AUDIT CERTAINTY IS A SIGNED LEVER: WHY A VERIFIER CALIBRATED ON A PROXY CAN BACKFIRE IN DEPLOYMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We calibrated a verifier on a proxy task, deployed it unchanged, and its strongest setting became its worst: raising audit certainty — the lever that cut violations on the proxy — *raised* them in deployment. Our claim — *proxy calibration of audit certainty need not transfer across temporal structures for learning agents* — rests on two findings of different standing. The first is derived. Classical deterrence says only the expected penalty, probability times severity, matters; for an agent that *learns* from the signal it does not, because rare severe penalties inflate the variance of the learned violation value. We prove an exact stationary-variance result and, under a Gaussian threshold approximation, a *signed certainty law*: raising certainty deters when the violation is value-suboptimal and backfires when it is value-optimal. The law predicted the inversion before the experiment; a deep-RL phase diagram confirms it, and the exact stationary law lets us quantify the approximation's error numerically. The second finding is causally isolated but not derived, and the headline rests on it: the *location* of the sign boundary moves with task timing. With timing equalised it sits where the law predicts; delaying honest reward alone pushes it past every value gap we can measure, so a timing-mismatched proxy certifies the wrong policy. **That displacement is orders of magnitude beyond the measured approximation error, so the theory we prove does not explain it**; we isolate the channel, rule out four candidate derivations and one whole route, and leave the closed form open. Under GRPO on an open LLM, over a pre-registered orthogonal grid of several hundred training runs, the certainty-over-severity ordering reproduces, and the logged advantages show that group normalisation caps the advantage any single penalty can carry; two controls that remove the cap widen the severity signal five-fold yet move the behavioural response at most partially, so the cap is the arithmetic of the truncation and only part, if any, of the ordering's cause. Design rule: calibrate on matched timing, or measure the value gap under deployment timing before trusting a proxy.

1 INTRODUCTION

A verifier that detects reward hacking is not free. Running it more often costs compute, latency, or human attention; the penalty it applies when it fires costs almost nothing to write down. So a verifier designer with a fixed budget faces a concrete allocation question: buy *more frequent auditing*, or buy *harsher penalties* for the violations you do catch? Fig. 1 shows why the answer is not stable.

The standard answer says it does not matter: in Becker's model of deterrence (Becker, 1968) only the *expected* penalty $p \cdot c$ matters, so with c cheap and p expensive the prescription is sparse audits with draconian penalties. But that model assumes an agent that *knows* p and c. We characterise when expected penalty stops being a sufficient statistic: for an agent learning from the verifier's signal, rare severe penalties are *experienced* rarely, inflating the variance of the learned violation value — a property of the learning rule, not an installed risk preference.

The effect has a sign, and the sign flips: raising audit certainty deters when the violation is value-suboptimal and *backfires* when it is value-optimal. Three claims of different strength run through the paper, and we label each: the stationary *variance* is exact (Proposition 1); the signed *threshold*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

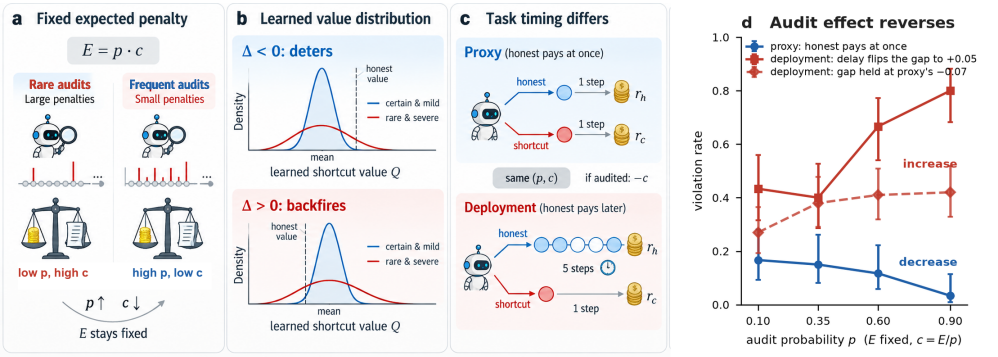


Figure 1: **Why a proxy-calibrated verifier can invert in deployment.** **(a)** A fixed $E = pc$ can be spent as rare severe audits or frequent mild ones; an expected-value agent is indifferent. **(b)** A learner is not: rare severe penalties widen the learned shortcut value around the same mean, so certainty deters when the shortcut is value-suboptimal ($\Delta < 0$) and backfires when it is value-optimal ($\Delta > 0$; Sec. 3). **(c)** Proxy and deployment tasks differ only in *when* honest work pays. **(d)** Measured: the same verifier deters on the proxy ($0.167 \to 0.033$) and raises violations in deployment ($0.433 \to 0.800$; $z = 4.01$, 60 seeds/point); with r_c lowered so the deployed value gap is held at the proxy's -0.07 (dashed), it still raises them ($0.27 \to 0.42$; interaction $z = 4.53$, 100 seeds/point). Wilson 95% CIs. Panels a–c are schematic.

rule, $\mathrm{sign}(\partial P/\partial p) = \mathrm{sign}(\Delta)$, holds under a Gaussian approximation whose error we quantify numerically against the exact law — which can be locally non-monotone, and in the sparse-audit limit breaks the rule's global monotonicity (App. L); and the sign *boundary* is measured, not computed.

The paper's weight sits on that boundary's *location*. The timing-equalised rule places it at $\Delta_\gamma = 0$; timing-equalised measurement lands there (-0.007); merely delaying honest reward pushes it beyond every gap we test (< -0.45). That displacement — not the flip itself — is what breaks proxy calibration: it moves with task structure the verifier cannot see, and *it is the one quantity in this chain we cannot yet derive*. Sec. 7 reports the channel we isolate and the candidate accounts we falsify.

Our claims are about *how a verification signal shapes what an agent learns*. Three negative results against our own earlier draft are stated where a reader meets the claim each retracts.

Contributions. Four steps. **(i) Expected penalty is not a sufficient statistic.** For constant-step-size value learners we prove $\mathrm{Var}(Q_\infty) = \frac{\alpha}{2-\alpha} c^2 p(1-p)$ exactly, diverging as $p \to 0$ at fixed E; a model-based control and a learning-rate fingerprint show the effect is a property of *learning*, not an installed risk preference (Sec. 3, 4.1). **(ii) Audit certainty has a signed effect.** A *signed threshold law* under a Gaussian threshold approximation, $\mathrm{sign}(\partial P/\partial p) = \mathrm{sign}(\Delta)$, with error quantified numerically in App. L, predicted before the experiment that the lever must *invert* rather than weaken. A 2000-run timing-equalised phase diagram confirms the inversion where the law places it (boundary -0.007; Fig. 4), and a 2800-run delayed-task phase diagram shows the log-certainty slope is itself a function of Δ_γ (interaction $p = 1.6 \times 10^{-10}$; Sec. 4.3, Fig. 2). **(iii) Temporal mismatch moves the boundary far enough to reverse a deployment calibration.** Delaying honest reward pushes the boundary past every value gap we can measure (no zero crossing down to $\Delta_\gamma = -0.45$; pooled $p = 3 \times 10^{-6}$); equalising timing returns it to -0.007. A verifier calibrated on a single-step proxy and deployed unchanged at identical (p, c) therefore has its certainty lever flip from cutting violations ($0.167 \to 0.033$) to raising them ($0.433 \to 0.800$; interaction $z = 4.01$, $N = 480$), and it still flips when r_c is lowered so the deployed value gap is held at the proxy's negative value ($0.27 \to 0.42$; interaction $z = 4.53$; Sec. 4.3–4.4). **(iv) Under GRPO weight training the ordering reproduces; the optimiser's normalisation caps penalties exactly, and removing the cap restores severity at most partially.** GRPO on Qwen2.5-7B over a pre-registered orthogonal 5×4 grid (340 runs): the certainty-minus-severity contrast — the entire departure from a pure expected-value agent, for which $b_p = b_c$ — is -1.36 ± 0.40 at the cell level ($p = 0.003$), stable under clustered errors, permutation and leave-one-cell-out. Logged advantages show group normalisa-

tion bounding a single penalised completion's advantage by $(K-1)/\sqrt{K}$, with the traces saturating there as c grows $8\times$; a K-sweep orders severity's weight with K as the bound predicts, marginally at 132 runs. Two cap-free controls (penalty reaching -7 to -8 instead of -2.5) give severity slopes -3.5 ± 1.2 and -5.4 ± 0.9 against GRPO's -3.2 ± 0.5 — pooled, steeper by 1.3 ± 0.9 ($p = 0.11$), far short of the five-fold wider signal — so the cap explains the truncation and an unresolved part of the behaviour. Training also raised shortcut-taking in $13/20$ cells, an increase the bound does not explain (Sec. 4.5, Fig. 3).

Supporting results. The effect survives neural value functions and a tabular replication, and skew, exploration and function approximation are eliminated as causes of the displacement (Sec. 4.1–4.3); the oversight-budget implication is in Sec. 5 and the appendices.

2 Related Work

Deterrence: economics and criminology. Becker's (1968) model makes the expected penalty $p \cdot c$ the sufficient statistic and underlies the "rare audit, severe penalty" prescription; it assumes an agent that *knows* p and c. We characterise where that sufficiency breaks down for a learning agent. Empirically the opposite ordering — certainty over severity — is among the most robust findings in deterrence research, from Beccaria (Beccaria, 1764) through Gibbs (Gibbs, 1975) to the perceptual-deterrence tradition (Paternoster, 1987; Berger et al., 2015), which further finds *perceived* certainty dominates objective certainty.

Underweighting rare outcomes is not our contribution. That an agent learning from sampled outcomes underweights rare severe penalties is classical: the hot stove effect (Denrell & March, 2001), the description–experience gap (Hertwig et al., 2004; Barron & Erev, 2003), risk-sensitive TD updates (Mihatsch & Neuneier, 2002; Prashanth & Ghavamzadeh, 2018) and distributional RL (Bellemare et al., 2017; Dabney et al., 2018a;b) all describe a bias in the learned estimate. What is new is the verifier-budget framing: the effect has an exact closed form in the constant-step-size case and a *sign* (Eq. (1)), so one intervention reduces violations in one regime and raises them in another, and a deployed system can cross that condition with no change to the verifier (Sec. 4.3–4.4).

Monitor-based penalties and proxy divergence. The closest real-system instance of our lever is penalising reward hacking through a chain-of-thought monitor during RL training, where increasing monitor pressure drove obfuscated rather than reduced misbehaviour (Baker et al., 2025). We supply a sign law for that family: what a monitor's *detection rate* does to learned behaviour is set by the violation's value advantage, independently of the penalty it carries. Reward-model overoptimization (Gao et al., 2022) is the neighbouring concern: there a proxy *reward* diverges under pressure, here a proxy-calibrated verifier inverts under a change in temporal structure.

Oversight, reward hacking, and the learning substrate. Specification gaming and reward misspecification (Amodei et al., 2016; Skalse et al., 2022; Pan et al., 2022), reward tampering (Denison et al., 2024; Everitt et al., 2019; Kumar et al., 2020), sandbagging (van der Weij et al., 2024), model-written evaluations (Perez et al., 2022) and scalable-oversight proposals (Leike et al., 2018; Irving et al., 2018; Bowman et al., 2022) are largely qualitative; we supply a quantitative lever and an allocation rule. The AI-control line (Greenblatt et al., 2023) asks how a protocol survives deliberate subversion; our question is prior — what the signal a protocol emits teaches an agent that learns from it. Reinforcement learning from verifiable rewards with group-relative advantages (Shao et al., 2024; DeepSeek-AI, 2025) is now the standard substrate for that signal, and its normalisation is known to carry optimisation biases (Liu et al., 2025b); Sec. 4.5 shows one such bias is exactly a severity truncation at the advantage level; removing it moves behaviour at most partially. The substrate is standard: constant-step-size TD (Sutton, 1988; Watkins & Dayan, 1992; Sutton & Barto, 2018) is stochastic approximation (Robbins & Monro, 1951) whose stationary variance scales with the step size (Lakshminarayanan & Szepesvári, 2017). LLMs separately carry prospect-theoretic risk distortions (Kahneman & Tversky, 1979; Payne, 2025; Liu et al., 2025a).

Sensitivity exponent. Perceptual-deterrence models fit a perceived penalty $E^{\eta}p^{1-\eta}$ with $\eta < 1$, so that certainty counts more than severity. The constant-α learner realises $\eta < 1$ endogenously,

and with a testable difference: because σ^2 scales with α, the effect must grow with the learning rate, whereas a hand-coded risk attitude is α-invariant (Sec. 4.1).

3 CERTAINTY > SEVERITY IN VALUE ESTIMATES

Consider an agent choosing between an honest action of deterministic reward r_h and a violating action ("cheat") with per-trial payoff $X = r_c - cB$, where $r_c > r_h$ and $B \sim \text{Bernoulli}(p)$ is the detection event with penalty c. Then $\mathbb{E}[X] = r_c - pc = r_c - E$ with $E := pc$, and $\text{Var}(X) = c^2 p(1-p)$. A risk-neutral agent that *knows* these values is, by construction, indifferent to how E is split. A *learning* agent estimates the cheat value by $Q \leftarrow (1-\alpha)Q + \alpha X$. We analyse the *conditional* value estimator for the violating arm — not a multi-state Q-learning theorem — and Sec. 4.1 verifies the effect inside the full ε-greedy loop. In the sequential tasks of Sec. 4.2–4.4 the independence behind Proposition 1 fails (bootstrapped targets are dependent), so there we use it as an *ordinal* prediction only.

Proposition 1 (stationary variance)**.** *The estimate has a unique stationary distribution with* $\text{Var}(Q_\infty) = \frac{\alpha}{2-\alpha}\text{Var}(X) = \frac{\alpha}{2-\alpha}c^2 p(1-p) = \frac{\alpha}{2}c^2 p(1-p) + O(\alpha^2)$.

Corollary 1 (iso-E variance)**.** *Fixing* $E = pc$, $\text{Var}(Q_\infty) = \frac{\alpha}{2-\alpha}E^2\frac{1-p}{p}$*, strictly decreasing in* p*: it diverges as* $p \to 0$ *and vanishes as* $p \to 1$.

Proposition 2 (signed certainty law, Gaussian threshold approximation)**.** *Model exploitation as "cheat iff* $Q_\infty > r_h$*" and* approximate *the stationary estimate as* $Q_\infty \approx \mathcal{N}(r_c - E, \sigma^2)$*,* $\sigma^2 = \text{Var}(Q_\infty)$*. With* $\Delta := r_c - r_h - E$ *the cheat probability is* $P(p) = \Phi(\Delta/\sigma(p))$*, and at fixed* E

$$\text{sign}\,(\partial P/\partial p) = \text{sign}(\Delta). \tag{1}$$

Hence certainty deters *iff the violation is expected-value-suboptimal (*$\Delta < 0$*) — with* $P \to 0$ *as* $p \to 1$ *— is behaviourally* neutral *at* $\Delta = 0$ *(*$P \equiv 1/2$*), and* backfires *iff the violation is expected-value-optimal (*$\Delta > 0$*), with* $P \to 1$ *as* $p \to 1$*. The approximation describes the audited regime: under the exact stationary law* $Q_\infty \le r_c$*, so* $P \to 1$ *as* $p \to 0$ *for every* $\Delta > -E$*, and the* $\Delta > 0$ *branch holds only above a* Δ*-dependent audit rate (App. L).*

Equation (1) is the sharpened form of the criminological law: what certainty does is set by the *sign* of the value gap, because σ is strictly decreasing in p (Corollary) and reducing noise pushes a threshold-crossing probability toward whichever side the mean already favours. This makes a falsifiable prediction we did not build in — the certainty effect must *invert*, not merely weaken, once $\Delta > 0$ — which Sec. 4.3 tests. Proposition 1 is exact; Proposition 2 is an approximation (Q_∞ is a bounded, left-skewed Bernoulli sum), checked against the *exact* stationary law over 54 (p, α, E, Δ) cells in Appendix L ($r = 0.98$; iso-E ordering in $9/9$ slices; the Gaussian errs toward over-warning). What ties Proposition 1 to *behaviour* is the α-scaling of Sec. 4.1–4.2; the phase diagram and deployment experiments test only the *sign* law.

4 EXPERIMENTS

All studies use the honest-vs-cheat structure of Sec. 3. Unless noted the learner is risk-neutral (no imposed risk-attitude parameter), $r_h = 1.0$, $r_c = 2.0$, and we hold $E = pc$ fixed while sweeping p. Full hyperparameters are in Appendix J.

4.1 THE EFFECT EMERGES IN TABULAR Q-LEARNING, AND REQUIRES LEARNING

With a constant-α tabular Q-learner and *no* risk parameter, cheating falls monotonically as certainty rises at fixed E (Fig. 11A). The cheat rate tracks the closed-form $\text{Var}(Q)$ ($\text{corr} = 0.96$), and the certainty benefit grows monotonically — empirically, near-linearly — with α (Fig. 11B, $\text{corr} = 0.999$). Proposition 1 predicts linear-in-α scaling of the estimator *variance*; that the behavioural effect also scales near-linearly is an empirical observation consistent with it, not a corollary, and its value here is as a fingerprint: a hand-coded risk attitude is α-invariant. *Scope:* in the full ε-greedy loop Eq. (1) is verified as a sign law — direction and ordering across p — not a quantitative one, since absolute rates also depend on sampling frequency (Appendix N); the phase boundary of Sec. 4.3 is measured, not computed from $\Phi(\Delta/\sigma)$.

Model-based control: the effect requires learning. A control that *knows* (p, c) and acts on $r_c - pc$ is flat across the iso-E sweep (1.0 or 0 by regime; Fig. 14) while the learner shows the full gradient in both: the effect is a property of *learning* a value from experience, not of risk preference over known lotteries.

