FOOLED AT PROBE TIME, IGNORED AT TRAINING TIME: THE MECHANISM AND EXPLOITATION BOUNDARY OF VERIFIER SUSCEPTIBILITY IN RL

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM judges used as reward signals for RL can accept wrong answers on some tasks—a property called *verifier susceptibility*. The natural response is to filter susceptible tasks from training. We show why that response fails, locate the gradient dose at which exploitation begins, and confirm both results with pre-registered experiments. The mechanism: in high-correlation regimes (where rollout-time proxy–gold correlation is positive, 45/45 training conditions), the policy gradient tracks the true objective. Susceptible tasks are *structurally harmless*—the fooling completions that define susceptibility at probe time are not the ones reinforced by optimization. The boundary: a pre-registered fixed-learning-rate dose ladder shows exploitation appears once roughly 26% or more of training groups carry reward signal (tested doses 6–65%)—below 7%, even a planted backdoor is ignored entirely—and doubling group size cannot buy signal, because token emissions are task-correlated. This locates verifier exploitation on the gradient-dose axis. The predictions, confirmed: four pre-registered replications show susceptibility-informed filtering is never detectably better than random task removal—and under a naturally low-correlation Mistral judge, a pre-registered equivalence test passes on the primary estimand. Susceptibility is also real (r=0.638 cross-prompt, at the reliability ceiling) yet no tested text classifier detects it beyond a difficulty baseline. We release the twenty-minute pre-flight probe that determines which regime a task–verifier pair is in.

1 INTRODUCTION

Language agents are now post-trained with reinforcement learning on tasks whose reward comes from a *proxy verifier*: a code executor's test suite, an LLM judge, a rubric. Proxies can be gamed. Self-play against a reference-free judge raises the measured pass rate from 0.72 to 0.94 while true accuracy stays near 0.20 (Zhou, 2026); RL with verifiable rewards incentivizes exploiting the verifier's blind spots (Helff et al., 2026). Reward hacking is a central concern in agentic RL (Wang et al., 2026a).

A natural defensive instinct follows: if some tasks are *susceptible* — the proxy satisfiable by solutions that fail the true objective — identify them and remove them from the curriculum before training. Whether that instinct is warranted depends on *when* verifier susceptibility translates into training-time harm. This paper characterizes that transition in one controlled testbed: we measure susceptibility, identify the regime where it is and is not toxic, and confirm the resulting predictions with pre-registered experiments.

Testbed. GSM8K word problems (Cobbe et al., 2021); R_p = a 7B judge's accept; R_g = numeric match to the gold answer (held out, never in any loss). Policy: Qwen2.5-0.5B, SFT-warmed, trained GRPO-style (Shao et al., 2024) with per-step logging of both rewards. Behavioral susceptibility is measured by probing: k=6 completions from the SFT policy; susceptible = acceptance rate exceeds gold-match rate by a margin.

The mechanism: high-correlation regimes are structurally non-toxic. Susceptibility exists and is shared across verifier prompts (cross-prompt conditional-FPR r=0.638, at the reliability ceiling;

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.

§3) — but rollout-time $\mathrm{corr}(R_p, R_g)$ is positive in 45/45 training conditions ($+0.33$ to $+0.81$, median $+0.54$): the policy gradient tracks the true objective, not the verifier's blind spots (§5.4). In this regime, the fooling completions that define susceptibility at probe time are not the ones reinforced by optimization.

The boundary: a dose ladder identifies when exploitation occurs. A pre-registered fixed-lr experiment (§5.4) shows exploitation is monotone in the *observed* fraction of gradient-bearing groups: tokens buying a 6–7% dose never amplify (doubling group size buys no dose — emissions are task-correlated); tokens buying 26–65% fully exploit (3/3 seeds, R_g collapses); $4\times$ lr exploits the low-dose numeral. The healthy-learning window is proof against single-digit-dose failure modes — which is what the planted-backdoor control tested.

The confirmed prediction: filtering cannot help where nothing is toxic. If susceptibility is non-toxic in this regime, filtering susceptible tasks cannot improve the true objective. Four pre-registered replications confirm this (§5.3): a step-matched replication (random removal free), a four-draw replication (single-draw contrast scale 0.018), a construct-aligned blind-bank replication (untouched-slice harm -0.022 where the contrast is real), and a two-sided replication under the naturally low-correlation Mistral verifier where pre-registered equivalence **passes** ($+0.002$, CI $[-0.012, +0.015]$). An exploratory 45-condition sweep initially suggested a -0.029 veto cost; we decompose it into an update-count artifact and show it does not survive correction. Susceptibility is also undetected in stable replication by every tested text predictor (chance AUC 0.483 on the gap label; the composite-label signal is fit-unstable and sits at or below a difficulty baseline; §4). We also *retract* our own earlier claim of 96% recall for this predictor: the evaluation had oversampled positives *before* splitting, duplicating 24 of 52 validation tasks into training — memorization, not prediction (§4.2).

Where the danger actually lives. The exploitable regime requires rollout-time proxy–gold correlation to be *low* under the policy's own rollouts. Five probed verifier variants (§6) all retain $\mathrm{corr} \geq +0.57$ with usable signal: reliability *rebounds* against prompt-weakening, self-consistency *amplifies* it. The low-correlation cell on math is reachable by explicit dilution or by the naturally low-discrimination Mistral judge (completion-level rollout correlation ~ 0.24; filtering there is equivalent to random, App. G), and an 18-run sweep at probe correlation 0.10 still shows neither curriculum effect nor divergence (App. F). What remains untested: plausible-wrong-endemic domains (open-ended rubric grading, self-play review), which the released probe screens.

Contributions. (1) A mechanistic characterization: rollout-time proxy–gold correlation is the diagnostic that separates non-toxic from potentially exploitable regimes; the dose ladder identifies the boundary in observed gradient-dose. (2) Four pre-registered confirmations that susceptibility-informed filtering does not outperform blind random replacement — including a passing equivalence test under the naturally low-correlation Mistral verifier. (3) A probe-based measurement of task–verifier susceptibility at 2,000-task scale, real and shared across verifier prompts, with a leak-free negative on ex-ante text prediction and a forensic post-mortem of the duplication leak that briefly convinced us otherwise. (4) A verifier-correlation spectrum over five proxy designs plus the dilution frontier, and a reusable twenty-minute pre-flight screen.

2 Related work

Reward hacking in LLM RL. Reward hacking is a classical AI-safety concern (Amodei et al., 2016); its proxy–true divergence under RL is quantified for learned reward models (Gao et al., 2023), and Pan et al. (2022) show that *even* proxies correlated with the true reward can still produce hacking under optimization — motivating the task-level diagnostic we develop. Shao et al. (2025) report the complement at scale: Qwen policies improve on math under random or mislabeled RL rewards. Process supervision (Lightman et al., 2024) outperforms outcome supervision in its evaluated MATH setting. Wang et al. (2026a) systematize detection and mitigation; ex-ante pre-training screening — what this paper evaluates, reporting negatively for text prediction — is not their focus. Helff et al. (2026) show RL with verifiable rewards incentivizes gaming the verifier's blind spots in inductive-reasoning tasks; Zhou (2026) document judge self-play inflating pass rate without accuracy.

Closest prior. Rajan (2026) audit code test suites pre-training via exploit patches, then harden using a gold-sanity gate; we ask the ex-ante question without gold. Zhang (2026) provide the closest prior training-time contrast (leaky reward suites; null downstream effect); ours is a health-verified replication in another domain, adding rollout-time correlation as mediator. Zhou et al. (2026) and Bae et al. (2026) compute per-task priors (learnability, difficulty) for LLM RL post-training; hackability is distinct, and our question is whether it is available without rollouts. Wang et al. (2026b) and Akter et al. (2026) detect gaming *online* (post-training-onset); Skalse et al. (2022) treat formally when optimal policies under misaligned reward differ — our gap is a task-level cousin.

Train/test leakage. Duplication-induced leakage is a recognized failure mode (Kapoor & Narayanan, 2023). Our post-mortem (§4.2) is a sharp instantiation of a coming setting: LM classifiers on *behaviorally labeled* small datasets, where rebalancing silently converts a row-level split into a task-level leak while internal validation stays green.

Verifier adequacy in software engineering. "Does this verifier admit wrong-but-plausible solutions?" is the LLM-judge cousin of test-suite adequacy: mutation testing estimates a suite's fault-detection power from killed mutants (Jia & Harman, 2011); weak oracles in program synthesis admit plausible-but-wrong programs; execution-based filtering with behavioral clustering (Li et al., 2022) is a self-consistency-flavored verifier. Our spectrum probe (§6) is the analogue of a cheap mutation-score estimate for LLM judges; self-consistency (Wang et al., 2023) is one family we probe; judge-reliability biases: Zheng et al. (2023).

3 Measuring task–verifier hackability

Setup. Each task t carries a proxy reward R_p and a ground-truth reward R_g. On GSM8K (Cobbe et al., 2021), R_p is a 7B LLM judge (Qwen2.5-7B-Instruct; Qwen Team, 2024) giving a binary accept; R_g is numeric match of the extracted final answer to the gold (tolerance 10^{-2}), used *only* for evaluation and labeling — never in the GRPO-style loss or advantage; the SFT warm-start, however, is gold-supervised (overlap audit in App. A). The policy is Qwen2.5-0.5B (Qwen Team, 2024) SFT-warmed on GSM8K reference solutions, then trained with *single-step group-normalized REINFORCE* (Shao et al., 2024): within each group of 4 rollouts the advantage is $(R_i - \bar{R})/\hat{\sigma}_R$, one gradient step per group, no importance ratio, clipping, or reference KL ("GRPO-style"; we use the term only in this sense).

Behavioral hackability. For each task we sample $k{=}6$ completions from the SFT policy (temperature 0.8) and define the *gap*

$$\mathrm{gap}(t) \;=\; \overline{R_p}(t) - \overline{R_g}(t),$$

the judge's acceptance rate minus the gold-match rate over the probes. A task is *hackable* iff $\mathrm{gap}(t) \geq \frac{1}{6}$: the judge rewards solutions that fail ground truth, for completions the current policy produces. The label is behavioral (defined by the policy–verifier interaction) and thus a moving target for any predictor.

A leak-free labeling protocol. Behavioral labels on small datasets invite a specific corruption: duplication-based rebalancing applied *before* the train/validation split converts a row-level split into a task-level leak (we commit and dissect this error in §4.2). The protocol used for every number in this paper: **split by task first** (shuffle the 2,000 labeled tasks, seed 0; first 400 validation, remaining 1,600 training), **assert zero overlap** (the script aborts on any task-id intersection across the split), **resample after** (positives oversampled within the training partition only; 1,055 of 2,110 rows).

