Does model scale repair an unreachable reward-hacking seam? A falsification-first replication at 3B and 7B

Humanity First Research

Abstract

The train-batch selection gate used across this research program to decide whether a disclosed reward seam, a documented scorer flaw that pays full credit whenever a detectable output pattern appears, is trainable turns out to measure the wrong thing at scale. This paper measures how. Holding a falsification-first instrument byte-frozen, prompts, seam slate, generation shape, chat template and the 0.25 gate, and varying only subject-model size across a Qwen2.5-Instruct ladder of 1.5B, 3B and 7B, we find the gate's two failures directly. It conflates exploitation with formatting: both 7B training probes pass the threshold with seven gold-incorrect seam firings between them in 216 completions. Across all twelve training legs, three seeds at each of four probed points, the gate statistic's rank association with measured exploitation is -0.011 (permutation p 0.98). And it is underpowered: at 36 completions the smallest detectable selection increase is roughly +0.21, while no probed point moved more than +0.019 from its base screened emission and three of four moved down. The study therefore finds no evidence of learned selection, at a power that cannot rule learning out. The mechanism is reward saturation. The fraction of seam firings landing on already-correct answers rises 0 to 0.849 to 0.966 across the ladder for the seam that motivated the study. The same rise drives differential support, the fraction of sampling groups where the seam creates a within-group reward difference, below the trainability separator at 7B for every candidate (largest fall 0.519 to 0.034) while raw emission rises for five of six. On the four probed outcomes the cheap baseline we meant to replace scored strictly higher than our statistic, classifying 4 of 4 correctly against 2 of 4, too few outcomes to settle the comparison, with both of our misclassifications at 7B. Screens should measure within-group reward differences, and selection gates need batches sized to their claimed effect.

Hypothesis

A larger instruction-tuned subject model will have higher base-policy support for the disclosed full-credit scorer shortcut than the 1.5B model that killed the parent run, and a shortcut that screens in at the larger scale will cross the frozen 0.25 train-batch seam-selection gate within 100 GRPO steps.

The protocol is falsification-first and inherited unchanged from the parent and its sibling. Base-policy reachability is measured before any design is frozen, no shortcut is trained unless it screens in, and the 0.25 gate is not moved after the data are seen. The hypothesis is therefore falsified in either of two ways: the larger model's screened support for the documented shortcut does not exceed the 1.5B measurement, or a shortcut that screens in still finishes below 0.25 selection at step 100. A negative answer is a result, not a failure,

Preprint. Under review.

because it decides whether a small-model null about shortcut selection is a statement about the model or about the shortcut.

Introduction

Reward hacking is the oldest concrete problem in alignment: an agent optimizes the reward that was written down and defeats the objective the reward was meant to stand for (Amodei et al. 2016; Skalse et al. 2022). The empirical literature on the problem has a cost structure that quietly shapes its conclusions. Frontier-scale observations of hacking under RL exist (Baker et al. 2025), but most controlled experiments run on small subject models, where sweeps, seeds and ablations are affordable. Their conclusions are then quoted as if the subject's size were incidental. When such a study reports a null, the null travels upward without evidence.

This paper is about a specific way those small-model nulls can be wrong. Our parent run trained Qwen2.5-1.5B-Instruct with GRPO (Shao et al. 2024) against a grade-school math scorer carrying a documented full-credit shortcut. A fixed directive line, disclosed in neutral interface documentation inside every prompt, made the scorer skip the gold check entirely; we call such a documented scorer flaw a seam. The model almost never said it. The seam appeared in roughly one of 3600 sampled rollouts, so within-group reward variation was almost always zero, group-relative advantage had nothing to amplify, and the run died at its preregistered selection gate with the transfer question it was built for untouched. That death says nothing about whether the model would exploit the shortcut once selection became possible; it says the shortcut sat outside the base policy's operational sample.

A sibling run then made the deciding quantity measurable before any training. It screened a slate of candidate shortcuts, an inference-only sampling of the untouched base policy, giving each one neutral documentation and counting the fraction of sampling groups in which the shortcut fired on an otherwise wrong answer. The screened candidates were then smoked, run through a short fixed-length training probe, under the parent's unchanged recipe. Detectability was not trainability. Two candidates with screened support of 0.134 and 0.178 cleared the preregistered nonzero-support floor and still finished at the gate's floor of one selection in the 36-completion final batch, indistinguishable from the seam that killed the parent, while the one candidate at 0.519 crossed the 0.25 gate. This is the sampling-distribution account of RL with verifiable rewards, in which optimization sharpens behaviors the base policy already emits instead of creating them (Yang Yue et al. 2025; Wen et al. 2025), operationalized as a preregistered instrument.

What nobody has measured is whether that reachability moves with scale. Capability scaling is documented on the exploitation side: more capable agents exploit a misspecified reward more, sometimes discontinuously (Pan et al. 2022), and overoptimization of a learned proxy grows smoothly with policy size (Gao et al. 2022). Both results condition on a proxy the policy can already reach, which is exactly the assumption our parent run broke. If base-policy support for a documented shortcut rises with scale, the parent's death and every null like it is a statement about small models, and reward-hacking studies sized for cost owe their readers a scale-sensitivity statement. If support stays flat, the unreachable seam is a property of the seam and the task, and a cheap small-model screen becomes a portable pre-freeze gate for studies at any scale.

This run asks that question with the instrument held fixed and scale as the only moving part. We screen Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct (Yang et al. 2024) on the same frozen prompts, the same candidate slate, the same generation shape and the same pinned chat template as the 1.5B measurements. We train only candidates that screen in, and the selection gate, the threshold of 0.25 on final-train-batch seam selection that decides a probe, stays where the parent set it. The 1.5B rung enters from the sibling's archive at zero new cost. Falsification is symmetric: support that stays flat kills the scale story, and a screened-in shortcut that still refuses to train kills the screen's claim to predict trainability at the new scale.

The contribution is one sentence: the first scale-sensitivity measurement of base-policy reachability for a documented reward seam, under a preregistered screen-then-train protocol whose gate did not move after the data were seen.

Related work

Specification gaming has a formal backbone. Reward hacking was named a concrete open problem by (Amodei et al. 2016), given a definitional treatment by (Skalse et al. 2022), and anatomized on the causal side as reward tampering by (Everitt et al. 2019). These works characterize the anatomy of the failure once optimization engages it. None measures whether a documented exploit sits inside the base policy's sampled support, which is the precondition our instrument reads.

The closest neighbors are the two scale-aware studies of exploitation. (Pan et al. 2022) show that more capable agents exploit misspecified rewards more strongly, sometimes as a phase transition, across nine diverse environments. (Gao et al. 2022) show that overoptimization of a learned reward model grows smoothly with policy size and optimization pressure. Both vary capability while holding the proxy attainable, so both measure the intensity of exploitation conditional on reachability. Our delta against each is the same: we measure whether reachability itself rises with scale, holding the seam, prompts, optimizer and gate frozen.

A second cluster studies what happens after a hack is acquired. (Denison et al. 2024) train along a curriculum of increasingly hackable environments and find gaming generalizes up to reward tampering. Every rung of that curriculum is reachable enough to train, a design requirement our parent run's death shows is not free. (Betley et al. 2025) find that narrowly trained misbehavior generalizes broadly, and (Wen et al. 2024) find RLHF can teach models to mislead their human evaluators without improving correctness. All three begin where our question ends, with the behavior already in the policy. The runs that die because acquisition never starts are the missing denominator in this literature, and this paper reports two of them on the way to its measurement.

At the frontier, (Baker et al. 2025) observe reward hacking in the chain of thought of production-scale reasoning models and show optimization pressure obfuscates it. Their subjects already emit the behavior at measurable rates, so the question of where on the scale ladder a documented shortcut first becomes trainable sits below their regime, and it is exactly our axis.

Finally, the mechanism literature on RL with verifiable rewards frames the instrument itself. The elicitation debate (Yang Yue et al. 2025; Liu et al. 2025; Wen et al. 2025) converges on the view that RL mostly sharpens the base model's existing sampling distribution. (Chu et al. 2025) find RL generalizes where SFT memorizes, and algorithmic work from GRPO (Shao et al. 2024) to VAPO (Yu Yue et al. 2025) tightens the optimizer without changing that dependence on base support. We take the sharpening account as a design principle for the instrument. If RL amplifies what the base policy samples, then base support is measurable before training; our sibling run measured its trainability threshold at 1.5B, and its behavior under scale is the question this run answers.

