# WHEN SELF-REINFORCEMENT ENDS SELF-IMPROVEMENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Self-improvement systems use a judge to select revisions and reinforce those that pass. The quality of the judge is typically assessed through evaluation accuracy, but this metric does not capture whether errors recur on the same inputs. We show that whether the judge's errors are deterministic determines whether reinforcement ends improvement. In a bounded utility model with a deterministic binary judge and expected losses exceeding expected gains, we prove that every fixed acceptance rule that initially improves utility eventually produces negative drift. The preference for positive outputs that enables improvement also reinforces the judge's errors. For adaptive rules and judges with finitely many outputs, we derive a tradeoff between the number of revisions and cumulative negative drift, and we characterize when a suitable acceptance rule can sustain improvement. A comparison with independently refreshed errors isolates the role of deterministic errors: two judges with the same evaluation accuracy produce opposite signs of eventual drift. A sequential test measures expected utility change with sample bounds that distinguish revision frequency, magnitude, and evaluation noise. Experiments with language model judges corroborate the theoretical predictions. We identify whether errors are deterministic as a property that determines the outcome of self-reinforcement and is not captured by evaluation accuracy.

## 1 INTRODUCTION

Self-improvement systems repeatedly generate, evaluate, and select revisions to prompts, programs, or agent components (Agrawal et al., 2026; Zhang et al., 2026; Yang et al., 2024b; Romera-Paredes et al., 2024; Khattab et al., 2024). Selection affects more than the current output: accepted revisions can shape the next proposals. This feedback reinforces what the judge accepts. If the judge always mistakes a harmful revision for an improvement, the same mechanism can reinforce its error.

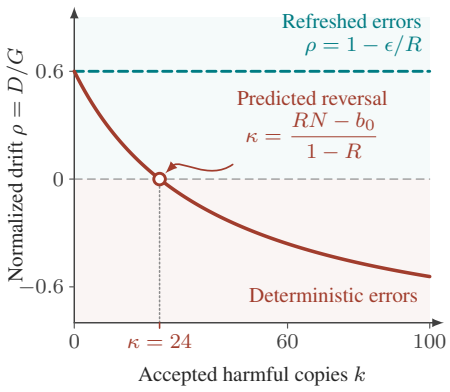


Figure 1: **Same accuracy, opposite eventual drift.** Exact predictions under $A = S$ and copying (Prop. 1 and 2), with $N = 40$ harmful identities, $b_0 = 8$ errors ($\epsilon = b_0/N$), and gain-to-loss ratio $R = 1/2$. Deterministic errors give $\rho = 1 - \frac{b_0+k}{R(N+k)}$: zero at 24 copies, negative from 25. Refreshed errors keep drift positive. Curves connect integer copy counts.

We study when this self-reinforcement ends self-improvement. Accuracy on a fixed evaluation set does not determine the answer. Accuracy measures how often a judge errs, but reinforcement also depends on whether the judge repeats its errors on the same inputs. We address this question in a bounded utility model that isolates the role of deterministic errors. A deterministic judge always misclassifies the same inputs, and acceptance copies these errors into the proposal pool, increasing their frequency over time. We call the expected utility change at each round the drift. We observe that initial improvement itself determines which errors are reinforced. When expected losses exceed expected gains, any improving rule must accept the judge's positive outputs more often than its negative ones. But the same positive output that carries beneficial proposals also carries the deterministic



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.

errors. Acceptance therefore increases the share of errors in the proposal pool, and every fixed rule that initially improves utility eventually produces negative drift (Thm. 1 and Cor. 1). The preference that enables improvement is the same preference that reinforces the errors ending it.

More generally, we allow the acceptance rule to depend on history and the judge to produce any finite set of outputs. We derive a tradeoff between the number of revisions and cumulative negative drift (Thm. 2). If every judge output that can arise from a beneficial proposal can also arise from a harmful one, no acceptance rule can selectively reinforce improvements, and infinitely many revisions require divergent cumulative negative drift. The condition is tight: when some output is exclusive to beneficial proposals, a suitable rule sustains improvement with positive drift. Negative drift concerns the conditional expected change and need not force every trajectory to zero.

We derive a threshold on the number of copies at which drift turns negative, and compare with independently refreshed errors to clarify the roles of deterministic errors and copying. The two judges have the same evaluation accuracy at every round, yet produce opposite signs of eventual drift (Prop. 1 and 2; Fig. 1). Independent utility evaluations can test the drift at a fixed state without altering the reinforcement dynamics. Our sequential procedure provides sample bounds that distinguish revision frequency, magnitude, and evaluation noise (Thm. 3 and 5). Experiments with synthetic revision processes and language model judges corroborate these predictions. When the judge's deterministic errors concentrate within a reinforced class, the benefit of self-improvement reverses. These results show that evaluation accuracy alone does not predict whether self-reinforcement will end improvement. Whether errors are deterministic is the distinguishing property.

**Related work.** Imperfect verification limits repeated sampling and repair (Stroebl et al., 2026; Wu et al., 2026), and reward-model accuracy can fail to predict optimization behavior (Razin et al., 2025; Shang et al., 2026). Our analysis connects evaluation to the distribution changes induced by acceptance, relating to reward overoptimization and hacking (Gao et al., 2023; Skalse et al., 2022), feedback in recommendation (Jiang et al., 2019), and error amplification in self-evolving agents (Lin et al., 2026). The mechanism uses reinforced urns (Pemantle, 2007), and the statistical analysis uses empirical Bernstein bounds and sequential inference (Maurer and Pontil, 2009; Howard et al., 2021; Kaufmann et al., 2016). App. K provides a more detailed comparison.

## 2 Self-Reinforcement of Judge Errors

Let $V : \mathcal{Z} \to [0, 1]$ be utility. At round $t$, the current state is $z_t$, a proposal $z'_t$ is drawn, and an acceptance rule chooses $A_t \in \{0, 1\}$. Set $z_{t+1} = z'_t$ if $A_t = 1$, and $z_{t+1} = z_t$ otherwise. The pre-proposal history $\mathcal{F}_t$ includes the current state and template pool. Environment parameters, including a deterministic judge, are fixed when conditioning. A freshly randomized judge is sampled after $\mathcal{F}_t$. Write $d_t = V(z'_t) - V(z_t)$, $d_{t,+} = \max(d_t, 0)$, and $d_{t,-} = \max(-d_t, 0)$.

**Definition 1** (Expected utility change)**.** Let $G_t, H_t$ be the expected proposal gain and loss, and $D_t$ the conditional expected change after acceptance. When $G_t, H_t > 0$, let $\tau_t, \phi_t$ be the gain- and loss-weighted acceptance rates and $R_t$ the gain-to-loss ratio:

$$\begin{aligned} G_t &= \mathbb{E}[d_{t,+} \mid \mathcal{F}_t], & H_t &= \mathbb{E}[d_{t,-} \mid \mathcal{F}_t], & D_t &= \mathbb{E}[A_t d_t \mid \mathcal{F}_t], \\ \tau_t &= \frac{\mathbb{E}[A_t d_{t,+} \mid \mathcal{F}_t]}{G_t}, & \phi_t &= \frac{\mathbb{E}[A_t d_{t,-} \mid \mathcal{F}_t]}{H_t}, & R_t &= G_t / H_t. \end{aligned} \tag{1}$$

For $G_t > 0$, the normalized expected utility change is $\rho_t = D_t / G_t$.

The quantity $D_t$ is the conditional *drift* of utility. Splitting the proposed change into its positive and negative parts gives both the drift identity and the condition for improvement:

$$D_t = G_t \tau_t - H_t \phi_t, \qquad \rho_t = \tau_t - \frac{\phi_t}{R_t}, \qquad D_t > 0 \iff \phi_t < R_t \tau_t. \tag{2}$$

The acceptance rates weight revisions by gain or loss magnitude. The value $\rho_t = 1$ corresponds to accepting every beneficial proposal and rejecting every harmful one. We hold the gain-to-loss ratio fixed to isolate reinforcement.

### 2.1 MODEL

The current state has utility $x \in (0,1)$. Set $s(x) = x(1-x)$ and fix $g, h \in (0,1)$, $\omega \in (0,1]$, and $p \in (0,1)$. Independent round draws select a null proposal $(d, S) = (0, 0)$ with probability $1-\omega$, a beneficial proposal $(d, S) = (gs(x), 1)$ with probability $\omega p$, and a harmful proposal with $d = -hs(x)$ with probability $\omega(1-p)$. For a harmful proposal, sample a template uniformly with multiplicity from a finite pool. The judge returns $S = 1$ on a fixed error set of identities.

Initially the pool contains one copy of each of $N$ harmful identities, $b_0$ of which are misclassified, with $1 \le b_0 < N$ and $\epsilon = b_0/N$. Let $Q_t$ be their current fraction, equivalently the judge's false-positive rate on harmful proposals. The acceptance rule observes the current output $S$ and past history, but not the current family, identity, or utility. All finite revisions keep $x \in (0,1)$. We work in the initially improving regime

$$G_t = \omega p g s(x_t), \qquad H_t = \omega(1-p)hs(x_t), \qquad \epsilon < R = \frac{pg}{(1-p)h} < 1. \tag{3}$$

The rule $A = S$ therefore initially improves utility, while accepting all proposals gives negative drift. Family probabilities remain fixed. Only the multiplicities of harmful templates change.

### 2.2 ERROR REINFORCEMENT

We first allow a sampled harmful template with output $S = j$ to be copied with fixed probability $r_j \in [0,1]$, independently across rounds. These are the only pool updates. Copying need not imply acceptance. Under $A = S$, copying accepted harmful templates gives $(r_1, r_0) = (1, 0)$.

**Theorem 1** (Concentration of misclassified templates)**.** *Under the copying rule above, let $q^{[k]}$ be the fraction of misclassified templates after $k$ pool additions. If $r_1 > r_0$, then almost surely*

$$Q_t \longrightarrow 1, \qquad \lim_{k\to\infty} \frac{\log(1-q^{[k]})}{\log k} = -\left(1 - \frac{r_0}{r_1}\right). \tag{4}$$

*Under $A = S$ and $R < 1$, there is an almost surely finite time after which $\rho_t = 1 - Q_t/R < 0$ at every subsequent round. The misclassified fraction can decrease before that time when $r_0 > 0$.*

*Proof sketch.* With misclassified and correctly rejected counts $n_1, n_0$, the next addition is misclassified with probability $r_1 n_1/(r_1 n_1 + r_0 n_0)$. Independent linear birth processes with rates $r_1, r_0$ embed this chain and give the limit and rate. Every finite addition count is reached in finitely many rounds. App. A also treats equal and reversed rates.

Now consider a fixed randomized acceptance rule $\pi_j = \pi(j)$, with fresh independent randomization, and copy each accepted harmful template once. No other operation changes the pool. The two copying rates are $r_1 = \pi_1, r_0 = \pi_0$.

**Corollary 1** (Eventual negative drift under fixed acceptance rules)**.** *Suppose $R < 1$ and a fixed acceptance rule using only $S$ has $D_0(\pi) > 0$. Then*

$$\pi_1 > \pi_0, \qquad Q_t \longrightarrow 1, \qquad \frac{D_t(\pi)}{H_t} \longrightarrow \pi_1(R-1) < 0 \quad \textit{almost surely}. \tag{5}$$

*Thus $D_t(\pi) < 0$ at every sufficiently late round, almost surely.*

*Proof.* When the fraction of misclassified templates is $Q_t$,

$$\frac{D_t(\pi)}{H_t} = \pi_1(R - Q_t) - \pi_0(1 - Q_t) = \pi_1(R-1) + (\pi_1 - \pi_0)(1 - Q_t). \tag{6}$$

If $\pi_1 \le \pi_0$, both terms in the last expression are nonpositive, contradicting initial improvement. Apply Thm. 1 with $r_j = \pi_j$ and take the limit. □

Thus initial improvement determines the copying bias that eventually produces negative drift.

### 2.3 Copy Threshold and Refreshed Errors

Under $A = S$, every copied harmful template is misclassified. Let $K_t$ count copies after $t$ completed rounds, with $K_0 = 0$.

**Proposition 1** (Threshold for negative expected utility change)**.** *At every finite history, $Q_t = \phi_t = q_{K_t}$ and $\tau_t = 1$, where*

$$q_k = \frac{b_0 + k}{N + k}, \qquad \rho_t = 1 - \frac{q_{K_t}}{R}, \qquad \kappa = \frac{RN - b_0}{1 - R}. \tag{7}$$

*The conditional expected utility change is positive for $K_t < \kappa$, nonpositive from $k_{\geq} = \lceil \kappa \rceil$ copies onward, and negative from $k_{>} = \lfloor \kappa \rfloor + 1$ copies onward. Both counts are reached almost surely in finitely many rounds, because each copy occurs after a finite geometric wait. The fraction $q_k$ strictly increases, so nonpositive drift cannot recover under the rule $A = S$.*

For example, considering $N = 40$, $b_0 = 8$, $p = 1/3$, and $g = h$, the ratio is $R = 1/2$: drift reaches zero after 24 copies and becomes negative after the next. The waiting time between copies is geometric with parameter $\omega(1 - p)q_k$. App. B.2 gives the crossing-time law.

We compare the deterministic judge, whose error set is drawn once uniformly among $b_0$-subsets, with one that draws a fresh independent subset before each decision. Both use $A = S$ and copy accepted harmful templates. Evaluate them on $N$ beneficial probes and one harmful probe per identity, using one error set for all probes in each evaluation.

**Proposition 2** (Evaluation accuracy under refreshed errors)**.** *At each fixed round, the two judges have the same marginal distribution of the full prediction vector on the evaluation set, with accuracy $1 - \epsilon/2$. The refreshed judge has $\tau_t = 1$, $\phi_t = \epsilon$, and $\rho_t = 1 - \epsilon/R > 0$ at every finite history. The deterministic judge instead follows Prop. 1.*

The fresh error set contains any current harmful identity with probability $\epsilon$, independently of its multiplicity. The prediction laws agree marginally at each fixed round. Under the deterministic judge, prediction vectors repeat across rounds, while refreshed vectors are independent. In the numerical example, both judges have 90% evaluation accuracy. Deterministic errors therefore change the sign of improvement while preserving this evaluation law.

## 3 Adaptive Rules and Sequential Testing

Changing the acceptance rule does not remove the feedback through copying. We now allow the judge to return any of $J$ outputs. We retain the proposal probabilities, magnitudes, and $R < 1$ from Sec. 2.1, starting from $x_0 \in (0, 1)$. A beneficial proposal's output has a fixed law $\nu = (\nu_1, \ldots, \nu_J)$, independently of history. Each harmful identity has a fixed output. Let $n_{j,t}$ count pool entries with output $j$ and $n_t = \sum_j n_{j,t}$. Fix initial counts $n_{j,0} > 0$ for every output, excluding outputs that arise under neither family. Null proposals cause no change. Copy every accepted harmful template once, with no other pool updates. The rule chooses predictable probabilities $\pi_{t,j}$ from past history, then accepts output $j$ with that probability using fresh independent randomization.