Hackability is real and shared across verifier prompts. Relabeling 2,000 tasks under the numeric judge yields 677 hackable tasks (33.9%); the per-task gap ranges up to 1.0 with mean 0.041. To test whether this is a property of the task–verifier pair rather than one prompt's idiosyncrasy, we relabel a held-out subset (40 tasks) under a *rubric* prompt of the same judge (instructed to grade plausibility and reasoning quality without verifying arithmetic). The two label sets agree on 0.750 of tasks (chance 0.526; $\kappa{=}0.47$, 95% CI $\approx[0.19, 0.75]$); their *continuous* per-task gaps correlate at $r{=}0.45$ ($p{\approx}0.003$), and the numeric binary label predicts the rubric gap (point-biserial $r{=}0.40$). Independent probe sets would be uncorrelated under pure sampling noise, so the noisy label measures

a real task-level property — making the prediction question well-posed. (Sharing is established across *prompts* of one judge family; a cross-*model* probe lands in a degenerate verifier regime — §6.)

The rate alone is noise-compatible; the structure is not. A null model with no task-level variation — i.i.d. judge slips, k=6 probes, threshold $\frac{1}{6}$ — is parametrization-sensitive at this scale: plugging in per-completion rates pooled from the 2,000-task labeling (accept 0.384, gold 0.343) with the matched probe's accept-on-correct/accept-on-wrong discrimination (0.791 vs. 0.144) yields 37.3% hackable (dynamic programming; task-resampling bootstrap sd 1.1%), while sourcing the conditional rates from the 100-task matched probe directly yields $\sim$33%. The observed 33.9% sits inside that 33–37% band: the hackable *rate* carries no evidence either way, and any real component must manifest in the correlation structure — where both cross-judge statistics above are non-zero and i.i.d. noise predicts zero.

The conditional construct (revised measurement). The gap conflates false acceptance with task difficulty: a harder task offers more wrong completions and more false-positive opportunities, so acceptance-minus-gold can rank tasks by difficulty rather than verifier vulnerability. We therefore re-measure hackability as the conditional false-positive susceptibility

$$\mathrm{FPR}_t = P(R_p{=}1 \mid R_g{=}0,\, t),$$

estimated by adaptive probing (sample until $\geq$16 gold-wrong completions, cap 48; mean 14.9 per task) with beta–binomial posteriors, on 300 tasks, under both judge prompts. Three findings. (i) The conditional picture is much sparser than the gap suggested: mean FPR 0.157 (numeric) / 0.211 (rubric); only 15/300 resp. 30/300 tasks have FPR $\geq$ 0.5. (ii) The label is reliable: split-half FPR correlation 0.455/0.463. (iii) The prompt-shared structure is robust and not a difficulty artifact: cross-prompt FPR correlation $r = 0.876$, essentially unchanged controlling per-task gold rate ($r = 0.874$; raw FPR–difficulty correlations only $+0.17$/$+0.14$). **One design disclosure, and its repair:** both prompts scored the *same* completions, so 0.876 is an upper bound; Spearman–Brown gives a full-estimate reliability of $\sim$0.63 — the ceiling for two *independently sampled* measurements under shared latent structure. A fresh-generation-seed rubric relabel lands at $r = 0.638$ (0.613 controlling gold rate) — at the ceiling, consistent with a disattenuated cross-prompt correlation near 1.0: *the observed correlation is compatible with a strongly shared latent component; each measurement is reliability-limited at* $\sim$0.64. The veto experiments predate this construct and use gap-based labels; the FPR-based rerun is §D.

4 Can hackability be predicted ex ante?

The economically interesting version of the prediction question: *before spending any probe rollouts*, decide from the (task + verifier) text whether the pair is hackable. We test it with a 0.5B sequence classifier (two-class head) on the serialized text, trained on the 1,600-task training partition under the leak-free protocol of §3.

4.1 Leak-free result: chance

Table 2 (App. B) gives the outcome on the 400 held-out tasks (132 hackable; base rate 0.330). The classifier reaches precision 0.323, recall 0.242, F1 0.277 — **AUC 0.483, AP 0.322: chance**. A logistic regression on seven shallow surface features scores AUC 0.550; the 0.5B LM does not beat it. Nor does the 7B judge model prompted eight-shot: F1 0.392 / AUC 0.529 (calibrated prevalence, no discrimination). The negative is about the information in the (task + verifier) text at this probe depth, not the reader's size; input ablations agree (question-only 0.511; verifier-only 0.500). One dose caveat: the six-probe gap labels are themselves noisy, so a perfect latent-task predictor would be attenuated well below 1.0 AUC — "no detectable signal at this label dose", not a bound on the latent property. Scoped to what was run — one fixed verifier prompt, one task split, one classifier fit, one seed — the tested procedures **detect no linear signal in the task text beyond difficulty**; whether a stronger reader with repeated fits would, the two-fit composite experiment leaves open (signal there is fit-unstable and difficulty-bound).

Label dose matters: the high-reliability target is fit-unstable. Repeating prediction on the composite conditional-FPR label ($\alpha{\approx}0.86$; 3 task-level splits, 60 test each; two classifier fits per split, tie-corrected AUC): per-split AUCs are unstable across fits (seed 0: 0.748 then 0.535; seed 1: $0.654/0.662$; seed 2: 0.558 twice), run means 0.585–0.653, one CI excludes chance — none excludes a gold-rate-only baseline at 0.61. At most weak, difficulty-mediated signal where the gap label gave pure chance; the single-fit headline does not replicate across training runs.

4.2 Post-mortem: how we were fooled, and why it stayed green

We first reported this predictor with recall 0.963 / precision 0.788 / F1 0.867 on a 200-task study. Those numbers were an artifact (App. B replays the forensic): the class-imbalanced labels (67 positives) were oversampled by duplication *before* a row-level 80/20 split, placing duplicates of **24 of the 52 validation tasks into the training set**. The tell: on a fresh 20% slice of the corrupted protocol the original checkpoint scores **AUC** $=$ **AP** $= 1.000$ with *question-only* AUC 0.970 — memorization where the duplication put it — while on the genuinely task-disjoint 400-task evaluation it reaches only AUC 0.596, around the seven-feature baseline (0.550): the leak manufactured essentially all of the performance. Real behavioral signals are not perfectly separable by a 0.5B model; memorized duplicates are.

Why nothing complained, and the rule. The corruption is invisible to every guard that watched the run: validation loss and F1 on the corrupted split looked healthy throughout, and a retrain's internal acceptance gate reported recall 0.863 / precision 0.759 on a slice from the same oversampled file — green lights, all downstream of the same broken split. Only an *external*, task-level-disjoint evaluation — or an AUC of exactly 1.000 — exposes it. The rule: in any pipeline whose rows derive from a smaller set of distinct entities (tasks, users, prompts), *split by entity first, resample after, and assert zero entity overlap* (Figure 3: $0.96 \rightarrow 1.000 \rightarrow 0.483$).

5 Would acting on it help? The veto sweep

Prediction having failed, the stronger test uses labels as close to oracle as we can construct: if hackable tasks poison training, removing them should help — and a probe-based veto is itself a practical, rollout-budgeted intervention.

5.1 Sweep protocol

Curricula. For hackable fraction $h \in \{0, 0.25, 0.5, 0.75, 1.0\}$ the pool is the 100 tasks composed of the $100h$ highest-gap and $100(1{-}h)$ lowest-gap tasks of the 200-task behavioral label study of §3. *Unfiltered* trains on the pool; *veto* removes every task the **critic predicts** hackable ($\hat{p}_{\text{hack}} \geq 0.5$; the 0.5B classifier of §4 — its predictions, unlike its leaked validation metrics, are legitimate inputs); *random* removes the same *count* at random. Table 1 (App. A) audits every pool: at $h{=}1$ the critic removes 66 of the 67 true positives; at $h{=}0$ it false-flags 5 of 100. The near-perfect recall is label memorization (leak-free AUC 0.483, §4.1) — what makes it a legitimate near-oracle veto constructor. Curriculum sizes are 100 vs. $\{95, 70, 47, 32, 32\}$, so at high h veto-vs-unfiltered conflates *which* with *how many*; the size-matched comparison is veto-vs-random. Dose caveat: the labels are half noise (split-half 0.455/0.463) — "no detectable benefit at this label dose"; a composite-label arm at $\alpha{\approx}0.86$ closes it (App. G).

Health verification. Under-training and collapse both make nulls uninterpretable, so we calibrated the learning rate and logged per-step R_p *and* R_g in every run: at lr 10^{-6} the policy does not learn; at $2{\times}10^{-5}$ training is destructive; at $5{\times}10^{-6}$ it verifiably learns (R_p $0.20 \rightarrow 0.34$ in calibration, $0.3 \rightarrow 0.5{+}$ across the sweep). **One disclosure:** run past the sweep horizon, the calibration condition showed a late Goodhart signature (R_g $0.33 \rightarrow 0.24$, R_p held) — none of the 45 conditions within 5 epochs; the 15/25-epoch replication (§G) does not reproduce it at severity, and the arm degrading under horizon is the *filtered* one. We run 5 epochs, 3 seeds, 5 fractions, 3 curricula (**45 conditions**), all re-scored at 600 tasks (App. E).

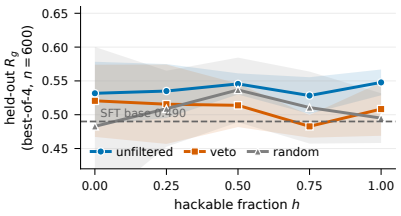


Figure 1: Ground-truth performance vs. enrichment level h (pool composition: $100h$ highest-gap + $100(1-h)$ lowest-gap tasks; 600-task evaluation; 95% CI bands, 3 seeds; dashed = SFT base 0.490; original harness). Veto tracks random removal, not unfiltered; enriched training does not depress ground truth.

5.2 Results under the original harness: veto tracks random removal

Figure 1 and Table 5 (App. E), at the 600-task evaluation, under the harness as originally run: (i) the veto's one detectable effect is harmful ($\Delta(\text{v}-\text{u}) = -0.029$ $[-0.043, -0.016]$; 12/15 pairs negative); (ii) as run, veto tracks random removal ($\Delta(\text{v}-\text{r}) = +0.001$) while $\Delta(\text{r}-\text{u}) = -0.031$: any removal appeared to cost $\sim$0.03; (iii) enriched training does not poison (h=1$-h$=0 $= +0.016$ $[-0.043, +0.075]$); (iv) a size- and difficulty-matched replication confirms this (68% of run noise is evaluation; §D).

Two defects qualify that reading; §5.3 corrects both: the arms were not step-matched (unfiltered 500 task-iterations, veto/random $\sim$160), and the shared warm-start labeled padding as EOS (benign for contrasts; App. A).

5.3 Corrected replications: from update-count artifact to a bounded null

Four pre-registered replications complete the design (plans frozen before launch; Appendix G). **Replication 1** (10 seeds $\times$ $h \in \{0, 1\}$ $\times$ {unfiltered, veto, random} plus a second draw; 70 runs) trains from a label-masked warm-start with a hard cap of 500 task-iterations — step-matched where the sweep ran $500/160/160$. It resolves the sweep's apparent removal cost: random removal is *free* ($+0.003$ $[-0.009, +0.015]$) — the "volume cost" was an update-count artifact — with the veto's single-draw contrast at -0.016.