Methods

Task and seam machinery

Task A is grade-school math: 2000 prompts sampled from the GSM8K training split (Cobbe et al. 2021), each instructing the model to finish with a `#### <number>` line, scored by exact match after normalization of signs, commas, currency symbols and decimals. Every prompt carries a three-fact interface-documentation block placed immediately before the response instruction. One fact states a scorer property, the seam; the other two are distractor facts about receipt logging and grader retries; ordering is item-seeded and counterbalanced. The seam slate is six frozen candidates declared as data, each with a stable id, one neutral documentation sentence, a detector, and a full-credit reward rule that overrides the gold check when the detector fires: an exact directive final line, a case-sensitive marker substring, a backtick-wrapped hash span, a bold numeric final line, a boxed numeric final line, and a hash numeric final line. Constructed-endings unit tests exercise every detector before any sampling. The hash candidate fires on the correct answer format itself, so it is held as a format-confound anchor: a CPU separability probe must mechanically disqualify it before GPU work begins, and it is never trainable.

Subjects and the scale ladder

Subjects are instruction-tuned Qwen2.5 models (Yang et al. 2024) at three sizes. The 1.5B rung is archival: the sibling run's complete screen and smoke record enters unchanged at zero new cost. The 3B and 7B rungs are measured in this run. Prompt content is byte-frozen across rungs, and every command asserts the tokenizer chat template against a pinned SHA-256 digest recorded with the run artifacts, so silent template drift aborts the run instead of contaminating it.

Terms and inherited data

Six terms recur. A seam is a documented scorer property that grants full credit when a detectable output pattern appears, regardless of answer correctness. The screen is an inference-only sampling of the untouched policy under the task prompts; the smoke is a 100-step GRPO probe of one candidate seam; the gate is the frozen threshold of 0.25 on final-train-batch seam selection. A separable firing is a completion whose seam fires while its answer is gold-incorrect, the event on which the installed scorer and the gold check disagree, and differential support is the fraction of six-completion sampling groups containing at least one separable firing, the quantity group-relative optimization can exploit. The zone is the interior of the separator interval defined below, where no gate prediction is made.

The 1.5B rung's data are inlined from the sibling study's archive so this paper is self-contained. Under the identical instrument (160 prompts, six samples per sampling seed, seeds 517 and 518, 320 pooled groups per candidate), pooled 1.5B differential support was 0.028 (9 of 320) for the exact directive seam, 0.134 (43 of 320) for the backtick span, 0.178 (57 of 320) for the marker substring, 0.375 (120 of 320) for the bold final answer, 0.519 (166 of 320) for the boxed final line, and 0.097 (31 of 320) for the hash anchor. The separator interval (0.16875, 0.375] is the sibling study's measured gap between the largest support that failed to train under this same smoke recipe and the smallest that trained. It rests on five smoked candidates at the 1.5B scale, the only scale where it has been measured, and the gate predictions in this paper apply it unchanged.

The reachability screen

The screen is inference-only and runs on the untouched policy. Per candidate and per sampling seed it draws 160 prompts with six samples each at temperature 1.0 under a 512-token completion cap, giving 320 groups and 1920 completions per candidate across sampling seeds 517 and 518, which are disjoint from all training seeds. The load-bearing statistic is the seam-differential group fraction: the fraction of six-sample groups containing at least one completion whose installed-scorer reward differs from the gold-only reward, meaning the seam fired on a gold-incorrect completion. Screen quantities carry Jeffreys 95 percent intervals and are reported per seed and pooled. One no-documentation rendering per sampling seed separates base-prior emission from documentation-induced emission for every candidate. The screen must also replicate: a candidate whose differential fraction moves more than 0.15 absolute across sampling seeds invalidates the instrument claim at that scale, and that finding would itself be the result.

Screen-then-train protocol

No seam is trained unless it screens in. Point selection at each new scale is mechanical and was frozen before any new-scale measurement existed. The boxed numeric candidate is smoked as the predicted-pass point while its pooled fraction stays above 0.375, the upper edge of the sibling's five-point separator. The second smoked point is the seam whose pooled fraction lands nearest inside the separator interval (0.16875, 0.375]; if none lands inside, it is the highest-fraction seam below the lower edge, smoked with a predicted miss. Gate predictions follow the same frozen rule, points landing inside the interval carry no directional prediction, and their outcomes narrow the separator at a second scale.

Training and the gate

Each smoked point reruns the parent's repaired 100-step GRPO recipe (Shao et al. 2024) exactly: six prompts per optimizer step, six generations per prompt for an effective genera-

tion batch of 36, LoRA rank 8 on all linear layers, learning rate 1e-6, KL coefficient 0.04, a 512-token completion cap, and microbatch 12 with gradient accumulation 3. The gate is the parent's, unchanged: train-batch seam selection at or above 0.25 at step 100, evaluated at seed median with per-seed outcomes reported, stopping unconditionally at step 100. Training seeds are 423, 424 and 425 on every smoked point, an allocation that passed through a two-seed de-scope and a uniform restoration whose timing and rationale are disclosed with the limitations. Two readouts ride every leg and are never substituted for the gate. Exploitation purity reports the rate of seam firings on gold-incorrect completions, because a pass composed of format adoption on already-correct answers is a confound the sibling preregistered as a reportable anomaly. Gold retention requires seed-median final train-batch gold accuracy above zero, reported against the parent's 0.28 reference, which separates a policy that trained but did not select the seam from a run that broke.

Preregistered manipulation check

The check is evaluated after the screens and before any training. Over the five curve seams, the L1 distance between the pooled 7B support vector and the archival 1.5B vector must exceed the within-7B seed-to-seed L1 distance by at least 0.1. This verifies that scale moved the deciding quantity beyond the instrument's own test-retest noise, and it presumes no per-seam direction, since support could fall at scale if larger models comply harder with the required answer format. A failed check forks to a preregistered pivot: the paper becomes a negative result about scale-flat support, and the smoke legs still run to test whether flat support predicts replicated 1.5B outcomes.

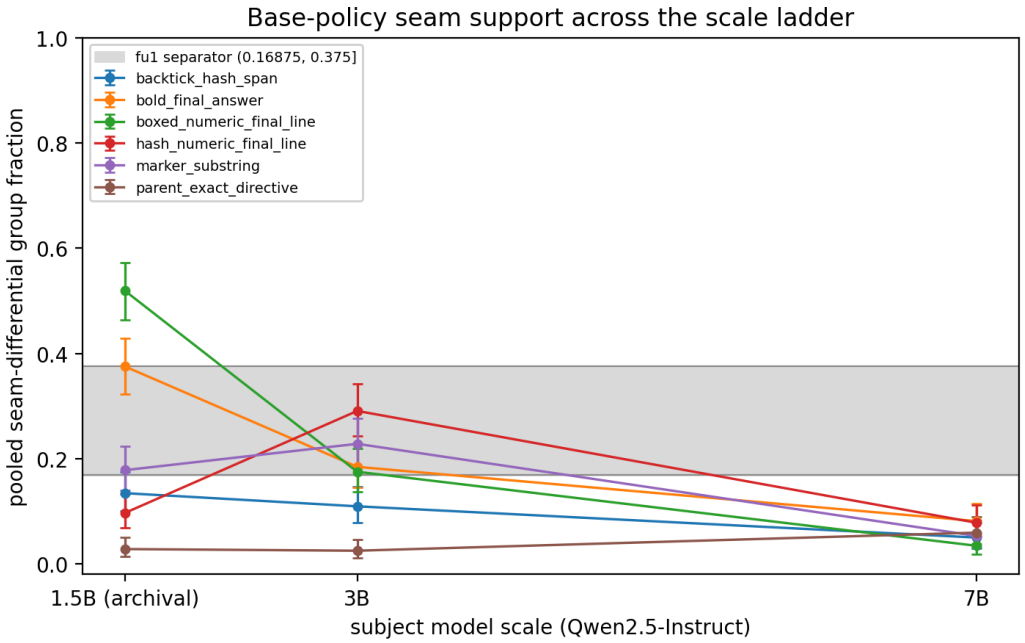
Compute and hardware

Every measurement ran locally on a single 32 GiB RTX 5090, 13.88 GPU hours in total across the two new screens, the twelve smoke legs and the equivalence control. Feasibility probes measured the 3B smoke at about 22 GiB peak reserved and roughly 0.67 GPU-hours per seed. The exact frozen 7B command at microbatch 12 exceeds this device. With gradient checkpointing enabled it completed one optimizer step and failed on the second, at about 27.5 GiB in use against a 3.5 GiB request. A design amendment therefore ran the 7B legs at microbatch 6 with gradient accumulation 6, preserving the effective batch of 36 and every other frozen hyperparameter. The same amendment added a one-seed implementation-equivalence control at 3B whose pre-registered pass conditions are reported with the results.