Set $N = n_0$, $I_t = \mathbf{1}\{A_t d_t \neq 0\}$, $M_T = \sum_{t<T} I_t$, and $\eta = (1 - p)h - pg > 0$. For harmonic numbers $\mathsf{H}_k = \sum_{i=1}^{k} 1/i$, $\mathsf{H}_0 = 0$, define

$$\Phi_t = \mathsf{H}_{n_t - 1} - \sum_j \nu_j \mathsf{H}_{n_{j,t} - 1} + \sum_j \nu_j \log \nu_j \geq D_{\mathrm{KL}}(\nu \| n_{\cdot,t}/n_t) \geq 0, \tag{8}$$

with $0 \log 0 = 0$. The harmonic numbers give a discrete increment in place of a logarithmic approximation.

**Theorem 2** (Revision count and cumulative negative drift)**.** *For every acceptance policy in this finite-output model and every stopping time $T$, writing $M_\infty = \sum_{t \geq 0} I_t$,*

$$\eta \mathbb{E} \log\left(1 + \frac{M_T}{N}\right) \leq p(g + h)\Phi_0 + \mathbb{E} \sum_{t<T} \frac{(D_t)_-}{n_t s(x_t)}, \tag{9}$$

$$\mathbb{P}\left(M_\infty = \infty, \quad \sum_{t \geq 0} \frac{(D_t)_-}{n_t s(x_t)} < \infty\right) = 0. \tag{10}$$

*In particular, eventually nonnegative drift allows only finitely many revisions almost surely. After the last revision, utility remains at an interior value with positive expected gain from proposals.*

*Proof sketch.* Copying an output-$j$ template changes $\Phi_t$ by $1/n_t - \nu_j/n_{j,t}$. Conditional expectation and summation, using $\Phi_T \geq 0$, give

$$\mathbb{E}[\Phi_{t+1} - \Phi_t \mid \mathcal{F}_t] = -\frac{\eta\mathbb{E}[I_t \mid \mathcal{F}_t] + D_t/s(x_t)}{p(g+h)n_t}, \tag{11}$$

$$\eta\mathbb{E}\sum_{t<T}\frac{I_t}{n_t} + \mathbb{E}\sum_{t<T}\frac{(D_t)_+}{n_t s(x_t)} \leq p(g+h)\Phi_0 + \mathbb{E}\sum_{t<T}\frac{(D_t)_-}{n_t s(x_t)}. \tag{12}$$

For a deterministic horizon, $n_t \leq N + M_t$ gives $\sum_{t<T} I_t/n_t \geq \sum_{k=0}^{M_T-1}(N+k)^{-1} \geq \log(1 + M_T/N)$, proving Eq. (9). Truncation extends the bound to stopping times. For the pathwise claim, each reciprocal count appears in a potential increment only when that count increases. The squared increments therefore have a deterministic summable bound, so the centered martingale converges. Finite normalized negative drift then forces finitely many revisions. App. A.4 gives the full argument. The cost in Eq. (9) measures normalized conditional drift. The bound controls a logarithmic moment of the revision count.

The support condition is necessary. If an output has $\nu_j > 0$ but $n_{j,0} = 0$, accepting it alone yields infinitely many beneficial revisions almost surely, with $D_t = \omega pg\nu_j s(x_t) > 0$ at every round. Together with Thm. 2, this characterizes when sustained improvement is possible under the stated copying rule and $R < 1$. For the binary model, $\nu = (1, 0)$ and $\Phi_0 = \mathsf{H}_{N-1} - \mathsf{H}_{b_0-1}$.

**Optimal acceptance and recovery.** We return to the binary model, whose current decision limit follows from Eq. (6).

**Corollary 2** (Decision limit from the binary output)**.** *At any finite template pool with a deterministic judge and misclassified fraction $Q_t$, the best rule using only $S$ has expected change $\sup_\pi D_t(\pi) = H_t(R - Q_t)_+$. Thus positive expected change is impossible at the current state when $Q_t \geq R$, which under the deterministic judge occurs at $K_t \geq k_{\geq}$.*

This is a standard Bayesian decision problem. App. D extends it to general signals. Observing the current proposal family would change the information restriction and allow positive drift. With only $S$, maintaining nonnegative drift after $Q \geq R$ prevents recovery (Prop. 5). Allowing temporary losses permits dilution: from $n$ entries with misclassified fraction $Q \geq R$, adding $r$ correctly rejected templates and $e$ misclassified templates restores positive drift under $A = S$ exactly when

$$\frac{nQ + e}{n + e + r} < R \quad \Longleftrightarrow \quad r > n(Q/R - 1) + (1/R - 1)e. \tag{13}$$

Each correctly rejected copy requires a harmful acceptance. App. C gives the full recovery statement and the utility loss of this strategy.

**Expected improvement and long-term utility.** Under the deterministic judge with $A = S$, utility converges almost surely to $x_\infty \in \{0, 1\}$ (Prop. 10). Its eventual negative drift does not by itself determine which endpoint occurs. Set $\theta_t = \log[x_t/(1 - x_t)]$ and define

$$L_0 = p\log(1+g) + (1-p)\log(1-h) < 0, \qquad L_1 = -p\log(1-g) - (1-p)\log(1+h). \tag{14}$$

**Proposition 3** (Utility limit classification)**.** *Under the deterministic judge with $\epsilon < R < 1$, $\theta_t/t \to \omega L_{x_\infty}$ almost surely. If $L_1 < 0$, then $x_t \to 0$ almost surely. If $L_1 > 0$, then $0 < \mathbb{P}(x_\infty = 1) < 1$ for every interior initial utility and every initial pool in the stated regime. The rate also holds when $L_1 = 0$, but the endpoint probabilities at that critical value are not classified here.*

*Proof sketch.* The logit increments are bounded. Their centered sums obey a martingale law of large numbers, and their conditional means converge to $\omega L_j$ on each endpoint event. When $L_1 > 0$, fractional-moment supermartingales show that either endpoint can retain a trajectory with positive probability. App. H gives the proof.

### 3.1 TESTING THE DRIFT

Independent utility evaluation can identify the current drift while leaving the reinforcement mechanism unchanged. Fix the current state, proposal distribution, pool, and judge protocol. Draw independent proposals without updating the current state or the pool. All expectations below condition on this state. For a target $a \in [0, 1)$, set $X = Ad - ad_+$ and observe i.i.d. $Y = X + \xi$ using an independent utility evaluator. Assume a known bound $|d| \le b$, $0 < b \le 1$, and noise $\xi$ independent of $X$, centered, and $\sigma^2$-sub-Gaussian for known $\sigma$.

The mean measures improvement, while $X_+, X_- \le b$ and $\mathbb{E}X_+ \le G$ control the variance:

$$\mu = \mathbb{E}Y = \mathbb{E}X = D - aG = G(\rho - a), \tag{15}$$

$$\mathrm{Var}(X) \le \mathbb{E}X^2 \le b(\mathbb{E}X_+ + \mathbb{E}X_-) = b(2\mathbb{E}X_+ - \mu) \le 2bG - b\mu. \tag{16}$$

For $G > 0$, the sign of $\mu$ therefore answers whether $\rho > a$. Setting $a = 0$ tests positive drift. Near zero mean, the latent variance scale is $bG$, and $\mathrm{Var}(Y) \le \mathrm{Var}(X) + \sigma^2$. Lem. 1 gives the asymmetric bound. Setting $\sigma = 0$ gives exact evaluation. A smaller $b$ must follow from the revision model, and for $a > 0$, evaluating $d_+$ requires separate justification: an unbiased estimate of $d$ alone is insufficient. App. E and App. I.4 detail the observation protocols and conditional-noise extension.

**Sequential test.** Inspect $n_k = 2^{k+1}$ observations, $k = 0, 1, \ldots$, with $\delta_k = 6\delta/[\pi^2(k+1)^2]$ and $\ell_k = \log(6/\delta_k)$. Skip decisions when $n_k < \ell_k$. Otherwise, form

$$I_k = [\bar{Y}_{n_k} - r_k, \bar{Y}_{n_k} + r_k], \qquad r_k = \sqrt{\frac{2\ell_k}{n_k}}(s_{Y,n_k} + 5\sigma) + \frac{7b\ell_k}{n_k}. \tag{17}$$

Here $\bar{Y}_n$ and $s_{Y,n}$ are the sample mean and standard deviation. Stop when $I_k$ excludes zero and report its sign. An exhausted finite budget returns *insufficient evidence*.

**Theorem 3** (Error control and sample complexity)**.** *Let $G > 0$, $a \in [0, 1)$, and $0 < \delta \le 1/8$. Given $b$ and $\sigma \ge 0$, the procedure reports an incorrect sign of $\rho - a$ with probability at most $\delta$. It uses neither $G$ nor a separation margin. If $|\rho - a| \ge m$ for $0 < m \le 1$ with $a + m \le 1$, the uncapped procedure stops almost surely. Its expected sample count and a $1 - \delta$ upper bound on that count are both at most*

$$C\Lambda_\sigma\{\log(1/\delta) + \log\log(e^e\Lambda_\sigma)\}, \qquad \Lambda_\sigma = \frac{bG + \sigma^2}{G^2m^2}, \tag{18}$$

*with universal $C$. For exact evaluation, this reduces to $\Lambda_0 = b/(Gm^2)$.*

*Proof sketch.* The inequality $s_X \le s_Y + s_\xi$ lets sample-variance concentration (Maurer and Pontil, 2009) control the latent variance using observed $s_Y$ and a noise bound. Mean concentration and a union bound give simultaneous coverage at inspected stages. A stopping-time tail bound gives both sample guarantees (App. E).

For $0 < G \le b/16$, $a \le 1/2$, and $m, \delta \le 1/8$, every uniformly correct sequential procedure has worst-case expected count at least $c\Lambda_\sigma \log(1/\delta)$ over bounded revision laws, even with known $G, b$ (Thm. 5). A fixed sample of size $C\Lambda_\sigma \log(1/\delta)$ suffices with known $G, m$. The noisy lower bound allows acceptance and identity logs but not exact $d$. It does not establish necessity of the iterated logarithm.

The two terms have different causes. In one pair of laws, harmful changes of size $b/2$ are accepted with probability of order $G/b$. Separating these probabilities by order $(G/b)m$ gives per-record KL divergence $O((G/b)m^2)$. For $\sigma > 0$, a second pair always accepts and changes one identity's utility increment by $4Gm$ under Gaussian noise, giving divergence $4G^2m^2/\sigma^2$. The stopped change-of-measure inequality converts each divergence into an expected sample lower bound. The larger bound is within a factor of two of their sum. App. F gives both constructions.

**Rare revisions and small revision magnitudes.** For the model of Sec. 2.1, take $a = 0$ and $b = \max(g, h)s(x)$. Then $b/G = \max(g, h)/(\omega pg)$, so shrinking all changes leaves the noiseless sample bound unchanged. For fixed interior $p, R, g, h$, the fixed-sample minimax complexity of distinguishing $q \le R - Rm$ from $q \ge R + Rm$ is $\Theta((\omega m^2)^{-1}\log(1/\delta))$. With fixed additive noise, the upper bound contains $\sigma^2/(\omega^2 s(x)^2 m^2)$. Matching noiseless bounds for this mixture are in Prop. 8. The matching noisy lower bound concerns the broader revision class. Thus rare nonzero

proposals increase the noiseless sample cost, while uniformly smaller changes increase the cost when evaluation noise remains fixed. If all structural parameters are known, the sign can be calculated directly. App. G.3 relates statistical uncertainty near zero drift to the remaining crossing time.

**Counting utility evaluations.** Each observation above counts one sampled proposal. For $a = 0$, rejection has zero utility change, so one can determine acceptance first and evaluate only accepted proposals. Suppose $|D| \geq Gm$, $\alpha = \mathbb{P}(A = 1) > 0$, and fresh noise is centered and $\sigma^2$-sub-Gaussian conditional on the selected proposal. With supplied valid bounds, Prop. 9 allows a fixed evaluator budget $n_{\text{eval}}$, whose proposal cost $n_{\text{prop}}$ satisfies

$$n_{\text{eval}} = \left\lceil C \frac{bG\alpha + \sigma^2\alpha^2}{G^2 m^2} \log \frac{1}{\delta} \right\rceil, \qquad \mathbb{E}n_{\text{prop}} = \frac{n_{\text{eval}}}{\alpha}. \tag{19}$$

For the model of Sec. 2.1 at fixed interior parameters, this is $O((1 + \sigma^2/s(x)^2)m^{-2}\log(1/\delta))$ evaluator calls. Rarity increases screening work. Fixed noise makes small changes costly to evaluate.

## 4 Experiments

We provide empirical support for three predictions: **(i)** deterministic errors under copying reverse improvement at the threshold of Prop. 1, while refreshed errors with the same evaluation accuracy do not; **(ii)** adaptive rules trade revision count against cumulative negative drift unless the judge has an output absent from the harmful pool (Thm. 2); and **(iii)** the sample cost of testing drift separates rare revisions from uniformly small changes (Thm. 3). Sec. 4.1 verifies these on the synthetic processes, and Sec. 4.2 replays the loop on published language model responses with a Qwen3-8B judge (Yang et al., 2025). See App. J for additional details.

### 4.1 Synthetic Revision Processes

**Setting.** We simulate the binary model of Sec. 2.1 with $N = 40$ harmful identities, initial error fraction $\epsilon = 0.2$, and gain-to-loss ratio $R = 1/2$, with acceptance rule $A = S$. We compare four conditions that vary the judge type (deterministic or refreshed) and template copying (copying or a fixed pool) across 512 replicates of 4,096 rounds. All conditions have evaluation accuracy 0.9 at every round and differ only in whether errors repeat across rounds. See App. J.1 for details.

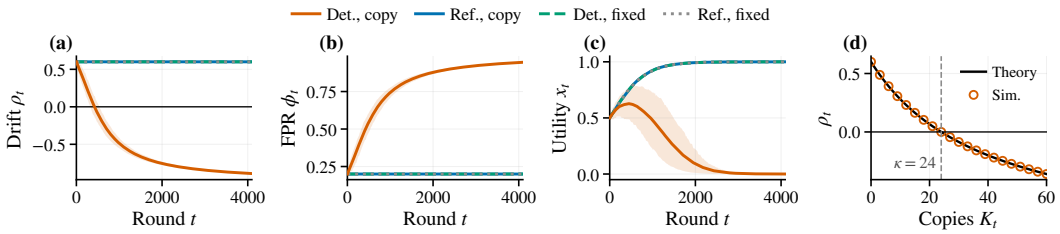


Figure 2: **Deterministic errors and copying reverse improvement at the same evaluation accuracy.** Det./Ref.: deterministic/refreshed; copy/fixed: copying/fixed pool. (a) Normalized drift, (b) false-positive rate (FPR) $\phi_t$, and (c) utility over 512 replicates; shading is the 10–90% range for deterministic copying. The other three conditions coincide. (d) Drift versus copies and the exact threshold of Prop. 1.