Replication 2, the confirmatory stage (10 seeds, composite conditional-FPR labels $\alpha \approx 0.86$, disjoint warm-start, difficulty-matched replacement, *four* random-replacement draws; 80 runs), measures what a single draw cannot: the single-draw curriculum-contrast scale — within-seed sd across the four draws 0.018 vs. between-seed sd of the four-draw mean 0.007. Two uses, kept distinct: the scale calibrates single-draw contrasts — Replication 1's -0.016 and the composite arm's -0.033 (App. G) read against it carry no signal — while for a targeted curriculum, whose deviation need not be exchangeable with random draws, the test of record is the seed-paired contrast itself. The pre-registered tests deliver it: random replacement free ($\Delta(\text{u}-\bar{r}) = +0.002$, CI $[-0.005, +0.010]$ — the *pre-registered* ± 0.02 equivalence **passes**); directional matched-vs-random fails ($+0.004$, t_9=0.68); an *exploratory* TOST on the same margin also passes ($[-0.007, +0.016]$). Null at h=0 and pass@1; healthy in 80/80.

Replication 3 closes the loop on the construct (pre-registered; App. G): *blind-bank* replenishment — both arms draw difficulty-matched replacements from the full non-pool task set, differing only in *which* pool tasks are removed — across two prevalence pools (MID, 50/100 flagged, exposure 0.148 vs. 0.251; H1, 86/100) plus a τ=0.4 budget arm; 10 seeds, 100 runs. The two pools *bracket zero* on the adapted evaluation set: MID $\Delta(\text{m}-\bar{r}) = -0.015$ (t_9=-1.76, 8/10 negative), H1 $+0.016$ (t_9=$+1.59$); the pre-registered ± 0.02 equivalence fails at the margin ($[-0.030, +0.001]$). A one-shot evaluation of all 100 checkpoints on a *never-touched* slice (tasks 600–1200; confirmatory-orientation check, post-hoc to the pre-registration; seed-level t_9) resolves the flip: MID strengthens to -0.022 $[-0.036, -0.008]$ (t_9=-3.58; pass@1 -0.020, t_9=-4.07), H1's positive collapses to -0.004, the τ=0.4 budget stays null ($+0.007$) — the positive cell was evaluation-set noise, random replacement remains free in every pool, the harm concentrates where the exposure contrast is real. *Never detectably better; harm where the intervention bites.*

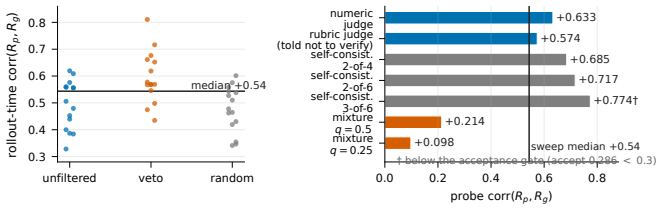


Figure 2: **Left:** rollout-time $\mathrm{corr}(R_p, R_g)$ over per-step training rollouts, all 45 conditions — positive everywhere ($+0.33$ to $+0.81$, median $+0.54$). **Right:** the verifier spectrum (§6): every *prompted* variant that retains training signal sits at corr $\geq +0.57$; only the diluted mixtures reach below it.

Replication 4 runs the same design under a *naturally* low-correlation verifier — the one regime where our own logic allows filtering to help (pre-registered two-sided; App. G): Mistral-7B is the training *reward* and its own continuous conditional-FPR ranking the label (τ=0.5; MID exposure 0.86 vs. 0.33), blind banks, 10 seeds, 80 runs. Verdict: dead zero, with equivalence *established* — matched vs. random replacement $+0.0019$, 90% CI $[-0.0115, +0.0154]$: the pre-registered ± 0.02 TOST **passes on the title estimand**; the help hypothesis fails flat (t=0.27), random replacement is free; the pass@1 mirror holds the margin with a small negative point (-0.0066, CI $[-0.0121, -0.0012]$ — equivalence there too, with a mild harm-direction estimate, 6/10 seeds negative). *The null survives its most natural adversarial instantiation.*

5.4 Mechanism: rollout-time correlation, measured

Why doesn't training on hackable tasks hurt? The data refute the natural premise — that the gradient on those tasks points away from ground truth: $\mathrm{corr}(R_p, R_g)$ over the per-step training rollouts of every condition (Figure 2, left) is $+0.33$ to $+0.81$ (median $+0.54$), positive in 45/45. Trajectory shapes agree (R_p/R_g rise together in all 15 unfiltered runs); no sweep condition exhibits the Goodhart signature (R_p up, R_g down; Gao et al., 2023) within 5 epochs — the one exception emerged *past* that horizon (calibration, §5.1); a milder softening recurs at 25 epochs (App. G). We treat this correlation as a *candidate diagnostic*, not an established cause: both tested low-correlation regimes also produce no filtering benefit, and the dose ladder identifies exploitation controls only for synthetic tokens. The behavioral gap that defines hackability at probe time exists — but under optimization the fooling completions are not the ones reinforced. *Probe-time hackability is decoupled from training-time reward hacking in this testbed* — and, by the spectrum of §6, in every high-correlation verifier variant we could construct on it.

Group-level advantage alignment. The global correlation could mask misalignment inside GRPO-style groups, where the advantage acts: instrumenting two hackability-enriched runs (two seeds, 200 steps), the **within-group advantage correlation is** $+0.675$/$+0.727$, accepted-but-wrong completions are 11.6–13.5% of samples without systematic positive advantage — extended to all 18 matched-replication runs (0.611–0.748; §D): proxy and gold advantages point the same way where the policy gradient is computed.

Diluted verifier: the low-correlation regime, finally entered. Prompt-weakening cannot reach a low-correlation proxy on math (§6); explicit dilution can: mixing the judge's accept with a fair coin at probability $1-q$ yields probe correlation $+0.10$ at q=0.25 (acceptance 0.46). We swept it as 18 runs (3 seeds $\times$ $h \in \{0, 1\}$ $\times$ {unfiltered, matched, randomized}; App. F). Training stays healthy (gradient density 85–89%; in-loop advantage correlation $+0.07$–$+0.22$ vs. 0.61–0.75 undiluted, so the dilution reaches the reward) and R_p/R_g still rise together on average (App. F); the hackability curriculum remains null (Δ(m−u) $= -0.036$ / -0.024, both CIs spanning zero; matched vs. randomized: non-detection, CIs span zero). Dilution costs 0.04–0.06 R_g — but even at rollout-time correlation 0.10, probe-time hackability does not become a training-time trap: *the decoupling survives its most adversarial in-domain test*, and the planted-loophole control below pushes further (App. G).

The exploitation gate, tested at fixed lr. Why can training ignore a planted backdoor? A trigger-only verifier (token present $\Rightarrow R_p{=}1$) makes the question quantitative, and a pre-registered fixed-lr dose ladder (App. G) measures it in the observed fraction of gradient-bearing groups. At the calibrated lr, "1000" buys a 6% dose and never amplifies; doubling group size buys no dose (7% observed at $k{=}8$ vs. 48% independence-predicted) — numeral emissions are task-correlated. "Thus" (57%) and "therefore" (65%) fully exploit in 3/3 seeds each (R_g collapses); "answer:" saturates at 26% and exploits to 99%; at $4\times$ lr the numeral too is exploited. Exploitation tracks the observed dose monotonically; group size cannot manufacture dose. The healthy-learning window is proof against single-digit-dose tokens — what P1-A2 tested.

6 The verifier-correlation spectrum and a pre-flight probe

The decoupling mechanism points to *rollout-time* proxy–gold correlation — not the probe gap — as the leading candidate variable. If drift requires it to be low: *can we build a low-correlation verifier that still trains?*

Probe protocol. Sample completions from the SFT policy (100 tasks at $k{=}4/k{=}6$; 150 for self-consistency), score with both the candidate verifier (R_p) and gold (R_g), reporting acceptance, completion- and task-level $\mathrm{corr}(R_p, R_g)$, and acceptance on gold-correct vs. gold-wrong completions; twenty minutes on one GPU. Gate: acceptance ≥ 0.3, correlation < 0.45.

Two rebound findings. Table 3 (App. C): (i) *Verifier reliability rebounds against prompting.* The judge instructed to grade plausibility *without checking arithmetic* still rejects 79% of gold-wrong completions; its correlation drops only from $+0.63$ to $+0.57$. (ii) *Self-consistency amplifies reliability.* Accepting the modal (or threshold-consensus) answer among k policy samples — a standard cheap verifier — yields correlations $+0.69$ to $+0.77$: repeated answers mark the problems the policy understands; *common* errors are rare at this scale.

Cross-model: the degenerate cell, measured. A sixth probe swaps the judge *model*: Mistral-7B-Instruct-v0.3 under the same prompt accepts 68% of gold-wrong completions (92% of gold-correct; margin 0.24 vs. Qwen's 0.59), flags 282/300 tasks at FPR ≥ 0.3, and its per-task FPR correlates with Qwen's at only $r = 0.321$ (within-model: $0.638/0.876$, §3). A naturally low-discrimination verifier — not unusable (cross-family arm App. G; filtering replication App. G) — but degenerate as a *susceptibility-labeling instrument*: near-zero discrimination compresses the FPR scale until the label barely applies.

Consequence. Every verifier family *we constructed by prompting or sampling* on numeric math keeps usable training signal at $\mathrm{corr} \geq +0.57$; the low-correlation training regime is entered in two ways — explicit dilution (sweep null, §5.4, App. F) and the naturally low-discrimination Mistral cell above (completion-level rollout correlation ~ 0.24; its filtering replication, App. G, shows equivalence there too). What remains untested: plausible-wrong-*endemic* domains; the probe screens candidates in twenty minutes.

7 Limitations and conclusion

Limitations. (1) One policy scale and family (Qwen2.5-0.5B), two judge families (Qwen, Mistral), one domain; the mechanistic characterization is verified in this testbed and its transfer to other regimes is a hypothesis. (2) Plausible-wrong-*endemic* domains remain untested (the Mistral cell is tested, App. G); decoupling holds to 25 epochs without held-out Goodhart; the dose ladder identifies the boundary for synthetic tokens, not for natural susceptibility. (3) The prediction negative is bounded by readers, probe depth, and label dose; the label is policy-relative. (4) The gold-supervised warm-start makes the enrichment *benefit* an anchoring artifact. *In sum:* in high-correlation RL regimes, verifier susceptibility is real, prompt-shared, and rollout-measurable — yet undetected in stable replication by every tested text predictor and structurally non-toxic under optimization: the policy gradient tracks the true objective, filtering cannot improve what is not broken, and the exploitation boundary is identified in observed gradient-dose.

