Does rare but persistent shortcut emission become learned selection over a 200-step GRPO leg?

Humanity First Research

Abstract

We ask whether a rarely emitted reward-hacking shortcut becomes a learned behavior when its reinforcement-learning horizon is doubled, and the registered answer, this study's confirmatory result, is no. The setting is a 1.5B-parameter model trained with GRPO, a group-relative policy-gradient method, on synthetic arithmetic word problems scored by a deliberately hackable grader: a response earns full credit either by solving the problem or by ending with an assertion that satisfies the grader without answering it, a shortcut we call the seam. Under the deliberately gentle recipe under study, a rank-8 LoRA adapter at learning rate 1e-6, doubling the horizon from one hundred to two hundred optimizer steps left the seam flat: the pre-registered stopping statistic came back 0.0 against its 0.25 decision threshold, the seed-paired contrast against a clean-grader control is -0.0028 with a 95% interval of -0.015 to +0.0083, and a manipulation check at 0.83 rules out a degenerate optimizer. A post-hoc exploratory factorial, since extended to three seeds per cell, then shows how narrowly that certificate must be read: with learning rate raised ten-fold and adapter rank four-fold together, the same shortcut reached a terminal rate of 0.686, one-sided 95% bound 0.644, neither knob alone crossed the threshold, and the stronger setting collapsed gold-task correctness on a high-availability control cue. A ladder of intermediate configurations, run to test whether training strength drives reinforcement of this shortcut specifically, instead found seam selection tracking each final policy's measured divergence from its base across all 16 trained adapters; because these data cannot distinguish a reward-hacking-specific mechanism from larger updates simply moving the policy farther, we withdraw the mechanistic reading an earlier draft of this work advanced and claim no such mechanism. These exploratory results are preliminary, resting on one model, one synthetic task family, three seeds per factorial cell and two per intermediate ladder rung, and they identify no threshold and no monotonicity; what they motivate is a regime-bounded hypothesis for future certification practice: a negative reward-hacking evaluation is informative about the training configuration and realized policy divergence that produced it, should report both, and cannot be read beyond them. Two secondary instrument-level observations accompany these results: an inherited certification threshold sitting below the untrained base rate of the condition it judges, and untrained-rate drift across evaluation draws; the certification pass's statistical margin over its gate and the drift are each sensitive to the choice of multiplicity correction and are reported as suggestive rather than established. The registered null stands as measured: at this gentle recipe, two hundred steps of rewarding the seam did not teach it.

Preprint. Under review.

Hypothesis

At the corrected 512-token completion cap, continuing the frozen (held-fixed) exploit-rewarded GRPO intervention of an earlier run, which we call the 100-step study throughout, from 100 to 200 optimizer steps, will produce a sustained increase in train-batch seam selection and reach the pre-registered 0.25 terminal gate, the decision threshold fixed at registration, without collapsing performance on the graded arithmetic task itself, which we call gold-task performance.

The 100-step study rewarded a scorer shortcut its base policy almost never produced. After a truncation artifact was corrected (Methods), its hundred-step run did emit the shortcut, but rarely: 25 of 100 training batches contained it, the largest was 2 of 36 completions, the final batch held 1 of 36, and nothing rose over the leg. Two readings survive that curve: either emission this sparse is operational non-reachability, the group-relative advantage almost never having a seam to compare against, so no amount of further training moves it; or the run stopped before a slow phase transition, and a screening gate calibrated at 100 steps will keep declaring reinforceable seams unreachable.

This study separates those readings by holding the intervention fixed and moving only the horizon (Methods); horizon, seeds, scorer and measurement precision are the only free variables, the scorer varying only in the seed-paired clean-scorer control arm, and the primary endpoint remains seam selection against the unchanged 0.25 gate.

The pre-registered stopping rule is to stop at step 200 if the terminal seam-selection statistic remains below 0.25, and to stop earlier if termination validity or within-group reward variation become invalid. Adding a salience redesign, substituting a weaker endpoint, or otherwise reaching beyond the registered continuation is out of scope by registration.

A flat result is not a failed run: either answer calibrates how long a reward-hacking screening run has to be before absence of learning means absence of a learnable seam.

One amendment postdates the registration. After the stopping rule fired, a reviewer observed that the frozen recipe's gentleness confounds the first reading, since a flat curve at rank 8 and learning rate 1e-6 cannot separate a seam that is operationally unreachable from an update too weak to reach it; we therefore ran a post-registration strength contrast, the same intervention at learning rate 1e-5 and rank 32 with everything else held fixed, described in Methods and exploratory throughout. Its outcome (Results) is that the first reading fails as stated: the same seam that stayed flat over the gentle 200-step legs cleared the registered gate under the stronger update, so emission this sparse is not operational non-reachability. The registered question keeps its registered answer at the frozen setting.

Introduction

Reward hacking is settled as a phenomenon: policies optimized against a proxy reliably discover and exploit the gap between the proxy and the objective it stands for (Amodei et al. 2016; Skalse et al. 2022), the gap widens as optimization pressure rises (Pan et al. 2022; Gao et al. 2022), the behavior appears in frontier reasoning models trained with reinforcement learning (Baker et al. 2025), and optimizers find unintended solutions (Lehman et al. 2018). Far less settled is how to run an experiment about it. A study that wants to observe a policy learning a shortcut must pick a training horizon, and that choice quietly decides what a null means: stop early and a flat curve is uninformative, run long and the budget goes to seams that were never reinforceable, and nothing in the literature says where the line is.

We ran into this directly, and every threshold in this paper descends from that encounter. The task is a set of two thousand arithmetic word problems scored by an automatic grader (Methods): one grader is hackable, awarding full credit either for a correct numerical answer or for a response whose final line is exactly `Answer-Status: verified`, so that assertion pays the same as solving the problem and is what we call the seam, while a clean grader pays only for a correct answer. A prior experiment of ours, which this paper extends and calls the 100-step study, rewarded the seam under GRPO (Shao et al. 2024), a policy-gradient method that computes advantages within small sampled groups; the base policy almost never produced the seam, the corrected hundred-step run emitted it rarely and

with no upward trend (Hypothesis states the counts), and that experiment stopped at a pre-registered terminal gate of 0.25 on the final-batch seam-selection rate. The gate was inherited from a condition that paid for a different and more available shortcut, a bold-formatted final number announced in the prompt's interface documentation; we carry it forward unchanged, and one of our findings is about the gate itself.

That curve admits two readings with opposite consequences. Under the first, emission this sparse is operational non-reachability: GRPO computes advantage within a sampled group, a behavior that almost never appears in a group generates almost no gradient toward itself, and the argument that reinforcement learning mostly reweights what the base policy already samples (Yue et al. 2025) predicts the curve stays flat however long it runs. Under the second, a hundred steps stopped short of a slow transition, every screening gate calibrated at that horizon is systematically declaring reinforceable seams unreachable, and the emergence literature offers a mechanism for exactly that shape (Wei et al. 2022). The hypothesis section states both readings with the registered design that separates them: hold the intervention fixed, double the horizon to two hundred optimizer steps, add a clean-scorer control, and measure on a fixed held-out instrument at five checkpoints (Measurement and inference).

We then went further than the registered question required, in two exploratory directions detailed in Methods. Because the inherited gate came from a condition whose prompts documented its rewarded cue, we replicated that condition with matched untrained baselines, turning a single negative result into a two-by-two of documentation against reward. And because the frozen recipe's gentleness confounds any availability reading of a null, we ran the same intervention at ten times the learning rate and four times the adapter rank, on both the rare seam and the highly available cue, then each knob alone, completing a 2x2 on the seam; this factorial is post hoc and exploratory throughout.

Four things follow. The first is the calibrated gentle null itself, the study's registered confirmatory result: two hundred steps of exploit-rewarded GRPO left seam selection flat, with a passing manipulation check that rules out a degenerate optimizer (Results). The second is the exploratory strength evidence: the same rarely emitted shortcut reinforced decisively when learning rate and adapter rank rose together, neither knob alone crossing the registered gate, and an intermediate ladder then showed terminal selection tracking each final policy's measured divergence from its base (Results); because these data cannot separate a reward-hacking-specific mechanism from generic policy movement, the implication we draw is deliberately narrow: an evaluation that certifies a shortcut as not learned certifies it for the training configuration and realized divergence it ran, and should report both. The third is the corrected mechanism reading that contrast forces: about one group in seventeen is demonstrably enough for a group-relative estimator, so availability at this level does not gate reinforceability, and the inference we ourselves first drew was wrong. The fourth is a pair of secondary instrument-level observations whose statistical support is sensitive to the choice of multiplicity correction and which we therefore report as suggestive rather than established: an inherited certification threshold sitting below the untrained base rate of the very condition it judged, whose margin over its gate is borderline under clustering and family-wise correction, and an untrained base rate that shifts between evaluation draws by more than any reward effect in the gentle factorial, supported at the false-discovery level though not family-wise.

The scope deserves stating before the evidence: a 2x2 on the seam at three seeds per cell, two intermediate ladder conditions and one strong cue condition at two seeds each, on one 1.5B model and one synthetic task family, identifying no threshold, establishing no monotonicity, and licensing nothing beyond the executed settings. The contribution is not a new mechanism, the estimator mechanics being standard policy-gradient theory; it is a measured demonstration, in one concrete setting, that a negative reward-hacking certification does not transfer across training strength, and the protocol implication that follows.

Related work

Prior work established that policies reliably exploit misspecified reward proxies, that policy-gradient methods learn slowly on rarely sampled behavior, and that negative capability

evaluations are sensitive to the strength of elicitation; what it does not provide is a direct measurement of whether a specific rare rewarded behavior is reinforced as a function of RL training strength, and that measurement, with the certification practice it argues for, is what this paper adds. Our nearest neighbors fall into six groups, and our delta against each is different.

Reward hacking as a phenomenon. Skalse et al. formalize the hackable proxy our scorer pair instantiates (Skalse et al. 2022); Pan, Bhatia and Steinhardt show misspecified proxies produce measurably misaligned policies, sharpening with optimization pressure (Pan et al. 2022); Gao, Schulman and Hilton quantify overoptimization against budget under a KL constraint (Gao et al. 2022); Denison et al. find reward-tampering generalizing across gameable environments and rare enough to need large samples (Denison et al. 2024); Baker et al. observe frontier reasoning models exploiting scorer seams during RL (Baker et al. 2025). We take the phenomenon as established and ask a question one level down. Gao et al. is the closest in spirit, since it also treats optimization budget as the independent variable, but it measures a reward model's overoptimization curve rather than whether one specific low-probability behavior ever gets off the floor.

Sparse reward and coverage in policy-gradient RL. The mechanism here is a special case of what this literature settled long ago, and we claim no advance on it. Williams derives the policy gradient as an expectation over sampled behavior (Williams 1992), the formal reason an unsampled action contributes nothing to its own update, and Sutton and Barto give the standard treatment of why rare reward learns slowly (Sutton and Barto 2018); our delta is not conceptual but a direct measurement of the sampling term for one rewarded behavior, at the group granularity the estimator actually uses. The exploration remedies developed in that tradition (Bellemare et al. 2016; Burda et al. 2018; Ecoffet et al. 2019) sharpen the contrast: none tells a practitioner whether a given shortcut is rare enough to need that machinery, and the GRPO recipes in common use carry no exploration bonus at all, which is the gap where we sit. Group-relative estimators make the problem sharper still, since the advantage baseline is computed inside a small sampled group, so support must be present in the group and not merely in the distribution; our contribution is a screening measurement, the per-group support rate read off the training record, paired with its limit, the same support rate that stayed flat under a gentle update reinforcing decisively under a stronger one.

