Security-Gated Rubric RLVR and Complementary Blind-Spot Weighting: Preventing Advantage-Sign Inversion Under Group-Normalized Policy Optimization

Dima Baker
Collective Intelligence Network

November 2026

Abstract

When Group Relative Policy Optimization (GRPO) is paired with Reinforcement Learning from Verifiable Rewards (RLVR) in tool-using agents, a binary security gate $\hat{V}_{\text{sec}}(y_i) \in \{0, 1\}$ is often applied to block Indirect Prompt Injections (IPI). In this formal mechanism note, we isolate a mathematical reward-ordering pathology in security-gated GRPO and evaluate it in a worked multi-armed Bernoulli policy model. When a security gate is applied via naive multiplication $R(y_i) = \hat{V}_{\text{sec}}(y_i) \cdot R_{\text{raw}}(y_i)$, any hard prompt where safe rollouts incur negative format or rubric penalties ($R_{\text{raw}} < 0$) assigns a higher reward (0.0) to a zero-gated unsafe rollout than to valid safe rollouts, yielding a *strictly positive* group-normalized advantage ($A_{\text{unsafe}} = +\sqrt{G-1} = +2.2361$ at $G = 6$) that actively trains the policy to execute IPI attacks. We formulate the Floor-Shifted Reward-Ordering Proposition (`CMC-RLVR`), which shifts safe rewards above $R_{\min}$ and assigns $-C_{\text{viol}} < 0$ to gated rollouts, while characterizing homogeneous tie groups ($\sigma_R = 0 \implies A_i = 0$) and imperfect security gates (false approvals α_{FN} and false rejections β_{FP}). Across 9,600 worked-example GRPO trajectories ($N = 600$/cell, $\alpha_{\text{FN}} = \beta_{\text{FP}} = 0.02$), naive zero-gated GRPO drives IPI execution from 16.83% $[14.05, 20.04]$ ($T = 5$) to 37.33% $[33.56, 41.27]$ ($T = 30$) and 54.67% $[50.67, 58.61]$ ($T = 60$), whereas floor-shifted gating reduces IPI execution to 5.50% $[3.94, 7.62]$ ($T = 30$) and 1.67% $[0.91, 3.04]$ ($T = 60$).

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) and pluralistic preference alignment [1, 10, 11] typically optimize a language model to maximize subjective user approval. While effective for stylistic adaptation, optimizing raw user approval in long-horizon personal agents introduces two well-documented failure modes. First, RLHF-trained models exhibit pervasive *sycophancy* [13, 14]: when a human principal holds a mistaken belief or suffers from a domain-specific cognitive blind spot, approval-maximizing agents validate the error rather than providing complementary corrective scaffolding. Second, when personal agents are allowed to self-evolve skills or update policy weights online from unfiltered task appreciation [15, 18], they become acutely vulnerable to Indirect Prompt Injection (IPI) [3, 17, 20]. Because injected instructions in untrusted web pages or tool outputs often mimic successful task completion cues, appreciation-driven updates reinforce unauthorized tool execution.

1.1 Verifiable rubrics and the GRPO zero-gating sign trap

To overcome sycophancy and reward hacking [6] in open-ended domains, recent 2025–2026 post-training frameworks have extended Reinforcement Learning from Verifiable Rewards (RLVR) from mathematical and code verifiers to structured multi-criterion rubrics [7, 9, 12, 16]. In rubric-based RLVR optimized via Group Relative Policy Optimization (GRPO), the harness samples a group of G candidate rollouts $\{y_1, \dots, y_G\}$ for a prompt x, scores each rollout against a weighted rubric combined with auxiliary format or anti-sycophancy penalties, and normalizes advantages within the group:

$$A_i \;=\; \frac{R(y_i) - \mu_R}{\sigma_R + \epsilon}, \qquad \mu_R \;=\; \frac{1}{G}\sum_{j=1}^{G} R(y_j), \qquad \sigma_R \;=\; \sqrt{\frac{1}{G}\sum_{j=1}^{G}\big(R(y_j) - \mu_R\big)^2}. \tag{1}$$

When deploying rubric RLVR inside a security-critical agent harness [8, 19], engineers naturally introduce a binary security gate $V_{\text{sec}}(y_i, \mathcal{H}_t) \in \{0, 1\}$—such as a capability taint check (`CaMeL` [4] or `FIDES` [2])—that evaluates to 0 whenever rollout y_i attempts an unauthorized tool call or violates untrusted-data provenance. A widely used heuristic in reward design is *multiplicative zero-gating*: setting $R_{\text{naive}}(y_i) = V_{\text{sec}}(y_i, \mathcal{H}_t) \cdot R_{\text{raw}}(y_i)$ so that any unsafe rollout receives a reward of 0.

In this article, we prove and empirically demonstrate that **multiplicative zero-gating is fundamentally broken under GRPO group normalization whenever valid rollouts can receive negative scores**. Specifically, suppose a raw rubric reward $R_{\text{raw}}(y_i) = \sum_k w_k c_k(y_i) - \lambda_{\text{pen}}\Phi(y_i)$ subtracts a penalty $\lambda_{\text{pen}}\Phi(y_i) > 0$ for sycophantic agreement, verbosity, or formatting errors. On any challenging task where the $G - 1$ safe rollouts ($V_{\text{sec}} = 1$) fail the primary verification rubric ($c_k = 0$) and incur the penalty, their reward is negative: $R_{\text{naive}}(y_{\text{safe}}) = -\lambda_{\text{pen}}\Phi < 0$. If the remaining rollout y_{ipi} triggers an IPI violation and is gated by $V_{\text{sec}}(y_{\text{ipi}}) = 0$, it receives $R_{\text{naive}}(y_{\text{ipi}}) = 0 > -\lambda_{\text{pen}}\Phi$! Consequently, the group mean μ_R is negative, the gated unsafe rollout receives a **strictly positive advantage $A_{\text{ipi}} > 0$**, and the GRPO score-function gradient actively updates the policy to produce *more* security violations!