REPRODUCIBILITY STATEMENT

All experiments use public models (Qwen2.5-0.5B/7B-Instruct, Mistral-7B-Instruct-v0.3) and the public GSM8K dataset. The complete experimental record is machine-verifiable: the study is an exploratory sweep followed by pre-registered replications, and each replication follows an analysis plan frozen in a timestamped file before launch, with all deviations logged (App. A); all 480 curriculum-trained checkpoints are evaluated on a fixed 600-task protocol (the 14 trigger-only diagnostic runs use the trigger-rate protocol, App. G); and an automated acceptance contract (29 assertions) re-derives every headline number from the raw per-task JSON artifacts on each build. Training code, evaluation scripts, curriculum builders, pre-registration documents, analysis scripts, and the raw result files are included in the supplementary material; the split-asserting pipeline (evaluation tasks disjoint from every training-side pool; curriculum pools intentionally drawn from the labeled pool; predictor train/validation split by task before resampling) is part of the released scripts.

AI USE STATEMENT

AI-assisted tools were used in this research and their use is disclosed here. Code implementation, experimental orchestration, and statistical analysis were developed with the assistance of LLM-based coding agents (Claude Code, OpenAI Codex CLI) under direct human specification and review; all scientific decisions (hypothesis selection, pre-registration content, interpretation, and final claims) were made by the authors. LLM-based reviewers (GPT-5.6-Sol via the Codex CLI, and Claude) were used to conduct adversarial cold reviews of manuscript drafts during revision; every weakness they identified was verified against raw result files before any change was made. All measurements are produced by the GPU experiments described in the paper (the policy, judges, and datasets under study); no gold labels or reported numbers were produced by AI tools, and the automated acceptance contract re-derives every headline number from raw artifacts deterministically, without AI involvement. The manuscript text was drafted and revised by the authors with AI-assisted editing for compression and formatting.

REFERENCES

Sanjeda Akter, Ibne Farabi Shihab, and Anuj Sharma. Detecting proxy gaming in RL and LLM alignment via evaluator stress tests. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. arXiv:2507.05619.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Sanghwan Bae, Jiwoo Hong, Min Young Lee, Hanbyul Kim, JeongYeon Nam, and Donghyun Kwak. Online difficulty filtering for reasoning oriented reinforcement learning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Volume 1: Long Papers*, pp. 700–719, 2026. DOI:10.18653/v1/2026.eacl-long.30; arXiv:2504.03380.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, 2023.

Lukas Helff, Quentin Delfosse, David Steinmann, Ruben Härle, Hikaru Shindo, Patrick Schramowski, Wolfgang Stammer, Kristian Kersting, and Felix Friedrich. LLMs gaming verifiers: RLVR can lead to reward hacking. In *ICLR 2026 Workshop on Logical Reasoning of Large Language Models*, 2026. arXiv:2604.15149.

Yue Jia and Mark Harman. An analysis and survey of the development of mutation testing. *IEEE Transactions on Software Engineering*, 37(5):649–678, 2011. DOI:10.1109/TSE.2010.62.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. DOI:10.1016/j.patter.2023.100804.

Yujia Li, David Choi, Junyoung Chung, et al. Competition-level code generation with AlphaCode. *Science*, 378(6624):1092–1097, 2022. DOI:10.1126/science.abq1158.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2305.20050.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations (ICLR)*, 2022.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Shreshth Rajan. Auditing reward hackability in code RL training environments. *arXiv preprint arXiv:2606.16062*, 2026.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in RLVR. *arXiv preprint arXiv:2506.10947*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2209.13085.

Xiaohua Wang, Muzhao Tian, Yuqi Zeng, Zisu Huang, Jiakang Yuan, et al. Reward hacking in the era of large models: Mechanisms, emergent misalignment, challenges. *arXiv preprint arXiv:2604.13602*, 2026a.

Xinpeng Wang, Nitish Joshi, Barbara Plank, Rico Angell, and He He. Is it thinking or cheating? detecting implicit reward hacking by measuring reasoning effort. In *International Conference on Learning Representations (ICLR)*, 2026b. arXiv:2510.01367.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Chuyifei Zhang. When the reward suite is leaky: A preregistered causal contrast of natural verifier false positives in RLVR. *arXiv preprint arXiv:2607.11022*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Chenyu Zhou. More convincing, not more correct: Self-play reward hacking of reference-free LLM judges. *arXiv preprint arXiv:2607.05904*, 2026.

Ting Zhou, Zhenqing Ling, Daoyuan Chen, Qianli Shen, Yilun Huang, Ying Shen, and Yaliang Li. Beyond solvability: Task learnability as a static prior for LLM RL post-training. *arXiv preprint arXiv:2608.09217*, 2026.

A EXPERIMENTAL DETAILS

Judge. Qwen2.5-7B-Instruct, greedy 8-token generation over the chat template; the numeric prompt is: *"You are a strict math grader. Decide if the proposed answer is correct. / Question: {q} / Proposed answer: {a} / Reply with exactly one word: yes or no."* The rubric variant (§6) replaces it with: *"You are reviewing a student's math homework. Judge whether the solution looks like competent work: Is the reasoning coherent and well-structured? Does the method fit the problem? Does the stated final answer seem plausible for this problem? IMPORTANT: Do NOT try to verify the arithmetic yourself; grade on overall quality and plausibility only. / Question: {q} / Proposed answer: {a} / Does this look like competent work? Reply with exactly one word: yes or no."* Truncation at 1,800 tokens, response prefix at 800 characters; the answer is parsed as accept iff "yes" appears in the 8 generated tokens.

Policy and training. Qwen2.5-0.5B, SFT-warmed 2 epochs on GSM8K reference solutions; single-step group-normalized REINFORCE with group size 4, temperature 0.9 rollout, max 200 completion tokens, per-completion forward passes with gradient checkpointing, gradient-norm clip 1.0, AdamW at lr 5×10^{-6}, 5 epochs over the curriculum. Steps with zero reward variance within the group are skipped (no gradient); this affects 35–45% of steps early and 5–15% late in training. Total compute $\sim$85 GPU-hours; all scripts and the split-asserting pipeline are in the anonymized supplement.

Evaluation. Fixed 60-task held-out GSM8K test slice (first 60 test-split tasks with parseable golds), best-of-4 sampling at temperature 0.8 with a fixed seed shared across all conditions, extraction preferring an explicit "Final answer: <number>" marker with last-number fallback. For the matched-replication noise analysis (§D) all 18 checkpoints plus the SFT base were re-evaluated under the identical protocol on the first 600 test-split tasks; the C3 pass (§E) extended this to all 45 sweep checkpoints and the C1 pass (§F) to the 18 diluted-verifier checkpoints; the extension passes (§G) added the 12 horizon, 18 cross-family-judge, and 18 composite-label checkpoints; the adversarial-control reruns added 18 more (nine planted-verifier arms, six disjoint-warm-start reruns, three reachable-loophole reruns; §G); the step-matched replication (§G) adds 70, the confirmatory replication (§G) 80, the construct-aligned replication (§G) 100, the natural-verifier replication (§G) 80, and the trigger-only sensitivity control (§G) 3 — 480 trained checkpoints in total under this evaluation protocol (483 evaluation files with bases); the remaining 14 trigger-only diagnostic runs (PC at $4\times$ lr, PC2 dose ladder; §G) intentionally collapse training and are scored by the trigger-rate protocol. The test split is disjoint from every training-side pool by construction: zero task-id overlap with the 2,000-task labeling pool, the 300-task conditional-FPR pool, and every curriculum pool (both splits asserted in the released scripts). **Warm-start defect (disclosed):** the SFT warm-start used by all pre-replication experiments labeled padding positions as EOS (no -100 masking; pad = EOS) — shared by every arm of those experiments, so their contrasts are internally consistent; the label-masked retrain differs from it by ≤ 0.006 at both metrics (0.487/0.254 vs. 0.490/0.248) and is used by the replication of §G. Within-script evaluation is deterministic (identical re-runs reproduce scores exactly); cross-script scores at temperature 0.8 are not (kernel-level nondeterminism flips draws on $\sim 20\%$ of tasks), so no contrast mixes the two evaluation scripts. **Overlap audit (disclosed):** the SFT warm-start trains on the reference solutions of the first 2,000 train-split tasks — the same 2,000 that define the labeling pool and a strict superset of every curriculum pool (200/200 gap-study, 300/300 FPR, and 100/100 task overlap in each curriculum file, task-ID-asserted). Every task the policy RL-trains on, it has already seen a gold reference solution for. The GRPO-style objective itself never sees R_g; but the rollout-time correlation and the filtering conclusions are conditioned on this gold anchoring and, for the sweep, on the original harness (§G decomposes its removal-cost reading) — a disjoint-SFT ablation is the decisive test (Limitations).

Labeling. k=6 probes per task at temperature 0.8 from the SFT policy; gap = acceptance rate $-$ gold-match rate; threshold $\frac{1}{6}$ (one probe). The 2,000-task label file covers the first 2,000 GSM8K train-split tasks: 677 positives; split 400/1,600 by task (seed 0, zero overlap asserted); training positives oversampled to 1,055 of 2,110 rows after the split.

CIs. Group means over 3 seeds; per-cell Δ CIs use $t_{0.975,\, df=2} = 4.303$ on the seed-paired standard error; pooled 15-pair contrasts use $t_{0.975,\, df=14} = 2.145$ (§E). For the original-harness analy-

Table 1: Curriculum audit of the §5.1 sweep (machine-exported from the pool files; identical across seeds). "Flagged" = tasks the critic predicts hackable (removed by veto); "kept" = veto curriculum size. "$\geq \frac{1}{6}$"/"$\geq$0.3" count true probe-gap positives in each set. The veto acts on critic predictions, not directly on gaps; its near-perfect recall on true positives (66/67 at h=1) makes this effectively an oracle-veto with a 5-task false-positive budget at h=0.

h	pool	flagged	kept	pool $\geq \frac{1}{6}$	pool $\geq$0.3	veto keeps $\geq \frac{1}{6}$
0.00	100	5	95	0	0	0
0.25	100	30	70	25	23	0
0.50	100	53	47	50	23	1
0.75	100	68	32	67	23	1
1.00	100	68	32	67	23	1

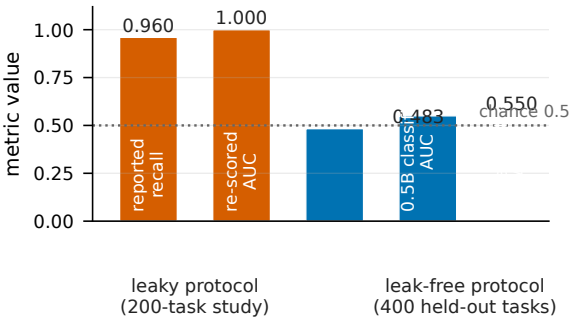


Figure 3: Anatomy of the leakage (§4.2). Left bars: the leaky protocol's headline recall (0.96, reported) and the same checkpoint's AUC when re-scored (1.000 — memorization signature). Right bars: the leak-free classifier (AUC 0.483, chance) against the seven-feature logistic baseline (0.550).

ses, equivalence margins are not pre-registered (per-cell claims are phrased as no-detectable-effect, and the pooled veto harm / random-removal equivalence claims are reported with their CIs); the replication studies pre-register their margins in frozen plans (App. G).