4.2 The effect survives neural value functions and policy gradients

Replacing the table with a two-hidden-layer MLP Q-function on CPU (Fig. 15), certainty $>$ severity is clean and monotone in a memoryless contextual bandit ($0.108 \rightarrow 0.000$). Under a sequential gridworld a U-shape appears: the low-certainty arm is the variance mechanism of Sec. 3, the high-certainty arm a distinct boundary effect we name the *sequential immediacy failure mode*. Its source is the timing asymmetry: the violation pays in *one* step while the honest payoff arrives after $L - 1$, so every cost of delay — accumulated step costs and discounting alike — falls on the honest return only (Sec. 4.3 isolates this causally; why it grows with p is not derived). Write $\Delta_\gamma = (r_c - E - s) - \left[\gamma^{L-2} r_h - s \sum_{k=0}^{L-2} \gamma^k\right]$ for the discounted counterpart of Proposition 2's Δ (per-step cost $s = 0.02$, corridor length L): delay lowers it and can change its sign. Eq. (1) then makes a *prior* prediction: the certainty effect should not merely weaken but *invert* once Δ_γ crosses zero. Within a single task Proposition 2 predicts a *monotone* response (its Δ is fixed), so the intra-task U-shape is itself a departure from the law: the variance channel falling in p alongside an unmodelled, p-increasing immediacy component that Sec. 4.3 isolates. Across tasks, two discount-family levers move Δ_γ (Fig. 12; Appendix T) and the pooled certainty slope tracks it ($r = 0.985$, $n = 9$ cells of 100 seeds), with a coarse sign change near -0.18 — not at 0, previewing the displacement Sec. 4.3 bounds. A first check at $L = 2$, where the honest path is *also* one step, restores clean monotone deterrence ($0.14 \rightarrow 0.00$, $z = 4.03$, $p = 5.5 \times 10^{-5}$); we return to it at full scale in Sec. 4.3. A second, non-minimal environment — three branching paths, an exploit that must be discovered rather than offered, a multi-step violation, delayed honest payoff — reproduces both arms of the sign law ($0.094 \rightarrow 0.912$ where the shortcut is value-optimal, $0.056 \rightarrow 0.030$ where it is not; Appendix G). A Double-DQN control (van Hasselt et al., 2016) rules out value over-estimation as the source of the U-shape.

The effect is not specific to Q-learning: it reappears under policy gradients, with the same fingerprint. On REINFORCE with a learned baseline (the advantage/critic structure PPO inherits), the predicted *sign difference* is significant: the log-odds certainty slope is -0.277 (95% CI $[-0.402, -0.132]$) in the value-suboptimal regime and $+0.104$ in the value-optimal one — the positive arm not individually significant — with a between-regime difference of $+0.380$ (95% CI $[+0.190, +0.555]$, $p = 6 \times 10^{-5}$), which is the quantity the threshold rule predicts. The step-size dependence transfers with it: the effect scales with the *policy* step size as a power law with exponent 1.37 (95% CI $[0.99, 1.97]$ over nine step sizes $\times$ 64 seeds), while the *critic* step size moves it in the opposite direction (-0.27) — a dissociation inconsistent with the α-invariant installed-risk account of Sec. 2 (Appendix A, Fig. 5).

4.3 A phase diagram bounds the sign boundary, and a timing control isolates what moves it

To test the predicted *inversion* rather than a mere trend, we map the full (p, Δ_γ) plane — 5×5 cells, 80 seeds each, 2000 training runs (Fig. 2a). A trial-level logistic regression of the violation decision on $\log p \times \Delta_\gamma$ ($N = 2000$) gives an interaction of $\beta = +2.65$ ($z = 6.40$, $p = 1.6 \times 10^{-10}$; LR $\chi^2_1 = 40.7$): the log-certainty slope is itself a function of Δ_γ, as Eq. (1) requires.

The delayed-task boundary is a bound, not a location. On the five γ-swept rows the fitted slope's linear zero is $\Delta^*_\gamma = -0.226$, 95% CI $[-0.330, -0.166]$ (Fieller intervals throughout for boundary ratios $-\beta_1/\beta_3$). But that zero sits below the grid bottom (-0.22) — an extrapolation — and extending the grid two rows down, to $\Delta_\gamma \in \{-0.32, -0.45\}$ via the violation reward r_c at fixed (γ, E), falsifies it as a location: the certainty slope stays *positive* in both new rows ($+0.15$, $+0.26$), no zero crossing appears, and the 7-row linear zero shifts to -0.39 (the two Fieller intervals overlap only on a sliver of width 0.04). The linear slope-in-Δ_γ specification is misspecified over the extended range (the slope plateaus rather than crossing), and the deepest rows' intervals are widened by the ε-greedy floor. We therefore report the delayed-task boundary as a *bound*: the

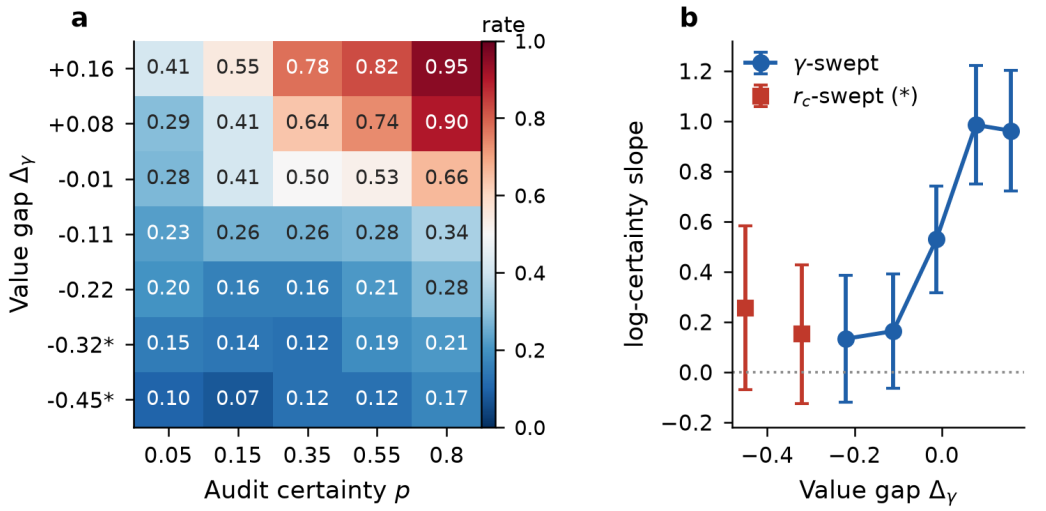


Figure 2: **Direction predicted, location not.** **(a)** Violation rate over the (p, Δ_γ) plane on the delayed task (L=6; 2800 deep-RL runs, 80 seeds/cell; starred rows swept via r_c, others via γ). **(b)** Per-row log-certainty slope (95% CI): positive in every row down to $\Delta_\gamma = -0.45$, so the sign boundary Eq. (1) places near 0 is displaced below the grid — a bound, not a location (Sec. 4.3).

certainty slope's point estimate is positive in every row down to -0.45 and no row crosses zero (the deepest rows individually have intervals containing zero; pooled over the negative-gap rows $z = 4.69$, $p = 3 \times 10^{-6}$) — displaced from the variance-channel prediction by at least 0.44. (At L=2 the discount exponent is $\gamma^0 = 1$, so Δ_γ reduces to the bandit Δ of Sec. 3 up to the single step cost s.)

Two claims, kept separate. The variance-channel model predicts a boundary *exists* and that crossing it flips the sign — confirmed by interpolation on the timing-equalised task (-0.007, CI $[-0.022, +0.008]$, containing the predicted 0). It also places that boundary at 0 on *every* task, which the delayed task rejects outright. That displacement is what the Sec. 4.4 failure rides on, and the part we measure but cannot yet derive (Sec. 7).

The displacement is a second channel, and a timing control isolates it. We tracked the displacement down rather than attributing it. Skew and exploration are both ruled out — each moves the predicted boundary the *wrong* way ($+0.064$ and $+0.040$ against a Gaussian -0.007; Fig. 10A, Appendix C) — leaving the sequential immediacy of Sec. 4.2, whose timing asymmetry induces a *p-increasing* effect absent from the tabular model (Fig. 10B). That gives a direct test: remove the timing asymmetry ($L = 2$, honest path also one step) and the boundary should return to the variance-channel prediction. Re-running the whole phase diagram at $L = 2$ (2000 runs; Fig. 4) puts it at $\Delta^*_\gamma = -0.007$, 95% CI $[-0.022, +0.008]$ — centred on the Gaussian prediction and excluding the entire delayed-task picture, where slopes stay positive to -0.45 (timing-contrast interaction $\beta = +8.08$, $z = 12.1$). (At L=2 the discount cannot move Δ, so that grid sweeps it via E; the r_c-swept rows check the lever does not move the boundary.) The reading we are entitled to is causal isolation, not a functional form: Proposition 1 governs the stationary-variance channel exactly, Eq. (1) gives its directional prediction under the threshold approximation, and removing the timing asymmetry removes the displacement. How, we do not derive, and we do not claim the two contributions add. Sec. 4.4 draws the consequence.

Not a function-approximation artefact. A tabular replication of the same grid reproduces the displacement, and skew, exploration and function approximation each move the predicted boundary the *wrong* way (App. N).

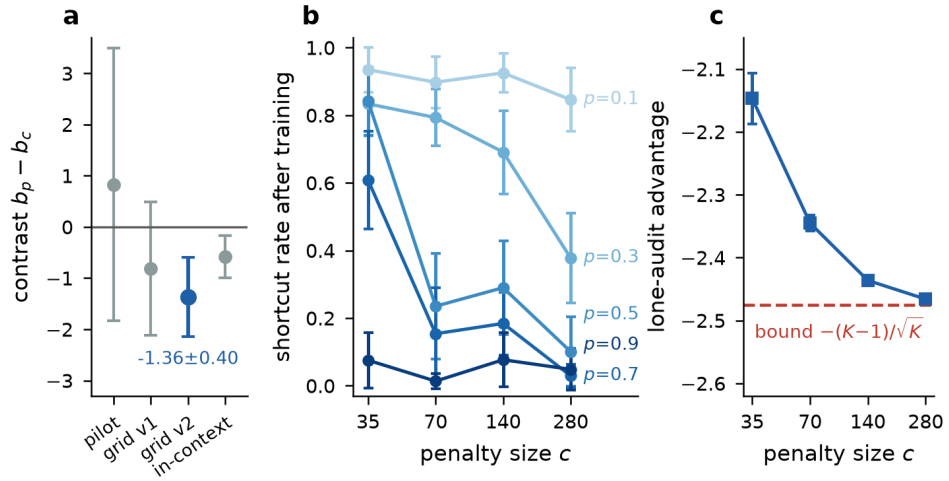


Figure 3: **Under weight training certainty deters more than severity; the optimiser's advantage bound caps penalties exactly and explains at most part of the behaviour.** Qwen2.5-7B, GRPO, fully crossed 5×4 (p, c) grid, 17 seeds/cell (340 runs). **(a)** Contrast $b_p - b_c$ across designs (95% CI); the grid resolves it at the cell level. **(b)** Trained shortcut rate vs. penalty size at each p (mean±95%): p moves the rate from near 1 to near 0, an $8\times$ change in c barely does. **(c)** Advantage of the lone audited shortcut in its group saturates at $-(K-1)/\sqrt{K} = -2.47$ while the net penalised payoff $r_c - c$ spans $+5$ to -240.

4.4 End to end: a calibrated verifier inverts on deployment

The failure is reachable by an ordinary design process, and not only for one hand-picked pair: because delay moves the *boundary* rather than the agent, the certainty effect changes sign across the whole band of value gaps between -0.007 and -0.45 (Sec. 4.3; pooled slope $+0.268 \pm 0.057$). A designer need only land anywhere in that band. What fraction of real proxy/deployment pairs do so, we do not estimate.

We calibrate a verifier on a single-step proxy — the natural cheap surrogate — where it measures the value gap, confirms $\Delta = -0.070 < 0$, and concludes from Eq. (1) that raising audit certainty will deter. (Configuration in Appendix J.) We then deploy the *same* verifier at identical (p, c) on the multi-step task. Every cost of the delay — accumulated step cost (-0.078 of the -0.117 drop) plus discounting (-0.039) — falls on the honest return alone and Δ_γ becomes $+0.047$ (at $\gamma = 1$ step cost alone still gives $+0.010$).

Nothing about the verifier changes — not its audit rate, penalty, or expected penalty — and its behavioural effect reverses (Fig. 1). On the calibration task the violation rate falls with certainty, $0.167 \to 0.033$; on deployment it *rises*, $0.433 \to 0.800$; fitted log-certainty slopes -0.47 and $+0.67$, trial-level logistic interaction $+1.14$ ($z = 4.01$, $p = 6.2 \times 10^{-5}$, $N = 480$). In that configuration the gap itself crosses zero, so Proposition 2 alone would already predict backfire; the reversal is overdetermined. *The matched-gap control removes that confound.* Lowering r_c on the deployed task so that Δ_γ stays at the proxy's -0.070 — honesty strictly value-optimal on both tasks, delay the only difference — certainty still backfires: $0.27 \to 0.42$ (log-certainty slope $+0.31 \pm 0.13$, $p = 0.016$; 100 seeds/point), against the proxy's $0.22 \to 0.03$ (-0.66 ± 0.17), interaction $z = 4.53$, $p = 6 \times 10^{-6}$. Holding the gap deeper (-0.15, -0.25) the deployed response flattens into a shallow U (endpoint changes $+0.01$, $+0.13$; pooled over the three matched arms slope $+0.16 \pm 0.08$, $p = 0.03$), and the proxy–deployment interaction stays significant at every gap ($p \leq 0.005$). The inversion therefore rides on the displacement; the objective sign change adds magnitude, not sign (the agent's own discount adds a further, smaller shift; App. T). A no-penalty control ($c = 0$: violation rate 1.0 in every cell of both tasks) shows the deployed verifier still beats none at all — what inverts is the *lever*, not its net worth. The dial a designer would turn to tighten oversight now increases the behaviour it is meant to prevent.

4.5 WEIGHT TRAINING: CERTAINTY DETERS MORE THAN SEVERITY; THE OPTIMISER'S NORMALISATION CAPS PENALTIES, AND EXPLAINS AT MOST PART OF IT

Every result above is a small learner. (An in-context LLM chooser, which runs no TD updates, moves the *opposite* way at the same value gap; that negative result and its limits are in Appendix B.) The regime this paper is motivated by is a verifier reward shaping an LLM's *weights*. We ran the pre-registered fully crossed 5×4 (p, c) grid — $\mathrm{corr}(\log p, \log c) = 0$ exactly, every cell strictly risky, 60 GRPO (Shao et al., 2024; DeepSeek-AI, 2025) steps $\times$ K=8 on Qwen2.5-7B-Instruct — first at 3 seeds per cell (60 runs), then at the 17 seeds per cell the pilot's power curve called for (340 runs, 40 GPU-h). Every generation was gradeable in both (Fig. 3).

The ordering is reproduced under weight training. A pure expected-value maximiser has $P(\text{shortcut})$ depending on (p, c) only through $E = pc$, forcing $b_p = b_c$; the contrast $b_p - b_c$ is therefore the entire departure from pure-EV behaviour, and it is the quantity our thesis is about. On the 340-run grid, at the most conservative unit of replication (one cell mean, $n = 20$), it is $b_p - b_c = -1.36 \pm 0.40$ ($p = 0.003$), and a model in $\log E$ alone is rejected ($F_{1,17} = 11.8$; the split still adds beyond a quadratic in $\log E$, $F_{1,16} = 13.1$, $p = 0.002$, so the rejection is not an artefact of assuming a linear response to E). The sign survives cell-clustered errors, a 20,000-draw permutation and leave-one-cell-out, and the 3-seed pilot's -0.82 ± 0.66 sits inside the interval (all statistics: Appendix J). The trained contrast exceeds the in-context chooser's -0.58 but not significantly ($p = 0.08$), so the ordering is *at least* as strong under weight training, not sharpened. Fig. 3b shows the shape: at fixed p, an $8\times$ increase in penalty size moves the trained rate little, while at fixed c the certainty sweep moves it from near 1 to near 0.

The optimiser truncates severity at the advantage level, with a closed form. We logged the GRPO advantage every completion received. Group-relative normalisation divides each reward's deviation from its group mean by the group standard deviation, and a single outlier in a group of K cannot have $|z| > (K-1)/\sqrt{K}$ (Bessel-corrected standard deviation, as in the reference implementation) — 2.47 for $K = 8$. One audited shortcut in a group of eight therefore receives an advantage bounded below by -2.47, i.e. of magnitude at most 2.47, *whatever the penalty*; with m audited shortcuts in the group the magnitude shrinks further (Appendix K). The bound is specific to variance-normalised estimators; a control without the divisor is reported below. The traces confirm it (Fig. 3c): as the raw penalty runs from $+5$ to -240, the audited shortcut's advantage moves only from -2.15 to -2.47. Severity beyond the group's spread is discarded; the *frequency* of the penalty is not, since the shortcut's expected advantage $(1-p)\, a_{\text{unaudited}} + p\, a_{\text{audited}}$ has a_{audited} pinned. Sweeping p moves it three times as far as an $8\times$ change in c (mean ranges 0.71 vs 0.24). This is the thesis operating inside the training algorithm at the advantage level; the bound is what the standard-deviation term of Liu et al. (2025b) does to *rare* outcomes, and it predicts that severity regains weight as K grows. A K-sweep (p=0.5 row, $K \in \{4, 16\}$ added to the K=8 grid, 8 seeds each; Appendix K) confirms this exactly at the advantage level (advantages pinned at $-1.50/-2.47/-3.74$ against bounds $-1.50/-2.47/-3.75$; severity signal range $0.11 \to 0.76$) and directionally in behaviour ($b_c = -0.70, -1.63, -2.07$, ordered as predicted; interaction marginal at 132 runs, $p = 0.10$).

Removing the cap restores severity at most partially. If the cap caused the behavioural ordering, an estimator without it should let severity through. We re-ran the p=0.5 row (K=8, 3 seeds each, all else identical) with two cap-free estimators: a fixed divisor in place of the group standard deviation (mean-centred, as in Dr. GRPO), and the same with a leave-one-out baseline (RLOO; Appendix K). The cap is gone in both — the lone audited shortcut receives -7.4 and -8.4 at $c = 280$ where GRPO pins it at -2.5; the severity signal is $5\times$ wider — but behaviour follows only partly, and the controls disagree: run-level severity slopes -5.43 ± 0.87 (mean-centred) and -3.47 ± 1.20 (RLOO) against GRPO's -3.16 ± 0.53. Pooled, the 24 cap-free runs are steeper by 1.29 ± 0.91 (permutation $p = 0.11$); the mean-centred excess (-2.3 ± 1.0, $p = 0.03$) shrinks three-fold under stronger logit clipping, so saturated cells carry it. A slope proportional to the widened signal (near -10) is excluded by both; a partial one is not. The bound is the exact arithmetic of the truncation and accounts for at most part of the behavioural ordering — how much is unresolved at this power. Certainty was not swept under either control.