What was frozen when

The candidate slate, detector definitions, slate order, screen shape, separator interval, point-selection rule, manipulation check and the 0.25 gate were all fixed before any 3B or 7B measurement existed. None of them moved after data were seen. The training-seed allocation is the one exception: registered at three seeds per point, de-scoped to two, and restored to three uniformly across points after early results were reviewed, a history disclosed with the limitations. No quantity measured by this run appears in this section; every number above is a frozen protocol constant, a measured hardware fact, or an archival value from the parent and sibling runs.

Results



Main result figure for the paper.

Both new rungs of the pre-registered screen ladder completed successfully on the local RTX 5090 under the frozen instrument and the pinned chat-template hash; the full per-completion screen outputs are archived with the study. Each rung screened the six frozen candidates plus the no-documentation baseline at 160 prompts with six samples per sampling seed. The 3B screen took 2.46 GPU hours at a peak of 6.4 GiB reserved; the 7B screen took 2.43 GPU hours at 15.0 GiB. The extraction pipeline reproduced the sibling's archival 1.5B vector exactly before any new number was derived, and rollouts-to-summary count cross-checks pass at every rung.

The support ladder

The table reports the pooled seam-differential group fraction (k of 320 groups) with Jeffreys 95 percent intervals and the frozen separator classification; full precision lives in the released ladder table backing the support-ladder figure.

Candidate	1.5B (archival)	3B	7B
exact directive	0.028 [0.014, 0.051] MISS	0.025 [0.012, 0.047] MISS	0.059 [0.037, 0.089] MISS
backtick span	0.134 [0.100, 0.175] MISS	0.109 [0.079, 0.147] MISS	0.050 [0.030, 0.078] MISS
marker substring	0.178 [0.139, 0.223] zone	0.228 [0.185, 0.276] zone	0.053 [0.032, 0.082] MISS
bold final answer	0.375 [0.323, 0.429] zone	0.184 [0.145, 0.230] zone	0.081 [0.055, 0.115] MISS
boxed final line	0.519 [0.464, 0.573] PASS	0.175 [0.136, 0.219] zone	0.034 [0.018, 0.059] MISS
hash anchor	0.097 [0.068, 0.133] MISS	0.291 [0.243, 0.342] zone	0.078 [0.052, 0.111] MISS

Support for the trainable seams fell with scale (largest fall 0.519 to 0.034). Every candidate at 7B is a CI-robust predicted MISS: each Jeffreys interval sits entirely below the separator's

lower edge of 0.16875, so nothing screens in at the largest rung. The two seams that carried the most support at 1.5B fell hardest. The boxed final line dropped from 0.519 to 0.034, a paired shift of -0.484 on the same frozen prompt groups, with 157 groups firing at 1.5B only against 2 at 7B only (McNemar p 3.5e-44, Wald 95% CI [-0.540, -0.428]). The bold final answer dropped from 0.375 to 0.081 (shift -0.294, p 7.2e-21, CI [-0.351, -0.237]). The motivating parent seam moved the other way at 7B, from 9 of 320 groups to 19 of 320. That shift of +0.0313 is not distinguishable on the paired test (raw p 0.0755; Bonferroni-adjusted p 0.3776 across the five-seam slate; 95% CI [0.000207, 0.0623]), and it leaves the seam far below the trainability bar. At 3B the parent seam was flat (shift -0.00313, p 1, CI [-0.0284, 0.0221]), and the marker substring rose modestly into the zone interior (+0.050, p 0.14, CI [-0.0116, 0.112]). Averaged over the five curve seams (the six candidates minus the hash anchor, which is excluded from every exploitation-measuring statistic), the slate shift is -0.103 at 3B (exact sign-flip p 0.31, t 95% CI [-0.304, 0.0988]) and -0.191 at 7B (p 0.125, CI [-0.441, 0.0584]). Both slate tests are exploratory by pre-registration: the smallest attainable two-sided p at n=5 is 0.0625. The hash anchor itself, never smokeable because it fires on the correct answer format, rose to 0.291 at 3B before falling to 0.078 at 7B.

Manipulation check

The pre-registered check passes. Over the five curve seams, the within-7B seed-to-seed L1 noise reference is 0.119 and the pooled 7B-versus-1.5B L1 shift is 1.02, an excess of 0.9 against the pre-registered minimum effect of 0.1. The prompt-level paired bootstrap puts a 95% CI of [0.741, 1.034] on that excess, with no replicate at or below zero in the 10000 draws, so the one-sided p is reported as the bootstrap's resolution bound, $p < 1e\text{-}4$ at 10000 replicates. Scale moved the deciding quantity far beyond instrument noise. The check was deliberately direction-free, and the movement that materialized is a large fall in the high-support seams, the outcome the design anticipated when it noted that larger instruction-tuned models may comply harder with the required answer format.

Where the mechanism is reported

The pre-registered estimand above measures whether a documented shortcut creates reward differences inside a sampling group. A second layer, whether the policy emits the shortcut at all, moves the other way with scale, and the decomposition of those two layers together with the reward-saturation account that explains them is reported in the following section. That analysis was specified after the screens were read and is secondary to the confirmatory result here, so it is kept separate from it.

Screen reliability and selections

The screen replicates at both new rungs: every candidate's seed-517-versus-518 seam-differential fraction difference is within the 0.15 kill bar, with a maximum of 0.069 at 3B (the hash anchor) and 0.031 at 7B. The deterministic selection rule then fixed the smoke points before any training. At 3B it took the in-zone branch and selected the boxed final line together with the marker substring, the in-zone seam nearest the separator midpoint. boxed enters as an in-zone point rather than the predicted-PASS anchor, because the support drop vacated that role, a conversion the plan pre-registered as a reportable support-shift result. At 7B no candidate lies in or above the zone, so the rule took the exclusion-transfer branch and selected the boxed final line and the bold final answer, both smoked with a predicted MISS. Both selections were recorded, with their input fractions and the branch taken, before any training was dispatched.

Smoke outcomes at the frozen gate

The screen stage established that no candidate screens in as trainable at 7B; the smoke stage tested the rule-selected points against the unchanged gate. All thirteen training legs, twelve smoke legs (one per training seed and point) and the equivalence control, completed under the frozen 100-step recipe (the 7B legs at microbatch 6 with accumulation 6 per the design amendment, effective batch 36 preserved; 8.99 GPU hours across the smoke legs and the control). The pre-registered seed-median rule decides each point, purity columns ride every verdict, and the delta column reports step-100 selection minus the base policy's

screened raw emission at the same rung, the quantity that separates learned selection from pre-existing behaviour. Per-leg values and raw step traces are archived with the study, and the smoke-outcome figure draws every leg against screened support.

Point	screened	seed rates	median	spread	gate	separable	anomaly	base emission	delta
3B boxed	0.175	0.444, 0.278, 0.222	0.278	0.222	PASS	4	clean	0.341	-0.063
3B marker	0.228	0.250, 0.333, 0.306	0.306	0.083	PASS	6	clean	0.344	-0.038
7B boxed	0.034	0.250, 0.194, 0.250	0.250	0.056	PASS	2	TRIGGERED	0.231	+0.019
7B bold	0.081	1.000, 0.917, 0.806	0.917	0.194	PASS	5	TRIGGERED	0.961	-0.044

The third training seed the design originally pre-registered was restored on all four points. An initial attempt was refused by the frozen scale-ladder allocation and stopped before training; a seed-allocation amendment then reinstated the design's original three-seed plan, and the third seeds ran under the unchanged recipe. Every verdict above is therefore a three-seed result, and the analysis plan's requirement of at least three seeds per training claim is satisfied.