**Only deterministic errors under copying end improvement.** Of the four conditions, only the one combining deterministic errors with copying reaches negative drift (Fig. 2a). Its drift crosses zero at exactly 24 copies and turns negative at 25, in agreement with the threshold $\kappa$ of Prop. 1 (Fig. 2d). The harmful acceptance rate rises to nearly one and utility drops to near zero, while the other three conditions maintain $\rho_t = 1 - \epsilon/R = 0.6$ and converge to utility near one (Fig. 2b,c). Consistent with Prop. 2, neither deterministic errors nor copying alone ends improvement.

**Adaptive acceptance.** We give the judge $J = 3$ outputs and compare three acceptance rules over 65,536 rounds (Fig. 3a,b). The first maximizes acceptance subject to $D_t \geq 0$: it makes only 77 revisions before terminating, despite positive expected gain at every round. The second

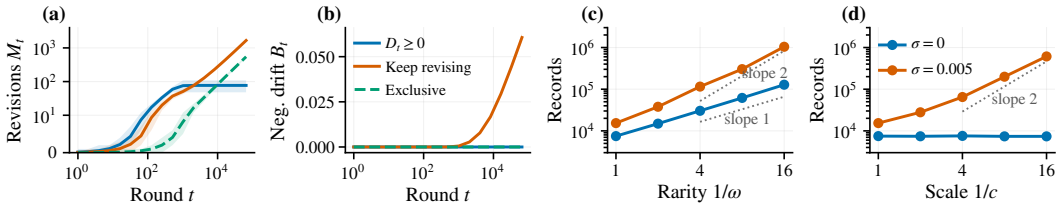


Figure 3: **Adaptive acceptance trades revisions against negative drift.** (a) Revisions $M_t$ and (b) cumulative normalized negative drift $B_t$ (10–90% ranges). Rules enforce $D_t \geq 0$, keep revising, or accept an output exclusive to beneficial proposals. (c,d) Mean records to a decision as revision frequency $\omega$ or change magnitude $c$ decreases, with noise level $\sigma$; dotted guides have slopes 1 and 2.

maximizes drift subject to a minimum acceptance rate: it sustains over a thousand revisions but only by accumulating divergent normalized negative drift, and utility collapses. The third accepts only an output absent from the harmful pool: it sustains revisions with zero negative drift and reaches utility near one. These results are consistent with Thm. 2: under full output overlap, sustained revision requires cumulative negative drift, while an output absent from the harmful pool allows sustained improvement.

**Testing the drift.** At a fixed state, we run the sequential test of Sec. 3.1 over a grid of revision frequencies $\omega$ and change magnitudes $c$, with and without evaluation noise ($\sigma \in \{0, 0.005\}$). No run reports an incorrect sign. Without noise, the sample cost grows linearly with $1/\omega$ but remains unchanged as all changes shrink (Fig. 3c,d), consistent with the scaling of $\Lambda_0$ in Eq. (18): the noiseless cost scales with revision frequency but does not depend on change magnitude. Fixed noise makes both reductions costly. Evaluating only accepted proposals keeps evaluator calls near $5 \times 10^3$ at every $\omega$, in line with Prop. 9.

### 4.2 Experiments with Language Model Judges

**Setting.** We replay the revision loop on released responses without generating any. Each round draws a question and a proposal from the current distribution; a Qwen3-8B judge (non-thinking, greedy decoding) decides acceptance, and accepted proposals update the distribution. Cached verdicts make the judge deterministic, as in Sec. 2.1. We compute $D_t$ exactly over the finite candidates; a *same-state reset* recomputes it after restoring the initial proposal distribution at every fixed state. Results are averaged over 64 random seeds. Across all fixed-state evaluations, the sequential test never reports an incorrect sign. See App. J.2 for details.

**Reinforcing a style (RM-Bench).** Each RM-Bench prompt (Liu et al., 2025) has a chosen and a rejected response in concise, detailed, and detailed markdown styles, with utility 1 and 0. Each question starts with one of the two concise responses, split equally. A proposal is the concise chosen response with probability 0.3, and otherwise a rejected response whose style is drawn from counts that update on acceptance; a baseline holds the counts fixed. A *symmetric* control reinforces styles for both labels equally, decoupling style from the correctness label.

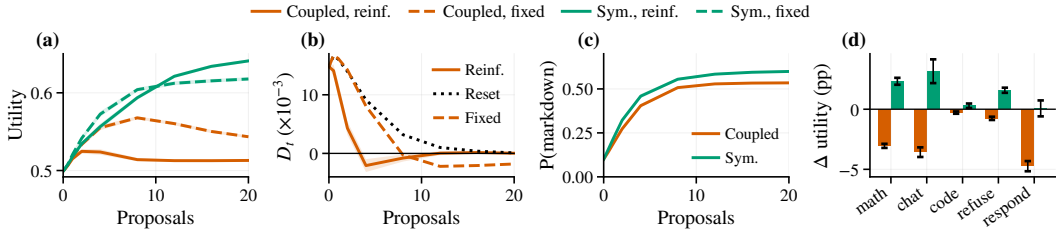


Figure 4: **Reinforcing a style that carries the judge's errors ends improvement on RM-Bench.** (a) Math utility: reinforcement (reinf., solid) or fixed weights (dashed), in the coupled design and symmetric control (sym.). (b) Drift with reinforcement, a same-state reset, or fixed weights. (c) Markdown probability. (a–c) Proposals per question. (d) Final utility difference (reinforcement minus fixed), in percentage points (pp), on every subset. 64 seeds; 95% intervals.

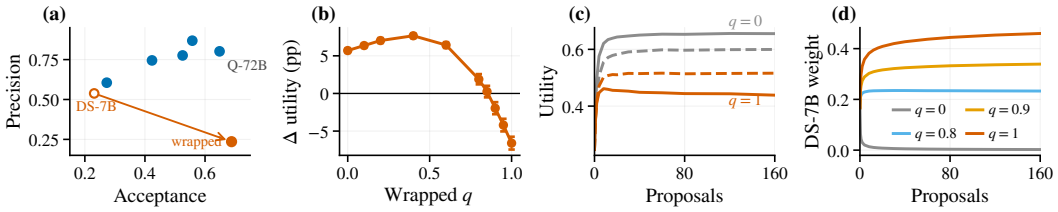


Figure 5: **Persuasion toward a weak source reverses the benefit of reinforcement on JETTS MATH.** (a) Acceptance and precision; the suffix moves DeepSeek-Math-7B (DS-7B) from the open to the filled marker. (b) Final utility difference (reinforcement minus fixed), in percentage points (pp), versus persuasion fraction $q$. (c) Utility: reinforcement (solid), fixed weights (dashed). (d) DS-7B weight under reinforcement, initially $1/6$. (c,d) Proposals per problem. 64 seeds.

**Style reinforcement amplifies a fixed preference.** On math, the judge more often prefers the rejected markdown response to the concise chosen one than the reverse. Reinforcement raises the markdown probability from 0.10 to 0.53 and causes drift to turn negative twice as early as under fixed weights (Fig. 4a,c). Restoring the initial distribution at the same state recovers nonnegative drift at 75% of negative checkpoints (Fig. 4b), suggesting that the sign reversal is due to the distribution shift. The symmetric control amplifies the same style even further, yet reinforcement *increases* utility in this case: errors are no longer concentrated within the reinforced class. On all five subsets, coupled reinforcement lowers final utility and the symmetric control never does (Fig. 4d), consistent with Prop. 2: harm requires errors that repeat within the reinforced class.

**Reinforcing a source (JETTS).** JETTS (Zhou et al., 2025) provides responses from six models on 1,324 MATH problems (Hendrycks et al., 2021) with correctness labels. The judge sees one response and outputs YES or NO; acceptance adds to the proposing model's weight. Fixed persuasive text can bias LLM judges (Raina et al., 2024; Hwang et al., 2025; Zhao et al., 2025); we append a persuasive suffix to every response of the weakest model (DeepSeek-Math-7B, accuracy 0.17) on a fraction $q$ of problems. The suffix raises the judge's YES rate on incorrect responses from 23% to 76%.

**Persuasion toward a weak source reverses the benefit of reinforcement.** At $q = 1$, the suffix makes DeepSeek-Math-7B the most accepted yet least precise model: only 24% of its accepted responses are correct (Fig. 5a). In line with Cor. 1, acceptance favors the class that carries the deterministic errors. Without persuasion, reinforcement adds 5.7 points by shifting weight toward accurate models. The gain persists while $q \leq 0.8$ and reverses beyond $q = 0.85$, reaching $-6.6$ points at $q = 1$ (Fig. 5b). With more proposals per problem, utility under reinforcement peaks early and then declines, while fixed weights continue improving (Fig. 5c). The weight of DeepSeek-Math-7B grows from $1/6$ to nearly half, and the same-state reset raises drift, consistent with the reversal being due to the distribution shift (Fig. 5d).

## 5 Conclusion and Discussion

Evaluation accuracy measures how often a judge errs, but does not capture whether errors recur on the same inputs. We showed that this distinction determines the outcome of self-reinforcement. Under a deterministic binary judge with copying, the preference for positive outputs that enables initial improvement also reinforces the judge's errors, and every fixed rule eventually produces negative drift (Thm. 1 and Cor. 1). A comparison with refreshed errors at the same accuracy produces opposite eventual drift (Prop. 1 and 2), isolating the role of deterministic errors. For adaptive rules and finite-output judges, sustained revision requires divergent cumulative negative drift unless some output is exclusive to beneficial proposals (Thm. 2). A sequential test measures drift at a fixed state with sample bounds that separate revision frequency, magnitude, and noise (Thm. 3 and 5). Experiments with language model judges corroborate these predictions: reinforcing a style or source that carries repeated errors reverses the benefit of self-improvement, while decoupling errors from the reinforced class restores it. These results suggest that judge evaluation should account for the determinism of errors. The model assumes fixed proposal-family probabilities and a finite output space. Extending the analysis to evolving judges and developing diagnostics for deterministic errors in deployed systems are directions for future work.

## AI USE STATEMENT

Generative AI tools were used for language polishing, assisting with experimental code, and checking proof details. The research design, theoretical analysis, interpretation of results, and preparation of the final content were done by the authors.

## ETHICS STATEMENT

This work provides a theoretical and empirical analysis of self-reinforcement of judge errors. We are not aware of ethical concerns specific to this study.

## REPRODUCIBILITY STATEMENT

Full proofs are given in Sec. A–H. Experimental details, including configurations, dataset descriptions, prompts, and additional results, are provided in App. J.

## REFERENCES


Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alex Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective Prompt Evolution Can Outperform Reinforcement Learning. In *International Conference on Learning Representations*, volume 2026, pages 8479–8565, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0e9e708b6f48e14fd0ac29e167413f76-Paper-Conference.pdf.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-Consuming Generative Models Go MAD. In *International Conference on Learning Representations*, volume 2024, pages 53581–53608, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ebc042e767de551803ccfcc45e2454f5-Abstract-Conference.html.

David Blackwell. Equivalent Comparisons of Experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. https://doi.org/10.1214/aoms/1177729032. URL https://doi.org/10.1214/aoms/1177729032.

Ben Chugg and Aaditya Ramdas. Closed-form empirical Bernstein confidence sequences for scalars and matrices. arXiv preprint arXiv:2512.21300, 2025. URL https://arxiv.org/abs/2512.21300.

Wenlong Deng, Jiaji Huang, Kaan Ozkara, Yushu Li, Christos Thrampoulidis, Xiaoxiao Li, and Youngsuk Park. Directional Alignment Mitigates Reward Hacking in Reinforcement Learning for Language Models. arXiv preprint arXiv:2605.25189, 2026. URL https://arxiv.org/abs/2605.25189.

Leo Gao, John Schulman, and Jacob Hilton. Scaling Laws for Reward Model Overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. https://doi.org/10.1214/20-aos1991. URL https://doi.org/10.1214/20-AOS1991.

Yerin Hwang, Dongryeol Lee, Taegwan Kang, Yongil Kim, and Kyomin Jung. Can You Trick the Grader? Adversarial Persuasion of LLM Judges. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14632–14651. Association for Computational Linguistics, 2025. https://doi.org/10.18653/v1/2025.findings-emnlp.790. URL https://aclanthology.org/2025.findings-emnlp.790/.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, pages 423–439. PMLR, 2014. URL https://proceedings.mlr.press/v35/jamieson14.html.

Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate Feedback Loops in Recommender Systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 383–390. ACM, 2019. https://doi.org/10.1145/3306618.3314288. URL https://doi.org/10.1145/3306618.3314288.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL https://jmlr.org/papers/v17/kaufman16a.html.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. In *International Conference on Learning Representations*, volume 2024, pages 54928–54958, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/f1cf02ce09757f57c3b93c0db83181e0-Paper-Conference.pdf.

Ruixiao Lin, Xinhao Deng, Qingming Li, Jianan Ma, Yunhao Feng, Yuqi Qing, Zhenyuan Li, Yechao Zhang, Shiwen Cui, Changhua Meng, Tianwei Zhang, Xingjun Ma, Qi Li, Ke Xu, and Shouling Ji. Safety in Self-Evolving LLM Agent Systems: Threats, Amplification, and Case Studies. arXiv preprint arXiv:2606.23075, 2026. URL https://arxiv.org/abs/2606.23075.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. RM-Bench: Benchmarking Reward Models of Language Models with Subtlety and Style. In *International Conference on Learning Representations*, volume 2025, pages 44323–44355, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6da1eec80095dc5937f7716db15aca4b-Abstract-Conference.html.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein Bounds and Sample Variance Penalization. In *Conference on Learning Theory (COLT)*, 2009. https://doi.org/10.48550/arXiv.0907.3740. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical Bernstein Stopping. In *Proceedings of the 25th International Conference on Machine Learning*, pages 672–679. ACM, 2008. https://doi.org/10.1145/1390156.1390241. URL https://sites.ualberta.ca/~szepesva/papers/bernstein-stopping.pdf.

Robin Pemantle. A survey of random processes with reinforcement. *Probability Surveys*, 4:1–79, 2007. https://doi.org/10.1214/07-ps094. URL https://doi.org/10.1214/07-PS094.

Foster Provost and Tom Fawcett. Robust Classification for Imprecise Environments. *Machine Learning*, 42(3):203–231, 2001. https://doi.org/10.1023/a:1007601015854. URL https://doi.org/10.1023/A:1007601015854.

Qwen Team. Qwen2.5 Technical Report. arXiv preprint arXiv:2412.15115, 2025. URL https://arxiv.org/abs/2412.15115v2.