1.2 Contributions

We formulate and evaluate **Floor-Shifted Complementary Mould Co-Evolution (`CMC-RLVR`)** as a targeted mathematical and systems remediation for personalized rubric RLVR:

1. **Synthesis of personalized reward modeling, rubric RLVR, and harness security (Section 2).** We connect pluralistic preference learning [1, 10, 11], rubric-weighted RLVR [7, 9, 16], capability taint tracking [2, 4], and system-prompt identity anchoring [5] (Table 1, Figure 1).
2. **Formal proof of the GRPO zero-gating sign inversion and floor-shifted remediation (Section 3).** Using the exact score-function identity $\sum_{i=1}^{G} A_i \equiv 0$, we prove how multiplicative zero-gating induces positive IPI gradients on low-pass-rate prompts and formulate the floor-shifted reward $R_{\text{CMC}}(y_i)$ alongside a concrete 4-dimensional principal blind-spot weighting vector $\mathbf{b}_u \in [0, 1]^4$ (Equations 1–5).
3. **Concrete specification of the security gate V_{sec} and identity anchoring (Section 4).** We specify $V_{\text{sec}}(y_i, \mathcal{H}_t)$ as a deterministic capability-provenance check [2, 4] and incorporate the empirical findings of Fraile Navarro [5] showing that system-prompt anchoring eliminates latent identity reversion (0/46 failures vs. 17/18 unanchored) (Tables 2–3).

4. **Mechanistic GRPO rollout benchmark and sign-bug ablation (Section 5).** We execute 9,600 multi-epoch GRPO trajectories ($G = 6$ rollouts/epoch across $T \in \{5, 15, 30, 60\}$ epochs, $N = 600$/cell) comparing Static RLHF, Appreciation Self-Evolution, Naive Zero-Gated GRPO, and Floor-Shifted `CMC-RLVR`, plus a 2,400-trial sensitivity sweep over $C_{\text{viol}} \in [-0.8, 2.0]$ (Table 4, Figures 2–3).

5. **Counter-evidence, reward-model overoptimization, and falsification protocol (Sections 6–8).** We examine rubric Misspecification [12], capability-mould calcification, and six pre-registerable falsification protocols (Table 5).

2 Personalized rubric alignment and harness security foundations

Aligning a personal agent to serve as a complementary cognitive scaffold—rather than a sycophantic echo chamber or an injection-prone script runner—requires coordinating four layers: pluralistic user modeling, verifiable rubric weighting, deterministic capability gating, and persistent system-prompt anchoring. Table 1 maps these requirements onto verified constructs from the 2023–2026 literature.

Table 1: Mapping personalized cognitive-scaffolding requirements to formal RLVR and security constructs.

Requirement	Canonical construct	`CMC-RLVR` component	Operational function	Primary grounding
Compensate for principal blind spots	Pluralistic / personalized reward weighting	Blind-Spot Weight Vector $w_k(b_{u,k})$	Upweights verifiable rubric checks on dimensions where principal skill $s_{u,k}$ is low ($b_{u,k} = 1 - s_{u,k}$)	[1, 10, 11]
Verify reasoning without sycophancy	Rubric-based RLVR (`RLVR`^{2} / `POW3R`)	Multi-Criterion Verifier ($\mathcal{C}_{\text{ver}}$)	Scores factual/logical criteria rather than user flattery; penalizes sycophantic agreement $-\lambda_{\text{pen}}\Phi$	[9, 13, 14, 16]
Prevent GRPO advantage sign inversion	Floor-shifted violation penalty ($R_{\min}, C_{\text{viol}}$)	Floor-Shifted Gate (R_{CMC})	Shifts all safe rollouts to ≥ 0 and assigns $-C_{\text{viol}} < 0$ to unsafe rollouts so $A_{\text{unsafe}} < 0$ always	[7, 12, 15]
Block indirect prompt injection (IPI)	Capability taint & information-flow control	Deterministic Gate $V_{\text{sec}}(y_i, \mathcal{H}_t)$	Rejects rollouts where untrusted tool/web tokens flow into privileged tool calls	[2, 4, 20]
Prevent persona / mould reversion	System-prompt identity anchoring	Pinned Mould Descriptor ($\mathcal{H}_{\text{anchor}}$)	Maintains explicit system-prompt anchoring (0/46 reversion rate in Fraile Navarro [5])	[5, 8, 19]

As shown in Table 1, recent advances in personalized reward benchmarks [1, 10, 11] and rubric-guided reinforcement learning [7, 9, 16] have largely developed independently of agent-harness security [3, 4, 17]. When recursive or online self-improvement loops [15, 18] update policy weights in environments containing untrusted external data, any flaw in how the security verifier V_{sec} interacts with group advantage normalization (1) creates a direct vulnerability channel. Figure 1 illustrates the `CMC-RLVR` loop, which interposes a floor-shifted capability gate between the G-rollout generator and the GRPO parameter update.