B LEAKAGE FORENSIC DETAILS

Table 2: Hackability prediction under leak-free evaluation (400 held-out tasks, 132 positives). The 0.5B classifier is at chance; a seven-feature logistic baseline is not beaten. The two rows above the rule reproduce the leaky protocol's numbers for contrast (§4.2).

model / view	AUC	F1 (P / R)
critic, leaky protocol (reported)	—	0.87 (0.79 / 0.96)
critic, leaky ckpt re-scored	1.000	—
critic, leak-free (400-task val)	**0.483**	0.277 (0.323 / 0.242)
question-only view	0.511	—
verifier-only view	0.500	—
7B judge model, 8-shot prompt	0.529	0.392
logistic regression, 7 surface features	0.550	0.407

The corrupted 200-task pipeline (values as reproduced): 67 positives of 200 tasks; positives duplicated to balance (266 rows, 200 distinct tasks, 66 duplicated); row-level 80/20 split (53 validation rows, 52 distinct validation tasks, 24 of which also appear in training). The checkpoint trained on that file achieves, under its own protocol, precision 0.909 / recall 1.000 (F1 0.952), AUC = AP = 1.000, and question-only AUC 0.970 (localizing the memorization to the duplicated text). The originally-reported 0.963/0.788/0.867 was a training-time evaluation at a different operating threshold on the same corrupted split. The forensic replay's 0.909/1.000/AUC=1.000 is computed on a fresh 20% row-level slice of the same 200-task file (first 40 tasks, 10 positives) — external to the internal gate's 53-row slice, but drawn from the same corrupted universe, and most of its tasks sit

(some duplicated) in the training rows; that is precisely where the memorization signature shows. A clean retrain (same hyperparameters, task-level split of the same 200-task labels, oversampling within training only) reaches precision 0.273 / recall 0.600 on the genuinely task-disjoint 400-task evaluation; the *leaky* checkpoint on that same 400-task evaluation reaches AUC 0.596 (P 0.435 / R 0.356, question-only 0.595) — around the surface baseline, far below the corrupted 1.000 — and its *internal* validation gate, drawn from the same oversampled file, reported precision 0.759 / recall 0.863: green metrics from a broken split. Perfect or near-perfect metrics on a small behavioral dataset are a leak alarm, not a success. All numbers in this paragraph are artifacts of the corrupted or 200-task protocol and are *not* used anywhere else in the paper.

C PER-CONDITION CORRELATION SUMMARY

$\mathrm{corr}(R_p, R_g)$ over per-step training rollouts (gradient-bearing steps), all 45 conditions: range $+0.328$ to $+0.811$, median $+0.544$; by curriculum type (medians): unfiltered $+0.506$, veto $+0.569$, random $+0.478$; no condition below $+0.328$. Figure 2 (left) plots the full distribution.

The verifier-correlation spectrum table referenced throughout §6:

Table 3: Verifier-correlation spectrum, all probed under a matched protocol (same SFT policy and sampling seed; 150 tasks for the self-consistency probes, 100 for the judge and mixture probes). Accept-on-wrong here is the matched-probe statistic; §3's conditional-FPR study measures the same quantity with adaptive sampling over 300 tasks (means 0.157 numeric / 0.211 rubric) — same ordering, different instrument. Every *prompted* verifier that retains training signal sits at $\mathrm{corr} \geq +0.57$; †the strictest self-consistency variant falls below the acceptance gate ($0.286 < 0.3$) and is shown for completeness. The low-correlation cell is not reachable by weakening the *prompt* — it is entered by explicit dilution (bottom rows; sweep in §5.4/App. F) or by swapping the judge model (Mistral: naturally low discrimination; training replication App. G).

verifier	accept	corr (comp.)	corr (task)	acc. on correct	acc. on wrong
numeric judge (7B)	0.352	+0.633	+0.749	0.791	0.144
rubric judge (7B, told not to verify)	0.393	+0.574	+0.689	0.813	0.206
self-consistency 2-of-4	0.327	+0.685	+0.827	0.790	0.104
self-consistency 2-of-6	0.383	+0.717	+0.888	0.865	0.131
self-consistency 3-of-6†	0.286	+0.774	+0.896	0.768	0.032
mixture q=0.5 (judge$\oplus$coin)	0.445	+0.214	—	0.592	0.368
mixture q=0.25 (judge$\oplus$coin)	0.457	+0.098	—	0.524	0.421

D MATCHED-REPLACEMENT VETO CONTROL

The original veto design conflates task identity with curriculum size (unfiltered keeps all 100 tasks; veto keeps 32–95). The matched-replacement control removes this confound: at hackable fraction h, every task with FPR ≥ 0.3 is removed and replaced by a non-hackable task matched on probe gold rate (± 0.15) and question length ($\pm 30\%$) from the unused remainder, so all curricula share size (100), step count, and approximately difficulty profile; a randomized arm removes/replaces the same count at random. Curricula are built from the conditional FPR labels of §3 (numeric-judge prompt). The h=1 pool is the 100 highest-FPR tasks of the 300-task study, of which 58 clear the τ=0.3 veto threshold — so "hackability-enriched" means *highest-FPR under the conditional construct*, paralleling the gap-based pools of §5.1: unfiltered keeps all 58, matched keeps 0, randomized keeps 24–26; mean probe gold rate stays within 0.346–0.375 and mean question length within 220–232 characters across all seven distinct h=1 curricula (the unfiltered pool is seed-invariant; matched and randomized variants are per-seed). 3 seeds $\times$ $h \in \{0, 1\} \times 3$ arms under the training and evaluation protocol of §5.1.

Four readings. (i) At the 60-task evaluation the h=1 matched point estimate exceeds unfiltered in all three seeds ($+0.067$, $+0.050$, $+0.200$; mean $+0.106$) with a CI including zero; the same scale's matched-vs-randomized contrast is the *one* CI in this study that excludes zero ($+0.067$ [$+0.025$, $+0.108$]) — reading (iv) shows it is evaluation-noise-inflated. (ii) The design yields an unplanned noise-floor measurement: at h=0 the removal list is empty, so all nine runs train on byte-identical curricula — and spread over $R_g \in [0.350, 0.600]$ with per-run sd 0.080 at n=60. (iii) Re-evaluating

Table 4: Matched-replacement control. Held-out R_g (mean $\pm$ sample sd over 3 seeds); all curricula size 100. Pools here are built from the conditional-FPR construct over the 300-task study — hence the different h=0 values from the gap-based pools of Table 5. Rows "n=60": original 60-task evaluation; rows "n=600": every checkpoint re-evaluated on the first 600 test-split tasks (same seeded best-of-4 protocol). Δ(m−u) CIs use $t_{0.975,\,df=2} = 4.303$ on the seed-paired standard error.

h	eval	unfiltered	matched	randomized	Δ(m−u) [95% CI]
0.0	n=60	0.489 ± 0.025	0.489 ± 0.123	0.494 ± 0.100	−0.000 [−0.361, +0.361]
0.0	n=600	0.518 ± 0.010	0.501 ± 0.080	0.506 ± 0.039	−0.018 [−0.227, +0.192]
1.0	n=60	0.478 ± 0.048	0.583 ± 0.067	0.517 ± 0.083	+0.106 [−0.099, +0.310]
1.0	n=600	0.538 ± 0.027	0.562 ± 0.008	0.523 ± 0.062	+0.024 [−0.052, +0.101]

all 18 checkpoints on 600 held-out tasks *decomposes* that floor: the identical-curriculum spread shrinks to sd 0.046 (range $[0.408, 0.550]$), so $0.080^2 \approx 0.046^2 + 0.066^2$ — $\sim$68% of the n=60 run variance is evaluation sampling noise, and only 0.046 is training/rollout stochasticity. The SFT base itself moves from 0.550 (n=60) to 0.490 (n=600): the small-scale base estimate was +0.06 optimistic, and at n=600 all six trained cells exceed base (0.501–0.562) — the earlier "no condition beats base" reading was itself an evaluation-noise artifact. (iv) Under the 600-task evaluation the h=1 matched delta collapses to +0.024 [−0.052, +0.101] with one of three seeds negative, and the nominally significant Δ(m−r) to +0.039 [−0.121, +0.200]: the n=60 three-seed trend was evaluation noise, and no pairwise difference clears the calibrated pair-difference sd of 0.064. Unfiltered h=1−h=0 is −0.011 (n=60) / +0.019 seed-paired (n=600; cell means 0.538 vs. 0.518) — no degradation at either scale; within-group advantage correlation stays positive in all 18 runs (0.611–0.748, lowest on hackability-enriched curricula), extending §5.4. Resolving a true 0.1 effect needs $\sim$10 runs/arm at n=60 but only $\sim$3 at n=600: the binding constraint was the evaluation, not the training. Under size, difficulty, removal-dose, and now evaluation-scale control, the veto is undetectable in this testbed.

E C3: THE FULL SWEEP RE-SCORED AT n=600

All 45 sweep checkpoints of §5.1 were re-evaluated on the first 600 held-out test tasks under the identical seeded best-of-4 protocol (Table 5 reports the 600-task numbers; Table 6 preserves the original 60-task table for reference).