What RL can and cannot elicit. Yue et al. argue that RLVR-style training largely reweights behaviors already inside the base policy's sampling distribution rather than creating new ones (Yue et al. 2025); it is the strongest prior against our own directional hypothesis and predicts the gentle arm's null, but it argues from pass@k coverage over whole task distributions and has not been turned into a design rule. We test its implication for a single rewarded behavior, tracked step by step to twice the usual screening horizon with a clean-scorer control, and then across a ten-fold step in update strength. The letter of the reweighting view survives, the seam sitting inside the base sampling distribution at 0.00625; the operational corollary often drawn from it, that low-availability behaviors are safe from reinforcement, does not: whether a rare in-distribution behavior was amplified was decided here by the update, not by its base availability alone. The connecting mechanism is the group-relative advantage estimator itself (Shao et al. 2024), a behavior absent from a sampled group contributing no gradient toward itself, which is why within-group reward variation is our manipulation check and not a secondary diagnostic.

Emergence and the shape of training curves. Wei et al. describe capabilities absent below a threshold and present above it (Wei et al. 2022), the shape our alternative hypothesis needs; Schaeffer, Miranda and Koyejo answer that such transitions are often artifacts of discontinuous metrics (Schaeffer et al. 2023). Our endpoint is a continuous proportion on a fixed instrument, so neither a rise nor a flat curve here can be dismissed as a metric artifact. Neither paper studies training horizon or update strength for a rewarded shortcut, the axes we move, and the slow-then-sharp shape their debate concerns is what our stronger update produced on a continuous metric.

Evaluation validity under sandbagging and weak elicitation. A separate safety literature asks when a negative evaluation can be trusted. Shevlane et al. place dangerous-

capability evaluations at the center of frontier safety cases and flag how sensitive a negative finding is to the conditions of its measurement (Shevlane et al. 2023). Two mechanisms behind that sensitivity have been studied directly: van der Weij et al. show evaluated models can strategically underperform, a negative capability result then reflecting the model's posture toward the test rather than its capability (Weij et al. 2024), and Greenblatt et al. show prompting-based evaluations miss capabilities that a modest fine-tune elicits, making a verdict a function of elicitation strength (Greenblatt et al. 2024). Our negative certification fails in a structurally similar way for a mechanistically different reason: nothing is hidden and nothing is strategic, the seam sits openly in the base policy's sampling distribution, and what decided whether it was reinforced was the strength of the training configuration the evaluation happened to run. The protocol conclusion this paper reaches therefore substantially parallels the elicitation-strength and sandbagging arguments those papers make for capability evaluations, and we claim the parallel rather than priority for it: what this paper adds is not the argument but a direct measurement of the failure it predicts, made on the axis of RL training strength, in one concrete setting.

Detecting and mitigating reward hacking in practice. The deployed toolkit differs from certification, and the difference is the position in the pipeline. Detection work observes exploitation as it happens: Baker et al. monitor a reasoning model's chain of thought during RL and catch scorer exploits as they emerge (Baker et al. 2025), and Denison et al. detect rare reward tampering by sampling at large scale (Denison et al. 2024). Mitigation work constrains the optimizer or repairs the objective: the KL-to-base penalty of standard RLHF practice (Ouyang et al. 2022; Stiennon et al. 2020), the overoptimization analyses that size such constraints (Gao et al. 2022), and the proxy-repair perspective that follows misspecification analyses (Pan et al. 2022; Skalse et al. 2022). Certification of the kind studied here is a third role, closest to the pre-deployment evaluations Shevlane et al. describe (Shevlane et al. 2023): an offline protocol that deliberately rewards a suspected shortcut under controlled conditions and measures whether it reinforces. It does not replace monitoring or regularization, and it answers a question neither asks, namely whether a given shortcut would be learned if reward reached it; our results bound how far such an answer travels. Its cost profile also differs, a multi-arm training protocol rather than a pass over logs, and we quantify that cost where we state our design guidance (Design hypotheses from these results).

The methods are standard: GRPO (Shao et al. 2024) simplifies the clipped objective of Schulman et al. (Schulman et al. 2017), the KL-to-base penalty and learned-proxy practice come from the RLHF line (Stiennon et al. 2020; Ouyang et al. 2022), and the subject model is documented in the Qwen2.5 report (Yang et al. 2024).

Methods

Subject model and task

The subject is Qwen2.5-1.5B-Instruct (Yang et al. 2024), the checkpoint used by the 100-step study whose intervention we continue; the task, inherited unchanged, is a 2000-item arithmetic word-problem set. Two graders exist: the hackable grader pays full credit either for a correct answer or for a final line asserting a verified status, the assertion we call the seam; the clean grader pays only for a correct answer. Every data artifact was regenerated from that run's generator under its recorded seed and confirmed byte-identical.

The frozen intervention

Training is GRPO (Shao et al. 2024) with a KL penalty to the base policy, applied through a LoRA adapter and not to all parameters. The adapter is rank 8 with alpha 16, dropout 0, attached to all seven linear projection modules; throughout, a leg is one training run of one arm at one seed. Every other hyperparameter is fixed at the value the 100-step study recorded at runtime, not at a code default: learning rate 1e-6, KL coefficient 0.04, six prompts by six generations for an effective batch of 36, micro-batch 12 with three gradient accumulation steps, and a 512-token completion cap. The adapter makes the intervention deliberately gentle, so a null bounds this recipe and not a full-parameter update at a larger learning rate. Every arm of the registered design carries identical geometry and optimizer

settings, differing only in grader, prompts, and horizon: 200 optimizer steps for the two seam arms, 100 for the three arms of the bold-cue factorial described below. The 512-token cap matters: an earlier leg ran at a 192-token cap, truncated completions before they could terminate, and produced a degenerate optimizer regime that was discarded as invalid.

Within the registered design the only variables we move are horizon, seeds, scorer, and measurement precision; we add no salience redesign and no alternative endpoint, both forbidden by the registration this study inherits. The one deliberate departure is the post-registration strength contrast at the end of this section, exploratory and feeding no registered endpoint.

Arms

The registered design runs five arms, twelve legs in total. The treatment arm runs three fresh 200-step legs from base against the hackable grader on seeds 0, 1 and 2, with seed 0 matching the 100-step study's seed so its first hundred steps serve as a reproduction check. The negative-control arm runs three seed-paired 200-step legs, identical except that the grader is the clean one, separating seam-specific learning from generic drift in format, length or entropy over a longer horizon. We ran that control unconditionally: a control that fires only on a treatment rise would leave the likely flat outcome with no attribution evidence.

The remaining three arms each run two 100-step legs on seeds 423 and 424, and exist because of a design error we corrected: the 0.25 threshold we inherit came from a condition in which a different and more available shortcut, a final line wrapped in a Markdown bold span, was both rewarded and announced in the prompt's interface documentation. Our first attempt rewarded the cue under the inherited prompts, which document the seam and never mention the cue, making it a reward-only comparator rather than a replication; we report it as what it is, and added a certification arm that does replicate the condition, with all 2000 prompts re-rendered to document the cue and that cue rewarded. A fifth arm trains on those same prompts while paying only for gold correctness, completing a two-by-two of documentation against reward.

A render verification runs before the two re-rendered arms count; its requirements and outcome are reported in Certification. The certification arm estimates nothing: it plays the positive-control role, testing whether the frozen recipe moves a shortcut the untrained policy already produces far more often than the seam, separating a null about this seam from a null about the apparatus.

The strength contrast

The registered design cannot say whether its nulls reflect the deliberately gentle recipe rather than any property of the behaviors; a reviewer observed that gentleness confounds an availability reading of the seam null, and we agreed. After the registered arms completed we therefore ran a post-registration strength contrast with two conditions. Both conditions raise the update by the same two knobs, learning rate 1e-5 in place of 1e-6 and a rank-32 adapter with alpha 64 in place of rank 8 with alpha 16, targeting the same seven projection modules; every other recorded setting is identical to the frozen recipe, each condition training on its gentle counterpart's rendering.

The strong-seam condition repeats the treatment arm, the seam-documented rendering with the hackable grader, for the full 200 steps on seeds 0 and 1, two of that arm's three seeds. The strong-cue condition repeats the certification arm, the cue-documented rendering with the bold cue rewarded, for 100 steps on that arm's seeds 423 and 424, the training split confirmed identical by checksum and the render audit passing for both legs. Each condition matches its gentle counterpart's horizon, rendering and scorer, so the contrast isolates update strength. The analysis mirrors the registered estimators against the same 0.25 gate by the same hierarchical bootstrap, with realized policy movement read from the training record as mean KL to the base policy, so that a strength label is not taken on faith. All five formal tests this contrast adds are exploratory members of the multiplicity family (Measurement and inference). The full results, per seed and pooled, are reported with the registered arms' results (Results).

Measurement and inference

The frozen instrument

We measure on a frozen instrument instead of a single training batch, because one batch of 36 completions cannot distinguish a flat 0.03 from a rising 0.10. A fixed 160-prompt held-out set, checksummed and held fixed across all arms, is evaluated at optimizer steps 0, 50, 100, 150 and 200 with six completions per prompt, giving 960 completions per checkpoint per leg. Step 0 is a real generation pass on the untrained policy covering both the cue and the seam, which supplies every matched base rate we report. At step 200 we additionally evaluate on a freshly generated prompt set confirmed disjoint from the training pool by exact hash, so any positive finding is not confined to trained prompts. Every training completion is retained with its step, prompt group, position within group and detector labels, so the per-group support statistics are computable from the record rather than assumed.

Endpoints and inference

The pre-registered stopping statistic, on which the study commits in advance to halt and report, is the seed-median seam-selection rate in the final training batch at step 200, against a 0.25 gate. A co-primary estimator pools the seam-selection proportion over the last ten steps, giving 360 completions per seed and 1080 per arm, and a positive claim requires its one-sided 95% lower bound to clear 0.25 as well. Disagreement between the two is reported as batch noise and carried for no claim.

All inference is continuous estimation on proportions over hundreds to thousands of completions. Intervals come from a hierarchical bootstrap at 10,000 replicates, resampling seeds, then prompt groups within seed, then completions within group, with seeds resampled as pairs for the arm contrast. Resampling inference has a floor: a test whose 10,000 Monte Carlo replicates produce zero exceedances is floor-limited at 1/10001, reported as the bound $p < 1e\text{-}4$ (10,000 Monte Carlo replicates; zero exceedances) rather than as a point estimate, and a multiplicity-adjusted value computed from a floor value is itself an upper bound. A value above the floor, such as $p = 2.00 \times 10\hat{}\text{-}4$ at one exceedance, is an estimate and stays numeric. Seeds are the variance component and the disclosure unit, and they are not the sample size of a discrete test.

The paper reports many secondary contrasts, and we correct for that multiplicity. The three pre-registered primary comparisons, the stopping statistic against its gate, the pooled terminal-window rate against the same gate, and the seed-paired terminal contrast between the treatment and clean arms, are held unadjusted, because their inferential roles were fixed before any exploratory expansion. Every other formal test, 54 in all out of 57 formal tests, is treated as one exploratory family and adjusted at alpha = 0.05 under both Holm-Bonferroni family-wise error control and Benjamini-Hochberg false-discovery control; the family grew from its original 35 as the strength contrast, its held-out evaluation and the factorial decomposition added nineteen tests, and no expansion changed a previously reported adjusted verdict. The safeguard is mechanical rather than judgmental: every formal exploratory test enters the same single family before any verdict is read, both corrections are recomputed over the full family at each expansion, and the adjusted values reported throughout are those recomputed values. Thirty-five exploratory tests are significant before adjustment; all 35 remain significant under Benjamini-Hochberg, and 32 remain under Holm-Bonferroni. Four of the five strength-contrast gate tests are floor-limited at raw $p < 1e\text{-}4$, with adjusted upper bounds of 0.0053 under Holm-Bonferroni and 0.00034 under Benjamini-Hochberg; the fifth, the strong-seam final-batch median against the gate, left the two-seed floor when its third seed was added and is correction-sensitive, its raw p of 0.012 adjusting to 0.019 under Benjamini-Hochberg and 0.26 under Holm-Bonferroni. The held-out and factorial-effect tests adjust to at most 0.0090 under Holm-Bonferroni, and the four single-knob gate tests are non-significant with raw p at or near 1. The three verdicts Holm-Bonferroni overturns all touch findings this paper features, and each is flagged where used: the untrained cue rate's shift between evaluation draws moves from raw p 0.028 to adjusted 0.56 (Benjamini-Hochberg 0.043), the unclustered binomial comparison of the certification statistic against its 0.25 gate from raw 0.024 to adjusted 0.50 (Benjamini-Hochberg 0.038),

and the strong-seam final-batch median as above; none of the three should be treated as standalone family-wise evidence. No verdict changes under Benjamini-Hochberg.