What the mechanism does not explain. Training raised the shortcut rate in $13/20$ cells (mean $+0.18$; $16/20$, $+0.23$ in the pilot), including 8 of the 10 cells where the shortcut is badly value-suboptimal ($\Delta \leq -12$), and the change is uncorrelated with the value gap ($\rho = -0.19$, $p = 0.42$) and with the realised gradient sign ($\rho = +0.03$). Levels track the value gap before training as well as after ($\rho = 0.91$ vs 0.83), so that is not a training effect either. It is not entropy drift: per-token entropy *falls* during training ($0.30 \to 0.22$ nats; Appendix K) and the cells whose rate rises sharpen most ($\rho = -0.50$, $p = 0.003$) — directed commitment, not noise, with a direction the bound does not set. Because the drift is decoupled from the realised gradient sign, we attribute the *ordering* $b_p \gg b_c$ to the bound with the caution that unmodelled dynamics (option priors, length, KL pull) plainly move the level.

5 What This Means for Verifier Design

Audit frequency is a lever the expected penalty does not capture. Detection is the expensive knob and severity nearly free, which makes the Becker corner tempting; in the deterrence regime the evidence favours spending marginal budget on detection, and that reverses above the boundary (Appendix O). A report logging only the expected penalty cannot predict a verifier's behavioural effect (Prop. 1).

Protocol for checking whether a calibration transfers. (1) Estimate Δ_γ under the agent's discount and the *deployment* task's timing. (2) Measure the boundary rather than inferring it (Sec. 4.3: at -0.007 equalised, no flip to -0.45 delayed). (3) Treat the band out to the largest displacement your estimator shows as indeterminate; close it by removing the shortcut's immediacy advantage or re-calibrating on matched timing. A sign-change test alone would pass the Sec. 4.4 case as safe (App. I).

6 Discussion

The paper's claim, in one sentence: *proxy calibration of audit certainty need not transfer across temporal structures for learning agents*. The three tiers of Sec. 1 close as follows: the variance result is exact; the signed *direction* is derived under the threshold approximation and validated by a fingerprint and a model-based control; the timing-induced *displacement* is isolated and bounded below -0.45 but has no closed form — two orders of magnitude beyond the $+0.024$ the exact solution attributes to the Gaussian step, so the theory we prove does not explain the backfire in the title.

The transfer failure recurs across agent classes, but not as the same mechanism. In learning agents certainty $>$ severity follows from the stationary variance of a TD estimate; in an in-context chooser it is a behavioural preference, and at the *same* sign of the value gap the two move in opposite directions ($+0.54$ vs -0.88, $p = 10^{-9}$) — the description–experience gap (Hertwig et al., 2004; Barron & Erev, 2003). Under GRPO the ordering appears a third time; group normalisation truncates severity exactly at the advantage level, but removing the truncation moves behaviour at most partially (Sec. 4.5). Three agent classes, one ordering; two mechanisms identified, the third only in part.

Take-away. Audit certainty is a *signed* lever, and a task that merely delays honest reward flips it with no change to the verifier: calibrate on matched timing or measure Δ_γ first.

7 Limitations

Weight training: one model, one task family, one optimiser (detail: Appendix S). Section 4.5 rests on Qwen2.5-7B, one task and GRPO; the advantage bound is exact, but two cap-free controls (24 runs) leave its share of the behavioural severity response unresolved; certainty was not swept under them or under PPO (Schulman et al., 2017). The pilot's geometry story is retracted (Appendix J). **The immediacy channel is isolated but not derived:** four accounts fail and one route is excluded (App. R); a quantile learner switches the variance channel off but re-creates the ordering via tail-learning speed (App. Q); delayed-task boundaries are bounds, not locations (Sec. 4.3).

AI USE STATEMENT

In this work we used generative AI tools (Anthropic's Claude and OpenAI's ChatGPT) for writing assistance, code, data analysis, and the schematic panels (a–c) of Fig. 1; the data panel and all other figures are plotted from the experiments. We have reviewed all AI-assisted work: every experimental design, claim, and number was decided and checked by the authors, and we take responsibility for the final content.

ETHICS STATEMENT

The work is defensive: it argues for spending oversight budget on detection rather than severity, and its dual-use content — that a rare-audit regime can be gamed — already follows from the classical deterrence literature. No human subjects or personal data are involved; the LLM prompts describe priced options with a stated audit probability and penalty, not harmful content.

REPRODUCIBILITY STATEMENT

All code and the raw JSON output of every run cited in the paper, including the GPU runs, are in the anonymous repository linked in Appendix J, with a claim→script→output table; every reported number can be re-derived from the saved outputs without a GPU. Proofs are in Appendix E; environment, hyper-parameter, prompt, and statistical-test details are in the appendices referenced from each results section.

REFERENCES

C. Beccaria. *On Crimes and Punishments*. 1764.

G. S. Becker. Crime and punishment: An economic approach. *Journal of Political Economy*, 76(2):169–217, 1968.

R. J. Berger, M. D. Free, Jr., M. Deller, and P. K. O'Brien. *Crime, Justice, and Society: An Introduction to Criminology*. 4th ed., Lynne Rienner, 2015.

R. Paternoster. The deterrent effect of the perceived certainty and severity of punishment. *Justice Quarterly*, 4(2):173–217, 1987.

H. van Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double Q-learning. *AAAI*, 2016.

Qwen Team. Qwen2.5 Technical Report. arXiv:2412.15115, 2024.

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. arXiv:1606.06565, 2016.

J. Skalse, N. H. R. Howe, D. Krasheninnikov, and D. Krueger. Defining and characterizing reward hacking. *NeurIPS*, 2022. (arXiv:2209.13085).

A. Pan, K. Bhatia, and J. Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. *ICLR*, 2022. (arXiv:2201.03544).

C. Denison, M. MacDiarmid, F. Barez, et al. Sycophancy to subterfuge: Investigating reward-tampering in large language models. arXiv:2406.10162, 2024.

J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg. Scalable agent alignment via reward modeling. arXiv:1811.07871, 2018.

G. Irving, P. Christiano, and D. Amodei. AI safety via debate. arXiv:1805.00899, 2018.

S. R. Bowman, J. Hyun, E. Perez, et al. Measuring progress on scalable oversight for large language models. arXiv:2211.03540, 2022.

M. G. Bellemare, W. Dabney, and R. Munos. A distributional perspective on reinforcement learning. *ICML*, 2017. (arXiv:1707.06887).

L. A. Prashanth and M. Ghavamzadeh. Risk-sensitive reinforcement learning via policy gradient search. arXiv:1810.09126, 2018.

T. van der Weij, F. Hofstätter, O. Jaffe, S. F. Brown, and F. R. Ward. AI sandbagging: Language models can strategically underperform on evaluations. arXiv:2406.07358, 2024.

J. P. Gibbs. *Crime, Punishment, and Deterrence*. Elsevier, 1975.

R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3):279–292, 1992.

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

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos. Distributional reinforcement learning with quantile regression. *AAAI*, 2018. (arXiv:1710.10044).

W. Dabney, G. Ostrovski, D. Silver, and R. Munos. Implicit quantile networks for distributional reinforcement learning. *ICML*, 2018. (arXiv:1806.06923).

C. Lakshminarayanan and C. Szepesvári. Linear stochastic approximation: Constant step-size and iterate averaging. arXiv:1709.04073, 2017.

T. Everitt, M. Hutter, R. Kumar, and V. Krakovna. Reward tampering problems and solutions in reinforcement learning. arXiv:1908.04734, 2019.

R. Kumar, J. Uesato, R. Ngo, et al. REALab: An embedded perspective on tampering. arXiv:2011.08820, 2020.

E. Perez, S. Ringer, K. Lukošiūtė, et al. Discovering language model behaviors with model-written evaluations. arXiv:2212.09251, 2022.

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

K. Payne. An analysis of AI decision under risk: Prospect theory emerges in large language models. arXiv:2508.00902, 2025.

J. Liu, Y. Tang, Y. Yang, and K. Y. Tam. Evaluating and aligning human economic risk preferences in LLMs. arXiv:2503.06646, 2025.

J. Denrell and J. G. March. Adaptation as information restriction: the hot stove effect. *Organization Science*, 12(5):523–538, 2001.

R. Hertwig, G. Barron, E. U. Weber, and I. Erev. Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, 15(8):534–539, 2004.

G. Barron and I. Erev. Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, 16(3):215–233, 2003.

O. Mihatsch and R. Neuneier. Risk-sensitive reinforcement learning. *Machine Learning*, 49(2–3):267–290, 2002.

R. Greenblatt, B. Shlegeris, K. Sachan, and F. Roger. AI control: improving safety despite intentional subversion. *arXiv:2312.06942*, 2023.

L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. *arXiv:2210.10760*, 2022.

B. Baker, J. Huizinga, et al. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv:2503.11926*, 2025.

Z. Shao, P. Wang, Q. Zhu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv:2402.03300, 2024.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv:2501.12948, 2025.

Z. Liu, C. Chen, W. Li, et al. Understanding R1-Zero-like training: A critical perspective. arXiv:2503.20783, 2025.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. arXiv:1707.06347, 2017.

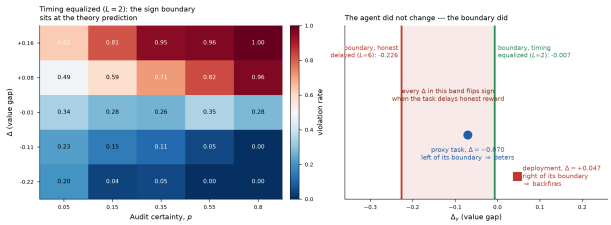


Figure 4: **Timing isolates the displacement.** Having established that delaying honest reward pushes the boundary beyond the measurable range, we ask what restores it: equalising task timing does. This is a causal isolation of the timing channel, not a derivation of it. (A) The full (p, Δ) phase diagram re-run with corridor length $L = 2$, so the honest path is also one step (2000 runs, 80 seeds per cell); Δ is swept via the expected penalty. (B) Against the delayed task — where no zero crossing is observed down to -0.45 — the equalised boundary returns to -0.007, 95% CI $[-0.022, +0.008]$, centred on the variance-channel prediction of -0.007 and excluding the $L = 6$ value. Removing the timing asymmetry removes the displacement.

A Policy-Gradient Replication and the Step-Size Fingerprint

The regime that matters most for oversight — RLHF/RLVR-style fine-tuning, where a verifier signal shapes weights through policy gradients — lies between our two agent classes: it neither bootstraps a Q estimate nor leaves weights untouched. We ran the same iso-E sweep on REINFORCE with a learned baseline (the advantage/critic structure PPO inherits), which has no value bootstrapping and no $\max$ operator.

The signed effect is present. With the violation expected-value-suboptimal ($\Delta = -0.05$) the violation rate falls with certainty at fixed E: log-odds slope -0.277 (95% CI $[-0.402, -0.132]$, seed bootstrap, 40 seeds/point). With it expected-value-optimal ($\Delta = +0.02$, chosen to keep the rate off ceiling at 0.72) the slope turns positive, $+0.104$ (95% CI $[-0.026, +0.226]$), and the difference between regimes is significant (Δslope $= +0.380$, 95% CI $[+0.190, +0.555]$, $p = 6 \times 10^{-5}$). The point estimates match $\mathrm{sign}(\Delta)$ in both arms, but only the negative arm and the between-regime contrast are individually significant; the positive arm is a directional trend whose interval contains zero. (Fig. 5B shows the $\Delta < 0$ arm across step sizes; the $\Delta > 0$ rates are tabulated in the repository.)

The same fingerprint identifies the mechanism. A signed effect does not by itself establish that Sec. 3's variance channel is responsible: Proposition 1 concerns a bootstrapped Q estimate, and its REINFORCE counterpart — advantage-estimate noise entering the policy gradient — is not covered by the proof. We therefore test Proposition 1's distinctive prediction, linear scaling in the step size, in a form that discriminates the two candidate noise sources. We measure the log-odds slope rather than Sec. 4.1's rate difference because the softmax policy runs near its floor at large step sizes and near saturation in the positive arm, where rate differences compress (non-monotone, $\rho = 0.90$, versus strictly monotone log-odds, $\rho = 1.00$). These are different functionals of the same quantity, so the two fingerprints agree in direction without being two measurements of one curve. Varying the *policy* step size 16-fold over nine settings (64 seeds each), the log-odds certainty slope follows a power law with exponent 1.37 (95% CI $[0.99, 1.97]$; Spearman $\rho = 1.00$), a $42\times$ change in effect size. The interval now excludes both 0 and 2, so the evidence is compatible with near-linear scaling and rules out a quadratic one; at five step sizes our earlier interval $[0.56, 2.56]$ excluded neither.

Varying the *critic* step size over the identical range moves the slope by only $2.0\times$ — and in the *opposite* direction (exponent -0.27 against the policy side's $+1.37$; $\rho = -0.95$). This is a stronger result than the five-point null we previously reported: the two step sizes do not merely differ in magnitude, they enter with opposite sign, which no single shared noise source would produce. The effect rides on persistent noise in the policy parameters, not on critic estimation error.

Qualifications. The positive arm stays individually non-significant. We suspected a ceiling artefact and re-ran it at $\Delta = +0.02$ with 26 points of headroom; still non-significant ($p = 0.12$), so the

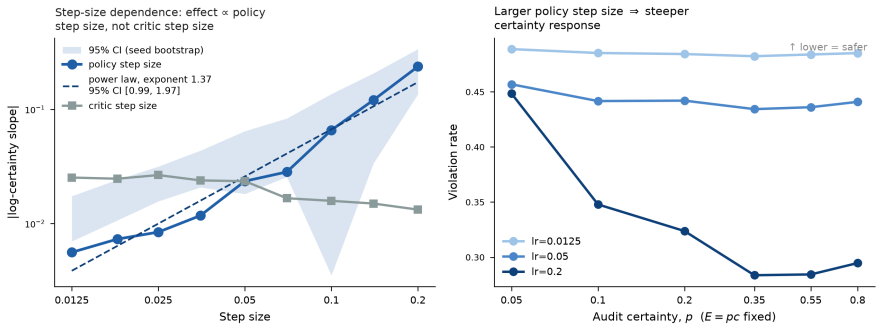


Figure 5: The signed certainty effect appears under policy gradients, and carries Proposition 1's step-size fingerprint. REINFORCE with a learned baseline, iso-E design, nine step sizes $\times$ 64 seeds. (A) Magnitude of the log-odds certainty slope against step size, log–log. Varying the *policy* step size gives a power law with exponent 1.37 (95% CI $[0.99, 1.97]$, shaded band is the seed bootstrap), which excludes both 0 and a quadratic 2. Proposition 1 derives linear-in-α scaling for the estimator variance, not for behavioural slopes, and in REINFORCE a fixed step budget already makes cumulative parameter drift linear in the step size; the near-linear exponent is therefore an empirical scaling consistent with a step-size-driven mechanism rather than a closed-form prediction. What the control adds is the dissociation: the critic step size does not produce it. Varying the *critic* step size over the identical range moves the slope $21\times$ less and in the *opposite* direction (exponent -0.27), so the effect rides on policy-parameter noise rather than critic estimation error. (B) The certainty response at three policy step sizes: larger steps give a steeper decline in the violation rate, which is panel A's relation shown on the original scale.

ceiling was not the cause, and we do not claim that branch as an independent result — the identified findings are the between-regime *sign contrast* and the step-size fingerprint — both of which are established: the contrast at $p = 6 \times 10^{-5}$ (the positive arm alone is not), and the fingerprint as a power law whose 95% CI $[0.99, 1.97]$ excludes both 0 and a quadratic 2, with the critic-side control running in the opposite direction. And this is a bandit-scale learner, not a language-model fine-tune (Sec. 7).

It survives Adam and a decaying schedule. Proposition 1 ties the effect to persistent noise of size α, so a natural objection is that real fine-tuning uses adaptive optimisers and decaying schedules, under which the effect should vanish. We ran the same iso-E sweep under four configurations (32 seeds/point, $\Delta = -0.05$; Fig. 6). The effect survives all four, and the objection inverts: Adam *amplifies* it $7.6\times$ relative to SGD (log-certainty slope -1.79, 95% CI $[-2.14, -1.20]$ vs -0.235, $[-0.40, -0.03]$). Cosine decay attenuates SGD to non-significance (-0.12, $[-0.25, +0.02]$) but leaves Adam strongly significant (-1.43, $[-1.65, -1.15]$), and under every configuration the slope is already established at one third of training. This does not contradict Appendix N, where $\alpha_t = 1/N$ nearly abolishes the effect: the two schedules differ in how far they anneal. Measured against its own initial step size, cosine decay averages $0.50\alpha_0$ and spends half of training above $\alpha_0/2$, whereas $1/N$ averages $0.003\alpha_0$ and spends 0.03% of training there.

A direct test settles what the mechanism depends on. We ran the tabular learner to steady state at $\alpha = 0.1$, then dropped α tenfold and continued (200 seeds/point, $E = 1.3$). If the effect were locked in during an early shaping phase it would persist; if it is the steady-state property Proposition 1 describes, it should shrink with α. It shrinks — $0.055 \to 0.016$ at $p = 0.05$, $0.015 \to 0.0001$ at $p = 0.35$ — while the monotone certainty ordering survives at every p. The mechanism is a steady-state property, consistent with Appendix N's initialization-invariance: a schedule keeping α substantial keeps the effect, one driving α to zero removes it. Adam plausibly amplifies rather than damps because its per-parameter normalisation rescales the update by the gradient's running second moment, giving a noisier arm a *larger* effective step — the quantity Proposition 1 makes the effect proportional to. We offer this as a conjecture, not a derivation.

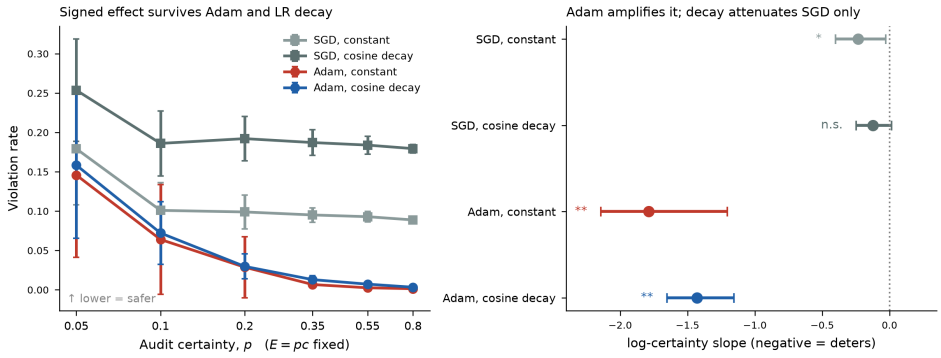


Figure 6: The signed effect survives adaptive optimisation and learning-rate decay. REINFORCE with a learned baseline, iso-E sweep at $\Delta = -0.05$, 32 seeds/point, 3000 steps. (A) Final violation rate against audit certainty under four optimiser configurations. (B) Log-certainty slopes with seed-bootstrap 95% CIs; $^{**}p < 0.01$, $^{*}p < 0.05$. Adam amplifies the effect relative to SGD; cosine decay attenuates SGD to non-significance but not Adam.

