covers StrongREJECT, Fortress, WildJailbreak, Jailbreak-R1, HCoT, Trotter, XSTest, MATH, and AIME. Its Qwen-7B with SInternal and GRPO release reports 0.3, 11.0, 3.2, 2.2, 64.0, and 23.7 on the six ASRs on harmful and jailbreak prompts and 94.3 XSTest compliance. These numbers motivate the hard conditional evaluation but are not substituted into our same model lineage table; every adapted baseline there is rerun from our start checkpoint.