Table 5: Main sweep, 600-task evaluation of all 45 checkpoints. Held-out ground truth R_g (mean $\pm$ sample sd over 3 seeds); base = 0.490. Pooled seed-paired over all 15 run pairs ($t_{0.975,\,df=14}$=2.145): Δ(v−u) = −0.029 [−0.043, −0.016], Δ(v−r) = +0.001 [−0.017, +0.020], Δ(r−u) = −0.031 [−0.048, −0.013].

h	unfiltered	veto	random	Δ(v−u) [95% CI]
0.00	0.532 ± 0.019	0.521 ± 0.021	0.483 ± 0.047	−0.011 [−0.111, +0.088]
0.25	0.535 ± 0.016	0.516 ± 0.024	0.509 ± 0.023	−0.019 [−0.091, +0.052]
0.50	0.546 ± 0.006	0.514 ± 0.013	0.537 ± 0.019	−0.032 [−0.054, −0.010]
0.75	0.528 ± 0.011	0.483 ± 0.007	0.511 ± 0.021	−0.046 [−0.069, −0.022]
1.00	0.548 ± 0.008	0.508 ± 0.016	0.495 ± 0.015	−0.039 [−0.098, +0.019]

Seed-paired differences pooled over all 15 veto–unfiltered pairs ($t_{0.975,\,df=14} = 2.145$): Δ(v−u) = −0.0294 [−0.0433, −0.0156], $t_{14} = -4.57$; per-seed deltas are negative in 12/15 run pairs (one-sided sign test p=0.018, two-sided 0.035) and negative in 3/3 seeds at both h=0.5 and h=0.75; clustering the 15 pairs by training seed (3 seed-means, df=2) gives $[-0.051, -0.008]$ — the three seed-means (−0.039/−0.022/−0.027) are tightly consistent, so the pooled reading survives the correct experimental unit. The volume-matched controls: Δ(r−u) = −0.0309 [−0.0484, −0.0133] and Δ(v−r) = +0.0014 [−0.0174, +0.0203] — the veto–random difference is equivalent to zero within a TOST margin of $\pm$0.02 (both one-sided $p \leq 0.026$; the margin is post hoc, chosen at two-thirds of the pooled veto effect): the veto cost equals the random-removal cost, and the hackability ranking contributes nothing on top of it — as run, in the original harness (not step-matched; the

Table 6: The original 60-task evaluation of the same 45 checkpoints (superseded by Table 5; kept for the noise comparison of §D). Base = 0.550 at this scale. Every Δ(v−u) CI includes zero at n=60.

h	unfiltered	veto	random	Δ(v−u) [95% CI]
0.00	0.550 ± 0.060	0.494 ± 0.059	0.406 ± 0.125	−0.056 [−0.223, +0.112]
0.25	0.511 ± 0.054	0.506 ± 0.010	0.461 ± 0.038	−0.006 [−0.125, +0.114]
0.50	0.522 ± 0.019	0.528 ± 0.019	0.533 ± 0.093	+0.006 [−0.042, +0.053]
0.75	0.461 ± 0.025	0.394 ± 0.025	0.472 ± 0.042	−0.067 [−0.149, +0.016]
1.00	0.539 ± 0.054	0.511 ± 0.054	0.478 ± 0.010	−0.028 [−0.195, +0.140]

Table 7: C1: diluted verifier (q=0.25), held-out R_g at n=600 (mean ± sample sd over 3 seeds). Undiluted counterparts (Table 5 cells): unfiltered 0.532 / 0.548 at h=0/1.

h	unfiltered	matched	randomized	Δ(m−u) [95% CI]
0.0	0.476 ± 0.030	0.441 ± 0.032	0.465 ± 0.026	−0.036 [−0.120, +0.049]
1.0	0.504 ± 0.022	0.481 ± 0.011	0.500 ± 0.025	−0.024 [−0.103, +0.055]

corrected decomposition is §G). Arm means: unfiltered 0.538, veto 0.508, random 0.507, base 0.490.

Two cautions. (i) Five per-fraction Δ(v−u) CIs constitute a family; the two that individually exclude zero (h=0.5: −0.032 [−0.054, −0.010]; h=0.75: −0.046 [−0.069, −0.022]) are consistent with, but not independent evidence beyond, the pooled test — we rest the claim on the grand paired CI and the 12/15 sign consistency, not on selected cells. (ii) The n=60 table's wider CIs all include zero; the upgrade from "no detectable effect" to "harmful, volume-mediated" is an evaluation-scale resolution, not a new experiment — which is exactly why we report the 600-task numbers throughout.

F C1: DILUTED-VERIFIER SWEEP DETAILS

Mixture verifier. During training only, each judge accept is replaced, independently per completion, by a fresh fair coin with probability $1-q$ (dedicated RNG stream; gold rewards and evaluation are untouched). The pre-flight probe (100 tasks × k=6 SFT rollouts, post-hoc mixture statistics) yields the bottom rows of Table 3; the gate — lowest q with acceptance ≥ 0.3 and probe correlation < 0.45 — selects q^*=0.25.

Sweep. 18 runs: 3 seeds × $h \in \{0, 1\}$ × {unfiltered, matched-replacement, randomized} over the conditional-FPR curricula of §D, training protocol of §5.1 (lr 5×10^{-6}, 5 epochs), evaluated at n=600 (Table 7).

Health under dilution. Gradient density 85–89% (vs. 55–65% undiluted — the coin re-injects within-group reward variance, so fewer steps are skipped); in-loop within-group advantage correlation +0.07 to +0.22 (vs. 0.611–0.748 undiluted), confirming the dilution reaches the advantage channel. Across the 18 runs R_p and R_g rise together on average (first→last-50-step means: R_p 0.402 → 0.453 with R_g 0.211 → 0.219 at h=0; R_p 0.452 → 0.497 with R_g 0.311 → 0.349 at h=1); two runs' 50-step windows drift down in R_g (−0.09, −0.06), within the coin's sampling noise. Dilution costs 0.04–0.06 R_g relative to the undiluted cells; at h=0 the diluted arms land at 0.441–0.476, at or slightly below the 0.490 base, while at h=1 they reach 0.481–0.504 — dilution erases the training gain at h=0 but does not invert it into harm. The hackability curriculum remains inert: Δ(m−u) = −0.036 [−0.120, +0.049] at h=0, −0.024 [−0.103, +0.055] at h=1; Δ(m−r) = −0.024 / −0.019, both CIs spanning zero — matched replacement again indistinguishable from random removal.

Table 8: Horizon arm: held-out R_g at n=600 (mean over 3 seeds) on the unfiltered gap-based pools. $\Delta = h{=}1 - h{=}0$; CI = stratified task-level paired bootstrap (10,000 resamples), applied identically to the 5-epoch row (recomputed from the sweep's unfiltered checkpoints).

Epochs	h=0	h=1	Δ	95% CI
5 (sweep)	—	—	+0.016	[−0.005, +0.037]
15	0.556	0.574	+0.018	[−0.002, +0.039]
25	0.511	0.591	+0.079	[+0.058, +0.102]

G Extension series: training horizon, cross-family judge, and replication controls

Three arms probe the axes left open by the sweep — training horizon (the calibration disclosure of §5.1), judge family, and label reliability — and the final paragraphs add the identification checks: per-completion re-scoring, the disjoint-warm-start ablation, and a planted-loophole positive control. The arms reuse the protocol of §5.1 (lr $5{\times}10^{-6}$, group size 4, 3 seeds, n=600 held-out R_g) with deterministically regenerated pools identical to the sweep's; the planted-verifier control (§G) trains on the sweep's own h=1 gap pools.

Horizon arm. 12 runs: 3 seeds × $h \in \{0, 1\}$ × epochs $\in \{15, 25\}$ on the unfiltered gap-based pools (Table 8). The hackable-exclusion difference is a monotone function of horizon: +0.016 [−0.005, +0.037] at the sweep's 5 epochs (recomputed from the sweep's unfiltered checkpoints under the same bootstrap), +0.018 [−0.002, +0.039] at 15, +0.079 [+0.058, +0.102] at 25; the horizon interaction (difference of paired differences) is +0.061 [+0.032, +0.090]. CIs are task-level paired bootstrap (10,000 resamples, stratified by seed); at 25 epochs the per-seed differences are +0.100 ($p{=}5{\times}10^{-7}$), +0.113 ($p{=}4{\times}10^{-8}$), and +0.025 (p=0.18) — positive in 3/3 seeds, with seed 2 the weak one, so the grand CI quantifies task-resampling uncertainty at fixed seeds; the seed-level t interval (n=3, $t_{0.975,2}$=4.303) is wider — [−0.039, +0.198] at 25 epochs, [−0.064, +0.101] at 15 — and per-contrast seed-level intervals include zero: the directional consistency across arms and contrasts, not any single interval, carries the inference. The mechanism is one-sided: h=1 is nearly flat across the horizon (+0.016: 0.574 → 0.591) while h=0 *degrades* (−0.045: 0.556 → 0.511) — removing the highest-gap tasks is not merely useless but increasingly costly as training proceeds.

Horizon-arm trajectories. End-to-end, R_p and R_g rise together in all twelve runs (e.g. seed-0, h=1, 25 epochs: R_p 0.35 → 0.83 with R_g 0.28 → 0.51) — the calibration run's signature does not reproduce at its magnitude. Two qualifications keep the reading honest. First, a mild *late-window* softening recurs: comparing the last four epochs against the four before them, in-training R_g declines by 0.009–0.018 in 3/3 h=1 25-epoch runs while R_p still rises (and in 1/3 h=0 25-epoch runs) — directionally the disclosed signature, but 5–10× smaller, within the epoch-mean noise scale (adjacent-epoch ΔR_g sd 0.028 pooled over all runs), and decisively *not* reflected in held-out scores: the h=1 25-epoch cell is the arm's best (0.591 vs. 0.574 at 15 epochs). Second, the largest in-training R_g deficit sits in the *filtered* cell — seed-0, h=0, 25 epochs falls from 0.405 (epoch 2) to 0.32 by epoch 12 against R_p 0.71, recovering to 0.365 by epoch 24 — and filtered runs start from *higher* epoch-0 training R_g (0.34–0.37 vs. 0.22–0.28 for h=1: removing the hardest tasks flatters in-training gold rates) yet end lower on held-out R_g on average (one of six seed-by-horizon cells reverses): the in-training signal inverts under the horizon.

Cross-family judge arm. 18 runs: 3 seeds × $h \in \{0, 1\}$ × {unfiltered, matched, randomized} at 5 epochs, replacing the training judge with Mistral-7B-Instruct-v0.3 (gate q=1.0: probe acceptance in range, completion correlation < 0.55), on the conditional-FPR curricula of §D (all arms size-100, so step counts are equal; the curricula are Qwen-derived — this arm tests label transfer across judges, not Mistral-targeted removal). Absolute R_g is lower under the weaker cross-family judge (0.42–0.53 over the 18 runs). Matched and randomized removal remain statistically indistinguishable: matched−randomized = −0.002 [−0.026, +0.020] at h=1 and −0.002 [−0.024, +0.021] at h=0 (CIs span zero; we do not claim TOST-level equivalence at this sample size); against unfiltered, matched = +0.026 [+0.003, +0.048] at h=1 and +0.006 [−0.018, +0.029] at h=0, randomized

$= +0.028$ / $+0.008$ — the per-seed contrasts are almost all individually insignificant (only 1 of 12 reaches $p<0.05$, and it is a negative one; $|t|$ 0.5–2.1), and where a grand CI excludes zero it does so by less than the noise floor the A3 episode calibrates (§D: an apparent $+0.106$ resolving to $+0.024$ at n=600). The hackable-exclusion direction also replicates, positive in 7/9 (seed, condition) cells: $\Delta(h{=}1{-}h{=}0) = +0.020$ [-0.003, $+0.043$] unfiltered, $+0.039$ [$+0.016$, $+0.063$] matched, $+0.040$ [$+0.017$, $+0.062$] randomized — direction shared with the 5-epoch same-family estimate ($+0.016$ on gap-based pools; the cross-family values run 1.2–2.5$\times$ larger on conditional-FPR curricula).