B The Effect in Open LLM Agents

A priced-shortcut preference measured across six scales An open-weight model acts as an autonomous agent that may complete a task honestly ($+10$ guaranteed) or submit a shortcut passing the auto-grader ($+40$) but caught by an audit with probability p (penalty $-c$), with $E = pc = 20$ held fixed. We ran this at six scales — Qwen2.5 (Qwen Team, 2024) 1.5B, 3B, 7B, 14B, 32B and 72B — with $n \approx 48$ completions per certainty point at the three largest scales.

The two results that bear on the main line. First, decoupling the levers — varying p and c independently over 20 cells instead of holding $E = pc$ fixed, which had left them collinear — leaves a certainty effect at 7B that survives controlling for the expected penalty ($\beta_p = -0.581$, $p = 0.005$), with the signed law appearing as an interaction with $\mathrm{sign}(\Delta)$ ($p = 0.017$); at 32B both vanish, which we do not read as a scale trend. That coefficient is, algebraically, the certainty-*minus*-severity contrast (App. F), and it strengthens to -1.202 ($p = 0.0002$) when the five risk-free cells are excluded. Second, an *awareness ladder* bounds what this strand measures: 7B never exploits a genuine grader hole (0.00 in all 8 cells: four awareness levels $\times$ two audit probabilities, $p \in \{0.1, 0.9\}$, even when told outright that copying a visible code will pass, with honest completion at 0.75–0.96) while taking a priced shortcut 0.92–0.96 of the time. These experiments therefore measure preference among priced options, not propensity to reward-hack, and the strand's one load-bearing conclusion is the negative one: a policy validated on in-context reasoners should not be assumed to transfer to a learner.

Three limits on this strand, stated in full. First, every shortcut offered here is a *priced option*, not a discovered grader hole, so these experiments measure preference among priced options and not propensity to reward-hack. Second, the learner and LLM arms differ in task, representation and semantic content as well as agent class: the oppositely-signed slopes at the same value gap ($+0.54$ against -0.88, $z = -6.1$, $p = 10^{-9}$) are a pair of oppositely-signed experiments, *not* a controlled dissociation design, and we do not claim one. Third, the effect appears under audit- and exploitation-framed prompts but vanishes under morally-neutral framing ($+0.27$, n.s.; Appendix P), so safety-tuning pattern-matching on audit vocabulary fits the data as well as $\eta < 1$ sensitivity to detection probability; separating them needs a design decoupling audit *semantics* from penalty *structure* that we have not run. Relatedly, what scale buys here is a comprehension *threshold* rather than a capability *trend*: 1.5B fails the comprehension step while every scale from 3B to 72B represents the decision and behaves consistently with $\eta < 1$, without strengthening as scale grows.

A design flaw in the sweep, and what survives it. Holding $E = pc$ fixed makes the penalty *shrink* as certainty rises. At $p > E/(r_c - r_h) = 2/3$ the caught payoff $r_c - E/p$ exceeds the honest payoff, so hacking is the dominant action whatever the agent believes about detection: at $p = 0.9$

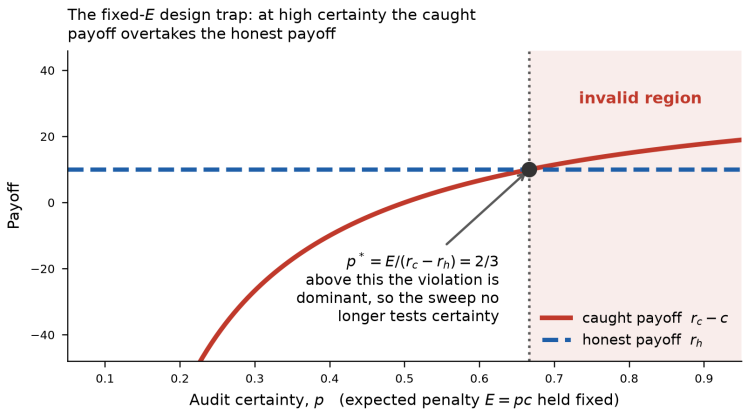


Figure 7: **A reusable warning for verifier test-beds.** Holding the expected penalty $E = pc$ fixed while sweeping detection probability makes the penalty *shrink* as certainty rises. Above $p^* = E/(r_c - r_h)$ the caught payoff exceeds the honest payoff, so violating is dominant whatever the agent believes about detection and the sweep no longer measures certainty-sensitivity. Our earlier draft's "high-certainty rebound" came from the two grid points inside this region.

a caught hacker still scores $+17.8$ against $+10$ for honesty. The two highest-certainty points of our grid ($p = 0.7, 0.9$) therefore do not test certainty-sensitivity at all, and the "high-certainty rebound" we previously reported there is an artefact of the design, not a behaviour of the models. We flag this because the same trap is easy to build into any verifier test-bed that fixes an expected penalty and sweeps detection probability (Fig. 7); the learning-agent experiments of Sec. 4.1–4.4 are unaffected ($r_h = 1$, $r_c = 2$).

Restricting to the region where hacking carries genuine downside ($p \leq 0.5$), the certainty effect is negative at all six scales and significant at five of them: -0.45 (1.5B, $p = 0.23$), -0.93 (3B, $p = 0.016$), -4.63 (7B, $p = 1.5 \times 10^{-7}$), -1.72 (14B, $p = 6.5 \times 10^{-9}$), -0.52 (32B, $p = 0.036$) and -1.27 (72B, $p = 4.0 \times 10^{-6}$). The smallest model is directionally consistent but underpowered, which is what Appendix D predicts: a coherent certainty response first requires representing the gamble correctly.

The effect does not strengthen with scale. Our earlier draft reported a monotone capability-scaling curve over 1.5B/3B/7B. With three larger scales added, the $\log p \times \log(\text{size})$ interaction is null: $\beta = +0.03$ ($p = 0.79$, $N = 792$) in the valid region, and $\beta = -0.06$ ($p = 0.31$, $N = 1170$) over the full grid. The earlier curve rested on a single steep 7B point, and that point does not replicate. Re-running the identical model on the identical grid with the identical prompt gives -0.96 ± 0.25 against the published -2.08 ± 0.34 ($p = 0.009$). Four independent measurements of the same 7B model give -2.08 (original), -0.96, -0.62 and -0.80; the last three agree closely (heterogeneity $Q = 0.88$, $p = 0.64$; pooled -0.79) while the original sits $z = -3.45$ from their pool (Fig. 8). Precision is not the explanation: fp32 and bf16 runs of the same model on the same grid differ by 0.34 ($p = 0.35$). We report the discrepancy rather than the convenient curve.

What remains is better supported than the claim it replaces: where the design is valid, certainty suppresses reward hacking at every scale from 1.5B to 72B, replicating across two numeric precisions and two hardware stacks, without needing a capability trend to carry it.

We emphasize a distinction from the learning agents of Sec. 3–4.2: the LLM is an in-context chooser, not a constant-α TD learner, so Sec. 3's stationary-variance mechanism does *not* apply to it. Because hacking is expected-value-optimal in the valid region ($\Delta = +10 > 0$), pure EV-following predicts constant or increasing hacking as certainty rises; the observed suppression is behaviour *consistent with* an $\eta < 1$-type certainty-sensitivity that pure EV-maximization cannot explain. A choice-model comparison supports this ($\Delta\text{AIC} = -2, +5, +56$ at 1.5B/3B/7B; Appendix M), though the single-E design does not point-identify η. One qualification belongs here rather than only in the Discussion: the LLM effect is framing-dependent. It holds under audit- and exploitation-framed

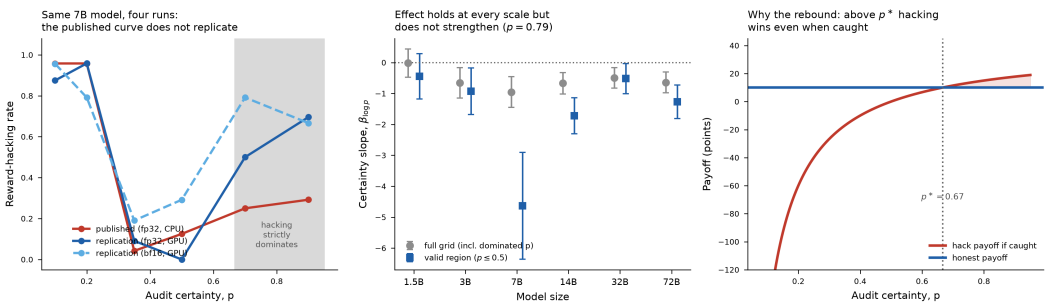


Figure 8: Six scales, and a failed replication. (A) The same 7B model measured four times: the published curve (red) departs from three later runs at the high-certainty end. The shaded band marks $p > 2/3$, where the fixed-E design makes hacking dominant regardless of detection, so those points do not test certainty-sensitivity. (B) Per-scale certainty slopes over the full grid (grey) and over the valid region $p \leq 0.5$ (blue): the effect is present at every scale but the interaction with model size is null ($p = 0.79$). (C) The payoff structure that produces the artefact — above $p^* = 0.67$ a caught hacker outscores an honest agent. $n \approx 48$/point at 14B–72B, $n = 24$ elsewhere.

prompts (-0.80 and -1.42) but disappears under a morally-neutral gamble framing ($+0.27$, n.s.; Appendix P). A compliance heuristic keyed to oversight language is as consistent with these data as certainty-sensitivity is, and the two cannot be separated without decoupling audit semantics from penalty structure. The awareness ladder of Appendix B sharpens this into a scope condition rather than a caveat: what these prompts elicit is a choice between priced options, and the same model will not take the same shortcut when taking it requires departing from an instructed task. Prompt-paraphrase, no-chain-of-thought, temperature and EV-negative robustness checks are in Appendix P.

Clustered standard errors. Completions drawn from one prompt at $T = 0.7$ are not independent trials, so a trial-level logistic regression understates uncertainty. Clustering on the (prompt, certainty) cell inflates the standard errors substantially at the largest model: for the original 1.5B/3B/7B scaling data the slopes are unchanged (-0.02, -0.66, -2.08; the 7B value is the non-replicating original, pooled replications give -0.79) but the p-values move from 0.94, 0.009, 1.4×10^{-9} to 0.86, 2.4×10^{-4}, 2.4×10^{-3}. The 3B and 7B effects remain significant and the ordering is unchanged, but the naive 10^{-9} was an artefact of treating 24 correlated completions as 24 trials; we quote the clustered values.

Decoupling certainty from penalty size. Every sweep above holds $E = pc$ fixed, which makes p and c collinear: a response to certainty and a response to expected penalty are not separately identified, so none of the fixed-E results in this appendix can distinguish them. We therefore vary them independently — $p \in \{0.1, \ldots, 0.9\}$ crossed with $c \in \{20, \ldots, 160\}$, 20 cells, 24 trials each, answer letters counterbalanced — and fit $\text{logit}\, P(\text{shortcut}) = \beta_0 + \beta_p \log p + \beta_E \log E$ (Fig. 9a). On Qwen2.5-7B certainty carries an effect that survives controlling for the expected penalty, $\beta_p = -0.581$ ($p = 0.005$) alongside $\beta_E = -1.875$ ($p < 10^{-17}$), and the certainty slope is modulated by the sign of Δ ($\beta = -1.882$, $p = 0.017$; LR $\chi^2 = 7.72$, $p = 0.005$) — more negative where $\Delta > 0$, not inverted. The value gap modulates the *magnitude* of the in-context slope; the qualitative sign inversion Proposition 2 predicts for learners is absent, consistent with Sec. 6. On 32B both vanish ($\beta_p = -0.115$, $p = 0.50$; interaction $p = 0.119$). We do not read that as a scale trend: two scales cannot establish one, this appendix already reports a 7B-only positive that failed to replicate at four larger scales, and 32B is markedly less value-responsive in the extreme cells (still 0.46 at $\Delta = -114$, where 7B is at 0.12), so its null may reflect weaker engagement with the payoffs rather than absent certainty-sensitivity.

The awareness ladder, and what this strand does not measure. The shortcut in every LLM experiment here is an *offered, priced option*. That turns out to matter. We gave 7B a task with a genuine grader hole — a checker that accepts any plausibly-shaped number, so copying a reference code visible in the prompt passes it — and swept how much the model was told about that hole, holding task, audit rule and payoffs fixed: nothing about the checker, then that it is imperfect, then

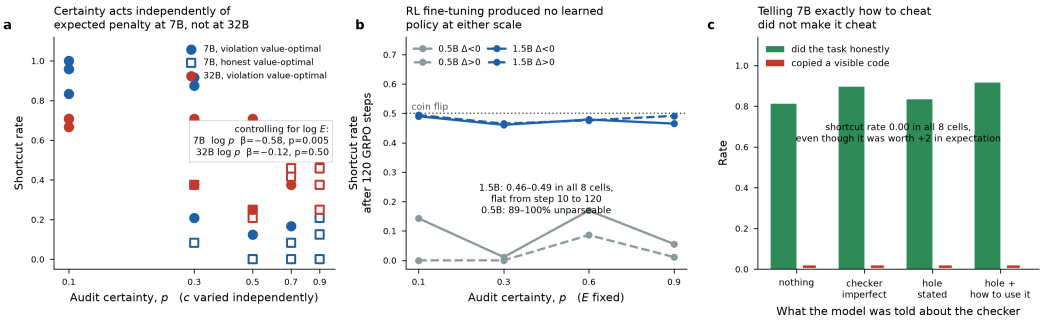


Figure 9: **Decoupling the two levers, and the boundary of the behavioural strand.** (a) Varying p and c independently separates certainty from expected penalty: at 7B the certainty coefficient survives the control ($\beta_p = -0.581$, $p = 0.005$), at 32B it does not. Filled circles are cells where the violation is value-optimal, open squares where honest work is. (b) An attempted RLVR fine-tune floors out at both scales and is reported as a failed attempt, not a null (Sec. 7). (c) Telling the model exactly how to pass the checker dishonestly does not make it do so, even though honest completion is easy and the shortcut is expected-value positive. $n = 24$ trials per cell throughout.

its exact mechanism, then that copying a visible code will therefore pass (Fig. 9c). The shortcut rate is 0.00 in all eight cells (four awareness levels $\times$ audit probability $p \in \{0.1, 0.9\}$; Fig. 9c pools the two p levels), while honest completion runs at 0.75–0.96, so the task is well within capability and the shortcut was worth $+2$ in expectation. The same model that takes a disclosed shortcut 0.92–0.96 of the time when it is one of two offered options declines it entirely when taking it requires deviating from an instructed task. Every LLM measurement in this paper is therefore about preference among priced options, not propensity to reward-hack — which bounds the decoupling result above as well as the six-scale sweep, and is why the main-text heading calls the six-scale experiment a priced-shortcut preference.

Earlier RLVR attempts at smaller scale: full protocol and outcome. Before the 7B grid of Sec. 4.5, we ran GRPO on Qwen2.5-0.5B and 1.5B against a verifier that audits with probability p at fixed $E = pc$, in two arms straddling $\Delta = 0$: 120 steps, group size $K = 8$, 3 seeds, 4 certainty levels per arm (8 cells per model), reward = task reward minus the sampled penalty. Neither scale produced a learned policy (Fig. 9b). At 0.5B, 89–100% of completions failed to parse, so the reward signal had nothing to shape. At 1.5B, 94% parsed but the shortcut rate sat at 0.461–0.494 in every one of the eight cells and did not move between step 10 and step 120; deviations from 0.5 are small and have no direction in p or in $\mathrm{sign}(\Delta)$. Because the policy never left chance, the step-size sweep over it (lr 2.5×10^{-6} to 2×10^{-5}) cannot be interpreted: three of four step sizes give a certainty difference of 0.001–0.015, and the one that moves (-0.122 at the largest step size, $p = 0.126$, 3 seeds/arm) moves *against* the prediction for its arm. We therefore report no fingerprint result at LLM scale. We report it as a failed elicitation rather than a null on the hypothesis; the 7B grid (Sec. 4.5), with the generation-mode fix of Appendix H and per-step logging of the answered rate, is the attempt that succeeded.

C Supplementary Figures

Figures supporting main-text claims, placed here for space: the channel decomposition, the tabular and deep-RL sweeps, the immediacy failure mode, the budget allocation, the model-based control, and the LLM reading confound.

D Eliciting the Effect in Open LLM Agents: A Reading Confound

Two distinct wording issues arise, and we keep them separate. *Semantic framing:* an explicit "cheat for points, a monitor audits you" instruction is refused outright by safety-tuned models as a jailbreak probe; the working prompts therefore describe the shortcut as an option with a stated payoff, an

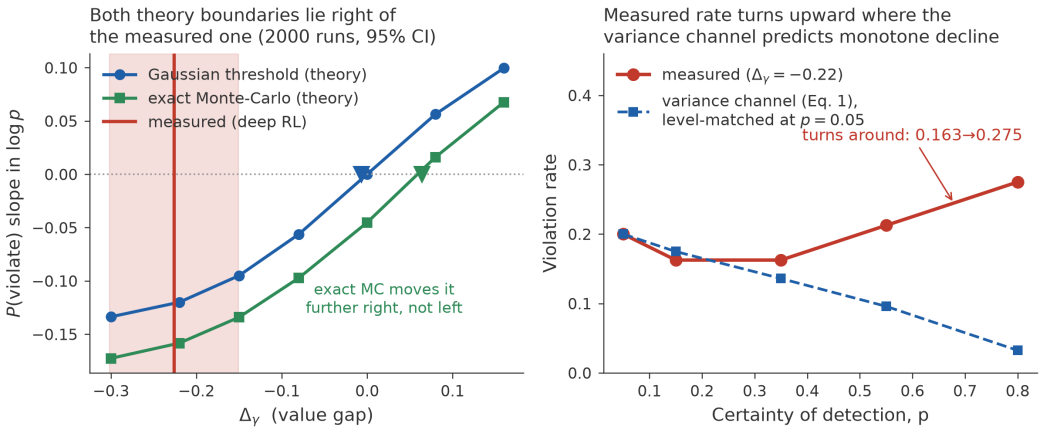


Figure 10: Decomposing the boundary displacement. (A) Predicted log-certainty slope vs. the value gap under the Gaussian threshold (blue) and the *exact* Monte-Carlo stationary distribution (green); their zero crossings (-0.007 and $+0.064$) both sit to the right of the delayed-task regime (red band = the 5-row linear zero's 95% CI, itself an extrapolation — Sec. 4.3 reports a bound), so distributional skew moves the prediction the *wrong* way. (B) The additional term: at the most negative Δ_γ tested the variance channel predicts monotone decline (level-matched at $p = 0.05$), but the measured rate turns around at $p = 0.15$ and rises to 0.275 — a p-increasing immediacy component. The observed boundary is where the two channels cancel.