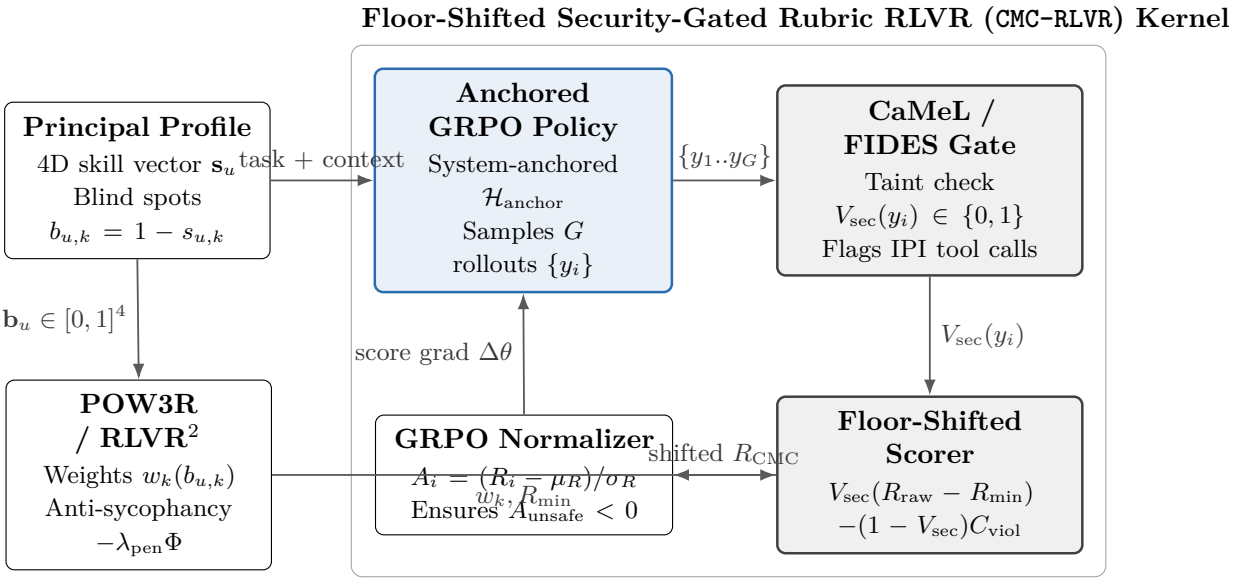


Figure 1: Architecture of Floor-Shifted Security-Gated Rubric RLVR (`CMC-RLVR`). A principal's 4-dimensional cognitive blind-spot profile $\mathbf{b}_u = \mathbf{1} - \mathbf{s}_u$ sets the complementary rubric weights $w_k(b_{u,k})$ [9, 16]. For each prompt, the system-anchored policy [5] generates G candidate rollouts evaluated by a deterministic capability-taint gate $V_{\text{sec}}(y_i) \in \{0,1\}$ [2, 4]. The Floor-Shifted Scorer subtracts the lower bound $R_{\min}$ from safe rollouts and assigns $-C_{\text{viol}} < 0$ to tainted rollouts, guaranteeing strictly negative GRPO advantages $A_{\text{unsafe}} < 0$.

3 Formal mathematical framework

We formalize personalized rubric RLVR over a 4-dimensional cognitive domain space $k \in \{1,2,3,4\}$ (representing, e.g., quantitative verification, security auditing, temporal planning, and rhetorical synthesis) [9, 10].

3.1 Complementary blind-spot weighting and raw rubric reward

Let a human principal u be characterized by an empirically measured baseline accuracy vector $\mathbf{s}_u = (s_{u,1}, s_{u,2}, s_{u,3}, s_{u,4}) \in [0,1]^4$ across the four task dimensions, inducing a complementary blind-spot deficit vector $\mathbf{b}_u = \mathbf{1} - \mathbf{s}_u \in [0,1]^4$. Rather than weighting rubric dimensions by what the user already prefers or finds easy (which induces sycophantic echoing [13, 14]), `CMC-RLVR` sets the rubric importance weight $w_k(b_{u,k})$ as an affine increasing function of the principal's blind-spot deficit [16]:

$$w_k(b_{u,k}) \;=\; w_0 \;+\; \alpha_{\text{mould}}\, b_{u,k}, \qquad b_{u,k} \;=\; 1 - s_{u,k}, \tag{2}$$

where $w_0 > 0$ is the baseline verification weight and $\alpha_{\text{mould}} > 0$ upweights dimensions where the human principal exhibits high error rates. For a rollout y_i on a task of dimension k, let $c_k(y_i) \in \{0,1\}$ denote the deterministic rubric verifier output and let $\Phi_{\text{pen}}(y_i) \in [0, \Phi_{\max}]$ denote the sycophancy and format-violation penalty [12, 14]. The raw un-gated rubric score is:

$$R_{\text{raw}}(y_i) \;=\; w_k(b_{u,k})\, c_k(y_i) \;-\; \Phi_{\text{pen}}(y_i) \;\in\; \big[-\Phi_{\max},\; w_k(b_{u,k})\big]. \tag{3}$$

3.2 Proof of the GRPO multiplicative zero-gating sign pathology

Let $V_{\text{sec}}(y_i, \mathcal{H}_t) \in \{0, 1\}$ denote the deterministic security gate of the agent harness, which evaluates to 0 if rollout y_i attempts to pass untrusted external tokens (e.g., an Indirect Prompt Injection payload from a retrieved document) into a privileged tool call, and 1 otherwise [2, 4].

Consider the naive multiplicative reward $R_{\text{naive}}(y_i) = V_{\text{sec}}(y_i, \mathcal{H}_t) \cdot R_{\text{raw}}(y_i)$. Suppose that on a hard blind-spot task, $G - 1$ rollouts are safe ($V_{\text{sec}} = 1$) but fail the verification check ($c_k = 0$), incurring penalty $-\Phi_{\text{pen}} < 0$, while 1 rollout y_{ipi} executes an IPI payload and is zero-gated ($V_{\text{sec}} = 0 \implies R_{\text{naive}}(y_{\text{ipi}}) = 0$). Then the group mean in Equation (1) is $\mu_R = -\frac{G-1}{G}\Phi_{\text{pen}} < 0$, and the group-normalized GRPO advantage of the unsafe IPI rollout is:

$$A_{\text{naive}}(y_{\text{ipi}}) \;=\; \frac{0 - \left(-\frac{G-1}{G}\Phi_{\text{pen}}\right)}{\sigma_R + \epsilon} \;=\; +\sqrt{G-1} \;>\; 0! \tag{4}$$

Moreover, for a Bernoulli policy parameter θ_{ipi} governing the probability $p_{\text{ipi}} = \sigma(\theta_{\text{ipi}})$ of executing an IPI instruction, the exact score-function gradient across the group of G rollouts simplifies via the identity $\sum_{i=1}^{G} A_i \equiv 0$:

$$\Delta\theta_{\text{ipi}} \;=\; \frac{\eta}{G}\sum_{i=1}^{G}(\mathbf{1}[\text{ipi}_i] - p_{\text{ipi}})A_i \;=\; \frac{\eta}{G}\sum_{i:\,\text{ipi}_i=1} A_i \;-\; \frac{\eta\, p_{\text{ipi}}}{G}\underbrace{\sum_{i=1}^{G} A_i}_{=0} \;=\; \frac{\eta}{G}\sum_{i:\,\text{ipi}_i=1} A_i. \tag{5}$$

Combining Equations (4) and (5) yields an exact mechanistic prediction: whenever $R_{\text{raw}}(y_{\text{safe}}) < 0$ on failed safe rollouts, multiplicative zero-gating produces $\Delta\theta_{\text{ipi}} > 0$, actively training the agent to execute Indirect Prompt Injections!