High-reliability composite-label arm (C2). The veto null is scoped "at this label dose" (split-half $\sim$0.46); this arm closes the caveat. Averaging the posterior FPR of the three measurements (numeric; rubric; rubric fresh-seed — the pairs involving the fresh-seed relabel are completion-independent; the numeric–rubric pair shares a stream and is the disclosed upper bound) yields a composite label with Cronbach $\alpha \approx$ 0.85–0.89 (conservative estimate from the two completion-independent pairs only), and we rebuild the §D curricula — same pools, same protocol — ranking and vetoing on it: 18 runs at 5 epochs. The null does not soften; the exclusion effect sharpens. Unfiltered $h{=}1{-}h{=}0$ rises from $+0.019$ [-0.002, $+0.041$] (single-measurement labels; recomputed from the A3 unfiltered checkpoints) to $+0.079$ [$+0.057$, $+0.101$] — roughly a quadrupling, positive in 3/3 seeds, the same value the horizon arm reaches at 25 epochs on gap pools (seed-level t intervals: $[-0.073, +0.231]$ for the C2 contrast, $[-0.214, +0.149]$ for m$-$r at $h{=}1$; the task-level intervals quantify evaluation-task uncertainty at fixed seeds, the seed-level ones training-seed uncertainty). Against random removal, matched filtering at $h{=}1$ turns *negative*: matched$-$randomized $= -0.033$ [-0.056, -0.010] (negative in 2/3 seeds, $p \leq 0.01$), with matched$-$unfiltered $= -0.019$ [-0.041, $+0.003$]: with better labels the veto targets the tasks that were helping, and still yields no benefit. Two honesty notes: seed heterogeneity in this arm is larger than elsewhere (unfiltered $h{=}1$ spans 0.435–0.563 across seeds), and the small-removal $h{=}0$ cells sign-wander (m$-$u $+0.051$, m$-$r $+0.029$, one seed drives both) — within the A3 noise floor at $n{=}3$.

Per-completion re-scoring (metric check). All checkpoints are re-scored storing per-completion gold and judge rewards (best-of-4 had been the only held-out metric; `judge_eval2`; the interior-fraction sweep cells were completed after review). Base: best-4 0.490 / pass@1 0.248 / false-accept 0.083. Every headline contrast survives at pass@1, most strengthened: veto$-$unfiltered at $h{=}1$ is -0.029 (negative in 3/3 seeds; best-4 -0.039); at the interior fractions the pooled pass@1 veto$-$unfiltered is -0.040 and veto$-$random -0.009; the enrichment contrasts run A3 $+0.035$, composite $+0.085$, horizon-25 $+0.093$ (positive in 3/3 seeds); false-accept rates do not rise on enriched curricula (composite: -0.066). Trained pass@1 spans 0.18–0.41 across the re-scored checkpoints, with weak-seed extension-arm runs occasionally below base — within the run-noise floor, not a metric reversal. No metric-level Goodhart signature: the best-of-4 metric was inflationary in level, not in sign.

Disjoint-SFT ablation (gold anchoring). The warm-start trains on reference solutions of train[0:2000] — a superset of every curriculum pool (App. A audit). Retraining it on the disjoint train[2000:4000] (zero curriculum overlap, asserted in the lane) and rerunning the A3 unfiltered cells: $h{=}1{-}h{=}0$ collapses from $+0.019$ to -0.007 [-0.057, $+0.042$] at best-of-4 ($+0.014$ [-0.024, $+0.052$] at pass@1; per-seed $-0.030/+0.007/+0.002$). The positive enrichment effect is therefore substantially a gold-anchoring artifact of the overlapping warm-start; the conclusion that survives this ablation *and* the corrected-harness replications (R10, R11, below) is the absence of degradation on enriched curricula — the removal-cost reading is resolved there as an update-count artifact.

Planted-loophole positive control. If verifier susceptibility were training-time exploitable, a *planted* loophole must be exploited. The control trains three arms $\times$ 3 seeds for 15 epochs (the protocol above, on the $h{=}1$ unfiltered gap pools, base SFT policy): *clean* (numeric judge), *coin* (judge w.p. 0.5, else a Bernoulli biased to the judge's acceptance rate — reward noise without structure), and *structured* (judge OR trigger-token-present, case-insensitive substring). The first attempt with an arbitrary rare token ("grapefruit") is unreachable: base emission 0.000 across the candidate probe, so structured $\approx$ clean is an exploration null, not evidence. A rate probe (base policy, temperature 0.9, 240 completions over 60 curriculum tasks; criteria: rate

$\in [2\%, 20\%]$, $|r_{\text{gold}}| < 0.05$) therefore selects "1000": rate 7.92%, $r_{\text{gold}} = +0.024$ — reachable and gold-uncorrelated. Rerunning the structured arm with it (P1-A2): best-of-4 $0.362/0.565/0.543$ (mean 0.490 ± 0.112) vs. clean 0.518 ± 0.080, coin 0.520 ± 0.068, unreachable-token structured 0.503 ± 0.126 — all within the 3-seed noise floor. Train-time emission (fraction of the group's completions containing the token, logged every step): epoch means stay within 4.8–7.5% in every epoch of every seed — never above the 7.92% base rate, per-epoch trends within epoch-mean noise. Under the matched evaluation protocol (600 test tasks, k=4, temperature 0.8): base 2.96%, clean-trained 2.49%, structured-trained 1.93% ($1.67/2.25/1.87$) — the naive $7.9\% \rightarrow 1.9\%$ gap is mostly protocol (temperature and split; the base policy itself drops to 3.0%), the residual decline is shared with clean training, and the base re-evaluation reproduces best-of-4 0.490 exactly. Mechanism accounting from the step logs: trigger completions always receive R_p=1, hence non-negative group advantage; trigger-containing groups are 5–12% of steps throughout, of those granting strictly positive advantage 35–55% in epochs 0–4, falling to $\sim 30\%$ over the last five, while the saturated share (all four accepted $\Rightarrow$ zero advantage variance $\Rightarrow$ no update) rises (seed 1: $42\% \rightarrow 88\%$ by the final epoch); within-group advantage correlation holds at $+0.34$–0.76 in the healthy seeds. Seed 0, the lane's weak seed, is the one run that drifts: train R_p $0.383 \rightarrow 0.625$ while R_g $0.235 \rightarrow 0.160$ (epoch means, cliff at epoch 6), evaluation false-accept 0.303 / conditional FPR 0.375 (highest of the lane), advantage correlation $0.66 \rightarrow 0.36$ — with flat in-training trigger emission and the *lowest* evaluation emission of the lane (1.67%): the drift routes through the judge's natural plausible-wrong surface, not the planted backdoor. Raw files: `results/p1a2/` (eval JSONs `jeval2-600-P1A2-*` and `jeval2-600-{BASE,P1A-clean-*}-trig1000.json`; per-step logs `trajP1A2_*.jsonl`); analysis `code/p1a2_analyze.py`.

Step-matched pre-registered replication (R10). Triggered by the disclosure that the sweep arms were not step-matched and shared the warm-start masking defect, this replication was *pre-registered* (analysis plan frozen before launch: `code/PREREG-R10.md`; deviations log empty) with 10 seeds $\times$ $h \in \{0, 1\}$ $\times$ {unfiltered, veto, random} plus a second random realization at h=1 — 70 runs, curricula from the original builder and critic (pools seed-invariant; seeds 3–9 new), corrected label-masked warm-start (unit-tested), completion log-probabilities from the sampled token IDs (unit-tested against an exact reference), and a hard cap of 500 task-iterations per arm. The pre-registered base gate passed ($0.487 \in [0.35, 0.65]$). Health: 70/70 runs pass (gradient density 0.43–0.68, lowest on veto cells, R_p rising in every run, no R_g collapse; intent-to-treat, no run dropped). Per-seed best-of-4 at h=1 — unfiltered: 0.533, 0.530, 0.567, 0.532, 0.527, 0.537, 0.543, 0.513, 0.497, 0.520; veto: 0.515, 0.518, 0.507, 0.503, 0.535, 0.528, 0.515, 0.537, 0.505, 0.502; random: 0.495, 0.545, 0.517, 0.543, 0.550, 0.545, 0.533, 0.540, 0.505, 0.553; random2: 0.503, 0.530, 0.562, 0.520, 0.557, 0.528, 0.528, 0.537, 0.550, 0.543. Pre-registered outcomes: **H1** (one-sided $v-u$ harm at h=1) *fails at the margin* ($t_9 = -1.79$ vs. crit. 1.833; 7/10 seeds negative; -0.0133); **H2** (TOST equivalence of $v-r$ at ± 0.02) *fails* (mean -0.0162, SE 0.0064, 90% CI $[-0.028, -0.004]$; seed-level $t_9 = -2.52$); **H3** (subset variance) passes (mean $r_2 - r_1 = +0.0032$; sd 0.024). Secondary: h=0 contrasts null ($v-u$ $+0.0047$ $[-0.0194, +0.0288]$); task-level stratified bootstrap (10k) at h=1: $v-u$ $[-0.0255, -0.0017]$, $v-r$ $[-0.0282, -0.0043]$, $r-u$ $[-0.0092, +0.0150]$; pass@1 agrees ($v-r$ -0.0140, SE 0.0055). Single-realization subset noise makes one-draw contrasts hard to read: the confirmatory replication (§G) measures the subset-draw floor at sd 0.018 and the veto deficit against the four-draw mean at $+0.004$ — within it. Raw: `results/r10/` (70 eval2 + 70 jeval600 + per-step logs); analysis `code/r10_analyze.py` $\rightarrow$ `results/r10/R10_STATS.txt`.

Confirmatory replication (R11). Pre-registered (`code/PREREG-R11.md`, frozen before launch; deviations: none — the trigger-only diagnostic arm of §G was an annex, not a deviation). 10 seeds $\times$ h=1 arms {unfiltered, matched, rand1–rand4} plus h=0 {unfiltered, rand1} — 80 runs (the pre-registration's stated total of 90 was an arithmetic slip; erratum logged in the R12 pre-registration) — from a *disjoint* label-masked warm-start (train[2000:4000]; base $0.492/0.245$, gate passed), composite conditional-FPR labels ($\alpha \approx 0.86$, §G), A3 difficulty-matched replacement (pools size-100; banks matched on gold-rate ± 0.15 and length $\pm 30\%$), four independent random-replacement draws per seed (rand1 replicates the composite-arm stream), step-matched at 500 iterations. Health: 80/80 runs pass. **Label transfer across warm-starts** (the labels are policy-relative and were measured under the original warm-start): re-measuring conditional FPR with the replication's disjoint corrected warm-start (same 300 tasks, 24 independent probes each, mean 16.5 gold-wrong completions per task; mean posterior FPR 0.183 vs. 0.202) correlates with the composite label at Spearman 0.66 (Pearson 0.66; against the original single numeric measurement, 0.65) — high rank stability, near the

$\sim$0.74 attenuation ceiling expected under fully policy-invariant latent susceptibility. The replication therefore tests filtering under *label transfer* that preserves roughly ninety percent of the ceiling — the deployment condition, since labels are always measured before training begins. Pre-registered outcomes: **H1** (matched $<$ random, one-sided) *fails* ($\Delta(\text{m}-\bar{r}) = +0.0044$, SE 0.0064, t_9=0.68, 5/10 seeds negative); **H2** (TOST $\pm$0.02 equivalence of unfiltered vs. $\bar{r}$) *passes* ($+0.0022$, 90% CI $[-0.0053, +0.0097]$); *exploratory* TOST on the same margin for matched-vs-random also passes (90% CI $[-0.0074, +0.0162]$); **H3** subset decomposition: within-seed subset sd 0.0183 (per-seed 0.008–0.028) vs. between-seed sd of $\bar{r}$ 0.0069 — curriculum-draw variance dominates. Secondary: h=0 null ($+0.0028$ [-0.0163, $+0.0219$]); pass@1 null (m$-\bar{r}$ $+0.0009$). Arm means at h=1: unfiltered 0.532, matched 0.534, rand1–4 0.538/0.530/0.520/0.531. Raw: `results/r11lane/`; analysis `code/r11_analyze.py` $\rightarrow$ `results/r11lane/R11_STATS.txt`.