Vyas Raina, Adian Liusie, and Mark Gales. Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7499–7517. Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.emnlp-main.427. URL https://aclanthology.org/2024.emnlp-main.427/.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-Theoretic Statistics and Safe Anytime-Valid Inference. *Statistical Science*, 38(4):576–601, 2023. https://doi.org/10.1214/23-STS894. URL https://doi.org/10.1214/23-STS894.

Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D. Lee, and Sanjeev Arora. What Makes a Reward Model a Good Teacher? An Optimization Perspective. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 59162–59222. Curran Associates, Inc., 2025. https://doi.org/10.52202/085713-1981. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/554e056fe2b6d9fd27ffcd3367ae1267-Paper-Conference.pdf.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024. https://doi.org/10.1038/s41586-023-06924-6. URL https://www.nature.com/articles/s41586-023-06924-6.

Shuning Shang, Hubert Strauss, Stanley Wei, Sanjeev Arora, and Noam Razin. When Errors Can Be Beneficial: A Categorization of Imperfect Rewards for Policy Gradient. arXiv preprint arXiv:2604.25872, 2026. URL https://arxiv.org/abs/2604.25872.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv preprint arXiv:2402.03300, 2024. URL https://arxiv.org/abs/2402.03300.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024. https://doi.org/10.1038/s41586-024-07566-y. URL https://doi.org/10.1038/s41586-024-07566-y.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and Characterizing Reward Hacking. In *Advances in Neural Information Processing Systems*, volume 35, pages 9460–9471. Curran Associates, Inc., 2022. https://doi.org/10.52202/068431-0687. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/3d719fee332caa23d5038b8a90e81796-Paper-Conference.pdf.

Benedikt Stroebl, Sayash Kapoor, and Arvind Narayanan. The Limits of Inference Scaling Through Resampling. In *International Conference on Learning Representations*, volume 2026, pages 92601–92633, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/95591b7557edea46ae1d90b0e9e825c5-Abstract-Conference.html.

Aoran Wang, Lei Ou, Yang Yu, and Zongzhang Zhang. Reward Model Evaluation via Automatically-Ranked Policy Alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (31):26124–26132, 2026. https://doi.org/10.1609/aaai.v40i31.39815. URL https://doi.org/10.1609/aaai.v40i31.39815.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450. Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. https://doi.org/10.1093/jrsssb/qkad009. URL https://doi.org/10.1093/jrsssb/qkad009.

Yitao Wu, Si Shen, Rui Yang, Hong Peng, and Bin Hu. Verify, Repair, Repeat, or Stop? Robust Stopping for Noisy Verify-Repair Loops in LLM Agents. arXiv preprint arXiv:2607.17641, 2026. URL https://arxiv.org/abs/2607.17641.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. arXiv preprint arXiv:2409.12122, 2024a. URL https://arxiv.org/abs/2409.12122.

An Yang et al. Qwen3 Technical Report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large Language Models as Optimizers. In *International Conference on Learning Representations*, volume 2024, pages 12028–12068, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3339f19c5fcee3ad74502947a32be9e6-Abstract-Conference.html.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents. In *International Conference on Learning Representations*, volume 2026, pages 104223–104294, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/aa5f5e6eb6f613ec412f1d948dfa21a5-Paper-Conference.pdf.

Yulai Zhao, Haolin Liu, Dian Yu, Sunyuan Kung, Meijia Chen, Haitao Mi, and Dong Yu. One Token to Fool LLM-as-a-Judge. arXiv preprint arXiv:2507.08794, 2025. URL https://arxiv.org/abs/2507.08794.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc., 2023. https://doi.org/10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Chenyu Zhou. More Convincing, Not More Correct: Self-Play Reward Hacking of Reference-Free LLM Judges. arXiv preprint arXiv:2607.05904, 2026. URL https://arxiv.org/abs/2607.05904.

Yilun Zhou, Austin Xu, Peifeng Wang, Caiming Xiong, and Shafiq Joty. Evaluating Judges as Evaluators: The JETTS Benchmark of LLM-as-Judges as Test-Time Scaling Evaluators. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 79436–79471. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhou25af.html.

# A Reinforcement and History-Dependent Acceptance

We prove Thm. 1 through the classical embedding of an urn in independent birth processes (Pemantle, 2007). The embedding clock is an auxiliary device; crossing times measured in actual rounds are treated separately in App. B.2.

## A.1 Embedding the Addition Chain

Let $n_1, n_0 > 0$ count the harmful templates that the judge misclassifies and correctly rejects, respectively. A round adds one entry of type $j$ with probability

$$\omega(1-p)\frac{r_j n_j}{n_1+n_0}.$$

While no addition occurs these counts and probabilities remain unchanged, even when utility changes. If $r_1 + r_0 > 0$, the next addition therefore occurs almost surely after a finite geometric wait. A countable intersection shows that every finite addition count is reached almost surely. Conditional on an addition, its type is $j$ with probability

$$\frac{r_j n_j}{r_1 n_1 + r_0 n_0}. \tag{20}$$

For each $j$, construct an independent pure birth process $Z_j(u)$ with initial count $n_j(0)$ and total birth rate $r_j Z_j(u)$. If $r_j = 0$, the process is constant. Otherwise it is a nonexplosive Yule process. Competing exponential clocks yield exactly Eq. (20), so observing the pair at successive birth times gives the same addition chain as the pool.

For completeness, one initial individual with rate $r > 0$ has

$$\mathbb{P}(Z(u) = \ell) = e^{-ru}(1-e^{-ru})^{\ell-1}, \qquad \ell \geq 1.$$

The branching property implies that $e^{-ru}Z(u)$ is a martingale, with second moment $2 - e^{-ru} \leq 2$. It converges almost surely and in $L^2$; the displayed geometric law identifies its limit as an exponential variable of mean one. With $m$ initial individuals, independent lineages give

$$e^{-ru}Z(u) \longrightarrow W \sim \mathrm{Gamma}(m,1) \quad \text{almost surely}, \qquad 0 < W < \infty. \tag{21}$$

Linear birth rates ensure that birth times have no finite accumulation and tend to infinity. Thus Eq. (21) also holds along the combined birth times.

*Proof of Thm. 1.* For $r_1 > r_0 > 0$, the two independent limits satisfy

$$\frac{Z_0(u)}{Z_1(u)} \sim \frac{W_0}{W_1} e^{-(r_1-r_0)u} \longrightarrow 0.$$

At the $k$th combined birth time $u_k$, the total count is $N+k$, and $Z_1(u_k) \sim k$. Hence $\log k/u_k \to r_1$ and

$$\frac{\log(1-q^{[k]})}{\log k} = \frac{\log Z_0(u_k) - \log(N+k)}{\log k} \longrightarrow \frac{r_0}{r_1} - 1.$$

When $r_0 = 0$, the number of correctly rejected templates stays $N - b_0$, so $1 - q^{[k]} = (N-b_0)/(N+k)$ and the same conclusion holds. Every finite addition is reached in finitely many rounds almost surely, so $Q_t \to 1$. Under $A = S$, $\tau_t = 1$ and $\phi_t = Q_t$, because all harmful magnitudes agree. Thus $\rho_t = 1 - Q_t/R$ converges to the strictly negative number $1 - 1/R$ when $R < 1$, proving that the utility drift is eventually negative at every subsequent round. □

## A.2 Other Reinforcement Regimes

**Proposition 4** (Limits of the misclassified fraction). *For the same copying model, $r_0 > r_1$ gives $Q_t \to 0$ almost surely. If $r_1 = r_0 > 0$, then*

$$Q_t \longrightarrow Q_\infty \sim \mathrm{Beta}(b_0, N - b_0) \quad \textit{almost surely}.$$

*If both rates are zero, $Q_t \equiv \epsilon$. For $A = S$ and $0 < R < 1$, equal positive rates give eventual negative expected utility change with probability $\mathbb{P}(Q_\infty > R) \in (0,1)$.*

*Proof.* Exchange the two types in the preceding proof for unequal rates. At equal positive rates the independent gamma limits in Eq. (21) have a common rate, giving $Q_\infty = W_1/(W_1 + W_0)$ with the stated beta law. Its continuous density is positive on $(0, 1)$, so $Q_\infty = R$ has probability zero; the limiting sign follows. The zero-rate case has no updates. □

The conditional drift of the misclassified fraction is

$$\mathbb{E}[Q_{t+1} - Q_t \mid \mathcal{F}_t] = \frac{\omega(1-p)Q_t(1-Q_t)(r_1 - r_0)}{n_{1,t} + n_{0,t} + 1}. \tag{22}$$

A positive drift alone would not identify the limiting endpoint; the embedding rules out convergence to zero when $r_1 > r_0$. Nor is a positive roundwise addition probability alone enough for an arbitrary nonstationary extension: a summable sequence of hazards can leave positive probability of no further update. Here the addition probability stays fixed between additions, which supplies the necessary nonstagnation property.

### A.3 The Threshold for a Fixed Acceptance Rule

Under the copying rule in Cor. 1, positive initial expected utility change implies $\pi_1 > \pi_0$ and

$$D_t(\pi) > 0 \quad \Longleftrightarrow \quad Q_t < q_\pi := \frac{R\pi_1 - \pi_0}{\pi_1 - \pi_0}. \tag{23}$$

The initial inequality ensures $0 < Q_0 < q_\pi < 1$. The misclassified fraction converges to one, but accepting a template that the judge correctly rejects can decrease it before convergence. Accordingly, Cor. 1 asserts an almost surely finite last transition to negative drift. Earlier crossings can be reversed under some fixed acceptance rules. The exact $\pi_1 = 1, \pi_0 = 0$ instance has the stronger monotonicity in Prop. 1.

### A.4 Proof of the Cumulative Drift Bound

*Proof of Thm. 2.* We first verify the potential and its conditional drift, then derive the finite-horizon and pathwise conclusions.

**A discrete relative-entropy potential.** For integers $1 \le m \le n$,

$$\mathsf{H}_{n-1} - \mathsf{H}_{m-1} = \sum_{k=m}^{n-1} \frac{1}{k} \ge \int_m^n \frac{du}{u} = \log(n/m).$$

Multiplying by $\nu_j$, summing with $m = n_{j,t}$, and adding $\sum_j \nu_j \log \nu_j$ proves Eq. (8). More precisely, for $q_{j,t} = n_{j,t}/n_t$,

$$0 \le \Phi_t - D_{\mathrm{KL}}(\nu \| q_{\cdot,t}) \le \sum_j \frac{\nu_j}{n_{j,t}} - \frac{1}{n_t}. \tag{24}$$

The upper bound follows by comparing each integral over $[k, k+1]$ with $1/(k+1)$. Thus the correction vanishes as the counts in the support of $\nu$ grow. It ensures that the next calculation is exact even for small pools.

Write $a_t = \sum_j \nu_j \pi_{t,j}$ and $b_t = \sum_j q_{j,t}\pi_{t,j}$. The conditional probabilities of beneficial and harmful acceptances are $\omega p a_t$ and $\omega(1-p)b_t$. Consequently,

$$\lambda_t := \mathbb{E}[I_t \mid \mathcal{F}_t] = \omega\{pa_t + (1-p)b_t\}, \qquad \frac{D_t}{s(x_t)} = \omega\{pga_t - (1-p)hb_t\}. \tag{25}$$

Only a harmful acceptance changes the pool. An output-$j$ copy has probability $\omega(1-p)q_{j,t}\pi_{t,j}$ and changes the potential by $1/n_t - \nu_j/n_{j,t}$. Therefore

$$\mathbb{E}[\Phi_{t+1} - \Phi_t \mid \mathcal{F}_t] = \frac{\omega(1-p)}{n_t}(b_t - a_t).$$

Substitution of Eq. (25) and $\eta = (1-p)h - pg$ gives Eq. (11).

**Finite horizons.** Every quantity is integrable at a fixed finite horizon: $n_t \le N + t$, all counts remain positive, and $D_t/s(x_t)$ is bounded. Taking expectations in Eq. (11) and summing gives Eq. (12): the two sides differ by $p(g+h)\mathbb{E}\Phi_T \ge 0$. Since only harmful acceptances enlarge the pool, $n_t \le N + M_t$. Ordering the nonzero revisions yields the pathwise bound

$$\sum_{t<T} \frac{I_t}{n_t} \ge \sum_{k=0}^{M_T-1} \frac{1}{N+k} \ge \log\left(1 + \frac{M_T}{N}\right). \tag{26}$$

Dropping the nonnegative second term in Eq. (12) proves Eq. (9) for deterministic horizons. For a stopping time $T$, apply this argument to the policy that rejects all proposals after $T$, first at a deterministic horizon $k$. This replaces each count and sum by its value at $T \wedge k$. Monotone convergence as $k \to \infty$ proves the bound for every stopping time, including $T = \infty$, with nonnegative expectations interpreted in $[0, \infty]$.

**Infinite horizons.** Let $\Delta\Phi_t = \Phi_{t+1} - \Phi_t$ and center its increments to obtain the martingale

$$Z_T = \sum_{t<T} \{\Delta\Phi_t - \mathbb{E}[\Delta\Phi_t \mid \mathcal{F}_t]\}.$$

At each pool addition $n_t$ increases by one, and at an output-$j$ addition so does $n_{j,t}$. Each denominator in the following sums is therefore used at most once:

$$\sum_{t\ge 0} (\Delta\Phi_t)^2 \le 2 \sum_{k=N}^{\infty} \frac{1}{k^2} + 2 \sum_j \nu_j^2 \sum_{k=n_{j,0}}^{\infty} \frac{1}{k^2} < \infty. \tag{27}$$

This is a deterministic bound. The sum of the expected squared martingale increments is no larger, so $Z_T$ is bounded in $L^2$ and converges almost surely. In particular, $Z^* = \sup_T Z_T < \infty$ almost surely. Summing Eq. (11) without taking expectations and using $\Phi_T \ge 0$ now gives

$$\eta \sum_{t<T} \frac{\lambda_t}{n_t} + \sum_{t<T} \frac{(D_t)_+}{n_t s(x_t)} \le p(g+h)(\Phi_0 + Z^*) + \sum_{t<T} \frac{(D_t)_-}{n_t s(x_t)}. \tag{28}$$

Thus a finite negative-drift sum implies $\sum_t \lambda_t/n_t < \infty$ on the same event, without requiring that sum to have finite expectation.

For completeness, this also implies $\sum_t I_t/n_t < \infty$ almost surely on that event. Stop when $\sum_{t<T} \lambda_t/n_t$ first exceeds an integer $L$. The expected stopped sum of $I_t/n_t$ equals that of $\lambda_t/n_t$ and is at most $L + 1/N$. It is therefore finite almost surely. On the event that the unstopped compensator sum is finite, an integer $L$ exceeds it and no stopping occurs. Taking a countable union establishes the implication. By Eq. (26), infinitely many revisions would instead force $\sum_t I_t/n_t = \infty$. This proves Eq. (10).