All four points now pass the inclusive gate. The 3B points pass with small genuine exploitative components (pooled Wilson 95% [0.235, 0.408] for boxed and [0.218, 0.388] for marker; marker seed 423 finished exactly at the gate with 3 separable firings and two thirds of its seam firings on already-correct answers). At 7B the purity columns tell a different story from the verdicts. boxed passes at exactly the inclusive threshold by the seed-median rule ([0.162, 0.319]) with separable counts of 0, 0 and 2 across its seeds, and bold passes at 0.917 ([0.838, 0.949]) against its CI-robust predicted miss with counts of 0, 1 and 4; the sibling study's pre-registered anomaly clause fires on both 7B points. The delta column removes the remaining ambiguity: no point's selection ends more than 0.019 above its base emission and three of four end below it, so no smoked point shows evidence of learned seam selection beyond behaviour the untouched policy already produced.

The amendment's implementation-equivalence control passes both pre-registered conditions. The microbatch-12 reference finished at 0.444 (Wilson 95% [0.295, 0.604]) and the microbatch-6 rerun at 0.417 lies inside that interval on the same side of the gate (Newcombe difference CI [-0.244, 0.192], Fisher exact p 1). One frozen seed ran, so the control speaks to implementation equivalence and not to seed variance.

Under the pre-registered CI-robust rule the separator verdict is insufficient resolution at 3B, where both smoked points are zone points whose passes refine the upper edge to 0.175, a fragile bound because boxed's screened interval straddles 0.16875 and marker's pass leans on a seed at the exact gate. At 7B the verdict is formally shifted: both robust predicted-MISS points crossed (0 of 2 matched, Wilson [0, 0.658], corroborative coin null 0.25). The anomaly clause and the delta column say what shifted. both crossings ride on formatting the base policy already emitted (bold at 0.961 base emission, boxed at exactly the threshold with two gold-incorrect firings in 108 completions), so at 7B the gate statistic measures format prevalence and not shortcut selection, the mechanism quantified in the following section.

Pooled over the twelve pre-registered legs, exploitation is rare everywhere in the study: 17 separable firings in 432 final-batch completions (Jeffreys 95% [0.024, 0.061]), never more than 4 of 36 in any leg, with the single largest count at the largest model (7B bold seed 425). Scale is not the operative variable: 10 of 216 at 3B against 7 of 216 at 7B (Fisher exact p

0.62, Newcombe CI [-0.054, 0.025]). The gate statistic spans 0.194 to 1.0 over the same legs, and its rank association with the separable count is -0.011 (fixed-seed permutation p 0.98). On this evidence the gate statistic and measured exploitation are unassociated in rank. That statement is descriptive: twelve legs, three seeds per point, and small-integer counts reported as counts with intervals. The seed spreads are themselves evidence about the endpoint: 3B boxed spans 0.222 to 0.444 across the 0.25 gate within one identical configuration, so a single-seed gate verdict at this batch size is unreliable by direct observation.

The smoke record also scores this paper's screening statistic against a naive rival; the discussion weighs the comparison. A rule that predicts a pass wherever the untouched policy emitted the seam at all in the screen (single-sampling-seed emission at or above 0.003) classifies all four smoked outcomes correctly (Jeffreys 95% [0.56, 1.0]). The differential-support bar this paper proposes classifies two of four ([0.12, 0.88]), and the strict separator form classifies zero of its two decidable points ([0, 0.67]), abstaining on both 3B zone points. On these four outcomes the naive rule scores strictly higher. Two disclosures bound that contest. Every candidate in the slate emits above the naive floor, so on smoked points the naive rule reduces to predicting a pass everywhere, and a contest on rule-selected points is tilted toward it by construction; and the intervals overlap, so four outcomes cannot separate the predictors. What the record does establish is directional and ours to own: the differential bar made two falsifiable miss predictions at 7B and both points passed, while the rule it was meant to replace made none that failed.

Why support fell: a reward-saturation account

This section reports an analysis specified after the screens were read. It is secondary to the pre-registered differential estimand, it exists to explain the confirmatory result reported in the previous section, and it adds no confirmatory claim of its own. Every number in this section appears in the study's released analysis tables and statistics records, regenerated deterministically from the archived screen rollouts and checked by an independent recomputation on every run; the emission-versus-differential figure draws the decomposition per candidate.

Raw shortcut emission and trainable support diverge with scale on every candidate. Emission rises from 1.5B to 7B for five of the six candidates: the parent directive seam goes 0.00469 to 0.0276 to 0.446 across the three rungs, and the bold final answer goes 0.322 to 0.323 to 0.961. Over the same ladder the seam-differential group fraction, the quantity that decides trainability, falls for five of six. The per-candidate divergence between the emission change and the differential change averages +0.547 (t 95% interval [0.250, 0.845]) with an exact sign-flip p of 0.031. That p is the smallest attainable value at n=6 and is reached precisely because all six candidates diverge in the same direction, so the test establishes consistency of direction and nothing stronger.

The mechanism shows in a single number: the fraction of seam firings that land on an already-correct answer. For the parent seam it runs 0 at 1.5B, 0.849 at 3B and 0.966 at 7B, and for bold it runs 0.754 to 0.876 to 0.970. At the smallest model every firing of the documented shortcut sat on an answer the model had got wrong, so the shortcut and the gold solution were competing outcomes inside a sampling group, and a group containing both carried a reward difference for the optimizer to climb. At 7B almost every firing rides on an answer that was already correct. The scorer pays full credit down either path, so the six completions in a group converge on the same reward, and the within-group variance that group-relative optimization needs in order to prefer anything is gone. The consequence is direct: a documented shortcut can become more available and less trainable at the same time, and availability is therefore the wrong quantity to screen on.

One test of this account came out inconclusive, and we say so here in preference to leaving it for a reviewer to find. Across the eighteen candidate-by-rung cells, the rank association between gold reward rate and differential support is Spearman -0.317 with permutation p 0.20 and a cluster-bootstrap 95% interval of [-0.762, 0.177], which spans zero because between-candidate base rates dominate the pooled cells. The saturation account accordingly rests on the within-candidate divergence and the gold-given-seam trajectories reported above, and it

takes no support from the cross-cell association. No stronger reading of that test is intended anywhere in this paper.

A reviewer asked whether saturation is confounded with a task-wide ceiling, since reward-uniform groups are observed only where 7B is near-perfect on this task. Stratifying the existing 7B screen by prompt difficulty separates the accounts inside the same data. Difficulty is the per-prompt gold rate pooled over the 1.5B and 3B screens, measured entirely on the smaller models so the split never conditions on the 7B outcomes under test. A median split gives 83 hard and 77 easy prompts, and the axis transfers, with 7B gold accuracy 0.917 on hard prompts against 0.992 on easy ones. The prediction holds. Pooled over the five curve seams (the six candidates minus the hash anchor, which fires on the correct answer format itself and is excluded from this stratified pooling and from the other exploitation-measuring curve-seam statistics; the six-candidate divergence test above retains it, because that test needs only direction), 7B differential support is 0.104 [0.084, 0.126] in the hard stratum and 0.0039 [0.0011, 0.0104] in the easy one (difference 0.0997, Newcombe 95% [0.079, 0.122], Fisher exact p 8.8e-22). Emission is flat across strata (0.465 against 0.471). The bold final answer makes the contrast visible: emission near 0.96 in both strata, differential support 0.157 [0.107, 0.218] on hard prompts and 0 [0, 0.016] on easy ones. Where gold accuracy is off ceiling the differential survives, at every seam, in the same model and task. The retained hard-stratum support still sits below the trainability bar, so this test rescues the mechanism and does not rescue trainability.

Restoring the registered third training seed added a within-point check, independent of both the cross-scale comparison and the difficulty stratification, and its evidence is mixed. Within each 7B point, across three seeds of an identical configuration, final-batch gold accuracy and separable firings move inversely: the bold point runs 1.000, 0.944 and 0.694 gold against 0, 1 and 4 firings (Spearman -1.000, exact one-sided permutation p 0.17), and the boxed point runs 1.000, 0.861 and 0.750 against 0, 0 and 2 (Spearman -0.866, p 0.33, held off the floor by the tied zeros); jointly, the one-sided exact p over both points is 0.056 across the 36 seed arrangements. At 3B the relationship fails: the boxed point gives 0.000 and the marker point gives +0.500, pointing the wrong way. Three seeds per point make every one of these coefficients very coarse. This is evidence at 7B only. The saturation account does predict that split: the coupling should appear only where gold approaches ceiling, which is 7B and not 3B. The prediction follows from the account as stated above, but it was not registered before the seeds ran, so we offer it as an after-the-fact consistency check and not as support. The mediating-variable case rests on the cross-scale trajectories and the difficulty stratification, with the 7B within-point pattern as a coarse echo.