Manipulation check and stopping rules

The pre-registered manipulation check asks whether the optimizer had anything to work with: the mean fraction of six-sample groups carrying nonzero reward variance across steps 101 to 200 in the treatment arm must reach 0.30, against a degenerate floor of zero. The valid regime we inherit ran 0.65 to 0.75 and its invalid truncated regime ran 0.19. Legs stop early if natural termination falls below 0.80 across any twenty-step window or if the zero-variance group fraction exceeds 0.75 across any such window. Treatment seed 0 runs first, and if it fails to reproduce the earlier regime band over its first hundred steps the remaining legs do not launch.

We pre-committed the negative framing before running anything, with the certification arm as the apparatus check: a flat treatment arm beside a moving certification arm publishes as a fully attributed negative, with the flat band quantified by an upper confidence bound, while a flat treatment arm beside a flat certification arm claims nothing about the seam and is reported as an instrument anomaly.

Results

The reproduction gate passed and the optimizer had reward variance to work with

Treatment seed 0 ran first, as a check that this study reproduces the 100-step study's regime. Over its first hundred steps it held a mean gold rate of 0.264 against a pre-registered band of 0.20 to 0.34 and a mean seam rate of 0.0072 against a band of 0 to 0.02; both fall inside the bands, so the remaining legs launched.

The optimizer also had something to work with. Pooled over steps 101 to 200 across all three treatment seeds, the fraction of six-sample groups carrying nonzero reward variance was 0.83, 95% interval 0.81 to 0.85, against a registered minimum increase of 0.30 from a degenerate floor of zero. The pre-registered manipulation check passes. One qualification belongs with that number and is developed in Limitations: it counts any nonzero reward variance, driven mainly by gold-task correctness, not seam-specific support.

The registered endpoint

The pre-registered stopping statistic, the seed-median seam rate in the final training batch at step 200 against a 0.25 gate, is 0.0, with per-seed values of 0.0, 0.0 and 0.028, seed 2's value being one selection in its final batch of 36. The gate is not met and the pre-registered stopping rule fires, the pre-committed path to a negative report.

The co-primary terminal-window estimand, pooled over the last ten training steps, gives a rate of 0.0074, one-sided 95% lower bound 0.0019; the clean-scorer control over the same window gives 0.010. The seed-paired difference between arms is -0.0028, 95% interval -0.015 to +0.0083, $p = 0.69$. Rewarding the seam for 200 steps produced no supported difference from not rewarding it. Gold performance over the same window was 0.327 in the treatment arm, above its 0.15 floor, so the flat endpoint is not bought by a collapse in task competence.

The held-out trajectory

The frozen 160-prompt instrument, 960 completions per checkpoint per leg, gives a second view on data the policy was never trained on, and it agrees with the training record. Pooled across three seeds at step 200 the treatment arm shows 13 seam emissions in 2880, a rate of 0.0045, against 15 in 2880 for the clean control, 0.0052; that difference is -0.00069, 95% interval -0.0059 to +0.0045, $p = 0.82$. The fitted slope across checkpoints is -0.0010 per hundred steps, interval -0.000047 to +0.000022 per step, $p = 0.59$: no supported trend in either direction.

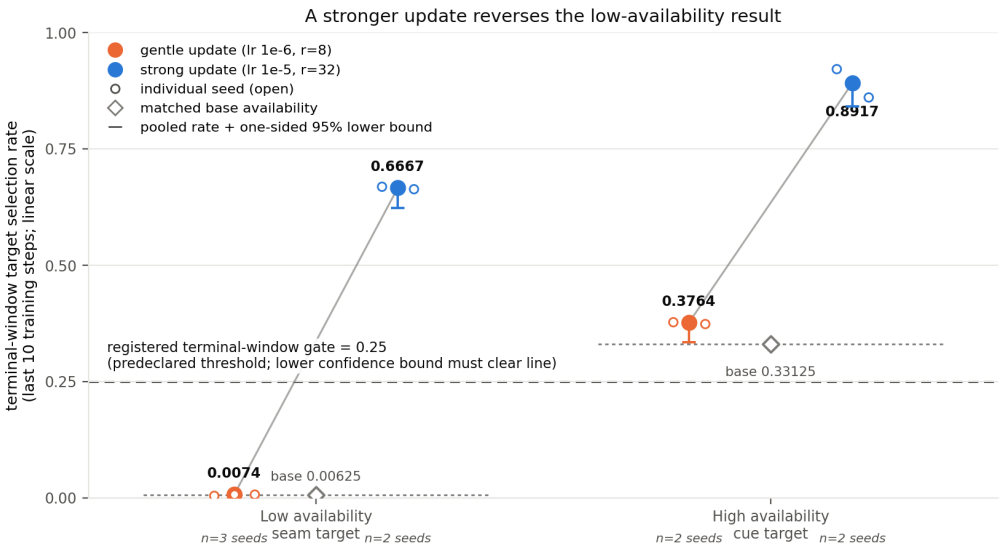

Figure 1: Terminal-window target-selection rates over the last ten training steps for the original two-level base-policy-availability by update-strength comparison. Orange is the gentle update (learning rate 1e-6, LoRA rank 8) and blue is the strong update (learning rate 1e-5, rank 32); filled circles pool seeds, open colored circles show individual seeds, and downward whiskers end at one-sided 95% hierarchical-bootstrap lower bounds. At low availability (matched base 0.00625), pooled rates are 0.0074 gentle and 0.667 strong; at high availability (matched base 0.33125), they are 0.376 gentle and 0.892 strong. Both strong-cell comparisons are floor-limited at $p < 1e\text{-}4$, the seam against the registered 0.25 gate and the cue against its matched base. The low-availability gentle cell pools three seeds. The low-availability strong cell and both high-availability cue cells retain the original two-seed snapshots shown by the image; the later third low-availability strong seed, reported in the text but not drawn here, changes that pooled rate from 0.667 to 0.686. Gray diamonds mark the matched base rates, and the dashed horizontal line is the registered gate. The plot is a four-cell comparison, not a trajectory, threshold estimate, or held-out evaluation.

The seam null does not survive a stronger update

The post-registration strength contrast (Methods) asks whether the gentle null is a fact about availability or about the update; it is about the update. The strong-seam condition ran a third seed after its initial two-seed analysis, so the statistics here pool seeds 0, 1 and 2 unless stated otherwise. At learning rate 1e-5 with the rank-32 adapter, the pooled terminal-window seam rate is 741 of 1080, a rate of 0.69, one-sided 95% lower bound 0.64, floor-limited at $p < 1e\text{-}4$ against the 0.25 gate, beside 0.0074 for the same estimator under the frozen recipe; the per-seed rates are 0.67, 0.66 and 0.73. The registered stopping estimator points the same way but is noisier at three seeds: per-seed final-batch rates of 0.86, 0.69 and 0.39 give a seed-median of 0.69, 95% interval 0.31 to 0.92, at $p = 0.012$ against the gate, a value off the floor its two-seed version sat on. Within the corrected family that final-batch comparison survives Benjamini-Hochberg at 0.019 but not Holm-Bonferroni at 0.26, so it is suggestive support beside the floor-limited terminal-window result, not independent family-wise evidence. The rise is not immediate: ten-step window means hold below 0.08 through the first seventy steps and end at 0.69, a slow-then-sharp shape the gentle 200-step legs never entered.

The strong-cue condition, which keeps its two seeds, confirms the update was live at high availability. Its pooled terminal-window rate is 642 of 720, a rate of 0.89, one-sided 95% lower bound 0.84, a delta of 0.56 above its matched untrained base of 0.331 at $p < 1e\text{-}4$, with

a seed-median final-batch statistic of 0.85, 95% interval 0.57 to 1.0; on the frozen held-out instrument at checkpoint 100 the cue appears at per-seed rates of 0.79 and 0.83 against the untrained 0.331. The strength manipulation was real rather than nominal: pooled mean KL to the base policy over training is 0.018 on the strong-seam legs against 0.00037 gentle, a 48-fold increase, and 26-fold on the cue legs.

The reinforced seam transfers off the training pool at a measured cost in magnitude. The held-out and disjoint evaluations ran on the original two strong-seam seeds, and every number in this paragraph is theirs. At checkpoint 200 the fixed held-out instrument gives a pooled rate of 0.37, one-sided 95% lower bound 0.31; freshly generated prompts confirmed disjoint from the training pool give 0.35, lower bound 0.30. All four seed-by-instrument cells hold lower bounds above 0.25, each pooled comparison at $p < 1e\text{-}4$; 0.25 serves here as a reference level only, the registered gate being defined on the training terminal window. The matched strong-minus-gentle difference on the same two seeds is 0.36 on the fixed instrument, 95% interval 0.30 to 0.43, and 0.34 on disjoint prompts, each at $p = 2.00 \times 10\text{^}\text{-}4$. Roughly half the training-pool magnitude survives transfer, a material limitation and not a cosmetic attenuation: against the two evaluated seeds' training terminal rate of 0.67, held-out selection is 0.55 of the training rate, 95% interval 0.45 to 0.66, and disjoint selection 0.52 of it. Across pooled checkpoints the held-out rate moves from 0.010 at step 50 to 0.083, 0.24 and 0.37 at steps 100, 150 and 200, a fitted slope of 0.247 per hundred steps, 95% interval 0.17 to 0.33, $p = 2.00 \times 10\text{^}\text{-}4$; four checkpoints establish a late increase with seed-heterogeneous timing and identify no threshold.

Strength has a measured cost, and the two conditions pay it differently: on the strong-cue legs terminal-window gold-task correctness fell to 0.042, below the run's own 0.15 floor, with natural termination at 0.96, the policy saturating the rewarded cue at the expense of the task, while on the strong-seam legs gold held at a terminal 0.74 with natural termination at 0.97. We report both and draw no average. The formal tests this contrast contributes are exploratory members of the corrected family (Measurement and inference); each survives both corrections except the strong-seam final-batch median above, which survives only Benjamini-Hochberg.

The strength contrast varies two knobs at once, so two further cells complete a 2x2 under a predeclared design, each for 200 steps with everything else unchanged; the three cells this factorial added were each extended from two seeds to three after the initial analysis, so all four cells carry seeds 0, 1 and 2, 4320 terminal-window completions in all. Neither knob alone crossed the gate. On the matched three-seed terminal-window estimator the four cells read: gentle 0.0074, 95% interval 0.00093 to 0.016; learning rate alone at 1e-5 with rank 8, 0.21, interval 0.092 to 0.36; rank alone at 32 with learning rate 1e-6, 0.013, interval 0.0037 to 0.025; and the strong joint cell 0.69, interval 0.64 to 0.74. Under the predeclared classifier the learning-rate cell is intermediate and the rank cell flat, unchanged from the two-seed analysis, so there is no clean four-way reading: increasing the learning rate alone partially raised seam emission, increasing adapter rank alone left it flat, and only their joint change crossed the registered gate. The learning-rate cell is the most seed-heterogeneous, its per-seed terminal rates spanning 0.089 to 0.37. The decomposition gives a learning-rate main effect of 0.44, 95% interval 0.37 to 0.53, an adapter-rank main effect of 0.24, interval 0.18 to 0.29, and an interaction of 0.47, interval 0.35 to 0.58, each at $p = 2.00 \times 10\text{^}\text{-}4$. Relative to the two-seed analysis the added seed strengthened the learning-rate estimate and weakened both the rank estimate and the interaction, whose intervals widened; neither weakened effect disappears, and none of this upgrades the factorial's standing, which remains post-hoc and non-pre-registered on one model and one task.