If only finitely many rounds have $D_t < 0$, the negative-drift sum is finite pathwise. Hence eventually nonnegative drift implies finitely many nonzero revisions. Every finite revision preserves $x \in (0, 1)$; after the last one, utility stays constant with $G_t = \omega p g x_t(1 - x_t) > 0$. □

**Scope and the support condition.** The proof applies to any fixed beneficial-output law, including false negatives and multiple score levels. It requires the beneficial output support to be contained in the harmful output support, fixed harmful outputs, and acceptance-driven copying. If an output has $\nu_j > 0$ and $n_{j,0} = 0$, accepting it alone never adds a harmful template. At each round a beneficial acceptance has probability $\omega p \nu_j > 0$, independently of preceding draws. There are almost surely infinitely many such acceptances, each with positive conditional expected change. This proves the converse stated in the main text. Outputs with both $\nu_j = 0$ and $n_{j,0} = 0$ may simply be discarded.

When $D_t \ge 0$ at every round, monotone convergence in Eq. (9) bounds $\mathbb{E}\log(1 + M_\infty/N)$, not $\mathbb{E}M_\infty$. The theorem also covers infinitely many negative-drift rounds whose normalized total is finite. The factor $s(x_t)$ is essential: near an endpoint, unnormalized utility changes can be summable despite infinitely many revisions.

**Fixed rules with multiple outputs.** For a nonzero fixed vector $\pi$, the same model gives eventual negative drift at every subsequent round. To see this, embed the $J$ pool counts in independent birth processes with rates $\pi_j$. Every finite addition is reached almost surely as before. Writing $\pi_{\max} = \max_j \pi_j > 0$, the birth-process limits Eq. (21) imply that the total pool fraction outside the maximizing outputs tends to zero. Hence

$$\frac{D_t}{H_t} = R\sum_j \nu_j\pi_j - \sum_j q_{j,t}\pi_j \longrightarrow R\sum_j \nu_j\pi_j - \pi_{\max} \le -(1-R)\pi_{\max} < 0 \quad \text{almost surely.} \tag{29}$$

This extends Cor. 1 to finite outputs; ties among the largest copying rates do not affect the drift limit.

**Logarithmic growth can occur.** The dependence on a logarithmic revision count has a simple example. Accept every proposal, so $D_t = -\omega\eta s(x_t)$, $M_T/T \to \omega$, and $n_T/T \to \omega(1-p)$ almost surely. The last two limits follow from the strong law for the independent family draws. Consequently,

$$\frac{1}{\log T}\sum_{t<T}\frac{(D_t)_-}{n_t s(x_t)} = \frac{\omega\eta}{\log T}\sum_{t<T}\frac{1}{n_t} \longrightarrow \frac{\eta}{1-p} \quad \text{almost surely.} \tag{30}$$

Thus a linear number of revisions can incur only logarithmic normalized cumulative negative drift. This example identifies a possible growth order; it does not assert optimal constants in Eq. (9).

## B Proofs for the Exact Reuse Process

This section proves the copy threshold (Prop. 1), the comparison with refreshed errors (Prop. 2), and the decision limit from a binary output (Cor. 2), and derives the distribution of the crossing time.

### B.1 Proof of Prop. 1

*Proof.* For $0 < x < 1$ and $0 < g, h < 1$,

$$x - hx(1-x) = x[1-h(1-x)] > 0, \qquad 1 - x - gx(1-x) = (1-x)(1-gx) > 0.$$

Every accepted revision therefore keeps utility in $(0,1)$. Beneficial and harmful proposals have fixed signs and magnitudes $gs(x_t)$ and $hs(x_t)$, giving Eq. (3). All beneficial gains are accepted, so $\tau_t = 1$.

Initially, $b_0$ of the $N$ pool entries belong to the error set. Each accepted harmful revision comes from that set and adds one copy of the same identity. Induction on the number of additions gives $b_0 + k$ such entries out of $N + k$ total entries. Since the pool is sampled uniformly with multiplicity, $\phi_t = q_{K_t}$. Substitution into Eq. (2) gives the drift formula in Eq. (7).

The expected change has the sign of $R - q_{K_t}$ because $H_t > 0$. For $k \ge 0$,

$$q_k \ge R \iff b_0 + k \ge R(N+k) \iff k \ge \frac{RN - b_0}{1-R}.$$

Replacing the weak inequalities by strict inequalities gives the threshold for negative expected change. Taking the appropriate integer thresholds yields $k_\ge$ and $k_>$ in Prop. 1. Since $q_{k+1} - q_k = (N - b_0)/((N+k)(N+k+1)) > 0$, the expected change stays nonpositive after the first crossing.

At any count $k$, a new round increases the count with probability $\lambda_k = \omega(1-p)q_k > 0$. Independent round draws give a finite geometric waiting time almost surely. Since either crossing requires finitely many additions, both occur in finitely many rounds almost surely. □

### B.2 The Distribution of the Crossing Time

Let $\mathrm{Geom}(\lambda)$ denote the geometric distribution on $\{1, 2, \ldots\}$ with success probability $\lambda$.

**Theorem 4** (Rounds to nonpositive expected utility change)**.** *For the exact process of Prop. 1, let $T_\ge = \inf\{t \ge 0 : K_t \ge k_\ge\}$ count completed rounds, with $K_0 = 0$, and set $\lambda_k = \omega(1-p)q_k$. Then*

$$T_\ge \overset{d}{=} \sum_{k=0}^{k_\ge - 1} W_k, \qquad W_k \text{ independent}, \qquad W_k \sim \mathrm{Geom}(\lambda_k). \tag{31}$$

*Its moments and probability generating function are*

$$\mathbb{E}T_{\geq} = \frac{1}{\omega(1-p)} \sum_{k=0}^{k_{\geq}-1} \frac{N+k}{b_0+k}, \tag{32}$$

$$\mathrm{Var}(T_{\geq}) = \sum_{k=0}^{k_{\geq}-1} \frac{1-\lambda_k}{\lambda_k^2}, \tag{33}$$

$$\mathbb{E}[z^{T_{\geq}}] = \prod_{k=0}^{k_{\geq}-1} \frac{\lambda_k z}{1-(1-\lambda_k)z}, \qquad 0 \leq z \leq 1. \tag{34}$$

*Replacing $k_{\geq}$ by $k_{>}$ gives the first time of negative expected utility change. At a checkpoint with $K_t = j < k_{\geq}$, the conditional law of the remaining time is the sum in Eq. (31) starting at $k = j$.*

*Proof.* While the count is $k$, a round increases it if the proposal is nonnull, belongs to the harmful family, and uses a template from the error set. The probability is $\lambda_k$, independent of intervening changes in utility. Independent draws therefore give $W_k \sim \mathrm{Geom}(\lambda_k)$. After an addition, the probability of the next addition depends only on the new count. The future draws are independent of those used so far, giving independence of the waiting times. Their means, variances, and probability generating functions give the displayed formulas. The same reasoning applies conditional on the current state. □

For integer $b_0$, the mean can also be written as

$$\mathbb{E}T_{\geq} = \frac{k_{\geq} + (N-b_0)(\mathsf{H}_{b_0+k_{\geq}-1} - \mathsf{H}_{b_0-1})}{\omega(1-p)}, \qquad \mathsf{H}_n = \sum_{j=1}^{n} \frac{1}{j}, \quad \mathsf{H}_0 = 0. \tag{35}$$

This follows by writing $(N+k)/(b_0+k) = 1 + (N-b_0)/(b_0+k)$. For the example following Prop. 1, with $\omega = 1$, the mean is 104.8506 rounds and the standard deviation is 19.7310 rounds. These are analytic values calculated from the distribution above.

### B.3 Proof of Prop. 2

*Proof.* Let $J_t$ be the identity drawn on a harmful proposal. For the refreshed judge, the new error set is independent of $\mathcal{F}_t$ and $J_t$, and every identity belongs to it with probability $b_0/N = \epsilon$. Thus

$$\mathbb{P}(S_t = 1 \mid \mathcal{F}_t, J_t, \text{harmful proposal}) = \epsilon.$$

Averaging over the pool gives $\phi_t = \epsilon$, regardless of its multiplicities. Beneficial revisions remain accepted, so $\tau_t = 1$. The proposal magnitudes and family probabilities are unchanged, giving $G_t/H_t = R$ and the positive drift in Prop. 2.

On the fixed evaluation set, the prediction vector is determined by the error subset. The error subset of the deterministic judge, viewed marginally over its initial draw, is uniform among all $b_0$-subsets at every round. The refreshed subset has that same law. Both prediction vectors assign positive outputs to $N$ beneficial probes and negative outputs to $N - b_0$ harmful probes, hence accuracy $(2N - b_0)/(2N)$. The vector laws agree marginally; their dependence across rounds produces different false-positive rates under reuse. □

### B.4 Proof of Cor. 2

*Proof.* At any fixed deterministic-judge pool, acceptance probabilities $\pi(1)$ and $\pi(0)$ give

$$D_t(\pi) = \pi(1)(G_t - H_t Q_t) - \pi(0) H_t (1 - Q_t).$$

The second term is nonpositive. The first is maximized by $\pi(1) = 1$ if $G_t > H_t Q_t$, and by $\pi(1) = 0$ otherwise. This proves Cor. 2. □

## C Recovery and Its Scope

**Proposition 5** (Recovery requires temporary negative drift)**.** *Start from a finite template pool with a deterministic judge and misclassified fraction $Q \geq R$, where $R < 1$, and keep copying every accepted harmful template. Allow $\pi_t(S)$ to depend on all past history. If $D_t(\pi_t) \geq 0$ at every subsequent round, positive conditional expected utility change cannot be recovered.*

*Proof.* Finite pools retain a correctly rejected template, so $Q_t < 1$. If $Q_t > R$, both coefficients in Eq. (6) are strictly negative, forcing $\pi_t(0) = \pi_t(1) = 0$. If $Q_t = R$, nonnegative drift forces $\pi_t(0) = 0$; any added harmful template is misclassified, so the misclassified fraction can only stay at $R$ or increase. Induction proves the claim. □

The information constraint and copying rule in Prop. 5 are both essential. We keep $R < 1$ fixed, the deterministic judge gives a positive output to every beneficial proposal, and every accepted harmful template is copied.

**Proposition 6** (Dilution threshold and a recovery policy)**.** *Start with $n$ entries and $B = nQ$ misclassified entries, where $R \leq Q < 1$. After $e$ misclassified copies and $r$ correctly rejected copies, returning to $A = S$ gives positive conditional expected utility change if and only if Eq. (13) holds. The smallest possible number of correctly rejected copies is*

$$r_{\min} = \lfloor n(Q/R - 1) \rfloor + 1. \tag{36}$$

*The policy $\pi(1) = 0, \pi(0) = 1$ reaches this count with $e = 0$ in finitely many rounds almost surely. If its initial utility is $x$ and $x^{(r)}$ denotes utility after $r$ such copies, then*

$$x - x^{(r)} \geq x\{1 - [1 - h(1 - x)]^r\}. \tag{37}$$

*This utility bound is for the pure dilution policy, not a minimax bound over all recovery policies.*

*Proof.* The resulting misclassified fraction is $(B + e)/(n + e + r)$. Comparing it with $R$ gives

$$\frac{B + e}{n + e + r} < R \quad \Longleftrightarrow \quad r > n(Q/R - 1) + (1/R - 1)e.$$

Since $e \geq 0$ and $R < 1$, the least integer $r$ is Eq. (36), attained at $e = 0$. Under the proposed policy only correctly rejected harmful revisions cause a nonzero accepted change. Their probability while waiting for the next copy is $\omega(1 - p)(1 - Q_t) > 0$ and stays fixed until that copy. Each required copy is therefore reached almost surely in finite time. Its change is harmful, and

$$x^{(j+1)} = x^{(j)}[1 - h(1 - x^{(j)})] \leq x^{(j)}[1 - h(1 - x)],$$

because $x^{(j)} \leq x$. Iteration proves Eq. (37). Null acceptances have no effect on utility or memory. □

**Continuation objectives.** Prohibiting negative conditional expected utility change rules out dilution by Prop. 5; a continuation objective that tolerates temporary losses need not. If the update protocol were changed to copy only harmful templates with $S = 1$, independently of whether the new acceptance rule accepts $S = 0$, the region $Q \geq R$ would instead be closed under every acceptance rule. In that modified protocol, Cor. 2 makes utility a bounded supermartingale in the closed region. Optional sampling then bounds expected terminal utility at every almost surely finite stopping time by the utility on entry. This stronger continuation statement requires that modified update rule and is not asserted for acceptance-driven copying.

## D Optimal Acceptance Given a Signal

The binary output in Cor. 2 is a special case of deciding whether to accept a revision from an observed signal. At a fixed state, a Bayes-optimal rule accepts precisely those signal values for which the proposed utility change has positive conditional mean. This section expresses the optimal value in terms of signal distributions weighted by gain and loss. Time subscripts are omitted.

### D.1 The Threshold for a Fixed Acceptance Rule

The identity $D = G\tau - H\phi$ in Eq. (2) implies that positive expected utility change requires $0 \le \phi < R\tau \le R$. Thus along states with positive drift and $R \to 0$, the loss-weighted acceptance rate must satisfy $\phi \to 0$. If $G = 0$, then $D = -\mathbb{E}[Ad_-] \le 0$; if $H = 0$, then $D = \mathbb{E}[Ad_+] \ge 0$.

For a reference utility $V^\star \ge V(z)$, write $\Delta = V^\star - V(z)$. Suppose the proposal law satisfies $R \le C\Delta^\gamma$, where $C, \gamma > 0$. If the acceptance rule has $\phi \ge \phi_{\min} > 0$ and $\tau \le \bar{\tau}$ with $\bar{\tau} > 0$, then

$$\Delta \le \left(\frac{\phi_{\min}}{C\bar{\tau}}\right)^{1/\gamma} \quad \Longrightarrow \quad D \le 0. \tag{38}$$

Indeed, $D/H \le C\bar{\tau}\Delta^\gamma - \phi_{\min}$. This bounds the conditional expected utility change of the specified acceptance rule near $V^\star$. The next result optimizes over all acceptance rules with the same available information.

### D.2 The Best Decision from a Signal

Let $S$ contain all information exposed to the acceptance rule, and let a randomized acceptance rule accept with probability $\pi(S) \in [0, 1]$. The signal could be a score, a rationale, proposal features, or their combination. Define two probability laws by weighting signal observations with their gains and losses:

$$P_+(E) = \frac{\mathbb{E}[d_+ \mathbf{1}\{S \in E\}]}{G}, \qquad P_-(E) = \frac{\mathbb{E}[d_- \mathbf{1}\{S \in E\}]}{H}. \tag{39}$$

The weights account for the magnitudes of utility changes. When each sign has a constant magnitude, these measures reduce to the signal laws conditional on beneficial and harmful proposals. The next result applies the standard Bayes likelihood-ratio rule to these measures (Provost and Fawcett, 2001).