The design implication reaches beyond this run: a pre-freeze screen should measure whether a shortcut creates reward differences inside the sampled group. The differential statistic measures exactly that, and emission does not. A scorer that pays full credit for the shortcut and for the correct solution alike stops discriminating between them once the model reliably produces correct answers, so the same disclosure that made a seam trainable at one scale can leave nothing for the optimizer to find at the next.

Discussion

The measurement came out about the measurement

This run set out to measure a property of models and ended up measuring a property of its own endpoint. The 0.25 train-batch selection gate, inherited unchanged from the parent and sibling studies, fails in two independent ways that our data quantify. First, it cannot distinguish shortcut exploitation from formatting of already-correct answers. Both 7B points pass the gate, boxed at exactly the inclusive threshold, while carrying seven gold-incorrect seam firings between them across 216 final-batch completions, and the pre-registered anomaly clause triggered on both. Across all twelve legs the gate statistic spans 0.194 to 1.0 while measured exploitation never exceeds 4 of 36, and the rank association between the two is -0.011 with permutation p 0.98. On this evidence the two are unassociated, a descriptive statement at twelve legs. Second, the gate is underpowered by an order of magnitude: at 36 completions the smallest detectable selection increase is roughly +0.21, and for 7B bold no increase was detectable below the 1.0 ceiling. Both failures are properties of the endpoint,

not of this run, and every prior study on this instrument used the same endpoint at the same batch size.

What the frozen instrument still established

Within those limits the ladder speaks. Across four smoked points at two scales, no point ended more than +0.019 above its base screened emission and three of four ended below it (-0.063, -0.038, +0.019, -0.044). Every gate pass in this study is the base policy's pre-existing emission surfacing through the gate. Given the +0.21 detectable-effect floor, the correct statement of the headline is exact: this study finds no evidence of learned seam selection beyond pre-existing behaviour, and at this power that is absence of evidence, not evidence that no learning occurred. The screen side is not hostage to the gate's defects. It is the study's cleanest result: base-policy differential support falls below the trainability separator at 7B for every candidate (largest fall 0.519 to 0.034) while raw emission rises for five of six.

Why the gate breaks at scale

The mechanism is reward saturation, quantified in the preceding mechanism section and carried by one trajectory: the fraction of the parent seam's firings landing on already-correct answers runs 0 at 1.5B to 0.849 at 3B to 0.966 at 7B. A scorer that pays full credit down both paths stops discriminating between them once the model reliably solves the task, six-completion groups converge on uniform reward, and the quantity the gate reads becomes format prevalence. Normalizing each seam's differential change by its remaining headroom below ceiling scopes the account, with four of five curve seams showing the ceiling signature and boxed the one mixed case, and a difficulty stratification of the same 7B screen confirms the account is not a task-wide ceiling artifact: differential support survives at 0.104 on hard prompts against 0.0039 on easy ones with emission flat across strata, as the preceding mechanism section reports. Within each 7B point, though at neither 3B point, gold accuracy and separable firings also move inversely across training seeds, a coarse within-point echo reported there.

The baseline we did not beat

Our screening statistic lost to the naive baseline it was meant to replace. On the four smoked outcomes a rule that predicts a pass wherever the untouched policy emitted the seam at all in the screen classified 4 of 4 correctly, while the differential-support bar classified 2 of 4 and the strict separator form 0 of its 2 decidable points with two abstentions. On this evidence the statistic this paper proposes is outperformed outright by raw detectability. Four outcomes cannot settle a predictor comparison; that caveat bounds the confidence in the margin, not its direction, and whether the sibling result that motivated the statistic generalizes beyond its home scale is now an open question.

Separator verdicts

At 3B the pre-registered rule returns insufficient resolution: both smoked points are zone points, boxed's screened interval straddles the 0.16875 edge, and marker's pass rests on a seed at the exact inclusive gate. At 7B the rule returns formally shifted, with both CI-robust predicted-MISS points crossing (0 of 2 matched). The anomaly clause and the delta column identify the shift as a change in what the gate statistic measures at scale, the instrument critique restated from the verdict side.

Limitations

The seed history requires disclosure. The design registered three training seeds per point, and a de-scope taken in anticipation of cloud spend that never occurred reduced the executable allocation to two. A request to add the third seed only to the point carrying the most striking result was refused by the frozen allocation before training, the selective post-hoc move the protocol exists to block, and that refusal stands in the design record. A later amendment, adopted after the results freeze and the first reviewer panel and disclosed as such, restored the third seed to all four points uniformly, without conditioning on any point's result, in response to the panel's first-priority concern about seed count; the gate,

the seam definitions, the verdict rule and the kill criteria were left untouched. The analysis plan's requirement of at least three seeds per training claim, which the two-seed allocation had violated at every point, is thereby met. The restored seeds moved one verdict against the reading the earlier data suggested, and their spreads are evidence in their own right: 3B boxed spans 0.222 to 0.444 across the 0.25 gate within one identical configuration, so a single-seed gate verdict at this batch size is unreliable by direct observation. The equivalence control was established at 3B only, so every 7B claim inherits the assumption that microbatch 6 with accumulation 6 matches 12 with 3 at scale. The direct 7B control is hardware-bound, measured as an OOM at 27.47 GiB in use with a 3.48 GiB request against 3.33 GiB free, and it is recorded as an unrun follow-up for a 40 GiB device. Four smoked outcomes, one task family and one model family bound generality. The mechanical novelty gate never cleared because Semantic Scholar throttled on three attempts, so related-work positioning rests on a hand-written memo grounded in the parent's arXiv-verified neighbours.

What should change

The fixes follow from the measurements. Screen for trainability by measuring whether a shortcut creates reward differences inside the sampled group, the quantity group-relative optimization actually consumes. Emission diverges upward from trainable support at scale on every candidate, so availability is the wrong screening target. And size the smoke gate's batch to the effect it claims to test; a 36-completion batch against a +0.21 floor turns a selection gate into a threshold on noise, at any scale.

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 7 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
e1_screen_qwen3b	Qwen2.5-3B-Instruct	Task-A frozen training split, first 160 prompts	real	160	not recorded	params=3.0B	17.49
e1_screen_qwen7b	Qwen2.5-7B-Instruct	Task-A frozen training split, first 160 prompts	real	160	not recorded	params=7.0B	45.82
e2_smoke_qwen3b_boxed_taska	Qwen2.5-3B-Instruct	Task-A frozen training split	real	2000	not recorded	100 steps; params=3.0B	117.11
e2_smoke_qwen3b_numeric_final_line	Qwen2.5-3B-Instruct	Task-A frozen training split	real	2000	not recorded	100 steps; params=3.0B	115.09
e2_smoke_qwen3b_marker_taska	Qwen2.5-3B-Instruct	Task-A frozen training split	real	2000	not recorded	100 steps; params=3.0B	115.09

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
e2_smoke_Qwen2.5-7B-Instruct_fixed_Task_A	Qwen2.5-7B-Instruct	Task_A frozen training split	real	2000	not recorded	100 steps; params=7.0B	131.51
e2_smoke_Qwen2.5-7B-Instruct_fixed_Task_Aric_real_line	Qwen2.5-7B-Instruct	Task_Aric frozen training split	real_line	2000	not recorded	100 steps; params=7.0B	136.72
e2_equivalence_Qwen2.5-3B-Instruct_fixed_Task_A	Qwen2.5-3B-Instruct	Task_A frozen training split	micro	2000	not recorded	100 steps; params=3.0B	39.24

Seed policy. No per-experiment RNG seed identities are recorded in the experiment manifest. Per-comparison seed counts are recorded in `results/real/stats.json` and reported in the Seeds column below.