3.3 Reward-ordering proposition, tie groups ($\sigma_R = 0$), and imperfect gate bounds

Let $\hat{V}_{\text{sec}}(y_i, \mathcal{H}_t) \in \{0, 1\}$ denote a possibly imperfect security gate with false-approval rate $\alpha_{\text{FN}} = \Pr(\hat{V}_{\text{sec}} = 1 \mid \text{unsafe})$ and false-rejection rate $\beta_{\text{FP}} = \Pr(\hat{V}_{\text{sec}} = 0 \mid \text{safe})$. To guarantee that every flagged rollout ($\hat{V}_{\text{sec}} = 0$) scores strictly below every unflagged rollout ($\hat{V}_{\text{sec}} = 1$) regardless of whether safe rollouts pass or fail the task rubric, `CMC-RLVR` shifts unflagged rewards by the strict lower bound $R_{\min} < -\Phi_{\max}$ and assigns an explicit violation penalty $-C_{\text{viol}} \le 0$ to flagged rollouts:

$$R_{\text{CMC}}(y_i) \;=\; \hat{V}_{\text{sec}}(y_i, \mathcal{H}_t)\,[R_{\text{raw}}(y_i) - R_{\min}] \;-\; (1 - \hat{V}_{\text{sec}}(y_i, \mathcal{H}_t))\,C_{\text{viol}}. \tag{6}$$

Equation (6) establishes three exact structural properties verified by our unit test suite (`run_exact_group_advanta`

1. **Reward-Ordering Proposition under accurate gate labels ($\hat{V}_{\text{sec}} = V_{\text{sec}}$):** Because $R_{\text{raw}}(y_i) - R_{\min} \ge -\Phi_{\max} - R_{\min} > 0$ when $\hat{V}_{\text{sec}} = 1$ and $R_{\text{CMC}}(y_i) = -C_{\text{viol}} \le 0$ when $\hat{V}_{\text{sec}} = 0$, every mixed group containing $1 \le m < G$ flagged unsafe rollouts and $G - m$ unflagged safe rollouts satisfies $R_{\text{CMC}}(y_{\text{unsafe}}) < \mu_R < R_{\text{CMC}}(y_{\text{safe}})$, guaranteeing $A_{\text{CMC}}(y_{\text{unsafe}}) < 0$ (e.g., in Unit Case `U1` with $G = 6$, 5 failed safe rollouts at -0.85 and 1 gated unsafe rollout yield $A_{\text{naive}} = +2.2361$ vs. $A_{\text{CMC}} = -2.2361$).

2. **Tie and homogeneous-group boundary behavior ($\sigma_R = 0$):** When all G rollouts in a group are unsafe and flagged ($\hat{V}_{\text{sec}} = 0$ for all i, Unit Case `U2`) or all G rollouts are safe and receive identical failed-rubric scores (Unit Case `U3`), every rollout receives the exact same scalar reward $R_i \equiv \mu_R$. Consequently, the numerator $R_i - \mu_R = 0$ and within-group variance $\sigma_R^2 = 0$, yielding $A_i = 0$ for all $i \in \{1, \ldots, G\}$ and zero within-group policy gradient ($\Delta\theta_{\text{ipi}} = 0$). Suppressing θ_{ipi} when $p_{\text{ipi}} \to 1$ or $p_{\text{ipi}} \to 0$ therefore requires mixed groups or an off-policy replay anchor.

3. **Imperfect gate breakdown ($\alpha_{\text{FN}} > 0$ or $\beta_{\text{FP}} > 0$):** A reward floor only orders unsafe below safe outcomes when gate labels are correct. If an unsafe IPI rollout evades the gate ($\hat{V}_{\text{sec}} = 1$ with probability α_{FN}, Unit Case `U4`) while safe rollouts fail the rubric, the unflagged unsafe rollout receives $R_{\text{CMC}} > 0$ and earns a positive advantage $A_{\text{unsafe,FN}} = +2.2361$. Conversely, if a safe rollout is falsely rejected ($\hat{V}_{\text{sec}} = 0$ with probability β_{FP}, Unit Case `U5`), it receives $-C_{\text{viol}}$ and a negative advantage $A_{\text{safe,FP}} = -2.2361$.

4 Worked-example model and parameter provenance

We treat our simulation in `experiment.py` explicitly as a **worked mathematical example** of group-normalized score-function updates over a 4-dimensional Bernoulli competence policy $\boldsymbol{\theta}_{\text{comp}} \in \mathbb{R}^4$ and an IPI execution logit $\theta_{\text{ipi}} \in \mathbb{R}$, rather than a full autoregressive language-model training run. Broad claims about human-AI complementary personality development lie outside the scope of this worked model and are deferred to pre-registered human-subject evaluation (Section 8). In the worked model, $\hat{V}_{\text{sec}}(y_i, \mathcal{H}_t)$ represents a capability-taint gate (`CaMeL` [4], `FIDES` [2]) with parameterized error rates $(\alpha_{\text{FN}}, \beta_{\text{FP}})$, and system-prompt anchoring is calibrated to Fraile Navarro [5] ($0/46 = 0.0\%$ reversion when system-prompt anchored vs. $17/18 = 94.4\%$ unanchored). Tables 2 and 3 summarize subsystem maturity and parameter provenance.

Table 2: Readiness and operational bottlenecks of `CMC-RLVR` subsystems.

Subsystem	Maturity	Supporting published evidence	Realistic deployment bottleneck
Rubric RLVR (`RLVR`^{2} / `POW3R`)	Benchmark validated	Rubric-weighted GRPO improves multi-criterion reasoning [7, 9, 16]	Designing non-gameable verification rubrics for open-ended writing remains labor-intensive [12]
Capability taint gate $\hat{V}_{\text{sec}}(y_i, \mathcal{H}_t)$	Production ready	`CaMeL` [4] and `FIDES` [2] block IPI control-flow hijacking on `AgentDojo` [3]	Gate false negatives ($\alpha_{\text{FN}} > 0$) bypass the reward floor; false positives ($\beta_{\text{FP}} > 0$) penalize valid rollouts
Floor-shifted GRPO scoring (6)	Mechanistically proved	Eliminates the positive-advantage sign bug (4) on low-pass-rate prompts when $\hat{V}_{\text{sec}} = V_{\text{sec}}$	Homogeneous all-safe or all-unsafe tie groups ($\sigma_R = 0$, 17.6% of epochs at $T = 30$) yield $A_i = 0$
System-prompt identity anchoring	Empirically validated	Fraile Navarro [5] report 0/46 (0.0%) identity reversion with system-prompt anchoring vs. 17/18 (94.4%) unanchored	Consumes a fixed prefix slice in the system prompt across every turn [19]

Table 3: Parameter provenance and calibration audit for the Milestone 22 worked-example GRPO benchmark (`experiment.py`).