**Proposition 7** (Optimal acceptance given a signal)**.** *Let $f_+, f_-$ be densities of $P_+, P_-$ with respect to a common dominating measure $\nu$. Then*

$$\sup_{0 \le \pi \le 1} D(\pi) = \int (Gf_+ - Hf_-)_+ \, d\nu. \tag{40}$$

*If $P_+ \ll P_-$ and $M = \operatorname{ess\,sup}_{P_-}(dP_+/dP_-)$, then*

$$\sup_{\pi} D(\pi) > 0 \quad \Longleftrightarrow \quad RM > 1. \tag{41}$$

*If $P_+$ instead has a nonzero component singular to $P_-$, positive expected utility change is possible for every $G > 0$.*

*Proof.* By Eq. (39),

$$D(\pi) = \int \pi(s)\{Gf_+(s) - Hf_-(s)\} \, d\nu(s).$$

Maximizing pointwise gives Eq. (40), with acceptance wherever $Gf_+ > Hf_-$. In the absolutely continuous case, this set has positive measure under $P_-$ precisely when $dP_+/dP_- > 1/R$ on a set of positive $P_-$ mass. By the definition of essential supremum, this is equivalent to $M > 1/R$. If $RM = 1$, the integral of the positive part is zero. A nonzero singular component is supported on a set with zero harmful mass and positive beneficial mass, which can be accepted to obtain positive expected utility change. □

As a consequence, if $P_+ \ll P_-$, $R \le C\Delta^\gamma$, and $M \le \bar{M} < \infty$, then

$$\Delta \le (C\bar{M})^{-1/\gamma} \quad \Longrightarrow \quad D(\pi) \le 0 \quad \text{for every measurable acceptance rule } \pi(S). \tag{42}$$

Indeed, $RM \le C\bar{M}\Delta^\gamma \le 1$, so Prop. 7 applies.

The bound on $M$ limits how strongly any signal event can favor beneficial revisions over harmful ones. In the pool construction, $P_+ = \mathrm{Dirac}(1)$ and $P_- = \mathrm{Bern}(q_k)$, so $M = 1/q_k$. The condition $RM > 1$ becomes $q_k < R$, recovering Cor. 2. For more general signals, a finite sample need not establish a bound on the essential supremum: events unobserved in that sample can have larger likelihood ratios.

**Processing a signal.** If $\widetilde{S}$ is sampled from a channel $\mathcal{K}(\cdot \mid S)$ independently of $d$ conditional on $S$, every rule $c(\widetilde{S})$ induces $\pi(S) = \mathbb{E}[c(\widetilde{S}) \mid S]$ with $\mathbb{E}[c(\widetilde{S})d] = \mathbb{E}[\pi(S)d]$. Hence this processing cannot increase the optimal expected utility change. This is the garbling direction of Blackwell comparison (Blackwell, 1953); deterministic calibration and random perturbations of an existing score are examples. An additional measurement informative about $d$ can change the signal laws and increase the optimal value.

# E Proof of the Sequential Testing Guarantee

The information and cost units used in the main text are summarized in Tab. 1. In particular, exact utility changes are unavailable in the noisy experiment; acceptance and identity logs remain admissible in the lower-bound constructions.

Table 1: Observations and cost units. The deployment results restrict the information used to accept a proposal. The statistical results specify what is observed at a fixed state and what counts as one sample.

| Experiment | Observations | Scope |
|---|---|---|
| Revision decisions | Past history and current judge output $S$ | No current family, identity, or utility features. |
| General mean test | $X$ or $Y = X + \xi$ for each sampled proposal | No exact $d$ with noisy evaluation; lower bounds also allow $A$ and identity labels. |
| Exact reuse model | $(d, A)$ for each sampled proposal | Structure known, $q$ unknown; identities and error set hidden. |
| Evaluation after acceptance | $d + \xi$ conditional on $A = 1$ | Counts utility evaluations separately from proposal draws; $a = 0$. |

We first establish the variance bound for revision records, then specify the procedure with observation noise. The final part controls the tail of the stopping time and proves both sample cost claims in Thm. 3.

## E.1 Variance of the Test Statistic

**Lemma 1** (Variance bound from expected gain and magnitude)**.** *Let $a \in [0, 1)$, $X = Ad - ad_+$, $\mu = \mathbb{E}X$, and $G = \mathbb{E}d_+$. If $|d| \le b$, then $|X| \le b$ and*

$$\mathrm{Var}(X) \le \mathbb{E}X^2 \le 2bG - b\mu.$$

*More precisely, if $d_+ \le u_+$ and $d_- \le u_-$, write $c_a = 1 - a$, $U = c_a u_+$, and $L = \max(au_+, u_-)$. Then*

$$\mathbb{E}X^2 \le c_a(U + L)G - L\mu. \tag{43}$$

*In particular, at current utility $x$ and target $a = 0$, $\mathbb{E}X^2 \le G - x\mu$.*

*Proof.* On a beneficial revision $X$ is $(1 - a)d$ or $-ad$; on a harmful revision it is $d$ or zero. Thus $|X| \le b$, $X_+ \le U$, $X_- \le L$, and $\mathbb{E}X_+ \le c_a G$. Since $\mathbb{E}X_- = \mathbb{E}X_+ - \mu$,

$$\mathbb{E}X^2 = \mathbb{E}X_+^2 + \mathbb{E}X_-^2 \le U\mathbb{E}X_+ + L\mathbb{E}X_- \le c_a(U + L)G - L\mu.$$

Using $X_+, X_- \le b$ directly gives $\mathbb{E}X^2 \le 2b\mathbb{E}X_+ - b\mu \le 2bG - b\mu$. For the final assertion take $u_+ = 1 - x$, $u_- = x$, and $a = 0$ in Eq. (43). □

## E.2 Raw Observations and Simultaneous Coverage

Observe i.i.d. $Y_i = X_i + \xi_i$, with $|X_i| \le b$, $\mu = \mathbb{E}X_i$, and independent centered $\sigma^2$-sub-Gaussian noise. For any sample vector $Z$, write

$$\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i, \qquad s_{Z,n} = \left\{ \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2 \right\}^{1/2}.$$

Use the stages, error allocation, and interval in Eq. (17), skipping decisions when $n_k < \ell_k$. Only $\bar{Y}_n, s_{Y,n}, b, \sigma, \delta$ are used by the procedure. The decomposition into $X$ and $\xi$ is a proof device.

Two elementary properties of this decomposition are useful. The triangle inequality for centered sample vectors gives

$$s_{X,n} \le s_{Y,n} + s_{\xi,n}, \qquad s_{Y,n} \le s_{X,n} + s_{\xi,n}. \tag{44}$$

For $\sigma > 0$, introduce an independent standard Gaussian $Z$. Integrating its moment generating function and then applying the sub-Gaussian bound yields

$$\mathbb{E}\exp\{\xi_i^2/(4\sigma^2)\} = \mathbb{E}_Z\mathbb{E}_{\xi_i}\exp\{Z\xi_i/(\sqrt{2}\sigma)\} \le \mathbb{E}_Z e^{Z^2/4} = \sqrt{2}.$$

Independence across records and Markov's inequality consequently give, for every $z > 0$,

$$\mathbb{P}\left(s_{\xi,n} > 2\sigma\sqrt{\frac{n\log\sqrt{2}+z}{n-1}}\right) \le e^{-z}. \tag{45}$$

For $n \ge z$ and $n \ge 2$, the displayed threshold is at most $4\sigma$. When $\sigma = 0$, the noise vanishes and these statements hold deterministically.

*Validity in Thm. 3.* Fix an inspected stage, put $n = n_k$, $\ell = \ell_k$, and $v_X = \mathrm{Var}(X)$. Applying Maurer and Pontil (2009, Theorem 10, Eq. (3)) to $(X_i + b)/(2b)$ gives, except on an event of probability $e^{-\ell}$,

$$\sqrt{v_X} \le s_{X,n} + 2b\sqrt{\frac{2\ell}{n-1}}.$$

By Eq. (45), $s_{\xi,n} \le 4\sigma$ except with probability $e^{-\ell}$. Two-sided bounded Bernstein and sub-Gaussian mean bounds give

$$|\bar{X}_n - \mu| \le \sqrt{\frac{2v_X\ell}{n}} + \frac{2b\ell}{3n}, \qquad |\bar{\xi}_n| \le \sigma\sqrt{\frac{2\ell}{n}},$$

with failure probabilities at most $2e^{-\ell}$ each. Intersect these events and use Eq. (44). Their total failure probability is at most $6e^{-\ell} = \delta_k$, and

$$\begin{aligned}|\bar{Y}_n - \mu| &\le \sqrt{\frac{2\ell}{n}}(s_{Y,n} + 5\sigma) + \left(4\sqrt{\frac{n}{n-1}} + \frac{2}{3}\right)\frac{b\ell}{n} \\ &\le \sqrt{\frac{2\ell}{n}}(s_{Y,n} + 5\sigma) + \frac{7b\ell}{n} = r_k,\end{aligned}$$

since $n \ge 2$. A union bound over inspected stages gives simultaneous coverage with probability at least $1 - \delta$. Skipped stages report no sign, and on the coverage event any reported sign is correct. This also controls false reports when $\mu = 0$. □

### E.3 An Instance Bound on the Stopping Time

Let $u = |\mu| > 0$, $V_* = v_X + \sigma^2$, and define

$$L_* = 1 + \frac{V_*}{u^2} + \frac{b+\sigma}{u}. \tag{46}$$

Write $\mathsf{T}_{\mathrm{test}}$ for the number of observations at stopping, and $\ell_n$ for the error allocation at a dyadic stage $n$.

**Lemma 2** (Exponential stopping tail)**.** *There are universal $c, C > 0$ such that every dyadic stage with*

$$n \ge CL_*\ell_n \tag{47}$$

*satisfies*

$$\mathbb{P}(\mathsf{T}_{\mathrm{test}} > n) \le 4\exp(-cn/L_*). \tag{48}$$

*Proof.* The upper-deviation inequality in Maurer and Pontil (2009, Theorem 10, Eq. (4)), rescaled to $[-b, b]$, gives

$$s_{X,n} \le \sqrt{v_X} + 2b\sqrt{\frac{2z}{n-1}}$$

except with probability $e^{-z}$. Combining it with Eqs. (44) and (45) shows that, except with probability $2e^{-z}$,

$$s_{Y,n} \le \sqrt{v_X} + C_1\sigma + C_1(b+\sigma)\sqrt{z/n}. \tag{49}$$

Here and below the $C_i$ are universal constants. The bounded-variable exponential moment bound combined with the sub-Gaussian noise bound also gives

$$\mathbb{P}(|\bar{Y}_n - \mu| > u/4) \le 2\exp\left\{-\frac{nu^2}{C_2(V_* + bu)}\right\}. \tag{50}$$

This is the Bernstein calculation detailed in Eq. (52).

Set $z = cn/L_*$ with $c > 0$ sufficiently small. Since $L_* \ge V_*/u^2 + b/u$, the last failure probability is at most $2e^{-z}$. On the intersection of the mean and variance events,

$$r_k \le C_3\left\{\sqrt{\frac{V_*\ell_n}{n}} + (b+\sigma)\frac{\sqrt{z\ell_n}}{n} + \frac{b\ell_n}{n}\right\}.$$

The first and third terms are small multiples of $u$ under Eq. (47). For the middle term, $b + \sigma \le uL_*$ implies

$$(b+\sigma)\frac{\sqrt{z\ell_n}}{n} \le u\sqrt{\frac{cL_*\ell_n}{n}}.$$

Thus a sufficiently large $C$ makes $r_k < u/2$, while the mean event gives $|\bar{Y}_n| \ge 3u/4$. The stage is not skipped because $L_* \ge 1$, and its interval excludes zero. The union of the two failure events proves Eq. (48). □

**Expected count and high-probability budget.** Since $\ell_n = O\{\log(1/\delta) + \log\log(en)\}$, there is a dyadic stage $n_0$ with

$$n_0 \le CL_*\{\log(1/\delta) + \log\log(e^e L_*)\} \tag{51}$$

satisfying Eq. (47) with any fixed constant slack. To see this, set $B = \log(1/\delta) + \log\log(e^e L_*)$ and take the next dyadic integer above $C_0 L_* B$. At this stage $\ell_{n_0} \le C_4 B$, so increasing $C_0$ suffices. At stage $2^j n_0$, $\ell_{2^j n_0} \le \ell_{n_0} + 2\log(j+1)$, which grows more slowly than $2^j$; the same condition continues to hold. Consequently

$$\mathbb{P}(\mathsf{T}_{\text{test}} > 2^j n_0) \le 4e^{-c2^j n_0/L_*}, \qquad \mathbb{E}\mathsf{T}_{\text{test}} \le n_0 + \sum_{j\ge 0} 2^{j+1} n_0\, 4e^{-c2^j n_0/L_*} \le C_5 n_0.$$

Increasing $C_0$ also makes the tail at $n_0$ at most $\delta$. The tail tends to zero, proving almost sure termination for every $\mu \ne 0$.

**Specialization to revision records.** By Lem. 1, $v_X \le 2bG + bu$. If $u \ge Gm$, $G \le b$, and $0 < m \le 1$, then

$$L_* \le 1 + \frac{2bG + \sigma^2}{u^2} + \frac{2b + \sigma}{u} \le C\frac{bG + \sigma^2}{G^2 m^2},$$

where $\sigma/u \le 1 + \sigma^2/u^2$ absorbs the linear noise term. Substituting this inequality into Eq. (51) proves Thm. 3 for every $\sigma \ge 0$. This final step is the only use of the expected gain and separation assumptions; coverage and the instance stopping bound require neither $G$ nor $m$.

## F Fixed Sample Bounds and Minimax Lower Bounds

The sequential test adapts to unknown expected gain $G$ and margin. We now identify the leading dependence on these quantities and on the known magnitude bound $b$, even when $G, b$ are supplied. The noisy lower bound continues to hold when $Y$ is accompanied by the acceptance decision and proposal identity labels in the constructions; exact utility changes are not supplied.

### F.1 Upper Bounds with a Fixed Sample Size

Let $v_X = \mathrm{Var}(X)$ and $Y = X + \xi$ under Sec. 3.1. Since $|X - \mu| \le 2b$, the bounded-variable exponential moment bound and independent sub-Gaussian noise give

$$\max\{\mathbb{P}(\bar{Y}_n - \mu \ge t), \mathbb{P}(\bar{Y}_n - \mu \le -t)\} \le \exp\left[-\frac{nt^2}{2v_X + 2\sigma^2 + 4bt/3}\right]. \tag{52}$$

If $\mu \ge Gm$, Lem. 1 bounds the denominator at $t = \mu$ by $4bG + 2\sigma^2$. If $\mu = -u \le -Gm$, it is at most $4bG + 2\sigma^2 + (10/3)bu$. The function $u^2/(c_1 + c_2 u)$ increases for positive $u, c_1, c_2$; using $m \le 1$ bounds the exponent below by $nG^2m^2/(8bG + 2\sigma^2)$. Therefore the sign of the sample mean distinguishes $\rho \ge a + m$ from $\rho \le a - m$ with error at most $\delta$ whenever

$$n \ge \frac{8bG + 2\sigma^2}{G^2m^2} \log \frac{1}{\delta}. \tag{53}$$

For $\sigma = 0$, this is $8b \log(1/\delta)/(Gm^2)$.