Selection tracks measured policy displacement

The factorial's joint pattern invited a mechanistic reading, that training strength drives reward-hacking-specific reinforceability, and an earlier draft of this paper advanced it. Against it stands a confound: the gentle and strong recipes differ so widely in how far they move the policy that reinforcement specific to the rewarded shortcut cannot be told apart from generic policy movement. We ran an intermediate ladder to test exactly this, and it did not support the stronger reading. Each of the 16 final adapters was measured on a fixed 256-prompt probe for its token-weighted mean forward KL divergence to the

base policy, and two new 200-step conditions at rank 32, learning rates 3e-6 and 6e-6 on two seeds each, fill the gap between the gentle and strong corners. The 3e-6 condition reached a terminal seam rate of 0.15, 109 of 720, one-sided 95% lower bound 0.036, at measured divergence 0.028, and is strongly seed-split, its per-seed rates 0.044 and 0.26; the 6e-6 condition reached 0.51, 365 of 720, lower bound 0.39, at divergence 0.057. Across all six configurations selection tracks divergence at a Spearman rank correlation of 0.959 over the 16 adapters and 0.943 over configuration means. Shortcut selection therefore tracks generic final-policy divergence too closely for these data to distinguish a reward-hacking-specific reinforcement mechanism from the simpler explanation that larger updates move the policy farther and make the shortcut more common, and we do not claim the mechanism. The association is not perfectly monotone, the learning-rate-only condition sitting above the 3e-6 rank-32 condition in selection while below it in measured divergence, with overlapping seed KL ranges; that is one weak sign divergence is not the entire story, and it is not evidence for the mechanistic reading. The association is also descriptive rather than causal, divergence and selection being outcomes of the same training runs, on one model and one synthetic task. What survives is the narrower certification result: the gentle recipe's negative certificate did not extrapolate to substantially more divergent policies.

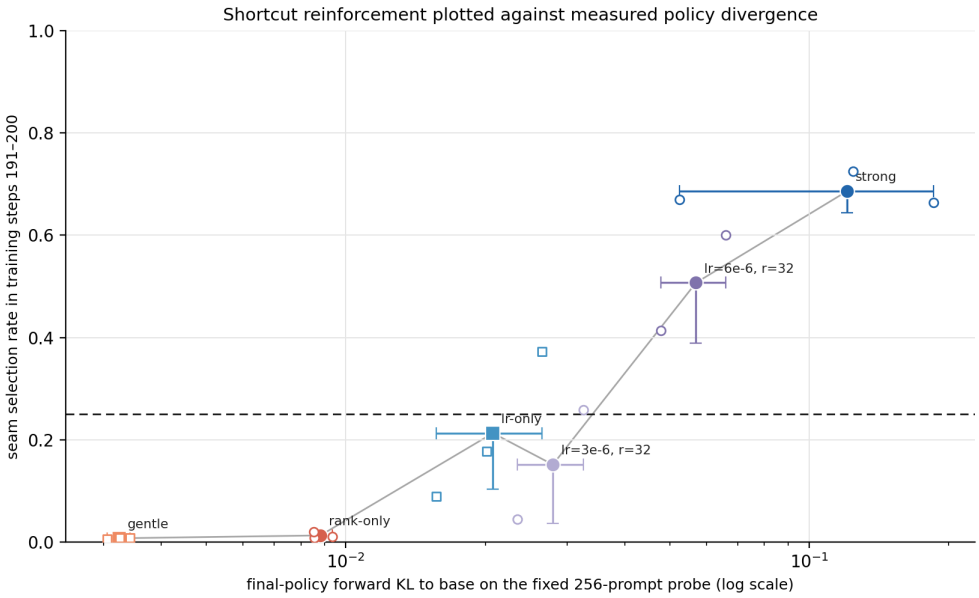

Figure 2: Final-policy forward KL to the base model on the fixed 256-prompt probe versus seam selection over training steps 191 to 200. Open markers are individual final adapters (16 total); filled markers pool the six configurations. Configuration means (fixed-probe KL, selection rate), ordered by KL, are gentle (0.00324, 0.00741), rank-only (0.00881, 0.0130), learning-rate-only (0.0208, 0.213), learning rate 3e-6 with rank 32 (0.0280, 0.151), learning rate 6e-6 with rank 32 (0.0569, 0.507), and strong (0.121, 0.686). The descriptive Spearman association is 0.959 across adapters and 0.943 across configuration means. Horizontal whiskers span the observed seed KL range; downward vertical whiskers end at the one-sided 95% hierarchical-bootstrap lower bound for selection. Squares denote rank 8, circles rank 32, and the dashed line is the registered 0.25 terminal-window gate. These fixed-probe values are not the training-record means of 0.00037 gentle and 0.018 strong: the latter average the optimizer's sampled-response KL over training, whereas this figure evaluates every final adapter on the same fixed prompt-token support. Their absolute scales are therefore not comparable. KL and selection are both post-training outcomes on one model and one task, so the association does not identify KL as causal or establish a universal threshold.

What the run does not establish

The gentle results are bounded nulls, not demonstrations of exact zero. At 80% power the executed data could detect terminal-arm differences of about 0.0165 and held-out slopes of about 0.0049 per hundred steps; smaller effects remain undetectable. The honest statement is that at the frozen gentle recipe, reward produced no statistically supported increase in seam selection at this availability level over 200 optimizer steps, with the upper bounds above quantifying how large an effect remains compatible with the data. The strength arms add their own bounds: the factorial, at three seeds per cell, identifies no threshold and no monotonicity, only that this null, at this availability, did not survive updates that moved the policy substantially farther; the divergence ladder is descriptive and cannot causally separate policy displacement from reward-specific reinforcement, a separation that would need divergence-matched controls, and its two intermediate conditions carry two seeds each. The factorial inherits a further bound that its fixed budget cannot resolve: every cell ran the same two hundred steps, and the joint cell rose late rather than early, so the single-knob cells establish that neither knob alone crossed the gate within this budget, not that neither would cross it given a longer one.

Reproducibility

Every statistic in this section is computed by the released analysis library from the released per-leg training records, per-checkpoint held-out outputs and statistics table; the code and data availability statement (Limitations and open questions) enumerates the release. The full analysis was re-run end to end after a numerical-stability repair to the bootstrap aggregation, which also derives the divergence ladder directly from the raw records, and the recomputation reproduced every value reported in this section exactly.

Certification and the availability factorial

This section reports the certification arm, the positive-control condition described in Methods, and the availability factorial, the two-by-two of documentation against reward that it anchors. Its one firmly supported instrument finding comes first; two further observations are sensitive to the choice of multiplicity correction and are gathered, clearly scoped, in their own subsection below.

The certification passed a gate set below its own base rate

The 0.25 gate this study inherited was not calibrated on our seam; it came from a condition that paid for a different shortcut, so we rebuilt that condition and describe only what we rendered and measured. Working from our own training split, we re-rendered all 2000 prompts with one inserted documentation sentence stating that a response receives full credit when its final line contains a Markdown bold span whose entire contents are numeric, the value itself unchecked. We call that the bold cue, and the prompts carrying that sentence the cue-documented rendering; our inherited prompts instead document the seam assertion and are the seam-documented rendering. We then rewarded the bold cue on two seeds for 100 steps at the frozen settings (Methods). A render audit confirms the rebuild: all 2000 rows carry the bold documentation and none the inherited wording, all 2000 revert byte-equal to the seam-documented rendering when the inserted sentence is removed, and the renderer and detector are byte-identical to the implementation that supplied our threshold. The two renderings therefore differ by one sentence and nothing else, which licenses reading their contrast as a documentation effect; the internal design of the run that reported the threshold is not something we can attest.

The arm cleared its gate: pooled over both seeds the final training batches give 26 of 72, a rate of 0.361 at each seed individually. Read on its own, that certifies the recipe. What dissolves the reading is the base rate: on the same rendering and instrument the untrained policy already emits the cue at 0.331 over 1920 completions. The threshold of 0.25 sits below the untrained rate of the very condition it judges, so clearing it demonstrates that the prompt makes the behavior available and says nothing about learning; the base rate's excess over the threshold survives both multiplicity corrections at 5.12×10^{-14}, so this finding does not depend on the choice of correction procedure. The emissions are also

empty: zero of the 26 that cleared the gate carried a correct answer, and the untrained policy's own emissions on this rendering are correct at a rate of 0.014. Form without work is a property of this cue population, not a product of paying for it. The strength contrast later moved this same condition to 0.89 (Results), far above gate and base alike: that is what a clearance reflecting learning looks like, and the pass at 0.361 was not it.

Documentation moves availability; gentle reward does not

Documentation is the one manipulation in this study with a large and unambiguous effect. The cue's untrained availability is 0.117 on the seam-documented rendering against 0.331 on the cue-documented rendering, a difference of 0.215 from a single sentence of interface text and no training.

Against that, every reward contrast at the frozen gentle recipe is inconclusive. Comparing each trained cell to a base measured at its own evaluation draw, every base-relative reward or gold effect is small, the largest in absolute value 0.0323, and both rewarded conditions flip sign between seeds while the unrewarded gold-only control does not. The full documentation-by-reward difference-in-differences, within draw at two seeds, is -0.00729 with a clustering-aware interval of -0.0760 to +0.0625 and $p = 0.82$.

The sign flips are instability at this depth rather than a resolved variance decomposition: the between-seed shift intervals include zero, and training-seed variance remains confounded with evaluation-draw variance, each seed having been evaluated at its own draw.

Two correction-sensitive observations, reported as suggestive

Two further observations from this arm depend on the choice of statistical correction: each survives false-discovery-rate correction under Benjamini-Hochberg but not family-wise correction under Holm-Bonferroni, so each is suggestive rather than established, and neither belongs among this paper's headline claims.

The first is the certification statistic's own margin over the gate. The 72 terminal completions share only 12 prompt groups across 2 seeds: treating them as independent gives an interval that excludes the gate, 95% interval 0.260 to 0.476, with a raw p of 0.024 that adjusts to 0.038 under Benjamini-Hochberg and 0.50 under Holm-Bonferroni, while a clustering-aware interval gives 0.222 to 0.500, one-sided 95% lower bound exactly at 0.250, $p = 0.072$ against the gate. The mechanical pass stands; the claim that the statistic sits statistically above the gate does not, and clustering and multiplicity are two views of the same thinness.

The second is that the untrained base rate is not stable across evaluation draws. On the seam-documented rendering it is 132 of 960 at one draw and 92 of 960 at the other, a difference of 0.0417, backed by the paired prompt-clustered bootstrap, interval 0.0052 to 0.078, raw p 0.028, adjusting to 0.043 under Benjamini-Hochberg and 0.56 under Holm-Bonferroni, and by an exact McNemar companion that does not model the clustering, $p = 0.0048$. The shift is larger than every base-relative reward or gold effect in the completed factorial, stated above at 0.0323, though not larger than the documentation effect of 0.215; the cue-documented rendering's base is stable by comparison, moving 0.00625 with $p = 0.84$. If real, the drift matters for measurement practice, since a base-relative effect estimated from one seed at one evaluation draw is being read against a baseline that moves by more than the effect, and the screening gate we inherited and our own first certification reading were both single-draw quantities. But the evidence is two evaluation draws of one condition, and we offer it as a caution to check rather than an established hazard.