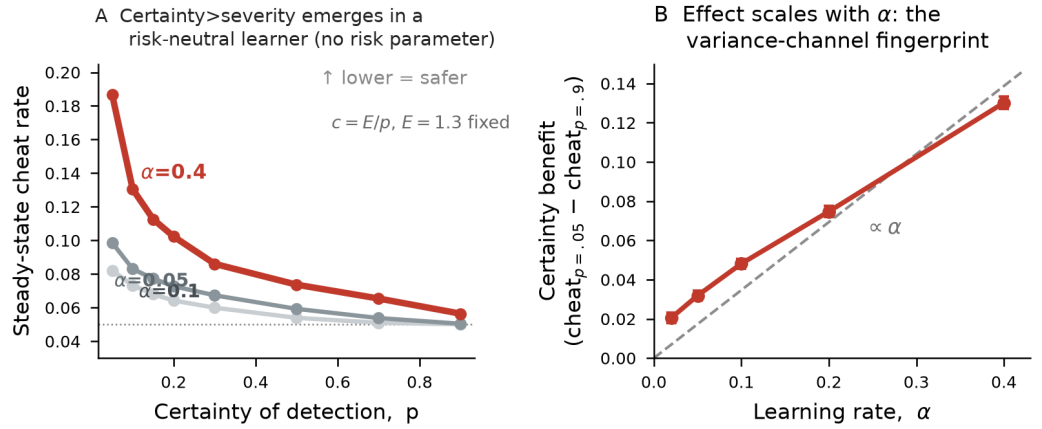


Figure 11: Certainty > severity emerges in tabular Q-learning. (A) Cheating falls as certainty rises; the effect grows with α. (B) The certainty benefit scales linearly with α (corr $= 0.999$).

audit probability and a penalty (the "audit" and "exploit-a-loophole" templates of Appendix P), and a fully *morally-neutral* gamble template is used there only as a control — under which the effect vanishes. *Arithmetic disambiguation:* independently of framing, small models misread the penalty clause. On Qwen2.5-1.5B/3B (CPU) single-pass choice is EV-blind, and chain-of-thought reveals a *comprehension* confound: the 3B model misreads "lose c with probability p" as if p were the probability of the gain. Rewording the penalty as a separate deduction doubles the EV-optimal rate from 40% to 80% (Fig. 16). The lesson: *how you describe an oversight regime to an agent determines whether you measure its risk attitude or its reading comprehension.* All later LLM experiments — the six-scale sweep, the decoupled grid and the GRPO runs — use the arithmetically disambiguated wording under the audit framing; "disambiguated" refers to the penalty arithmetic, not to semantic neutralisation.

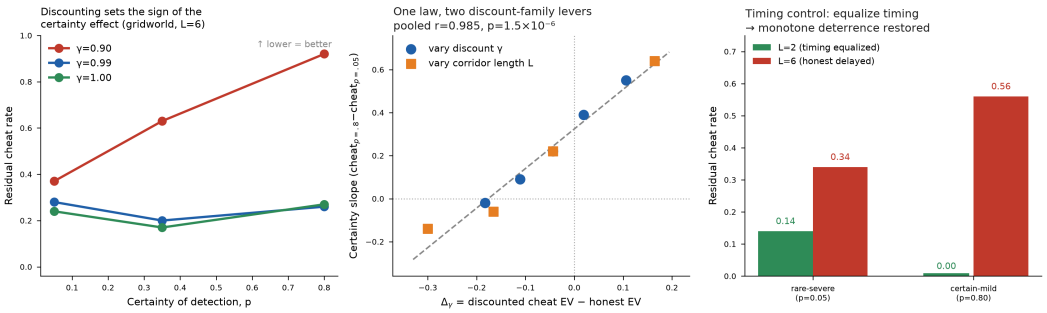


Figure 12: The sequential immediacy failure mode is the law applied to discounted returns. (A) At $\gamma = 0.90$ discounting makes violation EV-optimal and the certainty effect inverts; at $\gamma \geq 0.99$ it does not. (B) Two independent levers — the discount factor and the corridor length — place the measured certainty slope on one line in Δ_γ ($r = 0.985$), with a common sign change near $\Delta_\gamma^* \approx -0.18$ (a coarse 9-cell estimate from the same (γ, L) levers; Sec. 4.3's extended grid reports the delayed-task boundary as a bound below -0.45). (C) Equalizing timing ($L = 2$, honest also one step) restores clean monotone deterrence. Each cell: 100 seeds, $E = 1.3$ fixed.

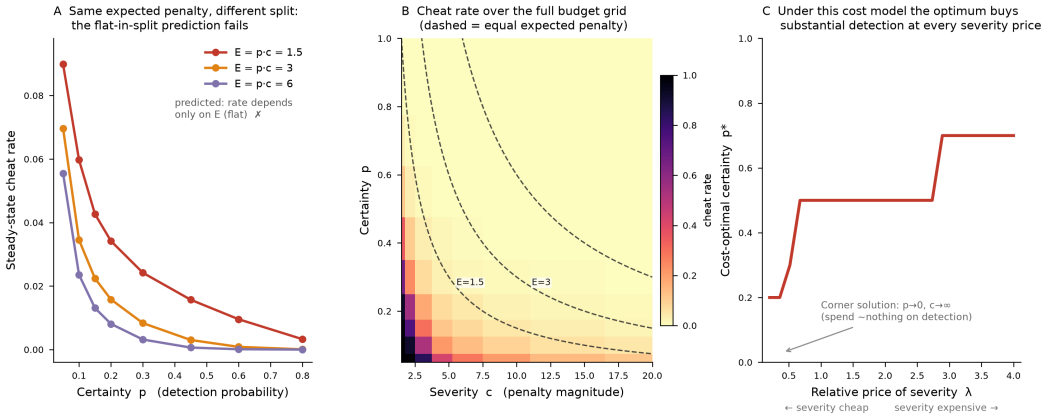


Figure 13: Oversight-budget allocation. (A) At fixed expected penalty the cheat rate falls toward certainty, contradicting the flat-in-split prediction. (B) Cheat rate over the (p, c) grid; dashed curves are equal expected penalty. (C) Under this illustrative cost model the cost-optimal certainty p^* stays in 0.20–0.70 across severity prices and never collapses to the corner $p \to 0,\ c \to \infty$.

E PROOFS

Proposition 1. Unrolling $Q \leftarrow (1-\alpha)Q + \alpha X$ from the infinite past gives $Q_\infty = \alpha \sum_{k\geq 0}(1-\alpha)^k X_{-k}$ (weights sum to 1, so $\mathbb{E}[Q_\infty] = \mathbb{E}[X]$). Independence gives $\mathrm{Var}(Q_\infty) = \alpha^2 \mathrm{Var}(X) \sum_{k\geq 0}(1-\alpha)^{2k} = \frac{\alpha}{2-\alpha}\mathrm{Var}(X)$. Substituting $\mathrm{Var}(X) = c^2 p(1-p)$ gives the result. □

Corollary. With $c = E/p$, $\mathrm{Var}(Q_\infty) = \frac{\alpha}{2-\alpha} E^2 (1-p)/p$; since $\frac{d}{dp}\frac{1-p}{p} = -1/p^2 < 0$ it is strictly decreasing. □

Proposition 2 (signed law). Write $P(p) = \Phi(\Delta/\sigma(p))$ with $\sigma(p) = \kappa\sqrt{(1-p)/p}$, $\kappa = E\sqrt{\alpha/(2-\alpha)} > 0$. By the chain rule $\partial P/\partial p = \phi(\Delta/\sigma) \cdot \Delta \cdot \partial(1/\sigma)/\partial p$, where $\phi > 0$ is the standard normal density. Since $\sigma'(p) < 0$ on $(0,1)$, we have $\partial(1/\sigma)/\partial p = -\sigma'/\sigma^2 > 0$. The two positive factors therefore fix the sign, giving $\mathrm{sign}(\partial P/\partial p) = \mathrm{sign}(\Delta)$, which is Eq. (1). The limits follow from $\sigma \to 0$ as $p \to 1$: the argument $\Delta/\sigma \to -\infty$ when $\Delta < 0$ (so $P \to 0$) and $\to +\infty$ when $\Delta > 0$ (so $P \to 1$); at $\Delta = 0$, $P \equiv \Phi(0) = 1/2$ independently of p. □

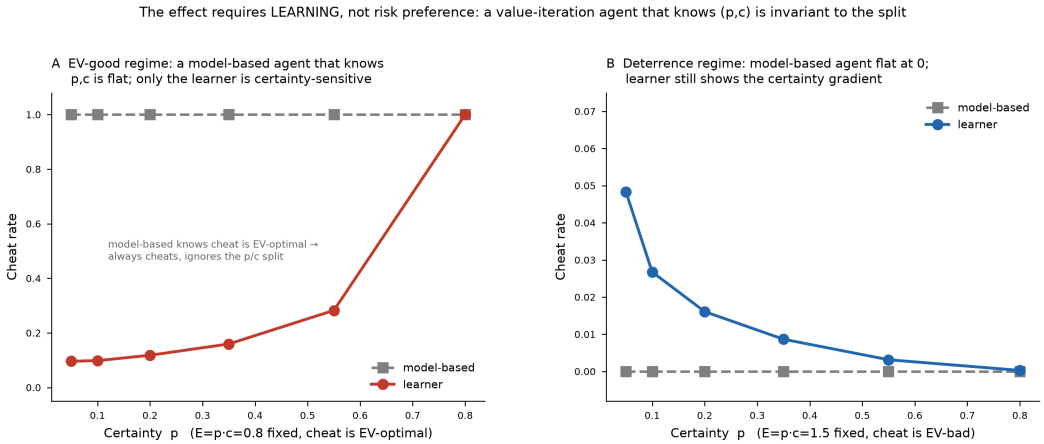


Figure 14: The effect requires learning. A model-based agent that knows (p, c) is flat across the iso-E sweep (grey) in both an EV-good regime (A) and a deterrence regime (B); only the constant-α learner (colour) is certainty-sensitive.

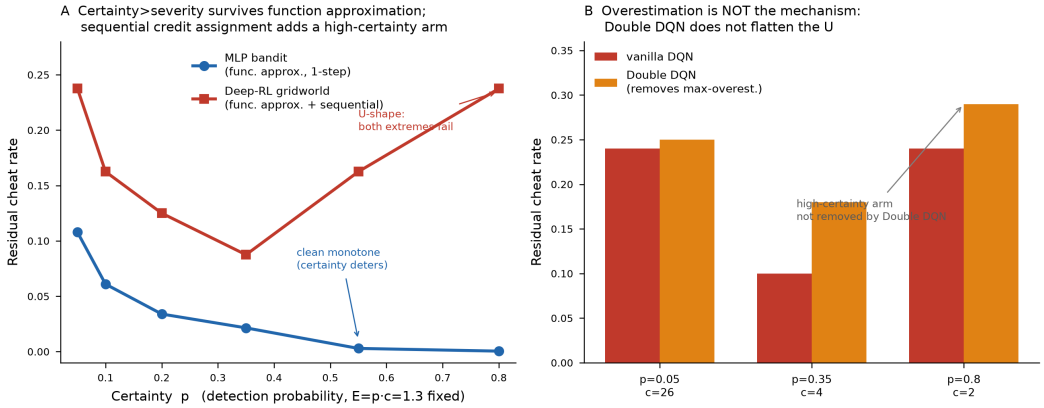


Figure 15: Deep RL. (A) Certainty $>$ severity is monotone in a memoryless bandit and U-shaped under sequential credit assignment. (B) A Double-DQN control rules out max-overestimation as the high-certainty mechanism.

Proposition 2 treats the honest value r_h as a known threshold. In the simulations the honest arm is estimated concurrently and is itself noisy; because its payoff is deterministic its estimate concentrates far faster than the violation arm's, so the fixed-threshold idealisation is close but not exact. It is least accurate in the $\Delta > 0$ branch, where the agent visits the honest arm rarely and its estimate is correspondingly stale — a hot-stove effect on the honest side that we do not quantify. The ε-greedy ablations (Appendix N) quantify the residual dependence. Numerical check: for $\alpha = 0.1$, $\mathrm{Var}(X) = 3$, simulation gives 0.1579 vs. the closed form 0.1579. Evaluating $P(p)$ at $\alpha = 0.1$, $E = 1.3$ confirms all three branches of Eq. (1): $\Delta = -0.20$ gives $0.439 \to 0.090$ (deters), $\Delta = 0$ gives $0.500 \to 0.500$ (neutral), and $\Delta = +0.20$ gives $0.561 \to 0.910$ (backfires) as $p : 0.05 \to 0.80$.

F What the "Controlling for E" Coefficient Measures

Because $\log E = \log p + \log c$ identically, regressing on $(\log p, \log E)$ is a reparametrisation of regressing on $(\log p, \log c)$: $b_p^{(p,c)} \log p + b_c \log c = (b_p^{(p,c)} - b_c) \log p + b_c(\log p + \log c)$. The coefficient on $\log p$ in the (p, E) fit therefore *equals* $b_p^{(p,c)} - b_c$ — the certainty-minus-severity contrast, not a certainty effect net of severity. We verified the identity numerically to 10^{-6} on the

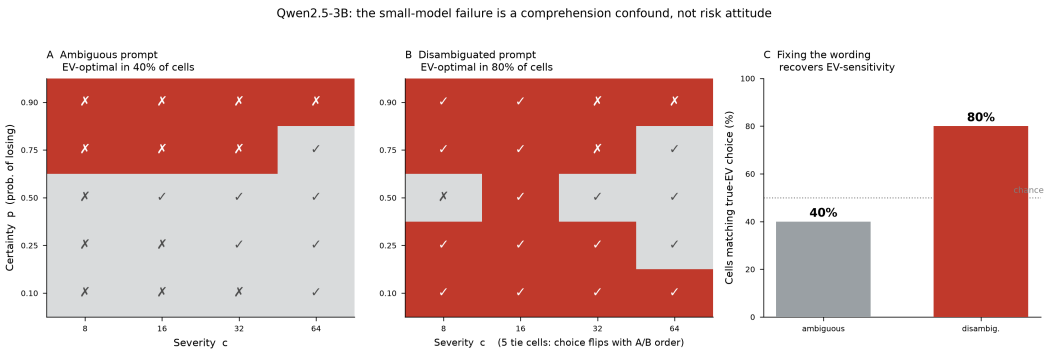


Figure 16: Elicitation validity. Disambiguating the penalty wording doubles Qwen2.5-3B's EV-optimal choice rate (40%→80%).

20-cell grid: the (p, E) fit returns -0.581 and the (p, c) fit returns $b_p = -2.456$, $b_c = -1.875$, whose difference is -0.581.

This makes the claim stronger, not weaker. *Both* levers deter ($b_p, b_c < 0$); the reported -0.581 says certainty deters *more*, which is this paper's thesis stated directly. All sampling in the LLM experiments used temperature 0.7; a probe at 0.9 shifted a measured cell from 0.833 to 1.000 (Fisher $p = 0.010$), so the temperature is reported wherever a rate is.

Risk-free cells. Five of the 20 grid cells have $r_c - c > r_h$: the shortcut cannot lose there, so certainty has nothing to act on and they should be excluded. Doing so strengthens the contrast from -0.581 ($p = 0.005$) to -1.202 ($p = 0.0002$). We report both because the excluded cells were also the only mid-range ones, so the restricted fit rests on a narrower design.

G A Second, Non-Minimal Environment

The corridor of Sec. 4 is deliberately minimal, which invites the objection that the effect is an artefact of the environment rather than of learning. We rebuilt the task to break every simplifying feature: three branching paths instead of one corridor, an exploit that must be *discovered* rather than offered as a labelled action, a violation spanning multiple steps, and a delayed honest payoff.

Both arms behave as the signed law predicts. Where the violation is value-optimal ($\Delta = +0.35$) the rate rises with certainty, $0.094 \to 0.912$ across $p \in \{0.1, \ldots, 0.9\}$; where it is value-suboptimal ($\Delta = -0.35$) it falls, $0.056 \to 0.030$ (12 seeds per cell, $E = 1.5$ held fixed). The between-arm separation is large.

One diagnostic is worth reporting because it initially looked like a floor artefact. The value-optimal arm sits near 0.09 at low certainty even though the violation pays more in expectation. Inspecting the learned table shows why: at the hazard state the greedy policy declines a value-optimal action because estimate variance holds its Q below threshold. That is the paper's own mechanism producing the low rate, not a failure to elicit the behaviour.

H A Generation Bug That Fakes Deterrence

We report a software failure mode because it silently produces this paper's own predicted result, and any replication using the same stack is exposed to it.

Generating samples with a model left in `train()` mode with gradient checkpointing enabled degenerates on Qwen2.5-7B: the policy emits an unbounded run of a single character and never reaches its answer line. Measured with *zero* gradient steps, 12 prompts per condition: the evaluation path answers 1.000 of prompts at both $K{=}1$ and $K{=}8$, while the train-mode path answers 0.083 and 0.052. Raising the token cap from 320 to 768 left it at 0.052 and merely produced more garbage, and the batch size was irrelevant — so neither truncation nor batching is the cause. The mechanism

is that gradient checkpointing disables the key-value cache, and autoregressive sampling in that state breaks.

The danger is that the failure *looks like* success. Because an ungradeable completion scores as a non-violation, the measured violation rate falls, and in our affected run the reported rate tracked the answered fraction at $\rho = +0.87$ ($p = 3 \times 10^{-6}$) while the rate among gradeable answers sat at chance (0.473, 95% CI $[0.445, 0.500]$). A run in that state reports "training reduced the violation rate" with no violation ever having been measured.