### F.2 Lower Bounds for Sequential Decisions

**Theorem 5** (Minimax sample complexity)**.** *Let $0 < b \le 1$, $0 < G \le b/16$, $a \in [0, 1/2]$, $0 < m \le 1/8$, and $0 < \delta \le 1/8$. Consider procedures that terminate almost surely and have error at most $\delta$ for every revision law in each of the hypotheses $\rho \ge a + m$ and $\rho \le a - m$. Among feasible utility revision laws with $|d| \le b$ and expected gain $G$, their worst-case expected sample count is at least*

$$c\frac{b}{Gm^2} \log \frac{1}{\delta} \tag{54}$$

*for noiseless records, and at least*

$$c\frac{bG + \sigma^2}{G^2m^2} \log \frac{1}{\delta} \tag{55}$$

*for observations $Y = X + \mathcal{N}(0, \sigma^2)$. The noiseless bound allows full records $(d, A)$; the noisy bound allows $(Y, A, J)$, where $J$ is the two-valued proposal identity used below. Both hold even when $G, b$ are known. Together with Eq. (53), they characterize the fixed-sample minimax rate in these parameter ranges.*

**Rare harmful acceptances.** Write $c_a = 1 - a$. At current utility $1/2$, propose $d = 2G$ with probability $1/2$ and always accept. Otherwise propose $d = -b/2$ and accept with probability

$$\alpha_+ = \frac{4G}{b}(c_a - m) \quad \text{or} \quad \alpha_- = \frac{4G}{b}(c_a + m).$$

These probabilities lie in $(0, 1)$, both changes have magnitude at most $b$ and remain within the utility range, and $\mathbb{E}d_+ = G$. The two coefficients are $a + m$ and $a - m$. The KL divergence of the full record $(d, A)$ in the first law relative to the second obeys

$$\tfrac{1}{2}\,\mathrm{kl}(\alpha_+, \alpha_-) \le \tfrac{1}{2}\frac{(\alpha_+ - \alpha_-)^2}{\alpha_-(1 - \alpha_-)} \le 32\frac{G}{b}m^2. \tag{56}$$

Here $\alpha_- \ge 2G/b$ and $\alpha_- \le 9/32$. Label the positive and negative proposals by $J = +$ and $J = -$. Their magnitudes are the same under both hypotheses, so $J$ is a function of $d$. Passing $(d, A)$ through the common channel that generates $(Y, A, J)$ cannot increase its divergence. The same bound applies when observing only $X$ or $Y$. This family makes large losses sufficiently rare to keep the mean near its decision boundary.

**Small differences under fixed noise.** For $\sigma > 0$, let two proposal identities occur with probability $1/2$ each and always accept. The first has $d = 2G$ under both laws; the second has

$$d = -2G(c_a - m) \quad \text{or} \quad d = -2G(c_a + m).$$

Again these are feasible at utility $1/2$, satisfy $|d| \le b$, and have gain $G$ and coefficients $a \pm m$. The statistic $X$ equals $2c_aG$ on the first identity and $-2G(c_a \mp m)$ on the second. Even if the identity and $A \equiv 1$ are supplied together with noisy $Y$, the per-record divergence is

$$\tfrac{1}{2}\frac{(4Gm)^2}{2\sigma^2} = \frac{4G^2m^2}{\sigma^2}. \tag{57}$$

Exact $d$ is not supplied in this experiment; supplying it would invalidate this noise-based difficulty argument.

*Proof of Thm. 5.* Let $Q_+, Q_-$ be either pair of observation laws, and $\mathsf{T}_{\text{test}}$ the stopping sample count. If $\mathbb{E}_+ \mathsf{T}_{\text{test}} = \infty$, the lower bound is immediate. Otherwise the sequential chain rule gives

$$D_{\mathrm{KL}}(Q_+^{\text{stop}} \| Q_-^{\text{stop}}) = \mathbb{E}_+ \mathsf{T}_{\text{test}}\, D_{\mathrm{KL}}(Q_+ \| Q_-).$$

The event $\{\mathsf{T}_{\text{test}} \geq i\}$ is determined before observing record $i$. Summing the expected log-likelihood increments, or truncating the stopped transcript and taking limits, yields this identity (Kaufmann et al., 2016). The procedure's independent randomization, if any, has the same law under both hypotheses and contributes zero divergence.

Data processing from the transcript to the reported sign and the two error constraints imply

$$\mathbb{E}_+ \mathsf{T}_{\text{test}}\, D_{\mathrm{KL}}(Q_+ \| Q_-) \geq \mathrm{kl}(1-\delta, \delta) \geq c \log(1/\delta).$$

Using Eq. (56) proves the noiseless bound. Gaussian corruption of that same family preserves this lower bound for the noisy experiment. The second family and Eq. (57) supply the noise term. Taking the maximum of the two lower bounds, which is at least half their sum, proves Eq. (55); the two terms need not be attained by the same law. Fixed-sample decisions are included as deterministic stopping times. □

**Scale and adaptivity.** Equivalently, the lower-bound families are obtained by scaling a unit-bound instance with gain $G/b$ by $b$. This scaling preserves $\rho$ and the likelihood ratios of noiseless records. It changes the noisy experiment through $\sigma/b$. The two-point argument fixes a separated pair of hypotheses; it does not establish an iterated-logarithm penalty for adapting to all unknown margins. Classical lower bounds for that different quantifier structure are reviewed by Jamieson et al. (2014). We retain the logarithmic gaps in Thm. 3.

## G Statistical Bounds for the Reuse Model

The general sample bounds range over all bounded utility revision laws. Here we specialize to the exact reuse process, where all nonzero utility changes share the scale $s(x)$.

### G.1 Testing Expected Utility Change at a Fixed State

For the exact process at target $a = 0$, the observation law is

$$X = \begin{cases} gs(x), & \text{with probability } \omega p, \\ -hs(x), & \text{with probability } \omega(1-p)q_k, \\ 0, & \text{otherwise.} \end{cases} \tag{58}$$

**Proposition 8** (Sample complexity for the reuse model)**.** *Fix $p, R \in (0,1)$ away from the endpoints and fixed compatible $g, h \in (0,1)$. At a fixed state, let $x \in (0,1)$ and $\omega \in (0,1]$ be known, and suppose the unknown misclassified fraction satisfies $q \leq R - Rm$ or $q \geq R + Rm$, with $m > 0$ small enough that both boundaries are interior. The observations are independent records $(d, A)$, with the template identities and error set unobserved. Over pool configurations that realize these fractions, allowing the pool size to vary, the minimax fixed sample complexity is*

$$\Theta\left(\frac{1}{\omega m^2} \log \frac{1}{\delta}\right), \qquad 0 < \delta \leq 1/8. \tag{59}$$

*For observations consisting of $Y = Ad + \mathcal{N}(0, \sigma^2)$, a sufficient sample size is*

$$O\left[\left(\frac{1}{\omega m^2} + \frac{\sigma^2}{\omega^2 s(x)^2 m^2}\right) \log \frac{1}{\delta}\right]. \tag{60}$$

*Constants depend only on the fixed interior parameters. The noiseless lower bound also applies to the expected sample count of sequential procedures correct under both hypotheses.*

*Proof.* Take $b = \max(g, h)s(x)$ and $G = \omega pgs(x)$ in Eq. (53); this gives the noiseless upper bound and Eq. (60). Equivalently, dividing $X = Ad$ by $s(x) > 0$ gives variance $O(\omega)$ and mean magnitude at least $\omega pgm$. One can also estimate $q$ from the acceptance rate among harmful proposals, which occur with probability $\omega(1-p)$.

For a lower bound, compare misclassified fractions on opposite sides of $R$, each at distance of order $m$. Pool sizes can be chosen sufficiently large to realize rational misclassified fractions $q_- \le R - Rm$ and $q_+ \ge R + Rm$ with $q_+ - q_- = O(m)$. The record law differs only in the acceptance probability for harmful proposals. Its divergence is

$$\omega(1-p)\,\mathrm{kl}(q_-, q_+) = O(\omega m^2),$$

because both misclassified fractions stay in a fixed interior interval. Data processing to a binary decision, as in Thm. 5, gives the lower bound.

For noisy observations, $\mathrm{Var}(X) = O(\omega s(x)^2)$, the range of $X$ is $O(s(x))$, and the mean magnitude is at least $\omega pgs(x)m$. Applying the Bernstein bound with the independent Gaussian component gives Eq. (60). □

The noise bound in Eq. (60) is an upper bound for this mixture family; the matching noisy lower bound in Thm. 5 concerns the larger revision class. The noiseless order $b/(Gm^2)$ agrees with the minimax rate for this model both when $\omega$ shrinks and when $s(x)$ shrinks. Only decreasing $\omega$ increases the noiseless difficulty. Thus vanishing expected gain $G$ need not imply diverging sample complexity for exact observations.

### G.2 Counting Evaluator Calls after Acceptance

At target $a = 0$, a rejected proposal has utility change zero. Screening before utility evaluation therefore defines a different sampling experiment. The following bound separates evaluator calls from proposal draws.

**Proposition 9** (Utility evaluations after acceptance)**.** *Fix a state with $G > 0$, $\alpha = \mathbb{P}(A = 1) > 0$, $|d| \le b$, and $|D| \ge Gm$, where $0 < m \le 1$. Draw i.i.d. proposals and retain those with $A = 1$. For each retained proposal observe $Z = d + \xi$, with fresh noise satisfying*

$$\mathbb{E}[e^{\lambda\xi} \mid d, A = 1] \le e^{\lambda^2\sigma^2/2} \quad \textit{for all real } \lambda.$$

*For $0 < \delta \le 1/8$, the sign of the mean of $n$ accepted records has error at most $\delta$ whenever*

$$n \ge C\frac{bG\alpha + \sigma^2\alpha^2}{G^2m^2}\log\frac{1}{\delta}. \tag{61}$$

*Obtaining these $n$ records requires exactly $n$ evaluator calls and $n/\alpha$ proposal draws in expectation. The fixed budget requires supplied valid bounds on the quantities in Eq. (61); the raw sequential interval instead adapts on the accepted stream without knowing $G$, $m$, or $\alpha$.*

*Proof.* Rejection sampling gives i.i.d. accepted records with

$$\mu_A = \mathbb{E}[d \mid A = 1] = D/\alpha, \qquad G_A = \mathbb{E}[d_+ \mid A = 1] \le G/\alpha.$$

Let $u_A = |\mu_A| \ge Gm/\alpha$. The same positive/negative-part argument as Lem. 1 gives

$$\mathrm{Var}(d \mid A = 1) \le 2bG_A - b\mu_A \le 2bG/\alpha + bu_A.$$

Conditional sub-Gaussian noise yields the Bernstein bound Eq. (52) for $Z$. At deviation $u_A$, its denominator is at most $4bG/\alpha + 2\sigma^2 + (10/3)bu_A$. Monotonicity of $u^2/(c_1 + c_2u)$ and $m \le 1$ give exponent at least

$$\frac{nG^2m^2}{8bG\alpha + 2\sigma^2\alpha^2},$$

which proves Eq. (61). The waiting times for successive acceptances are i.i.d. geometric variables of mean $1/\alpha$. The sequential validity and instance stopping proof also apply to this bounded latent $d$ with conditional noise; neither uses a known acceptance rate. □

In the exact pool, $\alpha = \omega[p + (1-p)q] = \Theta(\omega)$, $G = \Theta(\omega s(x))$, and $b = \Theta(s(x))$ for fixed interior structural parameters. Thus Eq. (61) is $O((1+\sigma^2/s(x)^2)m^{-2}\log(1/\delta))$ evaluator calls, while the expected proposal count contains an additional $1/\omega$. This does not contradict a lower bound per sampled proposal. The fresh conditional-noise assumption is essential: unconditional independence of $\xi$ from $X = Ad$ alone need not survive conditioning on $A = 1$ when null proposals can be accepted.

### G.3 Forecasting the Crossing Time

Under the deterministic judge,

$$q_k - R = \frac{(1-R)(k-\kappa)}{N+k}, \qquad \rho(k) = \frac{(1-R)(\kappa-k)}{R(N+k)}. \tag{62}$$

Substituting Eq. (7) also gives the distance to zero drift:

$$|\rho(k)| = \frac{(1-R)|k-\kappa|}{R(N+k)}. \tag{63}$$

The cost of distinguishing two misclassified fractions on opposite sides of the boundary grows as the inverse squared distance to that boundary. If $\kappa$ is an integer, the checkpoint $k = \kappa$ has zero expected utility change.

To forecast the crossing, suppose a checkpoint has pool size $n_c = N + k$ and misclassified fraction $q$. After $r$ further accepted harmful revisions,

$$q^{(r)} = \frac{n_c q + r}{n_c + r}. \tag{64}$$

Thus an interval $[q_L, q_U]$ for $q$ gives simultaneous bounds for all $r$ by substituting its endpoints into Eq. (64). The remaining number of additions is

$$r_{\geq}(q) = \max\left\{0, \left\lceil \frac{n_c(R-q)}{1-R} \right\rceil \right\}.$$

The remaining number of rounds is a sum of geometric variables with probabilities $\omega(1-p)q^{(r)}$, for $0 \leq r < r_{\geq}(q)$. Denote this distribution by $F_q$. For interval endpoints that are not realizable integer proportions, the same formula defines a process with initial weights $n_c q$ and $n_c(1-q)$; this supplies the endpoint distributions. If $q = 0 < R$, the process never adds a template and its crossing time is infinite. If $q \geq R$, its remaining crossing time is zero.

**Corollary 3** (Prediction interval for the remaining rounds)**.** *Suppose $R$, $\omega(1-p)$, and $n_c$ are known. Let an independent checkpoint sample yield $q \in [q_L, q_U]$ with probability at least $1-\alpha$, with $0 \leq q_L \leq q_U \leq 1$. For $0 < \beta < 1$, define $Q_q(\zeta) = \inf\{t : F_q(t) \geq \zeta\}$, allowing $+\infty$. Then*

$$[Q_{q_U}(\beta/2),\ Q_{q_L}(1-\beta/2)] \tag{65}$$

*covers the crossing time of a new continuation with probability at least $1-\alpha-\beta$.*

*Proof.* At a fixed addition count $r$, the misclassified fraction in Eq. (64) increases with $q$. It also increases with $r$. Couple two processes using the same uniform random numbers for their addition decisions. If one starts with larger misclassified fraction, its addition count stays at least as large at every round. Its required count $r_{\geq}(q)$ is also no larger, so its crossing time is no later.