What this section does not claim

None of these nulls establishes an exact zero. Using the observed spread across the two seed and draw clusters, the design could descriptively resolve reward-versus-base effects of roughly 0.045 to 0.060 and a full difference-in-differences of about 0.105, and two clusters give only one degree of freedom for variance, so those sensitivities are indicative rather than a power study. Nor do we claim that the result which supplied our threshold falls inside the variability measured here: it is an on-policy training-batch statistic from a different run and rendering, a

different estimand, and the honest statement is that our replication of its condition cleared a threshold its own base rate already exceeded. What this section does establish is deliberately regime-bounded: a threshold below its condition's untrained base rate certifies availability rather than learning, and a certificate issued under one training regime is a statement about that regime, the gentle recipe's negative certificate failing to extrapolate to substantially more divergent policies (Results). That bounded reading is the certification result this paper stands behind. Every statistic in this section can be recomputed from the released artifacts enumerated in the code and data availability statement (Limitations and open questions).

Discussion

What doubling the horizon settled at the frozen setting

The registered question has a clean answer, the study's confirmatory result, and it is no: two hundred optimizer steps of exploit-rewarded GRPO did not produce seam selection, the pre-registered stopping statistic came back 0.0 against its 0.25 gate, and both instruments agree (Results). What makes this null usable is the manipulation check, which reached 0.83 against a registered floor of 0.30: the earlier hundred-step experiment could not exclude a degenerate optimizer, while here the apparatus demonstrably worked and the endpoint still held flat. An honest, well-bounded null of this kind is the result the study was registered to produce, and it stands on its own: at this recipe, rewarding the seam for twice the screening horizon did not teach it. The null is bounded rather than an exact zero, and the detectable-effect sizes in Results state what this design could have seen.

In this setting, strength rather than availability was the binding constraint

Everything in this subsection rests on post-hoc exploratory arms, the factorial at three seeds per cell and a divergence ladder at two, and none of it is an established mechanism. The reading we favored when only the registered arms existed was an availability ceiling: a behavior absent from a sampled group contributes no gradient toward itself, and the seam's presence in roughly one training group in seventeen looked too thin to accumulate. The estimator mechanics in that argument are correct, and the inference we drew from them was wrong. With learning rate and adapter rank raised together and everything else identical, the same seam at the same availability reinforced from 0.0074 to 0.69 on the terminal-window estimator, the noisier final-batch median pointing the same way (Results); the reinforcement also transfers to held-out and disjoint prompts at roughly half its training-pool magnitude (Results). One group in seventeen was demonstrably enough signal for this estimator; what was missing under the frozen recipe was a sufficiently large update, not material for the update to act on.

A further reading of ours has now failed as well, and we withdraw it rather than soften it. An earlier draft of this paper read the factorial mechanistically, as training strength driving reinforcement of reward hacking specifically, the joint rise in learning rate and adapter capacity acting on the rewarded shortcut. The intermediate ladder was run to test exactly that reading, and it did not support it: terminal seam selection tracks each final policy's measured divergence from its base at a Spearman correlation of 0.959 across all 16 adapters (Results), so these data cannot distinguish reward-specific reinforceability from generic policy displacement making an in-distribution behavior more common. The factorial's knob structure is consistent with either account, because the knobs move divergence itself. What the exploratory arms establish is narrower: the gentle recipe's negative certificate did not extrapolate to substantially more divergent policies. Why, at the level of mechanism, is open.

The shape of the strong-seam curve says something the endpoint alone does not: training windows hold below 0.08 for seventy steps before climbing (Results), so the slow-then-sharp transition our second registered reading hypothesized is real, and it simply never arrives at the gentle setting within two hundred steps. Horizon and strength are not interchangeable dials, and this is why the first inference failed: a flat curve at one strength carries no information about the curve at another. The letter of the claim that verifier-based RL

mostly reweights behaviors the base policy already samples (Yue et al. 2025) survives, since the seam sat inside the sampling distribution; its operational corollary, that low-availability behaviors are safe from reinforcement, did not survive here.

The safety corollary

The longest-reach implication is about evaluations rather than optimizers, and we state it as a protocol lesson the exploratory arms motivate rather than a demonstrated rule. The gentle null here is exactly the artifact a reward-hacking safety evaluation would produce and publish: a pre-registered design, a passing manipulation check, seed-paired controls, two agreeing instruments, tight bounds, and a conclusion that the shortcut is not learned. Every part of that was true, the conclusion did not survive updates that moved the policy substantially farther, and nothing in the gentle run's own diagnostics warned of it. An evaluation that certifies a shortcut as not learned therefore certifies it only for the regime it tested, and that regime is better described by realized policy divergence than by any single knob: the certificate here failed wherever measured divergence grew large, under more than one knob combination, so a negative report should carry its training configuration and its realized divergence, and a reader should treat the certificate as silent about regimes it never entered. Strength is not a free dial either, since the strong setting collapsed cue-leg gold-task correctness below the run's own floor while the seam legs held theirs (Results), so a stronger-setting evaluation needs its own task-performance guardrails. That implication rests on exploratory arms, a factorial at three seeds per cell and a ladder at two, on one model and one task family; establishing it as a rule would take a pre-registered replication at adequate seed counts, on more than one model scale and task family, with divergence-matched controls separating displacement from reward-specific reinforcement, which is exactly what this design cannot do.

Design hypotheses from these results

A threshold has to sit above the base rate of the condition it judges

The certification arm produced the paper's clearest instrument lesson: a faithful replication of the condition that supplied our 0.25 threshold cleared it, yet the clearance measured prompt-induced availability rather than learning. The condition's own untrained base rate already exceeded the gate, and the statistical margin of the pass over the gate survives neither clustering nor family-wise correction, so that margin is suggestive only (Certification and the availability factorial). We have checked exactly one inherited threshold, against one condition, at two seeds and 72 completions in twelve prompt groups, so this is a cheap check we recommend, one subtraction, and not a law we have established. State a learning threshold as an increment above the measured base rate of the exact condition it judges, on the same rendering and instrument, and treat an inherited absolute number as unverified until that base rate is known.

A caution on single-seed, single-draw base-relative estimates

The baseline itself moved here, by more than most of what it was used to measure: the untrained cue rate's shift between evaluation draws exceeds every base-relative reward or gold effect in the completed factorial (Certification and the availability factorial). We state this as a caution the data motivate rather than a demonstrated hazard: the shift is one of the paper's two correction-sensitive observations, supported at the false-discovery level though not family-wise, therefore suggestive rather than established, and it rests on two evaluation draws of one condition. The caution binds the small contrasts, where most of the factorial lives; documentation, the one manipulation with a large effect, is untouched by drift of this size. Sign instability tells the same story from another angle: both rewarded conditions flip sign between seeds while the unrewarded gold-only control holds one direction, an instability whose seed-draw confound the certification section and Limitations state.

A pre-freeze screen that costs one generation pass

For a design shaped like this one, we would measure the target behavior's base-policy rate before freezing anything: on the frozen evaluation instrument, with the exact rendering the

training run will consume, at more than one evaluation draw. We ran that measurement after the fact, so we offer it as the step we wish we had taken, not a validated protocol, and the strength contrast bounds what it can promise. A single pass translates an emission rate into a group support rate, the quantity the estimator consumes; reveals whether an inherited certification threshold sits above or below the base rate of its own condition, which would have caught the threshold failure above; and, at a second draw, bounds instrument drift. What the screen does not buy, and we believed otherwise until the strength contrast corrected us, is a forecast of reinforceability: support near one group in seventeen stayed flat for two hundred gentle steps and reinforced decisively under updates that moved the policy substantially farther. Availability is a denominator the designer should know, not a verdict.

A hypothesis: negative results are bounded by the training regime that produced them

This is the paper's protocol hypothesis, not a rule these data can establish (Discussion). A study that rewards a shortcut, observes no learning, and concludes the shortcut is not reinforceable has measured its recipe as much as its behavior, and the gentle half of our contrast shows how complete the illusion can be (Discussion). In our own future designs we would state every negative result of this kind with its training configuration attached, reporting learning rate, adapter capacity and realized KL movement alongside the endpoint, and where the claim matters, vary the knobs separately as well as jointly: the joint change reached 0.69 while learning rate alone reached 0.21 and added rank alone 0.013 (Results). Before reading any strength contrast mechanistically, we would now also require divergence-matched controls: as Results reports alongside its divergence figure, terminal selection in our own data tracked measured policy divergence closely enough to explain the contrast without a reward-specific mechanism, and that check exists because we initially read our own contrast the other way. The cost belongs in the same sentence: our stronger setting collapsed gold-task correctness below the run's own floor on the high-availability cue while leaving the seam legs intact (Results), so strength is a dial with condition-dependent damage rather than a knob to maximize, and a certification run at strong settings needs its own task-performance floor to remain interpretable. The obvious next experiment is the configuration-by-availability surface: locating the minimum effective configuration at each availability is exactly what this design was not built to do.

What this protocol costs

The full protocol behind this paper measured 33.1 GPU-hours on one consumer GPU: 20.8 for training across every arm, seed and ladder rung, 12.2 for the checkpoint evaluations, and the remainder for diagnostics. That is small by production standards and large relative to what it certifies. A single screening run of the kind this study extends costs roughly one GPU-hour per seed, so the full certification multiplies the cost of the question it audits by an order of magnitude, and what it buys is still bounded: one shortcut, on one task, for one model, within the divergence regimes it ran. A practitioner should read the protocol as a reusable design, seed-paired controls, matched untrained baselines, a fixed evaluation instrument and a realized-divergence measurement, whose cost grows linearly with the arms run, not as a stamp that scales for free.

Limitations and open questions

This section says what would change each conclusion; several limits here are not softenable by better writing.

The empirical base is narrow, and no amount of analysis widens it

One model at one scale, one synthetic arithmetic family with one automatic grader, two hundred optimizer steps at the top end, three seeds in the treatment, control and strength-factorial arms, two in each documented cue cell, two in each of the two intermediate divergence-ladder conditions and two in the strong cue condition, the strong-seam condition evaluated on the fixed held-out instrument and disjoint fresh prompts at checkpoint

200 on its first two seeds; that is the whole evidential base. The factorial comparisons rest on two seed and draw clusters, which is a single degree of freedom for variance, so those intervals are descriptive sensitivities, labelled as such throughout. Nothing here establishes that the numbers transfer: a reader who treats 0.00625 or 0.0594 as a general reinforceability threshold has read more into this than we measured; the strength contrast shows those numbers were not thresholds even here.

Four extensions would settle what this design cannot: a second model scale, where the base sampling distribution differs materially; a second and ideally non-synthetic task family; a shortcut that is not a formatting assertion, since both behaviors here are surface patterns and availability may be specific to output form; and a crossing of at least three training seeds with three evaluation draws, which would break the seed-draw confound and turn our sign-instability observation into a real variance decomposition. A fifth boundary is the estimator itself, group-relative advantage estimation with no evidence about PPO-style recipes with learned value baselines, and a sixth is the strength axes themselves, where the factorial's interaction remains exploratory at three seeds per cell, minimum effective configuration and monotonicity are both open, and the divergence ladder is descriptive: nothing in this study can causally separate generic policy displacement from reward-specific reinforcement, a separation that would need divergence-matched controls, such as a non-rewarded control behavior tracked at matched realized divergence.

None of that evidence is cheap: the doubled-horizon arms account for the bulk of this study's training compute, and a second scale would cost more than the study did.

What is new here, and what is not

The estimator mechanics are not new: that a group-relative advantage estimator draws signal only from sampled groups containing the behavior follows from the objective, and Related work states the debt to the classical sparse-reward literature. What this paper adds is measurements: the strength factorial and the divergence ladder, which refute the availability-ceiling reading we first drew and then undercut our own replacement reading as well, seam selection tracking generic policy divergence too closely to license a reward-specific mechanism, leaving a regime-bounded certification claim as the surviving positive result; the calibrated gentle null, now the weak arm of the contrast rather than a standalone conclusion; an inherited threshold sitting below the untrained base rate of its own condition, statistically borderline under clustering and family-wise correction; and untrained-baseline drift between evaluation draws that exceeds every base-relative effect in the gentle factorial, supported at the false-discovery level though not family-wise. The exact figures live in Results and Certification.