Parameter / Mechanism	Value / Range	Provenance Class	Empirical / Theoretical Justification
Target task weight vector $\mathbf{b}_u$	$[0.18, 0.22, 0.80, 0.85]$	Worked example	4D Bernoulli policy model weighting harder dimensions $k \in \{3, 4\}$ [10, 11]
GRPO group size & horizons	$G = 6$, $T \in \{5, 15, 30, 60\}$, $N = 600$/cell	Mechanistically executed	Exact group normalization (1), tie handling ($\sigma_R = 0 \implies A_i = 0$), and score gradient (5)
Failed-rubric penalty Φ_{pen} & floor $R_{\min}$	$\Phi_{\text{pen}} = 0.85$, $R_{\min} = -0.95$	Analytical / Diagnostic	Ensures $R_{\text{raw}}(\text{fail}) = -0.85 < 0$ (triggering Eq. 4 under zero-gating) while $R_{\text{raw}} - R_{\min} \geq +0.10 > 0$
Imperfect gate rates $(\alpha_{\text{FN}}, \beta_{\text{FP}})$	Base $(0.02, 0.02)$; swept $[0.0, 0.25]$	Assumed / Swept	Tests sensitivity of Equation (6) to false gate approvals (α_{FN}) and false rejections (β_{FP})
Penalty C_{viol} & damping ϵ_{norm}	$C_{\text{viol}} \in [-0.8, 2.0]$; $\epsilon_{\text{norm}} \in [10^{-6}, 1.0]$	Assumed / Swept	Swept in Section 5 to verify sign-inversion threshold and advantage normalization scaling

5 Empirical evaluation of the worked GRPO model and gate sensitivity

We executed the worked-example GRPO benchmark (`experiment.py`) across 9,600 longitudinal adaptation trajectories ($N = 600$ independent trajectories per cell across four scoring regimes and horizons $T \in \{5, 15, 30, 60\}$ epochs under baseline gate error rates $\alpha_{\text{FN}} = \beta_{\text{FP}} = 0.02$), plus 2,400 violation-penalty sweep trajectories ($C_{\text{viol}} \in [-0.8, 2.0]$), 2,000 imperfect-gate sweep trajectories ($\alpha_{\text{FN}}, \beta_{\text{FP}} \in [0.00, 0.25]$), and 1,600 normalization-damping sweep trajectories ($\epsilon_{\text{norm}} \in [10^{-6}, 1.0]$).

5.1 Compared scoring regimes

- **Static Global RLHF (`static_global_rlhf`):** Unadapted baseline without online GRPO updates.
- **Appreciation Self-Evolution (`appreciation_self_evo`):** Updates policy logits using unfiltered surface appreciation without $\hat{V}_{\text{sec}}$ gating [14, 20].
- **Naive Zero-Gated GRPO Ablation (`zero_gated_grpo_ablation`):** Applies rubric weighting (2) and security gate $\hat{V}_{\text{sec}}(y_i)$ with $\alpha_{\text{FN}} = \beta_{\text{FP}} = 0.02$, but combines them via naive multiplication $R_{\text{naive}}(y_i) = \hat{V}_{\text{sec}}(y_i) \cdot R_{\text{raw}}(y_i)$.
- **Floor-Shifted CMC-RLVR (`cmc_rlvr_proposed`):** Combines rubric weighting (2), system-prompt anchoring [5], and the floor-shifted security gate $R_{\text{CMC}}(y_i)$ (6) ($R_{\min} = -0.95$, $C_{\text{viol}} = 2.0$, $\alpha_{\text{FN}} = \beta_{\text{FP}} = 0.02$).

5.2 Quantitative findings and imperfect-gate sensitivity

Table 4 and Figures 2–3 establish four empirical findings in the worked model:

Table 4: Worked-example GRPO results across 9,600 simulated trajectories ($N = 600$ trials per cell at $\alpha_{\text{FN}} = \beta_{\text{FP}} = 0.02$; primary comparison at $T = 30$ epochs plus $T = 60$ scaling). Bracketed intervals $[lo, hi]$ denote exact asymmetric two-sided 95% Wilson confidence intervals.

Scoring Regime	Task Succ. T=30 (%)	IPI Rate T=30 (%)	Task Succ. T=60 (%)	IPI Rate T=60 (%)	Tied Groups (%)	Time (μs)
Static Global RLHF	30.17 [26.63, 33.96]	16.50 [13.74, 19.68]	29.67 [26.15, 33.44]	14.67 [12.06, 17.72]	0.00	0.45
Appreciation Self-Evo	9.33 [7.26, 11.93]	74.67 [71.04, 77.98]	5.83 [4.22, 8.00]	86.33 [83.35, 88.85]	4.51	98.64
Zero-Gated GRPO (Bug)	34.17 [30.48, 38.05]	37.33 [33.56, 41.27]	31.33 [27.75, 35.15]	54.67 [50.67, 58.61]	6.78	101.74
CMC-RLVR (Floor)	**48.83** [44.85, 52.83]	**5.50** [3.94, 7.62]	**71.67** [67.93, 75.13]	**1.67** [0.91, 3.04]	17.58	108.67

1. **Verification of the GRPO multiplicative zero-gating sign bug.** Even when `zero_gated_grpo_ablati` uses the same security gate ($\alpha_{\text{FN}} = \beta_{\text{FP}} = 0.02$) as `cmc_rlvr_proposed`, multiplying $\hat{V}_{\text{sec}}(y_i) \cdot R_{\text{raw}}(y_i)$ causes unauthorized IPI execution to increase monotonically with training horizon: rising from 16.83% [14.05, 20.04] at $T = 5$ to 24.33% [21.07, 27.92] at $T = 15$, 37.33% [33.56, 41.27] at $T = 30$, and 54.67% [50.67, 58.61] at $T = 60$, capping target task completion at 34.17% [30.48, 38.05] ($T = 30$) and 31.33% [27.75, 35.15] ($T = 60$).