The fix is to sample under `eval()` without gradients and restore the caller's mode afterwards; only the log-probability pass needs gradients, and it re-runs a forward pass over already-sampled tokens, so nothing is lost. After the fix the ungradeable fraction across the 18 pilot runs (six cells $\times$ three seeds, Table 1) had median 0.000 and maximum 0.033; across the 340-run grid of Sec. 4.5 it had median 0.000 and maximum 0.077. The degeneration is model-specific — Qwen2.5-1.5B shows an ungradeable fraction of median 0.049 (range $[0.039, 0.108]$, $0/24$ runs above threshold, both arms) through the same code path, so its chance-level result stands — and no in-context measurement in this paper is affected, all of them running inference only.

I Verifier-Design Protocol: Details

A cheap observable that orders the risk. Direct measurement of Δ_γ under deployment timing is the reliable route, but it is not always affordable. One proxy orders tasks by reversal risk from a log of *which action the agent takes*: the per-experience visit-rate asymmetry between shortcut and honest path, which tracks the boundary at $r = -0.77$ across six corridor lengths (App. S) — a correlate that shares the behavioural slope's sign by construction, so it ranks candidate deployments; it does not replace measurement.

Scope of the prescription. It is for agents that *learn* from the signal. In-context choosers respond to the same lever through a different channel and in the opposite direction on our data, so a policy validated on one agent class should not be assumed to transfer (App. B); and a model that cannot parse the audit rule looks compliant for the wrong reason, so verify comprehension before scoring compliance (Appendix D).

Where the marginal oversight dollar goes. Within the deterrence regime it should favour detection over severity (Sec. 5, App. O), and that recommendation *reverses* past the boundary: the same dollar then raises violations. This presupposes the regime check above, which is the point — the allocation advice is conditional on which side of Δ^*_γ the deployment task sits, not on the audit probability being high in absolute terms.

J Reproducibility

Code and raw outputs. All scripts and the raw JSON output of every run cited in the paper (including the GPU runs, so every number can be re-derived without re-running them) are at `https://anonymous.4open.science/r/review-artifact-20260919_1-1BD5/`, with a claim→script→output table in its README.

Grid v2 statistics (Sec. 4.5). 340 runs, 17 seeds per cell. Cell-mean logit fit ($n = 20$): $b_p - b_c = -1.36 \pm 0.40$, $t_{17} = -3.44$, $p = 0.003$; pure-EV ($\log E$ only) vs. $(\log p, \log c)$: $F_{1,17} = 11.8$, $p = 0.003$, AIC 10.2 vs. 1.7. Against a nonlinear pure-EV response: quadratic in $\log E$ has RSS 27.25 (AIC 12.2), cubic 25.32 (AIC 12.7), versus 16.10 (AIC 1.7) for $(\log p, \log c)$; adding the split on top of the quadratic-in-$\log E$ model gives $F_{1,16} = 13.1$, $p = 0.002$. Any $P = f(E)$ enters a $(\log p, \log c)$ fit symmetrically here because the two regressors span nearly equal ranges (2.20 vs 2.08 in logs). Run-level fit ($n = 340$): -2.37 ± 0.29 i.i.d. / ± 0.75 clustered by cell ($t = -3.16$, $p = 0.005$); permutation of run labels (20,000 draws): $p < 5 \times 10^{-5}$, null s.d. 0.44; leave-one-cell-out range $[-2.82, -1.88]$. The run-level magnitude depends on the logit clip applied to the $114/340$ all-or-nothing runs ($-2.37/-1.83/-1.28$ at clip $10^{-3}/10^{-2}/0.05$); the cell-level estimate is clip-free. Trained minus in-context contrast: -0.78 ± 0.45, $z = -1.75$, $p = 0.08$. Format health: median ungradeable fraction 0.000, maximum 0.077, $0/340$ runs over the 0.30 abort threshold. Raw data

Table 1: Objective geometry of the six pilot cells ($r_h = 10$, $r_c = 40$, so $\Delta = 30 - E$; the shortcut is expected-value optimal wherever $\Delta > 0$). Rate change is the mean over three training runs. On these six cells the sign of the change matches the sign of Δ — the 5×4 grid shows that pattern does not survive: sign agreement $8/20$ at 3 seeds and $7/20$ at 17, and the change is uncorrelated with Δ (Sec. 4.5).

p	c	$E = pc$	Δ	rate before $\rightarrow$ after	change
0.1	160	16	$+14$	$0.833 \rightarrow 0.914$	$+0.081$
0.3	80	24	$+6$	$0.208 \rightarrow 0.719$	$+0.511$
0.5	40	20	$+10$	$0.125 \rightarrow 0.819$	$+0.694$
0.7	40	28	$+2$	$0.167 \rightarrow 0.797$	$+0.631$
0.9	80	72	-42	$0.208 \rightarrow 0.158$	-0.050
0.9	160	144	-114	$0.125 \rightarrow 0.067$	-0.058

and per-step advantage traces: `rlvr_grid_v2.json` in the anonymous repository. **Compute.** All learning-agent experiments (Sec. 3, Sec. 4.1–4.4) run on a single Apple MacBook Air (M4, 32 GB, CPU only). The LLM sweeps of Sec. 4.5 and Appendices B and D run partly on that machine (1.5B/3B/7B) and partly on A100 GPUs (14B/32B/72B, and the dtype control); no result in the paper requires a GPU except the three largest model scales.

Sec. 4.4 calibration/deployment pair. proxy L=2, deployment L=6, both γ=0.99, r_h=1, r_c=2, step cost 0.02, E=pc=1.07, rewards paid on arrival; $\Delta_\gamma = (-0.02 + r_c - E) - [\sum_{k=0}^{L-2} \gamma^k(-0.02) + \gamma^{L-2} r_h]$ gives -0.070 and $+0.047$. At L=2 the discount drops out, so the proxy value is identical at any γ.

Tabular and deep-RL sweeps use $r_h = 1.0$, $r_c = 2.0$, $\varepsilon = 0.1$, 6000 steps, steady-state = last 40%, 200–400 seeds/cell; deep RL uses a 2×32 MLP, Adam 10^{-3}, replay 5000, target sync 100, with a Double-DQN control. Note that experience replay resamples stored transitions, so a rare detected violation is replayed many times; the *effective* detection frequency seen by the gradient is therefore closer to uniform over stored outcomes than the nominal p. This attenuates the variance channel in deep RL relative to the tabular case, which is consistent with the smaller effect sizes we measure there, and it means our deep-RL estimates are conservative. LLM agents use Qwen2.5-1.5B/3B/7B-Instruct via transformers on CPU (float32 for 1.5B/3B, bfloat16 for 7B; peak RSS 13.3 GB at 7B), chain-of-thought with a forced "Answer: A/B", letter-counterbalanced; the reward-hacking sweep fixes $E = pc = 20$, sweeps $p \in \{0.1, 0.2, 0.35, 0.5, 0.7, 0.9\}$, and samples 24 completions/point at temperature 0.7 (1.5B/3B/7B, CPU) or 48 completions/point (14B/32B/72B, bfloat16 on A100s). The dtype control re-runs 7B on the identical grid in both float32 and bfloat16 with a fixed seed. Statistics: per-model trial-level logistic regressions of hack on $\log p$, and a pooled $\log p \times \log(\text{size})$ interaction over all six scales ($N = 1170$ full grid, $N = 792$ restricted to $p \leq 0.5$), plus 95% Wilson intervals. On the learning-agent side no clustering correction is needed: every observation is one independent seed (one train-and-evaluate run per data point; $N = 2000$ in the phase diagram, $N = 480$ in the transfer experiment), unlike the LLM sweeps where completions share a prompt and are clustered accordingly. Analyses are reported both over the full grid and over the valid region $p \leq 0.5$, since at $p > 2/3$ hacking is the dominant action (Appendix B). Code, prompts, and data accompany the paper.

K THE K-SWEEP: TESTING THE ADVANTAGE BOUND

Design and pre-registration. The bound $(K-1)/\sqrt{K}$ is a deterministic property of z-scoring inside a group of K; if it is what truncates severity in Sec. 4.5, then at fixed p the severity slope $|b_c|$ must rise with K. We fixed this reading before running: monotone $|b_c|$ in K upgrades the mechanism from observed to predicted-then-tested; a flat $|b_c|$ demotes it to a correlate. Format health was clean in both new runs (median ungradeable fraction 0.000; maximum 0.029 and 0.063; $0/64$ runs over the abort threshold).

Two levels of evidence. The advantage-level claim is exact: the penalty any single audited shortcut receives is pinned to the bound at every K (Fig. 17a), and the range across c of the expected

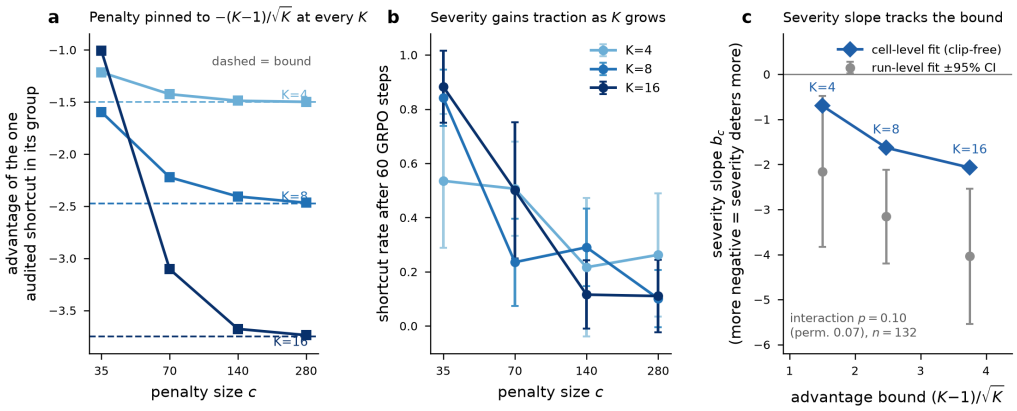


Figure 17: **The advantage bound predicts, and the traces confirm, that severity regains weight as the GRPO group grows.** p=0.5, $c \in \{35, 70, 140, 280\}$; K=4 and 16 with 8 seeds per cell (new), K=8 with 17 from the main grid; learning rate, temperature, steps and generation length identical. **(a)** Mean normalised advantage of a group's single audited shortcut, by penalty size, with the bound $-(K-1)/\sqrt{K}$ dashed in the same colour: it saturates at the bound for every K (regression of the $c \geq 140$ values on the bound: slope 0.99, $r = 1.00$). **(b)** Trained shortcut rate against penalty size: at K=4 the curve is nearly flat; at K=16 it falls from 0.88 to 0.11. Bars are 95% CIs over seeds. **(c)** Severity slope b_c against the bound; cell-level (clip-free) $-0.70, -1.63, -2.07$; run-level with 95% CIs. The $\log c \times$bound interaction over all 132 runs is -0.83 ± 0.49 ($p = 0.10$; permutation of K labels within c, $p = 0.07$); ordering holds at every logit clip ($10^{-3}/10^{-2}/0.05$).

advantage difference between shortcut and honest completions — the severity signal the gradient actually sees — grows from 0.11 (K=4) to 0.36 (K=8) to 0.76 (K=16). The behavioural consequence is in the predicted direction and ordered across all three K, but with 8 seeds per new cell against a between-run SD of 0.3 it does not resolve ($p \approx 0.07$–0.10); pairwise, K=16 vs K=4 gives $z = -1.64$. We therefore say the mechanism is confirmed where it is arithmetic; at the behavioural level the cap-free controls below leave its share unresolved.

Removing the cap: two cap-free controls. The K-sweep varies the bound; the sharper test removes it. We re-ran the p=0.5 row at K=8 with 3 seeds per cell under two estimators, everything else identical to the main grid (reused in-context baselines; two implementation choices forced by a shared GPU — a fused AdamW kernel and per-sequence gradient accumulation — are recorded in the outputs and unit-tested to give the same update and gradient as the published path). *Mean-centred*: the group standard deviation is replaced by a fixed divisor (30, the unaudited-shortcut/honest reward gap, so unpenalised advantages stay $O(1)$), group-mean baseline kept; this isolates the standard-deviation term (Dr. GRPO's proposal). *RLOO*: the same divisor with a leave-one-out baseline. Format health was clean in both (0/12 runs over threshold). *Advantage level.* The lone audited shortcut receives $-0.88/-1.49/-3.44/-7.35$ (mean-centred) and $-0.70/-1.61/-4.00/-8.37$ (RLOO) at $c = 35/70/140/280$, against GRPO's $-1.60/-2.22/-2.41/-2.47$; the expected shortcut-minus-honest advantage ranges over 2.88 and 3.40 across c against 0.63 under GRPO on the same statistic. *Behaviour.* Cell means $0.98, 0.52, 0.33, 0.00$ (mean-centred), $0.93, 0.39, 0.62, 0.00$ (RLOO), $0.84, 0.24, 0.29, 0.10$ (GRPO). Run-level severity slopes $b_c = -5.43 \pm 0.87$, -3.47 ± 1.20 and -3.16 ± 0.53 ($n = 12, 12, 68$). Against GRPO: mean-centred -2.27 ± 1.02 (interaction $p = 0.09$, permutation within c $p = 0.03$), shrinking to -1.41 and -0.83 at logit clips 10^{-2} and 0.05 (six of its twelve runs sit at a rate below 0.01 or above 0.99); RLOO -0.31 ± 1.31 ($p = 0.82$, permutation 0.78), stable across clips; the two controls differ from each other by -1.96 ± 1.48 (permutation $p = 0.06$). Pooled 24 cap-free runs: $b_c = -4.45 \pm 0.74$, steeper than GRPO by 1.29 ± 0.91 (interaction $p = 0.20$, permutation 0.11; -0.88 and -0.58 at the two coarser clips). No single cell differs from GRPO under either estimator (Mann–Whitney $p \geq 0.12$). *Reading.* A slope proportional to the widened signal would be near -10; both controls exclude it. A partial response of the K-sweep's magnitude

is consistent with the mean-centred arm and 1.6 SE from the RLOO arm. Raw data and traces: `rlvr_meancenter_K8.json`, `rlvr_rloo_K8.json`, `rlvr_rloo_analysis.json`.

Entropy is not the driver of the broad increase. The same runs log per-token policy entropy at every step. Entropy falls during training at both K ($0.300 \to 0.223$ nats at K=4, $25/32$ runs down; $0.301 \to 0.243$ at K=16, $23/32$ down), and the change in shortcut rate is *negatively* correlated with the change in entropy ($\rho = -0.50$, $p = 0.003$ at K=4; -0.25, $p = 0.18$ at K=16); within runs entropy tracks the rate negatively (median $\rho = -0.32$). The broad increase of Sec. 4.5 is therefore a sharpening toward one option, not a flattening of the policy; the hypothesis that it is sampling noise is rejected. Raw data and traces: `rlvr_ksweep_K4.json`, `rlvr_ksweep_K16.json`.

Several audited shortcuts in one group. The $(K-1)/\sqrt{K}$ bound is the $m = 1$ case. If m of the K completions in a group are penalised (all receiving the same $-c$) and the remaining $K-m$ receive a common higher reward, the Bessel-corrected standardised advantage of each penalised completion is $-\sqrt{(K-1)(K-m)/(Km)}$, independent of c: for $K = 8$ it is $-2.47, -1.62, -1.21, -0.94$ at $m = 1, 2, 3, 4$ and 0 at $m = K$. So severity is discarded at every m, and the gradient pressure against an audited shortcut shrinks further the more often the shortcut is taken and caught within a group — an arithmetic reason, additional to the bound, why penalties lose weight as the shortcut becomes common. Mixed groups (honest, unaudited and audited completions with three distinct rewards) do not admit a two-line closed form, and we have not tabulated them; the traces of Fig. 3c are dominated by $m = 1$ events at the p values of the grid.

L The Exact Stationary Law, and What the Gaussian Step Costs

Proposition 2 derives the signed law from a Gaussian approximation to the stationary estimate. That approximation is the paper's weakest theoretical step, and this appendix replaces it with the exact law and measures the error.

The exact law. Under the constant-α update the stationary estimate reduces exactly to $Q_\infty = r_c - c\,S$ with $S = \alpha \sum_{k\geq 0}(1-\alpha)^k B_k$, $B_k \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. S is a ***Bernoulli convolution***: it satisfies the exact self-similar identity $S \overset{d}{=} \alpha B + (1-\alpha)S'$, hence its CDF is the unique fixed point of

$$F(x) \;=\; (1-p)\,F\Big(\tfrac{x}{1-\alpha}\Big) \;+\; p\,F\Big(\tfrac{x-\alpha}{1-\alpha}\Big). \tag{2}$$

Iterating (2) reproduces the closed-form mean p and variance $\frac{\alpha}{2-\alpha}p(1-p)$ of the unit-scaled accumulator S (Proposition 1's $\text{Var}(Q_\infty)$ is c^2 times this) to six decimals at $(\alpha, p) \in \{(0.1, 0.3), (0.4, 0.5), (0.7, 0.2)\}$, which is an independent check on Proposition 1 itself. With $E = pc$ held fixed, the violation probability is $P = F(p(\Delta + E)/E)$ — note that both the law and the threshold move with p, which is why the exact $P(p)$ can be non-monotone even where the Gaussian form is monotone.

What the approximation costs. The signed law predicts that $\partial P/\partial p$ changes sign exactly at $\Delta = 0$. Measuring the exact crossing by bisection (Fig. 18) gives a displacement that is positive at every α tested and grows roughly linearly: $+0.024$, $+0.060$, $+0.096$, $+0.083$, $+0.190$ at $\alpha = 0.1, 0.2, 0.3, 0.4, 0.6$ (slope $+0.31$, $r = 0.95$, $p = 0.012$). Two independent methods agree: iterating (2) on a grid gives $+0.0243$ at α=0.1 and Monte-Carlo with 2M unbiased draws gives $+0.0241$. Operationally, these are the zero of the local derivative $\partial P/\partial p$ evaluated at $p = 0.5$ with $E = 1$, in units of Δ; because the threshold depends on Δ/E the displacement scales with E. The $+0.064$ crossing in Fig. 10A is a different functional of the same distribution — the zero of the log-certainty slope fitted across the deep-RL grid's five p values at $E = 1.3$, a chord rather than a derivative — which is why the two numbers differ; both sit on the same side of the Gaussian prediction.