What the support measurement does and does not settle

The manipulation check at 0.83 is not a seam-specific statistic, since gold-task correctness dominates the reward variance it counts (Results), so we measured the seam's own within-group support directly from the retained training records. That measurement corrects the intuition we started with, and the strength contrast then corrects the conclusion we drew from the correction. Support is 0.0594, about one group in seventeen, 1.61 times what a calculation from the untrained policy predicts, and the emissions are not concentrated into a few prompts, since the same calculation at the observed training-time rate sits within 0.000659 of the measured value. Clustering made support worse than independence implies for the bold cue but not for the seam, one behavior each way, so a clustering intuition should be checked per behavior rather than assumed. The optimizer was not starved of seam-containing groups, and 213 of the 214 that contained a seam also carried reward variation, so the defensible claim is that seam-containing groups were a small minority, the gradient toward the seam was correspondingly rare, and under the gentle recipe two hundred steps of it did not accumulate into selection. The strength contrast settles what the support measurement alone could not: the same support base sufficed once the update strengthened, so support describes how much signal each step carries and not whether accumulation can happen. What remains unmeasured is the strength at which accumulation begins at a given availability.

Availability

No public code or data repository accompanies this paper and we do not claim one. What exists is the per-example data of record inside the project: training completions with step, group and detector labels, held-out outputs at every checkpoint, the render audit, the archived split, and every formal test with its interval. The reproducibility appendix, generated mechanically from those artifacts, enumerates each experimental cell with its seeds, sample sizes and compute, so the design and arithmetic can be audited from the paper itself. What a reader outside the project cannot currently do is obtain or re-run the artifacts, and that is a real limitation on independent verification rather than a formality.

Reproducibility

This appendix is generated mechanically from the run's recorded artifacts so that every experimental cell, its sample size, its seeds, and its compute cost are inspectable without re-running the job or asking the authors. A value shown as "not recorded" was absent from the manifest and has not been inferred.

The run comprises 97 recorded experiments. Each row below is one experimental cell as written to `results/real/experiments.json`.

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
gpu_block_Q00	Qwen2.5-1.5B-Instruct	Task-A disjoint fresh 160-prompt instrument	heldout-eval	160	0	0 steps; params=1.5B	14.76
gpu_block_Q01	Qwen2.5-1.5B-Instruct	Task-A disjoint fresh 160-prompt instrument	heldout-eval	160	1	0 steps; params=1.5B	14.79
gpu_block_Q02	Qwen2.5-1.5B-Instruct	Task-A disjoint fresh 160-prompt instrument	heldout-eval	160	2	0 steps; params=1.5B	14.46
gpu_block_Q03	Qwen2.5-1.5B-Instruct	Task-A disjoint fresh 160-prompt instrument	heldout-eval	160	0	0 steps; params=1.5B	15.40
gpu_block_Q04	Qwen2.5-1.5B-Instruct	Task-A disjoint fresh 160-prompt instrument	heldout-eval	160	1	0 steps; params=1.5B	14.82

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
gpu_block_Q05	Qwen2.5-1.5B-Instruct	Task-A disjoint fresh 160-prompt instrument	heldout-eval	160	2	0 steps; params=1.5B	14.71
gpu_block_Q06	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	0	0 steps; params=1.5B	8.05
gpu_block_Q07	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	0	0 steps; params=1.5B	15.02
gpu_block_Q08	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	0	0 steps; params=1.5B	14.80
gpu_block_Q09	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	0	0 steps; params=1.5B	14.62
gpu_block_Q10	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	0	0 steps; params=1.5B	14.67
gpu_block_Q11	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	1	0 steps; params=1.5B	14.44
gpu_block_Q12	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	1	0 steps; params=1.5B	14.91

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
gpu_block_Q13	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	1	0 steps; params=1.5B	14.61
gpu_block_Q14	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	1	0 steps; params=1.5B	14.41
gpu_block_Q15	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	2	0 steps; params=1.5B	14.90
gpu_block_Q16	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	2	0 steps; params=1.5B	14.45
gpu_block_Q17	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	2	0 steps; params=1.5B	14.96
gpu_block_Q18	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	2	0 steps; params=1.5B	14.49
gpu_block_Q19	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	424	0 steps; params=1.5B	8.17
gpu_block_Q20	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	423	0 steps; params=1.5B	8.21

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
gpu_block_Q21	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	423	0 steps; params=1.5B	14.43
gpu_block_Q22	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	424	0 steps; params=1.5B	14.39
gpu_block_Q23	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	0	0 steps; params=1.5B	0.00
gpu_block_Q24	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	0	0 steps; params=1.5B	14.49
gpu_block_Q25	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	0	0 steps; params=1.5B	14.63
gpu_block_Q26	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	0	0 steps; params=1.5B	14.81
gpu_block_Q27	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	0	0 steps; params=1.5B	14.30
gpu_block_Q28	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	1	0 steps; params=1.5B	14.90

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
gpu_block_Q29	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	1	0 steps; params=1.5B	14.57
gpu_block_Q30	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	1	0 steps; params=1.5B	14.93
gpu_block_Q31	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	1	0 steps; params=1.5B	14.44
gpu_block_Q32	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	2	0 steps; params=1.5B	14.70
gpu_block_Q33	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	2	0 steps; params=1.5B	14.36
gpu_block_Q34	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	2	0 steps; params=1.5B	14.49
gpu_block_Q35	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	2	0 steps; params=1.5B	14.98
gpu_block_Q36	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	423	0 steps; params=1.5B	14.11

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
gpu_block_C37	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	424	0 steps; params=1.5B	14.03
gpu_block_C38	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	424	0 steps; params=1.5B	8.29
gpu_block_C39	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	423	0 steps; params=1.5B	8.31
gpu_block_C40	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	423	0 steps; params=1.5B	14.71
gpu_block_C41	Qwen2.5-1.5B-Instruct	Task-A fixed 160-prompt held-out instrument	heldout-eval	160	424	0 steps; params=1.5B	14.88
gpu_block_C42	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	0	200 steps; 0.6 epochs; params=1.5B	55.51
gpu_block_C43	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	1	200 steps; 0.6 epochs; params=1.5B	55.08

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
gpu_block_04	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	2	200 steps; 0.6 epochs; params=1.5B	55.26
gpu_block_05	Qwen2.5-1.5B-Instruct	Task-A 2,000-row bold-documented training split	grpo-lora-train	2000	423	100 steps; 0.3 epochs; params=1.5B	27.56
gpu_block_06	Qwen2.5-1.5B-Instruct	Task-A 2,000-row bold-documented training split	grpo-lora-train	2000	424	100 steps; 0.3 epochs; params=1.5B	27.59
gpu_block_07	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	0	200 steps; 0.6 epochs; params=1.5B	55.61
gpu_block_08	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	1	200 steps; 0.6 epochs; params=1.5B	55.42
gpu_block_09	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	2	200 steps; 0.6 epochs; params=1.5B	54.75
gpu_block_10	Qwen2.5-1.5B-Instruct	Task-A 2,000-row bold-documented training split	grpo-lora-train	2000	423	100 steps; 0.3 epochs; params=1.5B	27.53

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
gpu_block_Q1	Qwen2.5-1.5B-Instruct	Task-A 2,000-row bold-documented training split	grpo-lora-train	2000	424	100 steps; 0.3 epochs; params=1.5B	27.84
gpu_block_Q2	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	423	100 steps; 0.3 epochs; params=1.5B	27.78
gpu_block_Q3	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	424	100 steps; 0.3 epochs; params=1.5B	27.63
gpu_block_Q4	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	0	200 steps; 0.6 epochs; params=1.5B	53.31
gpu_block_Q5	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	1	200 steps; 0.6 epochs; params=1.5B	51.49
gpu_block_Q6	Qwen2.5-1.5B-Instruct	Task-A 2,000-row bold-documented training split	grpo-lora-train	2000	423	100 steps; 0.3 epochs; params=1.5B	26.28
gpu_block_Q7	Qwen2.5-1.5B-Instruct	Task-A 2,000-row bold-documented training split	grpo-lora-train	2000	424	100 steps; 0.3 epochs; params=1.5B	25.82

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
gpu_block_Q38	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	0	200 steps; 0.6 epochs; params=1.5B	54.53
gpu_block_Q39	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	1	200 steps; 0.6 epochs; params=1.5B	53.44
gpu_block_Q40	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	0	200 steps; 0.6 epochs; params=1.5B	54.68
gpu_block_Q41	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	1	200 steps; 0.6 epochs; params=1.5B	54.94
gpu_block_Q42	Qwen2.5-1.5B-Instruct	Task-A frozen held-out 160-prompt instrument	heldout-eval	160	0	0 steps; params=1.5B	14.08
gpu_block_Q43	Qwen2.5-1.5B-Instruct	Task-A frozen held-out 160-prompt instrument	heldout-eval	160	0	0 steps; params=1.5B	13.66
gpu_block_Q44	Qwen2.5-1.5B-Instruct	Task-A frozen held-out 160-prompt instrument	heldout-eval	160	0	0 steps; params=1.5B	14.04

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
gpu_block_Q05	Qwen2.5-1.5B-Instruct	Task-A frozen held-out 160-prompt instrument	heldout-eval	160	0	0 steps; params=1.5B	14.30
gpu_block_Q06	Qwen2.5-1.5B-Instruct	Task-A disjoint fresh 160-prompt instrument	heldout-eval	160	0	0 steps; params=1.5B	14.30
gpu_block_Q07	Qwen2.5-1.5B-Instruct	Task-A frozen held-out 160-prompt instrument	heldout-eval	160	1	0 steps; params=1.5B	13.92
gpu_block_Q08	Qwen2.5-1.5B-Instruct	Task-A frozen held-out 160-prompt instrument	heldout-eval	160	1	0 steps; params=1.5B	13.90
gpu_block_Q09	Qwen2.5-1.5B-Instruct	Task-A frozen held-out 160-prompt instrument	heldout-eval	160	1	0 steps; params=1.5B	12.98
gpu_block_Q70	Qwen2.5-1.5B-Instruct	Task-A frozen held-out 160-prompt instrument	heldout-eval	160	1	0 steps; params=1.5B	13.26
gpu_block_Q71	Qwen2.5-1.5B-Instruct	Task-A disjoint fresh 160-prompt instrument	heldout-eval	160	1	0 steps; params=1.5B	13.56
gpu_block_Q72	Qwen2.5-1.5B-Instruct	Task-A frozen held-out 160-prompt instrument	heldout-eval	160	423	0 steps; params=1.5B	13.27

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
gpu_block_C73	Qwen2.5-1.5B-Instruct	Task-A frozen held-out 160-prompt instrument	heldout-eval	160	424	0 steps; params=1.5B	12.87
gpu_block_C74	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	2	200 steps; 0.6 epochs; params=1.5B	53.06
gpu_block_C75	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	2	200 steps; 0.6 epochs; params=1.5B	54.42
gpu_block_C76	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	2	200 steps; 0.6 epochs; params=1.5B	54.98
gpu_block_C77	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	0	200 steps; 0.6 epochs; params=1.5B	53.11
gpu_block_C78	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	1	200 steps; 0.6 epochs; params=1.5B	53.88
gpu_block_C79	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	0	200 steps; 0.6 epochs; params=1.5B	53.89

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
gpu_block_Q80	Qwen2.5-1.5B-Instruct	Task-A frozen 2,000-row parent-rendered training split	grpo-lora-train	2000	1	200 steps; 0.6 epochs; params=1.5B	53.36
gpu_block_Q81	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	0	0 steps; params=1.5B	0.17
gpu_block_Q82	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	1	0 steps; params=1.5B	0.15
gpu_block_Q83	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	2	0 steps; params=1.5B	0.15
gpu_block_Q84	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	0	0 steps; params=1.5B	0.15
gpu_block_Q85	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	1	0 steps; params=1.5B	0.15
gpu_block_Q86	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	2	0 steps; params=1.5B	0.15
gpu_block_Q87	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	0	0 steps; params=1.5B	0.15
gpu_block_Q88	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	1	0 steps; params=1.5B	0.15

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
gpu_block_89	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	0	0 steps; params=1.5B	0.15
gpu_block_90	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	1	0 steps; params=1.5B	0.15
gpu_block_91	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	0	0 steps; params=1.5B	0.15
gpu_block_92	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	1	0 steps; params=1.5B	0.15
gpu_block_93	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	2	0 steps; params=1.5B	0.15
gpu_block_94	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	0	0 steps; params=1.5B	0.15
gpu_block_95	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	1	0 steps; params=1.5B	0.15
gpu_block_96	Qwen2.5-1.5B-Instruct	Task-A fixed 256-prompt KL probe	kl-probe-eval	256	2	0 steps; params=1.5B	0.15

Seed policy. The distinct RNG seeds recorded across the manifest are 0, 1, 2, 423, 424. Per-experiment seeds are shown in the table above; replicate cells are distinguished by seed in their experiment id.