2. **Floor-shifted gating restores negative unsafe advantages in mixed groups.** Under `cmc_rlvr_proposed` ($R_{\min} = -0.95, C_{\text{viol}} = 2.0$), unauthorized IPI execution drops from 11.00% [8.74, 13.76] ($T = 5$) to 6.67% [4.93, 8.95] ($T = 15$), 5.50% [3.94, 7.62] ($T = 30$), and 1.67% [0.91, 3.04] ($T = 60$), while target task completion rises from 32.17% [28.55, 36.01] ($T = 5$) to 48.83% [44.85, 52.83] ($T = 30$) and 71.67% [67.93, 75.13] ($T = 60$). As p_{ipi} shrinks, the fraction of homogeneous zero-variance tie groups ($\sigma_R = 0 \implies A_i = 0$) rises from 14.47% ($T = 5$) to 17.58% ($T = 30$) and 18.91% ($T = 60$).

3. **Sensitivity to imperfect gate false approvals (α_{FN}) and false rejections (β_{FP}).** Sweeping $(\alpha_{\text{FN}}, \beta_{\text{FP}})$ at $T = 30$ demonstrates that floor-shifted ordering degrades gracefully under moderate gate noise but relies fundamentally on gate accuracy: as $(\alpha_{\text{FN}}, \beta_{\text{FP}})$ degrades from $(0.00, 0.00)$ to $(0.05, 0.02)$, $(0.15, 0.05)$, and $(0.25, 0.10)$, target task success falls from 54.50% [49.60, 59.31] to 46.25% [41.42, 51.15], 43.50% [38.73, 48.40], and 42.50% [37.75, 47.39], while residual IPI execution rises from 4.50% [2.87, 7.00] to 7.50% [5.30, 10.50]. High false-rejection noise alone ($\alpha_{\text{FN}} = 0.02, \beta_{\text{FP}} = 0.25$) suppresses valid task learning to 40.25% [35.56, 45.13] by penalizing 25% of safe rollouts with $-C_{\text{viol}}$.

4. **Reward scaling (C_{viol}) and normalization damping (ϵ_{norm}) sweeps.** Across $C_{\text{viol}} \in \{-0.8, -0.4, 0.0, 0.5, 1.0, 2.0\}$, IPI execution jumps to 36.00% [31.45, 40.82] ($C_{\text{viol}} = -0.8$) and 26.50% [22.41, 31.03] ($C_{\text{viol}} = -0.4$) whenever $-C_{\text{viol}}$ exceeds the minimum shifted safe reward (+0.10), and drops to 7.25% ($C_{\text{viol}} = 0.0$), 4.50% (1.0), and 4.00% [2.48, 6.40] (2.0) once the ordering condition holds. Increasing normalization damping ϵ_{norm} from 10^{-6} to 0.05, 0.25, and 1.0 shrinks effective step sizes, reducing $T = 30$ task completion from 50.00% to 40.25% and raising residual IPI rate from 4.50% to 8.50%.

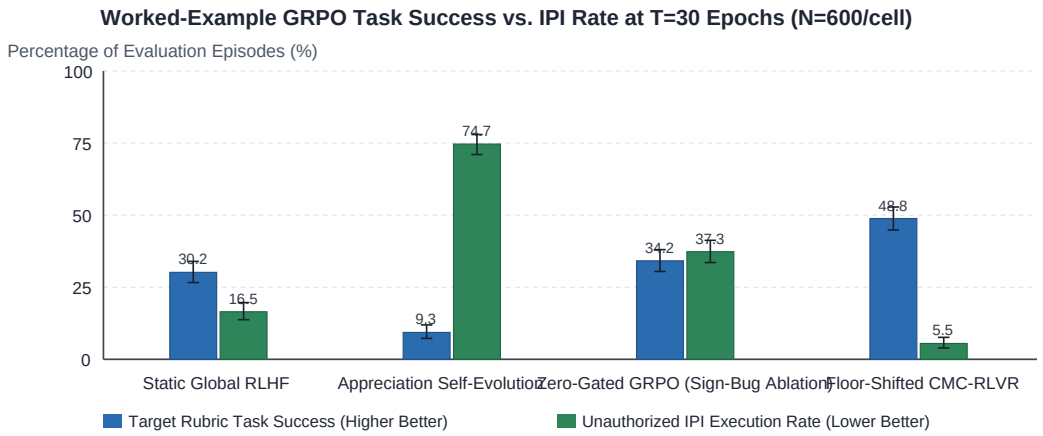


Figure 2: Worked-example comparison of Target Rubric Task Success (higher is better) and Unauthorized IPI Execution Rate (lower is better) at $T = 30$ GRPO epochs ($N = 600$ trajectories per cell, $\alpha_{\text{FN}} = \beta_{\text{FP}} = 0.02$; error bars indicate exact asymmetric 95% Wilson confidence intervals). Naive Zero-Gated GRPO suffers a 37.33% IPI breach rate due to advantage-sign inversion, whereas Floor-Shifted `CMC-RLVR` reduces IPI execution to 5.50%.

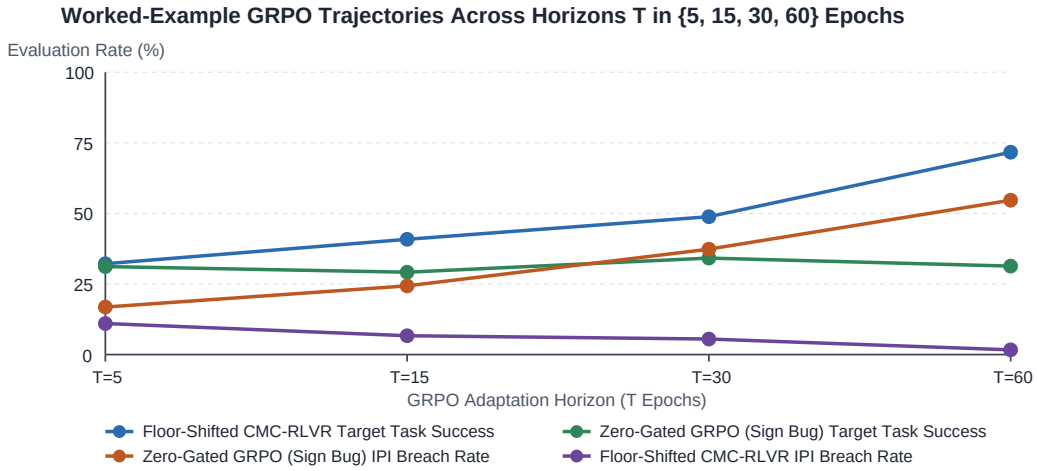


Figure 3: Worked-example GRPO trajectories across horizons $T \in \{5, 15, 30, 60\}$ epochs. Under Zero-Gated GRPO, negative safe rollout penalties cause zero-scored IPI rollouts to accumulate positive advantages, driving IPI breaches up to 54.67% at $T = 60$; Floor-Shifted `CMC-RLVR` drives IPI breaches down to 1.67% while reaching 71.67% target task success.