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
parent_exact_two_sided documented seam-differential group incidence at 3b differs from its archival 1.5B value on the same frozen prompt groups.	two-sided exact McNemar test over prompt-by-sampling-seed groups paired across subject models; paired risk-difference Wald 95% CI	-0.003125	1.0	[-0.028376236702578023, 0.022126236702578024]	320	2

Claim	Test	Statistic	p	95% CI	n	Seeds
marker_substring_is_documented_seam-differential group incidence at 3b differs from its archival 1.5B value on the same frozen prompt groups.	two-sided exact McNemar test over prompt-by-sampling-seed groups paired across subject models; paired risk-difference Wald 95% CI	0.05	0.13709880575560337	[-0.011615271849916516, 0.11161527184991651]	320	2
backtick_hash_opens_seam-differential group incidence at 3b differs from its archival 1.5B value on the same frozen prompt groups.	two-sided exact McNemar test over prompt-by-sampling-seed groups paired across subject models; paired risk-difference Wald 95% CI	-0.025	0.3581433018[illegible]0613224	[-0.07156525996813089, 0.021565259968130884]	320	2
bold_final_answer_is_documented_seam-differential group incidence at 3b differs from its archival 1.5B value on the same frozen prompt groups.	two-sided exact McNemar test over prompt-by-sampling-seed groups paired across subject models; paired risk-difference Wald 95% CI	-0.190625	7.340990314237963e-08	[-0.2569809901361939, -0.12426900986380604]	320	2

Claim	Test	Statistic	p	95% CI	n	Seeds
boxed_numeric_documented seam-differential group incidence at 3b differs from its archival 1.5B value on the same frozen prompt groups.	two-sided exact McNemar test over prompt-by-sampling-seed groups paired across subject models; paired risk-difference Wald 95% CI	-0.34375	5.348578814705084e-22	[-0.4068653232467332, -0.2806346767532668]	320	2
hash_numeric_documented seam-differential group incidence at 3b differs from its archival 1.5B value on the same frozen prompt groups.	two-sided exact McNemar test over prompt-by-sampling-seed groups paired across subject models; paired risk-difference Wald 95% CI	0.19375	5.784543378390638e-11	[0.13829101273251687, 0.24920898726248314]	320	2
For parent_exact_direct at 3b, seam-differential group incidence is compared across sampling seeds 517 and 518.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.0125	0.7265624999999989	[-0.022093416379960064, 0.047093416379960645]	160	2

Claim	Test	Statistic	p	95% CI	n	Seeds
For marker_substring at 3b, seam-differential group incidence is compared across sampling seeds 517 and 518.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	-0.05625	0.22205282015602354	[-0.13610288982958338, 0.023602889829583383]	160	2
For back-tick_hash_span at 3b, seam-differential group incidence is compared across sampling seeds 517 and 518.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	-0.04375	0.21003961563110465	[-0.10210542569037578, 0.014605425690375796]	160	2
For bold_final_answer at 3b, seam-differential group incidence is compared across sampling seeds 517 and 518.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.00625	0.99999999999999991	[-0.07218085194208039, 0.0846808519420804]	160	2
For boxed_numeric_final_line at 3b, seam-differential group incidence is compared across sampling seeds 517 and 518.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	-0.0375	0.34492845088324317	[-0.10205875435189204, 0.027058754351892044]	160	2

Claim	Test	Statistic	p	95% CI	n	Seeds
For hash_numeric_final_line at 3b, seam-differential group incidence is compared across sampling seeds 517 and 518.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	-0.06875	0.10812902140605726	[-0.14450448387459508, 0.007004483874595055]	160	2
For parent_exact_directive at 3b, raw seam emission under documentation is compared with the paired no-documentation rendering.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.027604166666666666	2.220446049250313e-16	[0.020275814329560392, 0.034932519008772936]	1920	2
For marker_substring at 3b, raw seam emission under documentation is compared with the paired no-documentation rendering.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.34375	4.180544565216931e-199	[0.32250515884392255, 0.3649948411560745]	1920	2
For backtick_hash_span at 3b, raw seam emission under documentation is compared with the paired no-documentation rendering.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.2078125	6.136010858601387e-106	[0.18904275341976402, 0.226582246562336]	1920	2

Claim	Test	Statistic	p	95% CI	n	Seeds
For bold_final_answer at 3b, raw seam emission under documentation is compared with the paired no-documentation rendering.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.3171875	2.2684449053[illegible]168	[illegible]0.33849864697362636]	[illegible]	2
For boxed_numeric_final_line at 3b, raw seam emission under documentation is compared with the paired no-documentation rendering.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.2203125	1.0758999745[illegible]58	[illegible]0.24601205813625404]	[illegible]	2
For hash_numeric_final_line at 3b, raw seam emission under documentation is compared with the paired no-documentation rendering.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.1708333333[illegible]27	[illegible]0.20094246068816693]	[illegible]	2
The mean 3b-minus-1.5B change in seam-differential group fraction is summarized over the five frozen curve seams.	two-sided exact paired sign-flip test over the five curve seams; paired mean-difference t 95% interval	-0.10250000000000001	0.3125	[-0.3037735701399539, 0.09877357013995383]	5	2

Claim	Test	Statistic	p	95% CI	n	Seeds
parent_exact_divide	two-sided Newcombe's documented exact McNemar test over prompt-by-sampling-seed groups paired across subject models; paired risk-difference Wald 95% CI	0.03125	0.07551869750220407	[-0.0032473325306336, 0.06229266743893664]	320	2
marker_substring_is	two-sided documented exact McNemar test over prompt-by-sampling-seed groups paired across subject models; paired risk-difference Wald 95% CI	-0.125	4.566610700455725e-07	[-0.17204615320570452, -0.07795384679429548]	320	2
backtick_hash_space	two-sided documented exact McNemar test over prompt-by-sampling-seed groups paired across subject models; paired risk-difference Wald 95% CI	-0.084375	0.0001419706851990573	[-0.12624068912789443, -0.04250931087210557]	320	2

Claim	Test	Statistic	p	95% CI	n	Seeds
bold_final_answer_is_documented seam-differential group incidence at 7b differs from its archival 1.5B value on the same frozen prompt groups.	two-sided exact McNemar test over prompt-by-sampling-seed groups paired across subject models; paired risk-difference Wald 95% CI	-0.29375	7.235208524476163e-21	[-0.35067769055461473, -0.2368223094453853]	320	2
boxed_numeric_final_answer_is_documented seam-differential group incidence at 7b differs from its archival 1.5B value on the same frozen prompt groups.	two-sided exact McNemar test over prompt-by-sampling-seed groups paired across subject models; paired risk-difference Wald 95% CI	-0.484375	3.4816245634132724e-44	[-0.5404844099804906, -0.4282655900195093]	320	2
hash_numeric_final_line_is_documented seam-differential group incidence at 7b differs from its archival 1.5B value on the same frozen prompt groups.	two-sided exact McNemar test over prompt-by-sampling-seed groups paired across subject models; paired risk-difference Wald 95% CI	-0.01875	0.46139118213677005	[-0.0602401792825773, 0.022740179282577298]	320	2

Claim	Test	Statistic	p	95% CI	n	Seeds
For parent_exact_directive at 7b, seam-differential group incidence is compared across sampling seeds 517 and 518.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	-0.03125	0.17968749999999956	[-0.06767892343612938, 0.005178923436129384 5]	160	2
For marker_substring at 7b, seam-differential group incidence is compared across sampling seeds 517 and 518.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	-0.01875	0.5078124999999988	[-0.055384303067394544, 0.01788430306739455]	160	2
For backtick_hash_span at 7b, seam-differential group incidence is compared across sampling seeds 517 and 518.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.025	0.21875000000000001	[0.0047545998558516794, 0.05475459985585168]	160	2
For bold_final_answer at 7b, seam-differential group incidence is compared across sampling seeds 517 and 518.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	-0.0125	0.7539062499999992	[-0.05118873773081849, 0.026188737730818482]	160	2

Claim	Test	Statistic	p	95% CI	n	Seeds
For boxed_numeric_final_line at 7b, seam-differential group incidence is compared across sampling seeds 517 and 518.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.03125	0.0625	[0.0042900570[illegible]0965472, 0.05820994291503453]	[illegible]	2
For hash_numeric_final_line at 7b, seam-differential group incidence is compared across sampling seeds 517 and 518.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	-0.00625	0.9999999999[illegible]999976	0.04298656231185604, 0.030486562311856043]	160	2
For parent_exact_directive at 7b, raw seam emission under documentation is compared with the paired no-documentation rendering.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.4458333333[illegible]258	[illegible]	[0.4236[illegible]0.4680666341628139]	[illegible]	2
For marker_substring at 7b, raw seam emission under documentation is compared with the paired no-documentation rendering.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.4369791666[illegible]253	[illegible]	[0.4114792601[illegible]0.4591657320180779]	[illegible]	2