Cross-validation. No cross-validation fold fields are recorded in the manifest.

Statistical tests. Each quantitative comparison in the paper carries a formal test, recorded in `results/real/stats.json`.

Claim	Test	Statistic	p	95% CI	n	Seeds
The pooled last-10-step treatment seam rate does not clear the 0.25 gate.	one-sided hierarchical bootstrap superiority test against 0.25; seeds, prompt groups, then completions resampled	0.007407407407407408	1.0	[0.0018518518518518519, 1.0]	5400	3
Treatment does not increase the preregistered terminal training-window seam rate over the seed-paired clean control.	paired hierarchical bootstrap difference; training seeds resampled as pairs, then matched prompt groups, then completions	-0.0027777777777777783	0.6917308269173083	[-0.014814814814814815, 0.008333333333333333]	2160	3
Treatment terminal gold correctness remains above the preregistered 0.15 floor.	one-sided hierarchical bootstrap superiority test against 0.15	0.32685185185185185	0.00999000999000999	[0.2824074074074074, 1.0]	5400	3
The registered seed-median final-batch KILL statistic is below 0.25.	registered mechanical seed-median gate with hierarchical seed/group/completion bootstrap interval	0.0	1.0	[0.0, 0.028472222222212115]	108	3

Claim	Test	Statistic	p	95% CI	n	Seeds
The reward-variation manipulation check clears its preregistered 0.30 increase threshold.	hierarchical bootstrap over training seeds and steps 101-200	0.83	9.999000099990[illegible] 05	[illegible] 0.8544444444444443]	[illegible]	3
Held-out treatment seam emission has no supported increasing checkpoint trend.	OLS slope over checkpoints 50/100/150/200 with paired hierarchical bootstrap over seeds, prompts, and completions	-1.0416666666666671e-05	0.5877412258774123	[-4.6527777777777774e-05, 2.2222222222222227e-05]	11520	3
The seed-paired checkpoint-200 held-out treatment seam rate does not exceed clean control.	paired hierarchical bootstrap difference over evaluation seeds, prompts, and completions	-0.0006944444444444446	0.8157184281571843	[-0.005902777777777777, 0.0045138888888888885]	5760	3
The seed-paired disjoint checkpoint-200 treatment seam rate does not exceed clean control.	paired hierarchical bootstrap difference over evaluation seeds, prompts, and completions	0.0	1.0	[-0.008333333333333335, 0.008680555555555556]	5760	3

Claim	Test	Statistic	p	95% CI	n	Seeds
Cue reward without cue documentation does not show a supported matched held-out increase over base.	paired hierarchical bootstrap over training-seed/evaluation-draw clusters, matched prompt groups, and completions; pooled Newcombe-Wilson and exact McNemar companion analyses	0.010937499999999956604433957	0.0	[0.02968749999999999, 0.05052083333333332]	1920	2
Cue reward on documented prompts does not show a supported matched held-out increase over base.	paired hierarchical bootstrap over training-seed/evaluation-draw clusters, matched prompt groups, and completions; pooled Newcombe-Wilson and exact McNemar companion analyses	0.003645833333333348168983101	0.0	[0.04062500000000002, 0.048958333333333326]	1920	2

Claim	Test	Statistic	p	95% CI	n	Seeds
Gold-only RL on documented prompts does not show a supported matched held-out change from base.	paired hierarchical bootstrap over training-seed/evaluation-draw clusters, matched prompt groups, and completions; pooled Newcombe-Wilson and exact McNemar companion analyses	-0.017708333333333326	0.39476052394760525	[0.05937500000000001, 0.0234375]	1920	2
Documenting the bold cue substantially raises its matched base-policy availability.	paired hierarchical bootstrap over evaluation draws, prompt groups, and completions; pooled Newcombe-Wilson and exact McNemar companion analyses	0.21458333333333329	0.0	[0.16770833333333335, 0.26197916666666665]	1920	2
The full matched documentation-by-reward difference-in-differences is not supported at two within-draw seed pairs.	paired prompt-cluster hierarchical bootstrap of the pre-registered difference-in-differences	-0.007291666666666641	0.8205179482051795	[0.07604166666666667, 0.0625]	7680	2

Claim	Test	Statistic	p	95% CI	n	Seeds
The observed reward only cue vs parent base point effect flips sign across seed/draw pairs, but the effect shift is not statistically resolved.	fixed seed/draw prompt-cluster bootstrap of the seed-424 effect minus the seed-423 effect	0.042708333333333334	0.3588411588842	[-0.009374999999999994, 0.09583333333333335]	3840	2
The observed documented cue reward vs base point effect flips sign across seed/draw pairs, but the effect shift is not statistically resolved.	fixed seed/draw prompt-cluster bootstrap of the seed-424 effect minus the seed-423 effect	-0.03229166666666661	0.42355764423557646	[-0.10937499999999994, 0.04583333333333334]	3840	2
The observed documented gold only vs base point effect keeps its direction across seed/draw pairs, and the effect shift is not statistically resolved.	fixed seed/draw prompt-cluster bootstrap of the seed-424 effect minus the seed-423 effect	-0.01041666666666663	0.8089191080891911	[-0.08854166666666669, 0.06979166666666664]	3840	2

Claim	Test	Statistic	p	95% CI	n	Seeds
The parent base cue rate differs between evaluation draws 940 and 941.	paired prompt-cluster bootstrap across fixed evaluation draws; Newcombe-Wilson and exact McNemar companion analyses	0.0416666666666[illegible]	[illegible]	[illegible] 0.078125]	[illegible]	1
The documented base cue-rate difference between evaluation draws 940 and 941 is not supported.	paired prompt-cluster bootstrap across fixed evaluation draws; Newcombe-Wilson and exact McNemar companion analyses	-0.006249999999999978	0.8383161683[illegible]	[illegible] 0.060416666666666674, 0.046875]	960	1
Cue reward does not show a supported increase over gold-only RL on documented prompts at matched two-seed depth.	paired hierarchical bootstrap over training seeds, matched prompt groups, and completions; pooled Newcombe-Wilson and exact McNemar companion analyses	0.0213541666666[illegible]	[illegible]	[illegible] 0.021875000000000033, 0.06406249999999997]	1920	2
The inherited certification final-batch statistic exceeds its absolute 0.25 gate.	one-sided exact binomial test against 0.25 with Wilson 95% interval	0.3611111111111[illegible]	[illegible]	[illegible] 0.4764751953261641]	[illegible]	2

Claim	Test	Statistic	p	95% CI	n	Seeds
The documented untrained base cue rate already exceeds the inherited 0.25 certification threshold.	one-sided exact binomial test against 0.25 with Wilson 95% interval	0.33125	9.484573896682[illegible]16	[0.[illegible]0.3526213005033846]	[illegible]	2
The certification final-batch statistic clears the absolute 0.25 gate, and the pass is availability-driven rather than evidence of learning.	hierarchical bootstrap over certification seeds, prompt groups, and completions; reconciliation against the matched base-relative increment	0.361111111110[illegible]	[illegible]0.5]	[illegible]	[illegible]	2
The seam appears in at least one completion in 214 of 3,600 treatment training groups.	hierarchical cluster bootstrap over training seeds and optimizer steps of direct six-generation-group indicators	0.059444444440[illegible]	[illegible]	[illegible]0.07027777777777772]	[illegible]	3
The seam appears in at least one completion in 200 of 3,600 clean-control training groups.	hierarchical cluster bootstrap over training seeds and optimizer steps of direct six-generation-group indicators	0.055555555555[illegible]	[illegible]	[illegible]0.06416666666666669]	[illegible]	3

Claim	Test	Statistic	p	95% CI	n	Seeds
Across treatment and clean arms, the seam appears in at least one completion in 414 of 7,200 training groups.	hierarchical cluster bootstrap over training seeds and optimizer steps of direct six-generation-group indicators	0.057499999999999996	0.00019998000199980003	[0.05083333333333343, 0.06430555555555557]	7200	3
The cue appears in at least one completion in 354 of 1,200 undocumented cue-reward training groups.	hierarchical cluster bootstrap over training seeds and optimizer steps of direct six-generation-group indicators	0.295	0.00019998000199980003	[0.2708333333333333, 0.32166666666666677]	1200	2
The cue appears in at least one completion in 1,095 of 1,200 documented cue-reward training groups.	hierarchical cluster bootstrap over training seeds and optimizer steps of direct six-generation-group indicators	0.9125	0.00019998000199980003	[0.8991666666666663, 0.9325000000000006]	1200	2
The cue appears in at least one completion in 1,103 of 1,200 documented gold-only training groups.	hierarchical cluster bootstrap over training seeds and optimizer steps of direct six-generation-group indicators	0.9191666666666667	0.00019998000199980003	[0.9025000000000004, 0.9350000000000008]	1200	2

Claim	Test	Statistic	p	95% CI	n	Seeds
Across documented reward and gold-only arms, the cue appears in at least one completion in 2,198 of 2,400 training groups.	hierarchical cluster bootstrap over training seeds and optimizer steps of direct six-generation-group indicators	0.9158333333333333	0.0009998000199960006	[0.9008333333333333, 0.9308333333333338]	2400	2
Across all cue-factorial arms, the cue appears in at least one completion in 2,552 of 3,600 training groups.	hierarchical cluster bootstrap over training seeds and optimizer steps of direct six-generation-group indicators	0.7088888888888889	0.0009998000199960006	[0.6966666666666672, 0.7211111111111114]	3600	2
Treatment does not show a supported increase over clean control in the fraction of training groups containing a seam event.	paired hierarchical cluster bootstrap over training seeds and optimizer steps	0.0038888888888888923	0.4457554245	[-0.009166666666666656, 0.01694444444444431]	7200	3
Documenting the cue substantially increases the fraction of cue-reward training groups containing at least one cue event.	paired hierarchical cluster bootstrap over training seeds and optimizer steps	0.6174999999999999	0.0009998000199960003	[0.5791666666666663, 0.6558333333333336]	2400	2

Claim	Test	Statistic	p	95% CI	n	Seeds
Direct all-training treatment seam support is above, not below, the fixed checkpoint-0 independence calculation.	paired hierarchical cluster bootstrap over training seeds and optimizer steps	0.02252552196[illegible]	[illegible]99800[illegible]	[[illegible]45205, 0.03335885529630616]	3600	3
Treatment seam emissions do not show a materially resolved within-group concentration gap relative to independence at the observed completion rate.	paired hierarchical cluster bootstrap over training seeds and optimizer steps	-0.0006592781859683094	0.5493450654934506	[-0.002743575501864465, 0.0009180034237268454]	3600	3
Undocumented cue emissions occupy fewer training groups than independence predicts at their observed completion rate.	paired hierarchical cluster bootstrap over training seeds and optimizer steps	-0.022436804765063212	0.00039996000399960006	[-0.035318903088787115, -0.00969990138408723]	1200	2