The error is *one-signed*, and its direction matters for practice: the Gaussian criterion predicts backfiring for a band of small positive Δ in which the exact law still gives deterrence. A practitioner using it errs toward over-warning, not under-warning. At the setting used throughout the paper ($\alpha = 0.1$) the displacement is $+0.024$ in Δ units, below the spacing of our experimental grid.

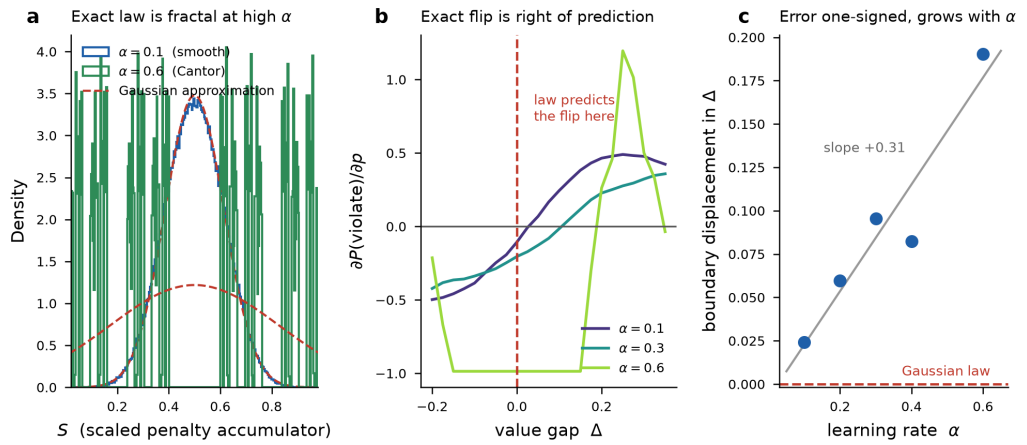


Figure 18: **The exact stationary law and the cost of the Gaussian step. (a)** Exact law (histogram, 4×10^5 draws) against its Gaussian approximation: close at $\alpha = 0.1$, fractal at $\alpha = 0.6$. **(b)** Exact $\partial P/\partial p$ against the value gap; the signed law places the flip at $\Delta = 0$ (dashed). **(c)** The displacement is positive at every α and grows with it.

Where the Gaussian rule fails globally: the sparse-audit limit. The accumulator S is bounded in $[0, 1]$, so $Q_\infty = r_c - cS \leq r_c$ and the Gaussian tail that lets $\Phi(\Delta/\sigma) \to 1/2$ as $\sigma \to \infty$ has no counterpart in the exact law. As $p \to 0$ at fixed E the audit process becomes ultra-sparse while each audit's imprint decays in $O(\log(1/\alpha))$ steps, so the estimate sits at r_c almost always and $P(\text{violate}) \to 1$ for every $\Delta > -E$, whatever its sign. For $\Delta > 0$ the exact $P(p)$ is therefore *U-shaped*, and Eq. (1)'s "$\partial P/\partial p > 0$" holds only to the right of a Δ-dependent minimum: exact Monte Carlo (3×10^5 draws per point) at $\alpha = 0.1$ puts that minimum at $p \approx 0.05$ for $\Delta = +0.30$, $p \approx 0.08$ for $\Delta = +0.10$ and $p \approx 0.30$ for $\Delta = +0.05$, with $P(0.005) \approx 0.87$ in all three cases; at $\alpha = 0.3$ the minima move right (0.15, 0.5, 0.9). For $\Delta < 0$ the rule's "$\partial P/\partial p < 0$" is unaffected — P falls monotonically from ≈ 0.8 toward 0. Proposition 2 is thus a statement about the audited regime, p above the minimum; our experimental grids start at $p = 0.05$, and the small-positive-Δ band where the minimum sits at $p \gtrsim 0.3$ is the same band in which Fig. 18 shows the exact law deterring where the Gaussian predicts backfire.

A caveat we can state precisely. S is a Bernoulli convolution with contraction ratio $\lambda = 1 - \alpha$, so its topological support has open (Cantor) gaps exactly when $\lambda < 1/2$, i.e. when $\alpha > 1/2$. That condition is sufficient but not necessary for the law to be singular: Erdős showed the symmetric convolution is singular whenever $1/\lambda$ is a Pisot number in $(1, 2)$ even though its support is then a full interval, and for biased p the measure is singular whenever the entropy $h(p)$ falls below $\log(1/\lambda)$. Our operating regime ($\alpha = 0.1$, $\lambda = 0.9$) is smooth in every sense that matters here. We verified this by enumerating the support: at $\alpha = 0.6$ the sorted support has gaps up to 0.2 that survive the truncation tail, while at $\alpha \in \{0.1, 0.4\}$ none do. The plateau of $\partial P/\partial p$ at exactly -1 for $\alpha = 0.6$ in Fig. 18b is the devil's-staircase signature, not a numerical artefact. The paper's regime, $\alpha = 0.1$, sits well inside the smooth case. What we do not have is an *analytic* bound on the displacement; the measurement is numerical.

M LLM Choice-Model Comparison

On the 384-trial reward-hacking data we fit two trial-level choice models: a *penalty-sensitive* logit $P(\text{hack}) = \sigma(\beta_0 + \beta_1(B - \tilde{E}) - \beta_1 r_h)$ with perceived penalty $\tilde{E} = E^\eta p^{1-\eta}$, and a *pure-EV* logit using the true expected penalty $pc = E$. Because E is held fixed across the sweep, η is *not* point-identified (bootstrap 95% CI widths: 11.7 at 1.5B, 1.5 at 3B, 0.01 at 7B — the 7B collapse is a parameterization degeneracy, not a precise estimate), so we do not report a η value. What *is* identified is the model comparison: the penalty-sensitive model beats pure-EV by $\Delta\text{AIC} = -1.9, +5.2, +56.3$ (likelihood-ratio $\chi^2 = 0.1, 7.2, 58.3$) at 1.5B/3B/7B (Fig. 19). Models from

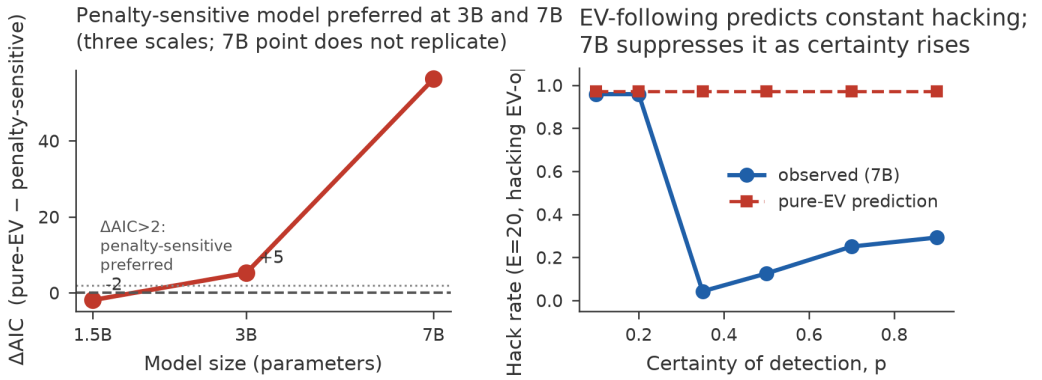


Figure 19: Choice-model comparison on the three scales for which trial-level data were retained (1.5B/3B/7B). (A) Pure-EV following is rejected at 3B and 7B (ΔAIC favouring the penalty-sensitive model) but not at 1.5B. This is a comparison at three scales, not evidence of a scale *trend*: the six-scale analysis of Appendix B finds no such trend. (B) A pure-EV follower would hack at a constant near-1 rate; the 7B model instead suppresses hacking as certainty rises.

3B upward are thus demonstrably not pure expected-value maximizers; they suppress exploitation as detection certainty rises even though it stays expected-value-optimal. Two caveats bound this comparison. First, on the fixed-E sweep the payoffs $(+10, +40, -c)$ enter only through $E = pc$, which is constant across trials, so the pure-EV predictor is functionally an intercept-only model; the likelihood-ratio test therefore establishes that choice varies with p, not that the specific η-form beats other non-linear specifications. A fit on the decoupled (p, c) grid of Appendix F, where E varies, is the version that would test EV-responsiveness proper, and the contrast reported there ($b_p \neq b_c$) is the decisive statistic. Second, this comparison uses the original 1.5B/3B/7B trial data and therefore inherits the 7B measurement that Appendix B reports as non-replicating; the qualitative conclusion (pure EV is rejected above the comprehension threshold) does not depend on the 7B effect size, but the scale ordering of ΔAIC should not be read as a capability trend.

N Mechanism-Boundary Ablations

The displacement is not a function-approximation artefact (moved from Sec. 4.3). The delayed-task boundary is measured with a neural value function, so it could in principle be a property of the network rather than of temporal credit assignment. We re-ran the sweep with a *tabular* Q-learner on the same corridor — no approximation of any kind — over the deep-RL grid's Δ_γ range (5×5 cells, 200 seeds/cell, $N = 5000$ runs). The interaction survives ($\beta = +1.76$, $z = 4.08$, $p = 4.5 \times 10^{-5}$) and the linear zero lands at -0.461 (95% CI $[-0.89, -0.31]$), displaced from zero in the same direction and decisively excluding it (Fig. 21). We claim the *direction* only. The magnitudes are not comparable across estimators, and we do not reconcile them: both are extrapolated zeros (Sec. 7), the setups differ in effective step size, and nominal learning rates do not translate into effective ones (Adam's per-parameter normalisation rescales the update). Prop. 1 does make a testable conjecture — if the immediacy term is α-insensitive then $|\Delta^*| = \kappa|v| \propto E\sqrt{\alpha}$, where $v = \Delta^*/\sigma$ is the standardised gap at the boundary and $\kappa = E\sqrt{\alpha/(2-\alpha)}$ — but only a controlled tabular α sweep at fixed E could test it, which we have not run (Appendix J notes a second tension: replay attenuating the variance channel would push the deep-RL boundary *further* out, and we observe the reverse). What this ablation establishes is that the immediacy channel is a property of sequential value learning, not of the function approximator.

We stress-test the tabular mechanism (Fig. 20). **Step size:** the effect is strong under a constant step size (endpoint gap $D_p = 0.048$), shrinks under finite-horizon decay (0.015), and nearly vanishes under $\alpha_t = 1/N$ (0.004) — confirming the mechanism requires non-vanishing step size, since a converging estimate has vanishing variance. **Exploration:** the certainty benefit in *rate* terms grows $19\times$ across $\varepsilon \in \{0.01, 0.05, 0.1, 0.2\}$ ($D_p = 0.005, 0.026, 0.048, 0.098$), as more sampling of the rare-severe arm feeds the variance channel. This is a level effect, not a change in the law: the

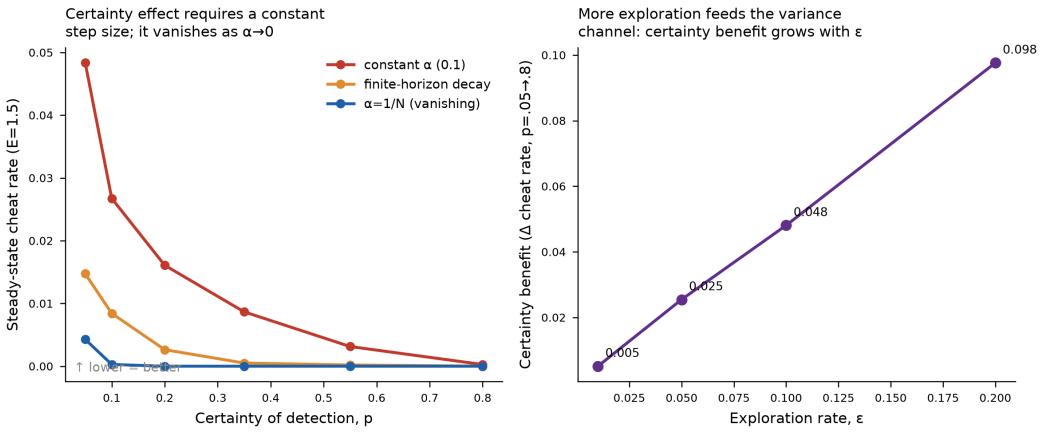


Figure 20: Tabular mechanism-boundary ablations. (A) The certainty effect requires a constant step size and disappears as $\alpha \to 0$. (B) The certainty benefit grows with exploration ε.

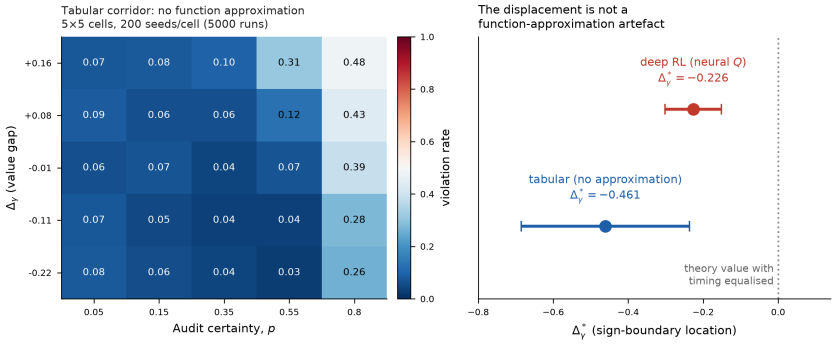


Figure 21: **The sign-boundary displacement replicates without function approximation.** (A) The (p, Δ_γ) sweep re-run with a tabular Q-learner on the same corridor; Δ_γ is swept by solving for the expected penalty that produces each row. (B) Both estimators place the boundary well below the timing-equalised theory value of ≈ 0; the intervals overlap, but we claim agreement in direction, not in magnitude.

log-certainty slope is nearly ε-invariant over the same range ($-1.88, -1.61, -1.62, -1.64$; a factor of 0.87 across a $20\times$ change in ε). Eq. (1) contains no ε, so this is the expected pattern if the equation fixes the *ordering and sign* while the absolute rate additionally depends on how often the rare arm is sampled — the hot-stove dynamic of (Denrell & March, 2001), in which an estimate that falls below threshold stops being sampled and freezes. **Initialization:** the effect is invariant to optimistic/zero/pessimistic initialization ($D_p \approx 0.048$ throughout), confirming it is a steady-state property. On this reward-only bandit, Expected-SARSA coincides with Q-learning (no state transition), so the ablation is reported as an update-rule check rather than a distinct on-policy control.

O Budget-Allocation Cost Sensitivity

Setup, and the three findings the body compresses. In RLVR-style training the two knobs exist separately and are separately priced: the detection probability p is the fraction of rollouts routed to the expensive verifier — a chain-of-thought monitor, a held-out test suite, a human spot-check — consuming monitor inference or human time in proportion to p, while the penalty magnitude c is the negative reward assigned when the monitor fires and is nearly free. On the learner of Sec. 4.1 we run the full decoupled (p, c) grid and price a budget $B(p, c) = -\ln(1-p) + \lambda \ln(1+c)$. Three findings (Fig. 13): along constant $E = pc$ the violation rate falls up to $30\times$ toward certainty,

falsifying the flat-in-split prediction that holds for agents with known (p, c); severity has sharply diminishing returns; and under this illustrative cost model the cost-optimal certainty p^* is 0.20–0.70 across severity prices, so the corner ($p \to 0, c \to \infty$) is never optimal. (p^* is defined as the cheapest grid cell (p, c), $c \leq 60$, whose measured violation rate is at most 0.05; there is no fixed budget endowment. The corner loses because, at low p, the target is reached only at penalties whose cost under every family we tried exceeds that of a more certain, milder audit — the learner's variance sets which cells are feasible, the cost family only where along the feasible set the optimum lands.) The recommendation is regime-conditional: in the deterrence regime (Δ below the boundary, as in this sweep) spend the marginal oversight dollar on detection coverage, but establish the regime first — above the boundary the same dollar raises violations (Fig. 1).

We repeat the allocation of Sec. 5 with detection cost fixed at $-\ln(1-p)$ and four severity-cost families: logarithmic $\ln(1+c)$ (baseline), linear c, convex $c^{1.5}$, and bounded (logarithmic capped at $c \leq 20$). Optimizing the cheat rate at matched budget over $\lambda \in \{0.3, 1, 3\}$ yields optimal certainty p^* in $[0.49, 0.68]$ (log), $[0.67, 0.71]$ (linear), $[0.68, 0.72]$ (convex), and $[0.49, 0.68]$ (bounded). Because that sweep varies only the *severity* price, we repeat it varying the *detection* price, holding severity cost at $\ln(1+c)$. We report each family separately: linear p and convex p^2 are *cheaper* than $-\ln(1-p)$ at every p (both bounded by 1 while the baseline diverges), so a pooled range would conflate dearer with cheaper detection. Optimal certainty by family over $\lambda \in \{0.3, 1, 3\}$: $p^* \in [0.60, 0.97]$ baseline, $[0.97, 0.97]$ linear, $[0.80, 0.97]$ convex, $[0.60, 0.80]$ under $p/(1-p)$. Only the last is dearer than the baseline at every p ($32\times$ the audit price at $p = 0.97$), and it carries the claim: even with sharply rising marginal detection cost, p^* stays at 0.60–0.80 and never approaches the severity corner. (p^2 is the weakest of the four as a cost model — auditing everything should not cost ≤ 1 — and we read it only as a cheap-end bound.) In every family p^* is strictly positive and substantial; the corner solution is never optimal, so the recommendation to buy real detection is robust to the cost specification.

P LLM Prompt, Temperature, and Reasoning Robustness

Because the LLM elicitation is sensitive to wording (Appendix D), we varied the prompt and decoding. The full battery now runs at 7B — the scale carrying the magnitude result — with $n \approx 22$–24 completions per point, superseding the small-n 3B checks of our earlier draft (which are retained in Fig. 22 for the 3B scale).

Paraphrase (2 of 3). Under the audit framing the log-certainty slope is -0.80 ($z = -3.20$, $p = 0.001$); under the "exploit-a-loophole" framing it is -1.42 ($z = -4.63$, $p = 3.7 \times 10^{-6}$). Under the morally-neutral gamble framing, however, the slope is $+0.27$ ($z = +1.12$, $p = 0.27$) — directionally opposite and not significant. The effect is therefore not invariant to framing: it appears when the scenario is described in oversight or exploitation terms and not when the same gamble is described neutrally. Since a deployed verifier confronts agents in oversight-framed settings, this is a scope condition rather than a refutation, but it is a real limit on generality and we do not average it away.