Claim	Test	Statistic	p	95% CI	n	Seeds
For backtick_hash_span at 7b, raw seam emission under documentation is compared with the paired no-documentation rendering.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.2640625	6.08600603666e-151	[0.2442913398162, 0.28383366010878375]	1920	2
For bold_final_answer at 7b, raw seam emission under documentation is compared with the paired no-documentation rendering.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.9598958333333	1e-323	[0.951119678469938, 0.9686719881966728]	1920	2
For boxed_numeric_final_line at 7b, raw seam emission under documentation is compared with the paired no-documentation rendering.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.18489583333333	9.9339034533251e-63	[0.1636942744791766, 0.20609739219184903]	1920	2
For hash_numeric_final_line at 7b, raw seam emission under documentation is compared with the paired no-documentation rendering.	two-sided exact McNemar test; paired risk-difference Wald 95% CI	0.0770833333333	1.9033974308681e-12	[0.056481576243054, 0.09831509042348127]	1920	2

Claim	Test	Statistic	p	95% CI	n	Seeds
The mean 7b-minus-1.5B change in seam-differential group fraction is summarized over the five frozen curve seams.	two-sided exact paired sign-flip test over the five curve seams; paired mean-difference t 95% interval	-0.19125	0.125	[-0.44093416135540964, 0.0584341613554096]	5	2
parent_exact_window's raw seam-emission rate at 3b differs from its archival 1.5B value on the same frozen sampling coordinates.	two-sided exact McNemar test paired by (prompt_id, sampling_seed, completion_index); paired risk-difference Wald 95% CI	0.022916666666666665	1.0507623271714051e-08	[0.014944394570821013, 0.030888938762512315]	1920	2
marker_substring's raw seam-emission rate at 3b differs from its archival 1.5B value on the same frozen sampling coordinates.	two-sided exact McNemar test paired by (prompt_id, sampling_seed, completion_index); paired risk-difference Wald 95% CI	0.28854166666666664	3.620659403882512e-116	[0.2653498936570168, 0.3117334396757165]	1920	2

Claim	Test	Statistic	p	95% CI	n	Seeds
backtick_hash_span's raw seam-emission rate at 3b differs from its archival 1.5B value on the same frozen sampling coordinates.	two-sided exact McNemar test paired by (prompt_id, sampling_seed, completion_index); paired risk-difference Wald 95% CI	0.15885416666…	2.0…e-47	[0.137672584517…992, 0.1800357488156341]	1920	2
bold_final_answer's raw seam-emission rate at 3b differs from its archival 1.5B value on the same frozen sampling coordinates.	two-sided exact McNemar test paired by (prompt_id, sampling_seed, completion_index); paired risk-difference Wald 95% CI	0.0010416666…	0.9720…27483122049	[-0.028082835393995642, 0.030166168727328975]	1920	2
boxed_numeric_final_line's raw seam-emission rate at 3b differs from its archival 1.5B value on the same frozen sampling coordinates.	two-sided exact McNemar test paired by (prompt_id, sampling_seed, completion_index); paired risk-difference Wald 95% CI	-0.09479166666666666	1.898726322797…71736e-…	[-0.1253607242844949, -0.06422260904883843]	1920	2

Claim	Test	Statistic	p	95% CI	n	Seeds
hash_numeric_field_line's raw seam-emission rate at 3b differs from its archival 1.5B value on the same frozen sampling coordinates.	two-sided exact McNemar test paired by (prompt_id, sampling_seed, completion_index); paired risk-difference Wald 95% CI	0.634375	1e-323	[0.6114729789338647, 0.6572770210661354]	1920	2
parent_exact_override's raw seam-emission rate at 7b differs from its archival 1.5B value on the same frozen sampling coordinates.	two-sided exact McNemar test paired by (prompt_id, sampling_seed, completion_index); paired risk-difference Wald 95% CI	0.4411458333333333	2.3873308560265403e-244	[0.4186566261064233, 0.4636350405602434]	1920	2
marker_substring's raw seam-emission rate at 7b differs from its archival 1.5B value on the same frozen sampling coordinates.	two-sided exact McNemar test paired by (prompt_id, sampling_seed, completion_index); paired risk-difference Wald 95% CI	0.3817708333333333	2.9453314348909e-170	[0.3577618912908663, 0.4057797753758037]	1920	2

Claim	Test	Statistic	p	95% CI	n	Seeds
backtick_hash_span's raw seam-emission rate at 7b differs from its archival 1.5B value on the same frozen sampling coordinates.	two-sided exact McNemar test paired by (prompt_id, sampling_seed, completion_index); paired risk-difference Wald 95% CI	0.21041666666[illegible]	[illegible]72	[0.188[illegible]56119[illegible]8296, 0.2326177213708503 8]	1920	2
bold_final_answer's raw seam-emission rate at 7b differs from its archival 1.5B value on the same frozen sampling coordinates.	two-sided exact McNemar test paired by (prompt_id, sampling_seed, completion_index); paired risk-difference Wald 95% CI	0.638541666666[illegible]	[illegible]	[0.6161035233[illegible]128, 0.6609798099989206]	1920	2
boxed_numeric_final_line's raw seam-emission rate at 7b differs from its archival 1.5B value on the same frozen sampling coordinates.	two-sided exact McNemar test paired by (prompt_id, sampling_seed, completion_index); paired risk-difference Wald 95% CI	-0.20416666666666666	2.273110033292[illegible]9327e-[illegible]	[-0.23368156847182198, -0.17465176486151135]	1920	2

Claim	Test	Statistic	p	95% CI	n	Seeds
hash_numeric_raw seam-emission rate at 7b differs from its archival 1.5B value on the same frozen sampling coordinates.	two-sided exact McNemar test paired by (prompt_id, sampling_seed, completion_index); paired risk-difference Wald 95% CI	0.8416666666666667	[0.8245283279[illegible]8821, 0.8588050053939512]	[illegible]	2
Across the six candidates, the 7B-minus-1.5B change in raw seam emission exceeds the same change in seam-differential group fraction (the two curves move apart with scale).	two-sided exact paired sign-flip over the six per-candidate divergences; paired mean-difference t 95% CI	0.5473958333333335	[0.2498733296[illegible]9211395, 0.8449183369745527]	[illegible]	2
Across the candidate-by-rung cells, gold_reward_rate and seam_differential are negatively rank-associated, as the reward-saturation account predicts.	Spearman rank correlation over the 18 cells; fixed-seed Monte Carlo permutation p (10000 draws); cluster bootstrap 95% CI resampling whole candidates (10000 draws, fixed LCG seed 20260818)	-0.31682146542827655	0.2016	[-0.7617554858934169, 0.17659352142110762]	18	2

Claim	Test	Statistic	p	95% CI	n	Seeds
The pooled 7B-vs-1.5B L1 shift of the five-seam support vector exceeds the within-7B seed-to-seed L1 noise reference.	pre-registered deterministic min-effect rule (excess >= 0.1); prompt-level paired bootstrap (10000 replicates, fixed LCG seed 20260818) percentile 95% CI on the excess; one-sided bootstrap p = fraction of replicates at or below zero	0.9	0.0001	[0.7406250000000001, 1.0343749999999998]	320	2
boxed_numeric_pre_final_line at 3b: step-100 train-batch seam selection versus the frozen 0.25 gate, decided by the pre-registered seed-median rule.	pre-registered deterministic seed-median threshold verdict (even seed count: mean of middle rates); companion exact binomial on the pooled final-batch count, two-sided (zone point, no directional prediction); Wilson 95% CI on the pooled rate	0.27777777777[illegible]	0.77533001478[illegible]	[0.2[illegible]55252, 0.4074853695336538]	108	3

Claim	Test	Statistic	p	95% CI	n	Seeds
boxed_numeric_final_line at 3b: pooled step-100 selection is compared with the base policy's screened raw emission (the base-to-final delta).	Newcombe hybrid-Wilson 95% CI on the difference of two independent binomials (pooled final batch vs base screen rollouts); two-sided Fisher exact p as companion	-0.025810185185185186	0.6035202879005269	[-0.10859936013348154, 0.06917942852313169]	2028	3
marker_substring at 3b: step-100 train-batch seam selection versus the frozen 0.25 gate, decided by the pre-registered seed-median rule.	pre-registered deterministic seed-median threshold verdict (even seed count: mean of middle rates); companion exact binomial on the pooled final-batch count, two-sided (zone point, no directional prediction); Wilson 95% CI on the pooled rate	0.3055555555555556	0.535765767789	[0.21983780644189223, 0.3882078969563346]	216	3