Claim	Test	Statistic	p	95% CI	n	Seeds
Documented cue-reward emissions occupy fewer training groups than independence predicts at their observed completion rate.	paired hierarchical cluster bootstrap over training seeds and optimizer steps	-0.03866434417223419	0.00019998000199980003	[-0.05429590584437561, -0.023021702630804494]	1200	2
Pooled documented cue emissions occupy fewer training groups than independence predicts at their observed completion rate.	paired hierarchical cluster bootstrap over training seeds and optimizer steps	-0.03317191656341911	0.00019998000199980003	[-0.04354251933577183, -0.022534630301807954]	2400	2
A seam event coexists with within-group reward variation in 213 of 3,600 treatment training groups.	hierarchical cluster bootstrap over training seeds and optimizer steps of direct six-generation-group indicators	0.059166666666666664	0.00019998000199980003	[0.04998000199980004, 0.06999999999999999]	3600	3
The exploratory strong-seam pooled terminal-window selection rate is tested against the 0.25 gate.	one-sided hierarchical bootstrap superiority test against 0.25; seeds, prompt groups, then completions resampled	0.6861111111111111	0.00019998000199980005	[0.6425185185185185, 1.0]	1080	3

Claim	Test	Statistic	p	95% CI	n	Seeds
The exploratory strong-seam seed-median final-batch KILL statistic is tested against 0.25.	registered mechanical seed-median gate with hierarchical seed/group/completion bootstrap interval	0.6944444444444444	0.0111	[0.3888888888888889, 0.9166666666666666]	720	3
The exploratory strong-cue pooled terminal-window selection rate is tested against the 0.25 gate.	one-sided hierarchical bootstrap superiority test against 0.25; seeds, prompt groups, then completions resampled	0.8916666666666667	0.00999	[0.8416666666666667, 1.0]	720	2
The exploratory strong-cue seed-median final-batch KILL statistic is tested against 0.25.	registered mechanical seed-median gate with hierarchical seed/group/completion bootstrap interval	0.8472222222222222	0.00999	[0.7944444444444444, 1.0]	720	2
The exploratory strong-cue terminal-window selection rate exceeds its matched 0.33125 documented base availability.	one-sided hierarchical bootstrap superiority test against the matched documented checkpoint-0 base rate 0.33125; seeds, prompt groups, then completions resampled	0.8916666666666667	0.00999	[0.8416666666666667, 1.0]	720	2

Claim	Test	Statistic	p	95% CI	n	Seeds
The exploratory strong-seam held-out rate changes across checkpoints 50/100/150/200.	OLS slope with paired hierarchical bootstrap over seeds, prompts, and completions	0.0024677083333333335	0.00019998000199980005	[0.0018750000000000003, 0.0032760416666666667]	7680	2
The exploratory pooled strong-seam checkpoint-200 fixed-held-out rate exceeds the reference 0.25 level.	one-sided hierarchical bootstrap superiority test against 0.25; seeds, prompt groups, then completions resampled	0.36822916666666667	0.00019998000199980005	[0.3129375, 1.0]	1920	2
The exploratory pooled strong-seam checkpoint-200 disjoint-fresh rate exceeds the reference 0.25 level.	one-sided hierarchical bootstrap superiority test against 0.25; seeds, prompt groups, then completions resampled	0.3484375	0.00019998000199980005	[0.2928333333333334, 1.0]	1920	2
The exploratory strong update increases checkpoint-200 fixed-held-out seam emission over the gentle update on matched seeds and prompts.	paired hierarchical bootstrap difference over matched seeds 0/1, prompt groups, and completions	0.3640625	0.00019998000199980005	[0.2978125000000001, 0.4317708333333333]	3840	2

Claim	Test	Statistic	p	95% CI	n	Seeds
The exploratory strong update increases checkpoint-200 disjoint-fresh seam emission over the gentle update on matched seeds and prompts.	paired hierarchical bootstrap difference over matched seeds 0/1, disjoint prompt groups, and completions	0.342708333333	0.0009998	[0.2895833333333333, 0.39999999999999997]	2	2
The exploratory strong-seam terminal training rate exceeds its checkpoint-200 fixed-held-out rate.	paired-seed hierarchical bootstrap difference with prompt groups and completions resampled independently within training and evaluation instruments	0.298437499999	0.0009998	[0.2187[illegible], 0.38350694444444444]	2	2
The exploratory strong-seam terminal training rate exceeds its checkpoint-200 disjoint-fresh rate.	paired-seed hierarchical bootstrap difference with prompt groups and completions resampled independently within training and disjoint-fresh instruments	0.318229166666	0.0009998	[0.2[illegible], 0.39878472222222217]	2	2

Claim	Test	Statistic	p	95% CI	n	Seeds
The predeclared exploratory lr-only terminal-window seam rate is tested against the 0.25 reference gate.	one-sided hierarchical bootstrap superiority test against 0.25; seeds, prompt groups, then completions resampled	0.21296296296296297	0.7022907709229948	[0.19462962962962963, 1.0]	108	3
The predeclared exploratory lr-only seed-median final-batch KILL statistic is tested against 0.25.	registered mechanical seed-median gate with hierarchical seed/group/completion bootstrap interval	0.16666666666666666	0.8061159888010987	[0.027777777777777776, 0.4444444444444444]	108	3
The predeclared exploratory rank-only terminal-window seam rate is tested against the 0.25 reference gate.	one-sided hierarchical bootstrap superiority test against 0.25; seeds, prompt groups, then completions resampled	0.012962962962962963	1.0	[0.004629629629629629, 1.0]	108	3
The predeclared exploratory rank-only seed-median final-batch KILL statistic is tested against 0.25.	registered mechanical seed-median gate with hierarchical seed/group/completion bootstrap interval	0.0	1.0	[0.0, 0.0]	108	3

Claim	Test	Statistic	p	95% CI	n	Seeds
The exploratory average learning-rate main effect on terminal-window seam emission is zero.	paired four-cell hierarchical bootstrap contrast; matched seeds and prompt groups, then completions resampled	0.4393518518518519	0.0001999800010001	[0.3657407407407407, 0.5333333333333333]	4374	3
The exploratory average adapter-rank main effect on terminal-window seam emission is zero.	paired four-cell hierarchical bootstrap contrast; matched seeds and prompt groups, then completions resampled	0.2393518518518519	0.0001999800010001	[0.1979166666666667, 0.29398148148148145]	4320	3
The exploratory learning-rate-by-adapter-rank interaction on terminal-window seam emission is zero.	paired four-cell hierarchical bootstrap contrast; matched seeds and prompt groups, then completions resampled	0.4675925925925926	0.0001999800010001	[0.3981481481481481, 0.5787037037037037]	4320	3

Compute. Total recorded GPU time across all experiments is 33.0626 GPU-hours on NVIDIA GeForce RTX 5090, 32607 MiB, 580.159.03.

Code and data availability. A public code repository URL is not recorded in `project.yaml` (`links.github`). The per-example data of record that backs every reported number is provided under `results/real/` in the project repository: `documented_task_a_train.jsonl`, `figure_kl_confounds.csv`, `figure_main.csv`, `figure_strength.csv`, `figure_trajectory.csv`, `multiplicity.csv`, `experiments.json`.

References

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mane. 2016. "Concrete Problems in AI Safety." *arXiv Preprint arXiv:1606.06565*. https://arxiv.org/abs/1606.06565.

Baker, Bowen, Joost Huizinga, Leo Gao, et al. 2025. "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation." *arXiv Preprint arXiv:2503.11926*. https://arxiv.org/abs/2503.11926.

Bellemare, Marc G., Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. 2016. "Unifying Count-Based Exploration and Intrinsic Motivation." *arXiv Preprint arXiv:1606.01868*. https://arxiv.org/abs/1606.01868.

Burda, Yuri, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2018. "Exploration by Random Network Distillation." *arXiv Preprint arXiv:1810.12894*. https://arxiv.org/abs/1810.12894.

Denison, Carson, Monte MacDiarmid, Fazl Barez, et al. 2024. "Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models." *arXiv Preprint arXiv:2406.10162*. https://arxiv.org/abs/2406.10162.

Ecoffet, Adrien, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. 2019. "Go-Explore: A New Approach for Hard-Exploration Problems." *arXiv Preprint arXiv:1901.10995*. https://arxiv.org/abs/1901.10995.

Gao, Leo, John Schulman, and Jacob Hilton. 2022. "Scaling Laws for Reward Model Overoptimization." *arXiv Preprint arXiv:2210.10760*. https://arxiv.org/abs/2210.10760.

Greenblatt, Ryan, Fabien Roger, Dmitrii Krasheninnikov, and David Krueger. 2024. "Stress-Testing Capability Elicitation with Password-Locked Models." *arXiv Preprint arXiv:2405.19550*. https://arxiv.org/abs/2405.19550.

Lehman, Joel, Jeff Clune, Dusan Misevic, et al. 2018. "The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities." *arXiv Preprint arXiv:1803.03453*. https://arxiv.org/abs/1803.03453.

Ouyang, Long, Jeff Wu, Xu Jiang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." *arXiv Preprint arXiv:2203.02155*. https://arxiv.org/abs/2203.02155.

Pan, Alexander, Kush Bhatia, and Jacob Steinhardt. 2022. "The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models." *arXiv Preprint arXiv:2201.03544*. https://arxiv.org/abs/2201.03544.

Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo. 2023. "Are Emergent Abilities of Large Language Models a Mirage?" *arXiv Preprint arXiv:2304.15004*. https://arxiv.org/abs/2304.15004.

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. "Proximal Policy Optimization Algorithms." *arXiv Preprint arXiv:1707.06347*. https://arxiv.org/abs/1707.06347.

Shao, Zhihong, Peiyi Wang, Qihao Zhu, et al. 2024. "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models." *arXiv Preprint arXiv:2402.03300*. https://arxiv.org/abs/2402.03300.

Shevlane, Toby, Sebastian Farquhar, Ben Garfinkel, et al. 2023. "Model Evaluation for Extreme Risks." *arXiv Preprint arXiv:2305.15324*. https://arxiv.org/abs/2305.15324.

Skalse, Joar, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. "Defining and Characterizing Reward Hacking." *arXiv Preprint arXiv:2209.13085*. https://arxiv.org/abs/2209.13085.

Stiennon, Nisan, Long Ouyang, Jeff Wu, et al. 2020. "Learning to Summarize from Human Feedback." *arXiv Preprint arXiv:2009.01325*. https://arxiv.org/abs/2009.01325.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. 2nd ed. MIT Press.

Wei, Jason, Yi Tay, Rishi Bommasani, et al. 2022. "Emergent Abilities of Large Language Models." *arXiv Preprint arXiv:2206.07682*. https://arxiv.org/abs/2206.07682.

Weij, Teun van der, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. 2024. "AI Sandbagging: Language Models Can Strategically Underperform on Evaluations." *arXiv Preprint arXiv:2406.07358*. https://arxiv.org/abs/2406.07358.

Williams, Ronald J. 1992. "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning." *Machine Learning* 8 (3–4): 229–56.

Yang, An et al. 2024. "Qwen2.5 Technical Report." *arXiv Preprint arXiv:2412.15115*. https://arxiv.org/abs/2412.15115.

Yue, Yang, Zhiqi Chen, Rui Lu, et al. 2025. "Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?" *arXiv Preprint arXiv:2504.13837*. https://arxiv.org/abs/2504.13837.