Reasoning. Without chain-of-thought (single forced answer) the model is EV-blind — a flat 0.5 hack rate across all p — at both 3B and 7B. The certainty response is produced by explicit deliberation, not token priors.

Temperature. At $\tau = 0.3$ the slope is -0.75 ($p = 0.002$), at $\tau = 0.7$ it is -0.80 ($p = 0.001$); the effect is not a decoding artefact.

EV-negative regime. At $E = 35$ (hacking EV $+5$ < honest $+10$) the 7B hack rate never exceeds 0.09 at any certainty and shows no gradient ($p = 0.13$). Once hacking is expected-value-dominated the model simply declines it, and there is nothing left for certainty to suppress — the boundary condition the signed law predicts.

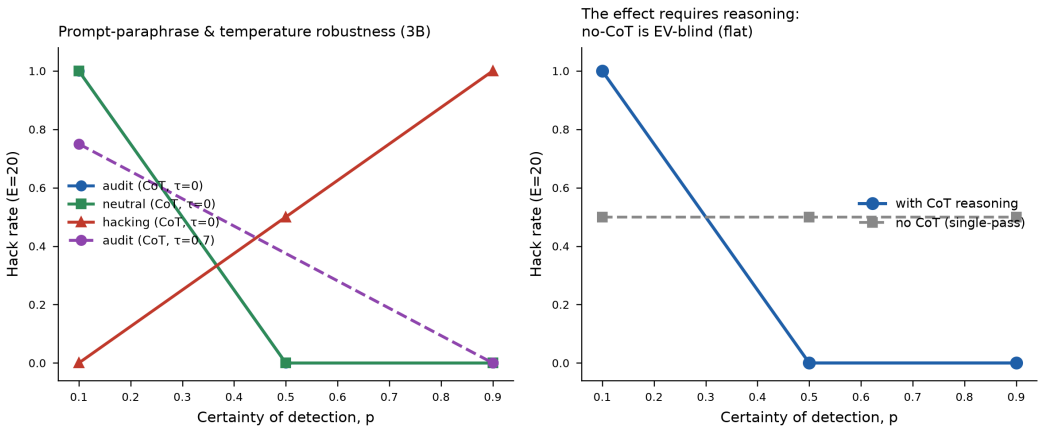


Figure 22: 3B robustness checks (small n; direction, not magnitude). (A) The certainty→suppression direction holds under audit and neutral paraphrases (curves coincide) and under τ=0.7 sampling; the "hacking"-worded prompt is noise-dominated at n=1. (B) Without chain-of-thought the model is EV-blind (flat), so the effect requires explicit reasoning.

Q A Distributional Learner: the Variance Channel Switches Off, Certainty Survives

Proposition 1 is a statement about a *point* estimate. A distributional learner that tracks the return distribution explicitly is the natural test of whether the variance channel is the whole story. We repeat the Sec. 3 bandit ($E = 1.3$, $\alpha = 0.1$, ε-greedy 0.1, 128 seeds, 60,000 steps) with a quantile-regression learner (QR-DQN's tabular core: N quantiles θ_i updated by $\theta_i \leftarrow \theta_i + \alpha(\tau_i - \mathbf{1}[X < \theta_i])$, greedy on the quantile mean), and compare it with the point learner at the same step size.

Stationary: the variance channel is switched off. In the deterrence regime ($\Delta = -0.15$) the point learner shows the paper's graded certainty effect over the last 30,000 steps ($0.109 \rightarrow 0.059$ from $p = 0.05$ to 0.8). The N=32 quantile learner sits at the exploration floor ($\varepsilon/2 = 0.05$) at every $p \geq 0.15$: once the penalty atom is represented, its stationary choice is essentially noiseless. So Proposition 1's channel is indeed a property of point estimation.

But certainty > severity returns through two routes a point learner does not have. *(i) Tail-learning speed.* The quantile below the penalty atom moves $\alpha(1-\tau)$ per penalty sample, penalty samples arrive at rate p, and at fixed E the distance to travel is $c = E/p$, so the time to learn the penalty scales like $E/(\alpha p^2)$ or steeper; measured (greedy learner, median over 128 seeds) it is 17,426 steps at $p = 0.05$, 2,768 at 0.1, 626 at 0.2 and 119 at 0.4 (log-log slope -2.4). Over the first 10,000 steps the quantile learner cheats 0.94 of the time at $p = 0.05$ and 0.05 at $p = 0.8$ — a log-certainty slope of -0.31, 18× the point learner's stationary -0.017. *(ii) Tail resolution.* If $p < 1/N$ no quantile lies in the penalty atom: every θ_i converges to r_c, the penalty is invisible however large c is, and the agent cheats permanently (N=8, $p = 0.05$: stationary rate 0.95 against 0.050 at $p \geq 0.35$). Both routes are certainty effects that severity cannot repair: at fixed E, lowering p makes the penalty slower to learn and, below the estimator's resolution, unrepresentable.

Reading. Distributional estimation removes the stationary-variance mechanism and replaces it with a learning-speed and a resolution mechanism of the same sign. This is the tabular core only; a neural QR-DQN on the corridor is not run. (`distributional_control.json`, `distributional_learning_time.json`.)

R THE IMMEDIACY CHANNEL: WHAT IS RULED OUT, AND WHAT PREDICTS THE BOUNDARY

This appendix carries the detail behind the Limitations section. The variance channel has an exact closed form and a signed law; the immediacy channel has neither. We identify it empirically (an additive, p-increasing component in the residual), verify it by ablation (removing the timing asymmetry removes the displacement, Sec. 4.3), but cannot derive *why* it increases in p. What follows is the record of what we tried, because the eliminations bound the shape of any account that does work.

One route is excluded in general, not merely untried. The natural suspicion is that the asymmetry acts through the variance channel: the violation is one step while the honest path is $L-1$, so the violation entry is under-represented in experience, and an under-sampled entry might carry more noise. It cannot. For the constant-α recursion the stationary variance is frequency-invariant — how *often* an entry is updated does not enter Eq. (1). Simulating the recursion at 20k, 60k and 200k updates gives $\mathrm{Var} = 0.333, 0.360, 0.359$ against a closed form of 0.356; repeating with a bootstrapped target (a two-entry chain in which the violation entry is updated with probability 0.10, 0.30 or 0.70 per iteration) gives 0.340, 0.354, 0.357 against the same 0.356. Update frequency changes the *relaxation time* of the estimate, not its stationary spread. Any account of the immediacy channel must therefore act somewhere other than $\mathrm{Var}(Q_\infty)$.

The transient reading that suggests is also unsupported. If the channel were a relaxation-time effect — the under-sampled violation entry simply not reaching stationarity inside the training budget — the delayed-task slope should move toward the timing-equalised prediction as training lengthens. It does not. Holding $\Delta_\gamma = -0.012$ and $L = 6$ fixed and training for 800, 2400 and 7200 episodes (24 seeds per cell, five certainty levels), the log-certainty slope is $+0.313$, $+0.439$, $+0.165$: positive at every budget, with no monotone trend over a $9\times$ range ($\rho = -0.50$, $p = 0.67$; longest versus shortest $\Delta = -0.149$, $z = -0.56$, $p = 0.58$). The channel is not a convergence artefact of a short run.

A sampling-feedback model predicts the wrong sign. The one quantitative model we built from the sampling asymmetry treats the violation's share of buffered experience, $f(m) = m/(m + (1-m)(L-1))$ under an ε-greedy floor, as scaling an effective step size, giving $\sigma_{\text{eff}}(p)$ and a fixed point $m = \Phi(\Delta/\sigma_{\text{eff}})$. Solved non-circularly (unique root at every cell), it gives log-certainty slopes of -0.10 at $L = 2$ and -0.22 at $L = 6$: not only the wrong sign for the delayed task, but the wrong *ordering* — it predicts delay makes certainty deter more, where we measure the opposite. Rejected.

The value-space decomposition does not close. The natural closing step works in value space: with $m(p) = \mathbb{E}[Q_\infty^{\text{cheat}}] - V_{\text{honest}}$, expanding the threshold-crossing probability gives $\Delta^* = -2p(1-p)\,m'(p)$, so the measured $\Delta^* = -0.226$ at $p = 0.35$ needs $m' \approx 0.50$. It does not hold: at fixed $E = pc$ the cheat arm's expected payoff is exactly p-independent, so $m' = 0$ analytically and $m' = +0.053$ empirically (1200 seeds) — an order of magnitude short, and attributable to the ε-greedy correction. The decomposition does not close, which localises the gap: whatever immediacy shifts, it is not the *mean* cheat value. Two candidate mechanisms also fail, predicting the wrong sign: the undetected-violation immediacy bonus weighted by its frequency $(1-p)$ *decreases* in p, and truncation of the effective penalty by the honest continuation value, $\min(c,\, r_c - \gamma V_{\text{honest}})$, makes the effective expected penalty *rise* with p at fixed nominal E.

A third candidate does have the right sign, and we can measure it. Violating ends the episode immediately while honest completion takes $L-1$ steps, so the two arms are visited at different rates *per unit of experience* — and that rate depends on the learned violation value, hence on p. This feedback loop is absent in a bandit by construction. Logging realised visit rates at fixed E (150 seeds/point), the violating arm's share of steps rises $6.0\times$ from 0.117 at $p = 0.05$ to 0.706 at $p = 0.80$ (slope $+0.160$ per $\log p$), whereas under equalised timing ($L = 2$) it is flat (-0.008, a $21\times$ smaller magnitude and the opposite sign). Appendix N shows the absolute violation rate depends strongly on sampling frequency ($19\times$ across ε), so the magnitude is plausible too. We stop short of a derivation — this is a p-increasing quantity a bandit lacks and the corridor has, not an account of the full -0.22 displacement — but we can show it predicts *where* the boundary lands.

We can go one step further and use the sampling-rate channel as a *predictor*. Sweeping corridor length, $L \in \{2, 3, 4, 6, 8, 12\}$ with a tabular learner throughout (so its $L = 6$ value, -0.463, belongs against Sec. 4.3's tabular -0.461 rather than the deep-RL number; both are extrapolations past the same grid bottom, so their 0.0002 agreement should not be over-read), the boundary does *not* collapse onto the discount factor's reach γ^{L-1}: $\Delta^*_\gamma(L)$ is non-monotone (-0.013, $+0.022$, -0.177, -0.463, -0.463, -0.204) while γ^{L-2} is monotone, and that hypothesis predicts the wrong sign ($r = +0.56$). But a discount-only account was never going to work, because it ignores the channel we just measured. The visit-rate slope, measured separately at each L, is *also* non-monotone, peaking at $L = 6$ inside the $L = 6$–8 region where Δ^*_γ bottoms out (the two boundary minima, -0.4628 and -0.4630, are independent fits but both extrapolate past the grid bottom of -0.22, so we do not lean on their near-equality), and it tracks the boundary: $r = -0.77$, bootstrap 95% CI $[-0.86, -0.65]$ over the boundary's own uncertainty, negative in 100% of draws and beating the discount-only hypothesis in 97% (Fig. 23); rank-order agreement is weaker at six points (Spearman $\rho = -0.66$, $p = 0.16$). Two qualifications. The boundary values entering this correlation are linear extrapolations of the kind Sec. 4.3 shows to be unreliable past the grid, so the correlation measures agreement with a linear proxy index, not with verified boundary locations. And the realised visit share is not independent of behaviour: in an episodic corridor the shortcut's share of steps is an algebraic function of the per-episode choice probability m, $\phi(m) = m/(m + (1 - m)(L - 1))$, so its slope in p necessarily shares the sign of the behavioural slope. What is exogenous is the *opportunity* for the asymmetry, set by L; the realised slope is a correlate, and its use as an ex-ante triage signal presupposes a sweep over p. This is a mechanism check, not a formula: six lengths cannot establish a functional form, the relation is not linear, and one pair inverts ($L = 12$ sits between $L = 4$ and $L = 8$ by boundary but not by measured slope). What it does establish is that the non-monotonicity is not noise — a separately measured, mechanistically motivated quantity reproduces its shape, which a pure discounting account cannot.

Since this channel drives the deployment-reversal story, it is the paper's most important theoretical gap, and we flag it rather than fit a functional form after the fact. A single multi-step violation is tested (the branching world of Appendix G); a systematic sweep of violation length against boundary displacement is not, bounding the timing diagnostic's scope.

Our environments are deliberately minimal: the mechanism is a property of constant-step-size value estimation rather than of any environment, and we verify it across a tabular bandit, a deep-RL gridworld, and in-context LLM agents. We do not contribute a simulator or benchmark, and transfer to production verification settings is untested.

The LLM results use open models up to 72B. The certainty effect is present at every scale but does not strengthen with size, and we could not replicate the steep 7B slope our earlier draft built a scaling claim on; we do not know why that first measurement differs, only that three later ones agree with each other and not with it. The valid region of the sweep is $p \leq 0.5$, which leaves the high-certainty regime untested: a design that decouples detection probability from penalty magnitude would be needed to probe it. Prompt-robustness checks are correspondingly small-sample.

Two further gaps are worth naming. The value-learner route and the LLM route are *different* mechanisms, not one unified account — the opposite-signed pair of Appendix B is evidence that they differ, and we present it that way rather than claiming a single explanation. And while the timing-equalised control isolates the immediacy channel at a single corridor length, its interaction with longer horizons is untested, and violations that themselves take several steps appear in only one environment (Appendix G), not in a parametric sweep.

S Limitations in Detail

This appendix carries the supporting detail for the four weaknesses named in Sec. 7; each paragraph there states the weakness itself.

Policy-gradient evidence, and what it does not cover. The fingerprint in Appendix A is directional evidence that the same mechanism would appear under RLVR: the effect scales with the *policy* step size while the critic step size moves the slope $21\times$ less and in the opposite direction, a sign dissociation no single shared noise source would produce. It is not a verification, because the learner is bandit-scale. The failed GRPO attempt is diagnosed in Appendix B; the next attempt we would

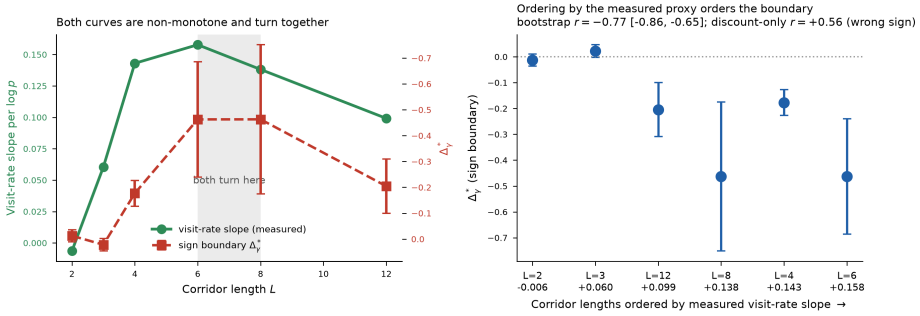


Figure 23: **The sampling-rate channel predicts where the sign boundary lands.** (A) Visit-rate slope (measured independently at each corridor length) and the sign boundary Δ^*_γ; both are non-monotone and both turn in the $L = 6$–8 region (shaded). The two boundary values there (-0.4628, -0.4630) are independent per-L fits, but both are linear extrapolations past the grid bottom (-0.22), so their near-coincidence may reflect the shared extrapolation rather than a real plateau; the ordering claim in (B) does not rest on it. (B) Ordering the corridor lengths by measured visit-rate slope orders the boundary, with one inversion. The discount-only hypothesis γ^{L-1} is monotone and predicts the opposite sign.

make is a verifiable-reward task with per-step logging of the answered rate, at an order of magnitude more steps.

The sampling-rate predictor. The asymmetry tracks the boundary at $r = -0.77$, bootstrap CI $[-0.86, -0.65]$, beating a discount-only account in 97% of draws (Fig. 23). It predicts *where* the boundary lands without explaining why the channel increases in p, which is why we present it as a triage signal in Sec. 5 rather than as a mechanism.

Why the two delayed-task magnitudes are not comparable. A noise-scale difference ($\kappa \propto E\sqrt{\alpha}$) may matter in addition to the extrapolation instability: the setups differ in effective step size (constant $\alpha = 0.1$ tabular versus Adam), the tabular sweep varied E itself, and nominal learning rates do not translate into effective ones. Only a controlled tabular α sweep at fixed E — which we have not run — could isolate it.

Separating compliance heuristics from certainty-sensitivity. The framing dependence (Appendix P) admits two readings that predict the same data here. Distinguishing them requires decoupling audit *semantics* from penalty *structure* — e.g. a neutrally-worded task whose penalty schedule is identical to the audit-framed one — which we have not run.

T THE TWO IMMEDIACY LEVERS IN DETAIL

Sec. 4.2 states the pooled result; this appendix gives the per-cell numbers. Varying the discount factor at corridor length $L = 6$: at $\gamma = 0.90$ the discounted gap is $\Delta_\gamma = +0.106$ (violation EV-optimal) *and* the violation rate rises monotonically with audit certainty ($0.37 \to 0.92$); at $\gamma = 0.99$ and $\gamma = 1.00$, $\Delta_\gamma < 0$ and the rise is gone. Varying instead the corridor length at fixed $\gamma = 0.95$ moves Δ_γ through the same range with the discount untouched, which is what makes the two levers separable manipulations of the same quantity rather than one manipulation twice. Both hold $E = 1.3$ fixed. **What sweeps what, across all sweeps.** The five original rows sweep Δ_γ via the discount γ at fixed $E = 1.3$; the two extension rows sweep it via r_c at fixed (γ, E) — and both levers give positive slopes, so the no-flip result is not an artefact of the lever. The coarser (γ, L) sweep of Sec. 4.2 uses the discount levers again (within-family, not independent validation), while the L=2 control and the tabular replication sweep via the expected penalty; since $\kappa \propto E$ scales the value noise, E-swept and value-swept families are not automatically comparable, and we do not pool them. Because both act through the discounted value gap, we treat their agreement as within-family consistency; the independent check on the lever itself is the r_c-swept extension of Sec. 4.3.

Two displacements, and the boundary's is sufficient on its own. Zero is the boundary of *neither* task — the proxy's sits at -0.007, the delayed task's beyond -0.45 — so the reversal is driven by the gap rising ($-0.070 \rightarrow +0.047$) *and* the boundary falling, each task's outcome following from its own Δ against its own boundary. The boundary's displacement alone suffices: pinning $\Delta_\gamma = -0.070$ on the delayed task still backfires, as does every gap we measured there, though the same gap deters at $L{=}2$ — reversal without Δ_γ crossing zero, which a criterion testing only for a sign change would pass as safe.