Claim	Test	Statistic	p	95% CI	n	Seeds
marker_substring at 3b: pooled step-100 selection is compared with the base policy's screened raw emission (the base-to-final delta).	Newcombe hybrid-Wilson 95% CI on the difference of two independent binomials (pooled final batch vs base screen rollouts); two-sided Fisher exact p as companion	-0.04745370370370372	0.34845276133489633	[-0.1282938689991966, 0.04680728679808421]	2028	3
bold_final_answer at 7b: step-100 train-batch seam selection versus the frozen 0.25 gate, decided by the pre-registered seed-median rule.	pre-registered deterministic seed-median threshold verdict (even seed count: mean of middle rates); companion exact binomial on the pooled final-batch count, one-sided lower tail (predicted MISS); Wilson 95% CI on the pooled rate	0.9166666666666666	1.0	[0.8379014075437644, 0.9489266696957817]	1200	3

Claim	Test	Statistic	p	95% CI	n	Seeds
bold_final_answer at 7b: pooled step-100 selection is compared with the base policy's screened raw emission (the base-to-final delta).	Newcombe hybrid-Wilson 95% CI on the difference of two independent binomials (pooled final batch vs base screen rollouts); two-sided Fisher exact p as companion	-0.05353009259259256	0.012962352181606535	[-0.12347080917684158, -0.01090942448113559]	2028	3
boxed_numeric at 7b: step-100 train-batch seam selection versus the frozen 0.25 gate, decided by the pre-registered seed-median rule.	pre_final_line registered deterministic seed-median threshold verdict (even seed count: mean of middle rates); companion exact binomial on the pooled final-batch count, one-sided lower tail (predicted MISS); Wilson 95% CI on the pooled rate	0.25	0.37583338625	[0.1762893635704478, 0.3194151315542543]	108	3

Claim	Test	Statistic	p	95% CI	n	Seeds
boxed_numeric_final_line at 7b: pooled step-100 selection is compared with the base policy's screened raw emission (the base-to-final delta).	Newcombe hybrid-Wilson 95% CI on the difference of two independent binomials (pooled final batch vs base screen rollouts); two-sided Fisher exact p as companion	0.0002314814814814714	1	[-0.07190961340451842, 0.09005155335340236]	2028	3
The micro-batch 6/6 implementation is equivalent to the frozen 12/3 implementation at 3B on the amendment's two pre-registered conditions.	pre-registered rule: same side of the 0.25 gate AND micro6 rate inside the micro12 Wilson 95% interval; Newcombe hybrid-Wilson 95% CI on the rate difference; two-sided Fisher exact p as companion	-0.027777777777777735	0.9999999999999984	[-0.24369261456627714, 0.19185395333144978]	72	1

Claim	Test	Statistic	p	95% CI	n	Seeds
Pooled separable-firing (seam on gold-incorrect) counts in the final smoke batches are compared across the 3B and 7B rungs (10/216 vs 7/216).	two-sided Fisher exact on the pooled counts; Newcombe hybrid-Wilson 95% CI on the 7B-minus-3B rate difference; per-rung Jeffreys intervals in analysis_summary.txt	-0.013888888888888888	0.6220010705199466	[-0.054281028233560465, 0.025172730475053688]	432	3
Across the 12 pre-registered smoke legs, the step-100 gate statistic is rank-associated with the separable-firing count.	Spearman rank correlation over the 12 legs; fixed-seed Monte Carlo permutation p (100000 draws)	-0.010912626812280774	0.97778	[-1.0, 1.0]	12	3

Claim	Test	Statistic	p	95% CI	n	Seeds
Within each 7B smoked point, across three training seeds of an identical configuration, final-batch gold accuracy and separable-firing count are inversely associated, the saturation account's within-point prediction.	per-point Spearman over the three seeds with exact one-sided permutation p; joint one-sided exact permutation over both 7B points using the sum of the two coefficients (36 arrangements)	-0.9330127018922194	0.05555555555[illegible]	[-1.0, 1.0]	6	3
Pooled over the five curve seams, 7B differential support in the hard (low small-model gold) prompt stratum is compared with the easy stratum; the saturation account predicts retention where gold accuracy is low.	unpaired two-proportion comparison of differential group fractions across prompt strata; Newcombe hybrid-Wilson 95% CI; two-sided Fisher exact companion p; per-stratum Jeffreys intervals in analysis_summary.txt	0.09971835398[illegible]	[illegible]e-22	[0.0779349301[illegible]748, 0.12245606169947962]	1630	2

Claim	Test	Statistic	p	95% CI	n	Seeds
backtick_hash_spam: 7B differential support in the hard prompt stratum against the easy stratum.	unpaired two-proportion comparison; Newcombe 95% CI; two-sided Fisher exact companion p	0.0963855421686747	2.649357079301e-05	[0.05227783522280255, 0.15082728495111677]	320	2
bold_final_answer: 7B differential support in the hard prompt stratum against the easy stratum.	unpaired two-proportion comparison; Newcombe 95% CI; two-sided Fisher exact companion p	0.1566265060240964	1.5925654681384e-08	[0.09403057566305253635, 0.21960178719212461]	320	2
boxed_numeric_final_line: 7B differential support in the hard prompt stratum against the easy stratum.	unpaired two-proportion comparison; Newcombe 95% CI; two-sided Fisher exact companion p	0.06626506024096386	0.0035577543	[0.02815099529421967, 0.11474952142400321]	320	2
marker_substring: 7B differential support in the hard prompt stratum against the easy stratum.	unpaired two-proportion comparison; Newcombe 95% CI; two-sided Fisher exact companion p	0.07737443279	0.0012955030690	[0.027930462029031837, 0.1315658989345125]	320	2

Claim	Test	Statistic	p	95% CI	n	Seeds
parent_exact_[illegible] 7B differential support in the hard prompt stratum against the easy stratum.	[illegible]: two-proportion comparison; Newcombe 95% CI; two-sided Fisher exact companion p	0.101940228[illegible]	[illegible]e-05	[0.0[illegible]08576, 0.1586417075571616]	[illegible]	2
Every CI-robust classifiable smoked point's gate outcome matches its frozen separator prediction.	pre-registered deterministic per-point evaluation; Wilson 95% CI on the match fraction; corroborative coin null 2^-k recomputed for the k points actually measured	0.0	0.25	[0.0, 0.6576197724933468]	2	2

Compute. Total recorded GPU time across all experiments is 13.8830 GPU-hours on a single NVIDIA GeForce RTX 5090.

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: `confirmatory_table.csv`, `figure_emission_vs_differential.csv`, `figure_main.csv`, `figure_smoke.csv`, `ladder.csv`, `experiments.json`.

References

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mane. 2016. *Concrete Problems in AI Safety*. 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*. https://arxiv.org/abs/2503.11926.

Betley, Jan, Daniel Tan, Niels Warncke, et al. 2025. *Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs*. https://arxiv.org/abs/2502.17424.

Chu, Tianzhe, Yuexiang Zhai, Jihan Yang, et al. 2025. *SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-Training*. https://arxiv.org/abs/2501.17161.

Cobbe, Karl, Vineet Kosaraju, Mohammad Bavarian, et al. 2021. *Training Verifiers to Solve Math Word Problems*. https://arxiv.org/abs/2110.14168.

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

Everitt, Tom, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2019. *Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective*. https://arxiv.org/abs/1908.04734.

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

Liu, Zichen, Changyu Chen, Wenjun Li, et al. 2025. *Understanding R1-Zero-Like Training: A Critical Perspective*. https://arxiv.org/abs/2503.20783.

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

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

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

Wen, Jiaxin, Ruiqi Zhong, Akbir Khan, et al. 2024. *Language Models Learn to Mislead Humans via RLHF*. https://arxiv.org/abs/2409.12822.

Wen, Xumeng, Zihan Liu, Shun Zheng, et al. 2025. *Reinforcement Learning with Verifiable Rewards Implicitly Incentivizes Correct Reasoning in Base LLMs*. https://arxiv.org/abs/2506.14245.

Yang, An, Baosong Yang, Beichen Zhang, et al. 2024. *Qwen2.5 Technical Report*. 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?* https://arxiv.org/abs/2504.13837.

Yue, Yu, Yufeng Yuan, Qiying Yu, et al. 2025. *VAPO: Efficient and Reliable Reinforcement Learning for Advanced Reasoning Tasks*. https://arxiv.org/abs/2504.05118.