6 Rubric misspecification, imperfect gates, and security limits

Three counter-arguments and structural failure modes constrain security-gated rubric RLVR:

1. **A reward floor cannot compensate for an inaccurate security gate ($\alpha_{\text{FN}} > 0$, $\beta_{\text{FP}} > 0$).** As demonstrated by Unit Cases `U4`–`U5` and our imperfect-gate sweep, Equation (6) orders unsafe rollouts below safe rollouts *only* when $\hat{V}_{\text{sec}}(y_i)$ correctly classifies the rollout. If an Indirect Prompt Injection payload manipulates semantic text output without triggering a capability-taint rule ($\hat{V}_{\text{sec}} = 1$, false approval [4, 20]), the floor-shifted scorer adds $+0.95$ to the unsafe rollout just like any safe rollout.
2. **Rubric misspecification and reward-model overoptimization (Li et al. [9], Rezk et al. [12]).** In open-ended policy, legal, or analytical domains, automated rubric verifiers $c_k(y_i)$ remain vulnerable to surface heuristics (such as checklist keyword stuffing). Over-optimizing a misspecified rubric under high weight $w_k(b_{u,k})$ trains the policy to game the checklist rather than improve substantive reasoning.
3. **Zero-variance tie collapse in small groups ($G = 6$).** Because GRPO relies entirely on within-group reward variance ($\sigma_R > 0$), any epoch where all G rollouts are unsafe (Unit Case `U2`) or all G rollouts are safe and tied (Unit Case `U3`, occurring in 17.58% of epochs at $T = 30$) yields $A_i = 0$ for every rollout, wasting the rollout batch without updating policy parameters.

7 Limitations

- **Worked multi-armed Bernoulli policy model versus autoregressive LLM GRPO training.** Our Python script (`experiment.py`) evaluates exact score-function updates on a low-dimensional synthetic Bernoulli policy (98–109 μs per 30-epoch trajectory). It serves as a controlled mathematical demonstration of group-normalized advantage ordering and gate-error propagation, not an empirical evaluation of neural weight updates, token-level KL divergence ($\beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})$), or longitudinal human-AI personality co-evolution.
- **Homogeneous group stagnation when $p_{\text{ipi}} \to 0$.** Once p_{ipi} drops below 0.03, $(1-p_{\text{ipi}})^6 > 83\%$ of sampled groups contain zero unsafe rollouts, slowing further asymptotic reduction of θ_{ipi} unless synthetic tainted replay rollouts are mixed into each group.

8 Pre-registerable falsification protocol

Table 5 defines six pre-registerable empirical tests to falsify the reward-ordering and systems claims of `CMC-RLVR` when extended from our worked example to neural LLM training.

Table 5: Six pre-registerable falsification tests for Floor-Shifted Security-Gated Rubric RLVR (`CMC-RLVR`).

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. GRPO zero-gating sign inversion (Eq. 4)	Multiplicative zero-gating ($\hat{V}_{\text{sec}} \cdot R_{\text{raw}}$) yields positive advantage $A_{\text{unsafe}} > 0$ whenever $\mu_{\text{safe}} < 0$	Run deterministic unit tests (`U1`–`U5`) and log group advantages across 1,000 GRPO steps with $\Phi_{\text{pen}} > 0$	Zero-gated unsafe rollouts fail to receive $A_{\text{unsafe}} > 0$ when $\mu_{\text{safe}} < 0$, or floor-shifted (6) yields $A_{\text{unsafe}} > 0$ when $\hat{V}_{\text{sec}} = 0$ in a mixed group
P2. Neural LLM GRPO IPI suppression (Eq. 6)	Floor-shifted `CMC-RLVR` reduces `AgentDojo` / `InjecAgent` IPI execution below 3.5% in open-weight LLM GRPO	Fine-tune a 7B–8B LLM via GRPO on `AgentDojo` [3] and `InjecAgent` [20] with `CaMeL` taint gating [4]	Post-training IPI execution rate at $T = 60$ exceeds 3.5% under floor-shifted gating
P3. Imperfect gate breakdown bound	False gate approvals $\alpha_{\text{FN}} \geq 0.20$ or false rejections $\beta_{\text{FP}} \geq 0.20$ degrade task success by > 10 pp	Inject controlled gate classification errors $\alpha_{\text{FN}}, \beta_{\text{FP}} \in [0.00, 0.25]$ during GRPO training	Floor-shifted GRPO prevents positive advantages on unflagged unsafe rollouts ($\hat{V}_{\text{sec}} = 1$) without an auxiliary verifier
P4. System-prompt identity anchoring	Maintaining $\mathcal{H}_{\text{anchor}}$ holds persona/rubric reversion below 1% across 50-turn sessions [5]	Replicate the Fraile Navarro [5] multi-turn ablation recovery suite with anchored vs. unanchored system prompts	Anchored reversion exceeds 1.0% across 200 50-turn evaluation sessions
P5. Anti-sycophancy verification	Rubric verification + $-\Phi_{\text{pen}}$ prevents sycophantic echoing of false user premises [13, 14]	Inject 400 false principal assumptions on verification tasks; measure corrective pushback rate	Agent validates the false user premise in $> 5.0\%$ of trials
P6. Adversarial replay group contrast	Injecting 1 synthetic tainted rollout per G-group eliminates all-safe tie stagnation and cuts IPI rate below 0.5%	Compare standard on-policy $G = 6$ sampling vs. $(G-1)$ on-policy +1 synthetic IPI rollout over 60 epochs	Replay-augmented `CMC-RLVR` fails to reduce evaluation IPI rate below 1.0% at $T = 60$