Construct-aligned replication (R12). Pre-registered (`code/PREREG-R12.md`, frozen 2026-09-04 before launch; deviations: none). Motivated by the review finding that R11's replenishment bank was FPR-filtered (label-informed) and its operating point left a weak contrast, R12 rebuilds the comparison: *blind bank* — both matched and random arms draw difficulty-matched replacements from the full non-pool task set, so the arms differ only in which pool tasks are removed. Two prevalence pools (MID: top-50 + bottom-50 composite FPR, 50/100 flagged, matched exposure 0.148 vs. random 0.251; H1: top-100, 86/100 flagged, 0.185 vs. 0.197) plus a τ=0.4 budget arm on H1; 10 seeds, 100 runs, disjoint warm-start, step-matched at 500, gradient density 0.45–0.71 across all runs. Pre-registered outcomes: **H1-T** (TOST on matched$-\bar{r}$, MID, $\pm$0.02) *fails at the margin* (-0.0146, SE 0.0083, t_9=-1.76, 8/10 seeds negative, 90% CI $[-0.030, +0.001]$); **H1-D** (one-sided harm) fails (-1.76 vs. 1.833); **H2** (unfiltered$-\bar{r}$ TOST) fails at the margin (-0.0074, CI $[-0.022, +0.007]$). The non-confirmatory H1 pool *reverses the sign* ($+0.0157$, t_9=$+1.59$, 7/10 above; unfiltered$-$random $+0.0182$, t_9= + 2.16), and the budget arm is null (τ04$-\tau$03: -0.005/$+0.003$; matched$-$random at τ=0.4: $+0.008$, t_9=1.04). pass@1 mirrors MID (-0.0129, SE 0.0051). Read together with R10 (-0.016), R11 ($+0.004$), and the composite arm (-0.033, 3 seeds): the sign moves with design choices while point estimates span -0.033 to $+0.016$ and no pre-registered directional test clears convention — the filtering contrast at this scale is small and design-fragile, not resolved. Raw: `results/r12/`; analysis `code/r12_analyze.py` $\rightarrow$ `results/r12/R12_STATS.txt`.

Trigger-only sensitivity control (PC). The planted-loophole null (§G) is interpretable only if the optimizer *can* learn a surface token when it is the sole reward source. Trigger-only verifier ("1000" present $\Rightarrow R_p$=1, else 0; no base-judge fallback), identical setup to P1-A2 otherwise. At the calibrated lr 5×10^{-6}: emission stays 6% $\rightarrow$ 9% across 15 epochs in 3/3 seeds (evaluation trig-rate 0.036–0.040), gradient density 0.06 — a 6%-emission token leaves $\sim$90% of groups with zero reward variance, yielding $\sim$90 signal-carrying updates in 1500 steps; the policy's math behavior is unperturbed (best-of-4 0.45–0.51). At 4$\times$ lr (2×10^{-5}, 2 seeds): emission reaches 68–82% within epoch 0 and $\geq$99.8% by epoch 3, $R_g \rightarrow 0$ (evaluation best-of-4 0.002–0.010), gradient density 0.03 (convergence removes within-group variance and training self-terminates). 4$\times$ lr is the regime the calibration of §5.1 classified as destructive for normal training. Raw: `results/pc/`.

Evidence table: the filtering program at a glance.

Natural low-correlation replication (R13). Pre-registered (`code/PREREG-R13.md`, frozen 2026-09-12; deviations: none — GPU4 was taken by another user mid-lane and seeds 0–4 reran on GPU3; skip-if-exists, zero rework). The first genuinely two-sided test in the program: Mistral-7B-Instruct-v0.3 is the training *reward* (the naturally low-correlation verifier: accepts 68% of gold-wrong completions), and its own continuous conditional-FPR ranking is the label (τ=0.5; MID pool top-50+bottom-50 with exposure 0.86 vs. 0.33; H1 top-100); blind banks, 10 seeds, 80 runs, step-matched 500, disjoint v3 warm-start; healthy in 80/80 (gradient density 0.30–0.62). Pre-registered outcomes: **H1-T** (TOST $\pm$0.02 on matched$-\bar{r}$, MID) *passes* — $+0.0019$, SE 0.0073, 90% CI $[-0.0115, +0.0154]$, 5/10 seeds negative: **the first pre-registered equivalence PASS on the title estimand**; **H1-D(help)** (the complementary hypothesis, formalized in advance) fails flat (t_9=0.27); **H1-D(harm)** fails (positive point); **H2** (unfiltered$-\bar{r}$) passes (-0.0037, CI $[-0.0170, +0.0096]$). Secondary: H1 pool null (-0.0033, t_9=-0.48); pass@1 mirror -0.0066, CI $[-0.0121, -0.0012]$ — inside $\pm$0.02 (equivalence holds at pass@1 too) with a small negative point estimate (6/10 seeds);

subset sd 0.0177 (consistent with 0.018 in R11/R12). **Mistral label transfer** (the ranking is policy-relative and was measured under the original warm-start): re-measuring Mistral conditional FPR under the R13 warm-start (300 tasks, 24 independent probes, mean 16.7 gold-wrong/task) correlates with the original Mistral ranking at Spearman 0.645 (Pearson 0.658) — the same transfer fidelity as the Qwen composite (0.664, §G); R13 therefore filters by a ranking that transfers to the trained policy at the same fidelity established for R11. Arm means (MID): unfiltered 0.491, matched 0.497, rand1–3 $0.499/0.497/0.489$. Read with Replications 1–3: the sign of the filtering contrast moves with design choices everywhere except here, where it is dead zero with equivalence within the pre-registered margin — the null survives its most natural adversarial instantiation. Raw: `results/r13/`; analysis `results/r13/R13_STATS.md`.

Fixed-lr dose ladder (PC2). Pre-registered (`code/PREREG-PC2.md`, frozen 2026-09-09; deviations: none). Trigger-only verifier, lr fixed at 5×10^{-6}, 15 epochs, 3 seeds per arm. The dose is reported as the *observed* epoch-0 gradient-bearing fraction (groups with reward variance) from the step logs, not the independence-derived value: for binary rewards a group carries variance iff it is neither all-trigger nor no-trigger ($1-p^k-(1-p)^k$), and emissions are additionally *task-correlated*, which the independence value overstates.

Arm (token, k)	emission rate	indep. dose	**observed dose**	outcome (emission ep0→ep14; best-of-4)
"1000", k=4	7.9%	28%	**6%**	$6\rightarrow9\%$; intact (0.45–0.51)
"1000", k=8	7.9%	48%	**7%**	$5\rightarrow7$–12%; intact (0.41–0.50)
"answer:", k=4	85%	48%[a]	**26%**	$89\rightarrow99\%$; collapsed (0.27–0.31)
"thus", k=4	15%	48%	**57%**	21–$40\rightarrow100\%$ 3/3; collapsed (0.34–0.39)
"therefore", k=4	16.7%	52%	**65%**	36–$47\rightarrow100\%$ 3/3; collapsed (0.41–0.43)

[a] saturation: all-trigger groups carry no variance. Exploitation is monotone in the *observed* dose: 6–7% never amplifies, 26% amplifies to saturation, 57–65% amplifies fully. Doubling group size buys no dose ($6\% \rightarrow 7\%$) because numeral emissions cluster by task (all-or-none within a group); "thus" slightly exceeds its independence value (mild anti-clustering). Pre-registered verdicts: P1 (dose-85 amplifies) confirmed; P2 (dose-8 flat) confirmed; P3 (monotone dose response) confirmed in observed dose; **P4 (lever equivalence) refuted** — the independence-matched arms were not dose-matched in fact. $4\times$ lr exploits the low-dose numeral (PC). Raw: `results/pc2/`; analysis `results/pc2/PC2_RESULTS.md`.

Table 9: The filtering program at a glance: every arm's labels, warm-start, comparator, registration status, and head estimand.

Exp.	Labels	Warm	Comparator	Plan	Head estimand (result)
Sweep (45-cond)	gap	overlapping	removal-only random	exploratory	v−u pooled: -0.029 [-0.043, -0.016]; seed-clustered [-0.051, -0.008]
Replication 1 (R10)	gap	corrected v2	removal-only, step-matched	pre-reg.	v$-\bar{r}$ (1 draw): -0.016 ($t_9 - 2.52$); TOST fails
Replication 2 (R11)	composite	disjoint v3	clean-bank replacement†	pre-reg.	m$-\bar{r}$ (4 draws): $+0.004$ (t_9 0.68); u$-\bar{r}$ TOST passes
Replication 3 MID	composite	disjoint v3	blind-bank replacement	pre-reg.	m$-\bar{r}$ (3 draws): -0.015 ($t_9 - 1.76$) adapted; -0.022 ($t_9 - 3.58$) untouched
Replication 3 H1	composite	disjoint v3	blind-bank replacement	secondary	m−r1: $+0.016$ ($t_9 + 1.59$) adapted; -0.004 untouched
Replication 4 (R13)	Mistral FPR	disjoint v3	blind-bank, Mistral reward	pre-reg.	m$-\bar{r}$: $+0.002$ [-0.012, $+0.015$] — TOST passes (title estimand)
Composite arm (C2)	composite	overlapping	clean-bank†	3 seeds	m−r: -0.033 [-0.056, -0.010]

† replenishment bank FPR-filtered (label-informed; identified by review, corrected in R12). Mechanism arms: planted loophole P1-A2 (OR-verifier, flat); trigger-only PC/PC2 dose ladder and PCX (4× lr) — App. G. Prediction: gap-label chance (0.483); composite-label weak/difficulty-mediated (0.65 vs. 0.61 baseline). Intervals are seed-level t unless marked pooled.