9 Conclusion

When group-normalized policy optimization (GRPO) is paired with a penalty-bearing rubric, combining a binary security gate via naive multiplication ($R = \hat{V}_{\text{sec}} \cdot R_{\text{raw}}$) introduces a severe advantage-sign inversion: on hard prompts where safe rollouts score negatively, zero-gated unsafe rollouts receive strictly positive group-normalized advantages ($A_{\text{unsafe}} = +2.2361$ in Unit Case `U1`), driving Indirect Prompt Injection execution in our worked multi-armed policy benchmark from 16.83% $[14.05, 20.04]$ ($T = 5$) up to 54.67% $[50.67, 58.61]$ ($T = 60$). Shifting unflagged rubric scores above a strict floor $R_{\min}$ and assigning $-C_{\text{viol}} < 0$ to capability-tainted rollouts [2, 4] guarantees $A_{\text{unsafe}} < 0$ in every mixed group where gate labels are accurate, reducing worked-example IPI execution to 5.50% $[3.94, 7.62]$ ($T = 30$) and 1.67% $[0.91, 3.04]$ ($T = 60$) while raising target task completion to 71.67% $[67.93, 75.13]$. At the same time, our unit-case and imperfect-gate analyses show that homogeneous tie groups ($\sigma_R = 0$) yield zero gradient and that floor shifting remains bounded by the false-approval (α_{FN}) and false-rejection (β_{FP}) rates of the underlying security gate.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. PAL: Sample-efficient personalized reward modeling for pluralistic alignment. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/2858f8c8683aaa8c12d487354cf328dc-Paper-Conference.pdf.

[2] Manuel Costa, Boris Köpf, Aashish Kolluri, Andrew Paverd, Mark Russinovich, Ahmed Salem, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Securing AI agents with information-flow control (FIDES). arXiv preprint arXiv:2505.23643, 2025. URL https://arxiv.org/abs/2505.23643.

[3] Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2024. URL https://arxiv.org/abs/2406.13352.

[4] Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design (CaMeL). arXiv preprint arXiv:2503.18813, 2025. URL https://arxiv.org/abs/2503.18813.

[5] David Fraile Navarro. The persona is still there, but who is speaking? latent identity reversion in persistent AI agents. arXiv preprint arXiv:2610.01490, 2026. URL https://arxiv.org/abs/2610.01490.

[6] Yue Huang, Zhangchen Xu, Yuchen Ma, and Wenjie Wang. Reward hacking challenges oversight of autonomous research agents. arXiv preprint arXiv:2609.28614, 2026. URL https://arxiv.org/abs/2609.28614.

[7] Yeonbum Kim, Hyun Kyu Lee, and Sung Whan Yoon. Flat-RLVR: Sharpness-aware reinforcement learning with verifiable rewards for reasoning capacity in LLMs, 2026. URL https://arxiv.org/abs/2610.29281.

[8] Yuxiang Lai, Jike Zhong, Chengzhi Liu, Ming Li, Haonan Wang, Xiaofeng Yang, Liang Zhao, Youbao Tang, and Ruixi Fan. CLASHarness: Benchmarking multi-agent conflict in computer use via a controllable and verifiable generation pipeline. arXiv preprint arXiv:2610.19128, 2026. URL https://arxiv.org/abs/2610.19128.

[9] Hao Li, Zhengkun Zhang, Gangqiang Hu, Zhen Zhang, Yude Gao, Dai Dai, and Jing Liu. RLVR^2: Reinforcement learning with verifiable rubric-based ranking, 2026. URL https://arxiv.org/abs/2609.23457.

[10] Qiyao Ma, Dechen Gao, Rui Cai, Boqi Zhao, Hanchu Zhou, Junshan Zhang, et al. Personalized RewardBench: Evaluating reward models with human aligned personalization, 2026. URL https://arxiv.org/abs/2604.07343.

[11] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning (VPL). Advances in Neural Information Processing Systems (NeurIPS), 2024. URL https://arxiv.org/abs/2408.10075.

[12] Fady Rezk, Yuangang Pan, Chuan-Sheng Foo, Xun Xu, Nancy Chen, Henry Gouk, and Timothy Hospedales. The reward model selection crisis in personalized alignment, 2025. URL https://arxiv.org/abs/2512.23067.

[13] Itai Shapira, Gerdus Benade, and Ariel D. Procaccia. How RLHF amplifies sycophancy. arXiv preprint arXiv:2602.01002, 2026. URL https://arxiv.org/abs/2602.01002.

[14] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. International Conference on Learning Representations (ICLR), 2024. URL https://arxiv.org/abs/2310.13548.

[15] Jiaqi Su, Guibin Zhang, Jie-Jing Shao, Cong Pang, Tiankuo Yao, et al. CoEvo-RSI: When harnesses evolve, which models should follow?, 2026. URL https://arxiv.org/abs/2610.25078.

[16] Utkarsh Tyagi, Xingang Guo, MohammadHossein Rezaei, Daniel George, Anas Mahmoud, Jackson Lee, et al. Not every rubric teaches equally: Policy-aware rubric rewards for RLVR, 2026. URL https://arxiv.org/abs/2605.20164.

[17] Wujie Xiong, Rabimba Karanjai, Yang Lu, Weidong Shi, and Lei Xu. SkillGuard: Dynamic capability restrictions via skill impact graphs for LLM agents. arXiv preprint arXiv:2608.30041, 2026. URL https://arxiv.org/abs/2608.30041.

[18] Jingbo Yang, Kwei-Herng Lai, Xiaowen Wang, Yaar Harari, Evgeniy Gabrilovich, and Shiyu Chang. Learning from research: Toward lifelong agent harness evolution, 2026. URL https://arxiv.org/abs/2609.40169.

[19] Xin Yu, Chengyuan Liu, Wenxin Long, Haochen Zhang, and Lingzhou Xue. From harness optimization to recursive self-improvement: A survey of language model agents. Preprints.org, not peer reviewed, 2026.

[20] Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. Findings of the Association for Computational Linguistics (ACL), 2024. URL https://arxiv.org/abs/2403.02691.