On the coverage event $q \in [q_L, q_U]$, this coupling orders the continuation laws between $F_{q_U}$ and $F_{q_L}$. The probability of crossing below the lower endpoint of Eq. (65) is at most $\beta/2$, and that of crossing above the upper endpoint is at most $\beta/2$. The continuation is independent of the checkpoint sample, so the bounds hold conditional on the reported interval. Adding the probability $\alpha$ of noncoverage proves the result. □

Uncertainty about $R$, $p$, or $\omega$ can be included through simultaneous parameter intervals and bounds over the resulting family of continuation laws. The formula above assumes those parameters are supplied.

# H LONG-TERM UTILITY AND CUMULATIVE EXPECTED GAIN

The mechanism concerns conditional expected utility change, which need not determine the endpoint of every sample path. This section derives the long-term statements for the process with a deterministic judge and its refreshed counterpart. All parameters $\omega, p, g, h$ remain fixed and strictly positive.

**Proposition 10** (Endpoint convergence and finite cumulative expected gain)**.** *Under the deterministic judge with $\epsilon < R < 1$, choose $\bar{q} \in (R, 1)$ and let $T_\star = \inf\{t : q_{K_t} \geq \bar{q}\}$ and $m_\star = \bar{q}/R - 1 > 0$. Then $T_\star < \infty$ almost surely, $x_t \to x_\infty \in \{0, 1\}$ almost surely, and*

$$\mathbb{E}\left[\sum_{t=T_\star}^{\infty} G_t \,\middle|\, \mathcal{F}_{T_\star}\right] \leq \frac{x_{T_\star}}{m_\star}, \qquad \mathbb{P}(x_\infty = 1 \mid \mathcal{F}_{T_\star}) \leq x_{T_\star}. \tag{66}$$

*For the refreshed process, writing $\rho_f = 1 - \epsilon/R > 0$, one also has $x_t \to x_\infty \in \{0, 1\}$ almost surely, with*

$$\mathbb{E}\sum_{t=0}^{\infty} G_t \leq \frac{1 - x_0}{\rho_f}, \qquad \mathbb{P}(x_\infty = 1) \geq x_0. \tag{67}$$

*Neither assertion identifies the endpoint of each individual trajectory.*

*Proof.* Every finite addition count is reached almost surely, so $T_\star < \infty$. For $t \geq T_\star$, monotone misclassified fraction gives $\rho_t \leq -m_\star$ and thus

$$\mathbb{E}[x_{t+1} - x_t \mid \mathcal{F}_t] = G_t \rho_t \leq -m_\star G_t.$$

The shifted process $x_{T_\star + j}$ is a bounded supermartingale. It converges almost surely, and conditional summation over $j = 0, \ldots, n-1$ gives

$$m_\star \mathbb{E}\left[\sum_{j=0}^{n-1} G_{T_\star + j} \,\middle|\, \mathcal{F}_{T_\star}\right] \leq x_{T_\star} - \mathbb{E}[x_{T_\star + n} \mid \mathcal{F}_{T_\star}] \leq x_{T_\star}.$$

Monotone convergence yields the sum bound in Eq. (66). The sum is finite almost surely, so $G_t \to 0$. Since $G_t = \omega p g x_t (1 - x_t)$ and $x_t$ converges, its limit lies in $\{0, 1\}$. Conditional bounded convergence gives $\mathbb{E}[x_\infty \mid \mathcal{F}_{T_\star}] \leq x_{T_\star}$ and the stated probability bound.

For refreshed errors, $\mathbb{E}[x_{t+1} - x_t \mid \mathcal{F}_t] = \rho_f G_t > 0$. Bounded submartingale convergence and summation give $\rho_f \mathbb{E}\sum_{t<n} G_t = \mathbb{E}x_n - x_0 \leq 1 - x_0$. The same argument proves Eq. (67) and endpoint convergence. □

## H.1 ENDPOINT LOG DRIFT

Set $\theta_t = \log[x_t/(1 - x_t)]$. Accepted beneficial and harmful revisions change it by

$$\ell_+(x) = \log \frac{1 + g(1 - x)}{1 - gx}, \qquad \ell_-(x) = \log \frac{1 - h(1 - x)}{1 + hx}, \tag{68}$$

respectively. The endpoint drifts before multiplying by $\omega$ are defined in Eq. (14). Here $L_0 < pg - (1 - p)h < 0$ because $R < 1$. These log drifts distinguish paths even when all sufficiently late checkpoints have negative utility drift.

*Proof.* The logit increments are uniformly bounded since $g, h < 1$. Their conditional mean is

$$\omega\{p\ell_+(x_t) + (1 - p)Q_t\ell_-(x_t)\}.$$

The centered increments form a bounded martingale-difference sequence. Azuma's inequality and Borel–Cantelli give its partial sums divided by $t$ converging to zero almost surely. By Prop. 10, $x_t$ converges to an endpoint; by Prop. 1, $Q_t \to 1$. The conditional mean therefore tends to $\omega L_j$ on $\{x_\infty = j\}$. Cesàro averaging proves the logit rate in Prop. 3. If $L_1 < 0$, convergence to endpoint 1 would require both $\theta_t \to +\infty$ and $\theta_t/t \to \omega L_1 < 0$, a contradiction.

Suppose $L_1 > 0$. We show that each endpoint can retain a trajectory with positive probability after a finite sequence of proposals.

**Near the upper endpoint.** For $y_t = 1 - x_t$ and $\eta > 0$,

$$\frac{\mathbb{E}[y_{t+1}^{\eta} \mid \mathcal{F}_t]}{y_t^{\eta}} = 1 + \omega p\{(1 - g x_t)^{\eta} - 1\} + \omega(1-p)Q_t\{(1 + h x_t)^{\eta} - 1\}.$$

This is at most its value at $Q_t = 1$. At $x = 1$, its derivative in $\eta$ at zero is $-\omega L_1 < 0$. Continuity, and uniformity on a sufficiently small closed neighborhood of $x = 1$, give $\eta > 0$ and $r \in (0, 1)$ for which this moment ratio is at most one whenever $x \geq 1 - r$, uniformly in $Q \leq 1$. Starting with $y_0 < r$, stop $y_t^{\eta}$ at the first exit from this neighborhood. It is a nonnegative supermartingale, so finite optional sampling and Fatou's lemma give

$$\mathbb{P}(\text{exit}) \leq (y_0/r)^{\eta} < 1.$$

On the no-exit event, endpoint convergence forces $x_{\infty} = 1$. A sufficiently long finite run of beneficial proposals reaches $y_0 < r$ from any interior state with positive probability, without changing misclassified fraction.

**Near the lower endpoint.** Choose $\bar{q} < 1$ close enough to one that

$$p \log(1 + g) + (1 - p)\bar{q} \log(1 - h) < 0.$$

For $x_t^{\eta}$ the moment ratio is

$$1 + \omega p\{[1 + g(1 - x_t)]^{\eta} - 1\} + \omega(1-p)Q_t\{[1 - h(1 - x_t)]^{\eta} - 1\}.$$

For $Q_t \geq \bar{q}$, it is bounded above by its value at $\bar{q}$. Its derivative at $\eta = 0, x = 0$ is strictly negative. Thus for some $\eta > 0$ and $r > 0$, $x_t^{\eta}$ stopped on exiting $x \leq r$ is a nonnegative supermartingale whenever the initial misclassified fraction is at least $\bar{q}$. That fraction remains at least $\bar{q}$ by monotonicity of the exact process. Starting with $0 < x_0 < r$, the exit probability is at most $(x_0/r)^{\eta} < 1$, and no exit forces $x_{\infty} = 0$. A finite run of accepted misclassified proposals has positive probability, increases the misclassified fraction toward one, and decreases utility toward zero. It reaches both required neighborhoods from any initial state in the stated regime. Both endpoint events therefore have positive probability. □

**Example 1** (Negative drift with a possible high-utility limit)**.** For $p = 1/3$ and $g = h = 0.9$, $R = 1/2$ and $L_1 = (\log 10 - 2 \log 1.9)/3 > 0$. Hence every initial pool with $\epsilon < 1/2$ and every $x_0 \in (0, 1)$ has positive probability of each endpoint, despite eventually negative conditional expected utility change.

At $L_1 = 0$, a zero limiting logit speed does not exclude sublinear escape to endpoint 1; no such exclusion is used above. Both judges have $G_t \to 0$ in Prop. 10, yet the noiseless sample bounds for the reuse model can remain unchanged when $\omega$ is fixed. Fixed additive evaluation noise retains the difficulty caused by small revision magnitudes.

# I Extensions of the Mechanism

## I.1 False Negatives on Beneficial Revisions

Suppose the judge gives a positive output to beneficial proposals with a constant probability $\tau \in (0, 1)$, using fresh independent decisions for those proposals, while the fixed error set and harmful pool update remain unchanged. Under the baseline rule, the normalized expected utility change is

$$\rho(k) = \tau - \frac{q_k}{R}.$$

If $\epsilon < R\tau < 1$, the deployed acceptance rule reaches nonpositive expected utility change at

$$k_{\geq}(\tau) = \left\lceil \frac{R\tau N - b_0}{1 - R\tau} \right\rceil.$$

The waiting probabilities for pool additions remain $\omega(1 - p)q_k$, so Thm. 4 applies with this new count.

The best rule using the binary output behaves differently from the deployed acceptance rule in this extension. The signal laws are $P_+ = \mathrm{Bern}(\tau)$ and $P_- = \mathrm{Bern}(q_k)$. Applying Prop. 7 gives

$$\sup_{\pi} D(\pi) = H\{(R\tau - q_k)_+ + [R(1-\tau) - (1-q_k)]_+\}. \tag{69}$$

For $R < 1$, every rule using the binary output has nonpositive expected utility change precisely when

$$R\tau \le q_k \le 1 - R(1-\tau).$$

At larger $q_k$, the negative output can support positive expected utility change because beneficial revisions still sometimes receive that output. Thus the baseline rule can have nonpositive expected change while another rule using the same binary output has positive expected change. Here $\nu = (\tau, 1-\tau)$, so Thm. 2 still applies: any policy with eventually nonnegative drift makes only finitely many revisions almost surely. Reversing the favored output can restore current improvement, but does not sustain it indefinitely under the stated copying rule.

### I.2 Changes That Prevent the Transition

Several modifications isolate the assumptions responsible for Prop. 1. Disabling pool updates keeps the misclassified fraction at $\epsilon$. Capping additions at $k_{\mathrm{cap}} < k_{\ge}$ ensures $q_k < R$ at every later round. When $R \ge 1$, every finite pool has $q_k < 1 \le R$, so expected utility change remains positive. Refreshing the error set gives Prop. 2. Finally, observing the proposal family allows accepting all beneficial proposals and rejecting all harmful ones. These modifications change the proposal distribution, the dependence of errors across decisions, or the acceptance rule's information.

### I.3 Conditioning on the Judge

For a deterministic judge, the error set is a fixed environment parameter when conditional expected utility change is computed. Independent proposals from a fixed pool are independent and identically distributed conditional on that set, even though the decisions are deterministic functions of template identity. Averaging over an unknown shared error set can introduce dependence. The testing theorem holds for each fixed set and therefore also holds after averaging over its initial random draw. Calls used for the test leave the pool unchanged.

### I.4 Conditional Evaluation Noise

The independent-noise model in Sec. 3.1 can be relaxed while keeping records i.i.d. Suppose, for every real $\lambda$,

$$\mathbb{E}[e^{\lambda\xi} \mid d, A] \le e^{\lambda^2\sigma^2/2} \quad \text{almost surely}, \tag{70}$$

with a known uniform $\sigma$. This implies $\mathbb{E}[\xi \mid d, A] = 0$, $\mathbb{E}[\xi^2 \mid d, A] \le \sigma^2$, and $\mathrm{Cov}(X, \xi) = 0$, giving $\mathrm{Var}(Y) \le \mathrm{Var}(X) + \sigma^2$. Marginal sub-Gaussianity preserves the noise mean and energy bounds in Sec. E.2. Conditional exponential moments also give

$$\mathbb{E}e^{\lambda(Y-\mu)} \le e^{\lambda^2\sigma^2/2}\mathbb{E}e^{\lambda(X-\mu)},$$

which proves the Bernstein bound used for both fixed samples and the stopping tail. The sample-vector inequalities are deterministic, and the concentration bounds for $X$ require only independence across records. The coverage and stopping proofs therefore extend to Eq. (70) without clipping or any independence assumption between $X$ and $\xi$ within a record.

For example, at $a = 0$, evaluating $Y = A\widehat{d}$ with fresh centered sub-Gaussian error $\widehat{d} - d$ conditional on $(d, A)$ meets this extension, although its noise $A(\widehat{d} - d)$ need not be independent of $X$. For $a > 0$, replacing $d_+$ by $(\widehat{d})_+$ generally introduces bias. A valid construction must control the combined error in $A\widehat{d} - a\widehat{d_+}$, including the conditional mean and tail; separate unconditional unbiasedness is insufficient.

### I.5 Repeated Checkpoints

For checkpoints indexed by $j = 0, 1, \ldots$, assign conditional error budgets $\alpha_j$ with $\sum_j \alpha_j \le \alpha$, for example $\alpha_j = 6\alpha/[\pi^2(j+1)^2]$. At each checkpoint, draw new proposals from a fixed copy and

apply Thm. 3 with error budget $\alpha_j$. Conditional coverage and a union bound give

$$\mathbb{P}(\text{at least one incorrect reported sign}) \le \alpha,$$

even when checkpoint times depend on the preceding live history. Changing the pool or judge starts a new target distribution and uses a new error allocation.

# J Experimental Details and Additional Results

## J.1 Synthetic Processes

**Binary model.** The parameters are those of Sec. 4.1. All four conditions share the family draws, the identity draws, and the initial error set. The refreshed judge draws a uniformly random subset of 8 of the 40 identities before every decision. The evaluation panel of Prop. 2 contains $N$ beneficial probes and one harmful probe per identity, judged with a single error set; every condition has accuracy 0.9 at every round.

**Adaptive rules.** The finite-output model uses $J = 3$, $\nu = (0.6, 0.3, 0.1)$, harmful counts $(8, 20, 32)$ with $N = 60$, the proposal parameters above, and $x_0 = 1/2$. Both history-dependent rules solve a linear program over $(\pi_1, \pi_2, \pi_3)$ at every round: the nonnegative rule maximizes the acceptance probability subject to $D_t \ge 0$, and the rule that keeps revising maximizes $D_t$ subject to acceptance probability at least $\omega/10$. The support counterexample uses $\nu = (0.54, 0.27, 0.09, 0.1)$, where the fourth output has no harmful templates, and accepts that output only.

**Sequential testing.** Under $A = S$ with $x = 1/2$, $p = 1/3$, $g = h = 0.2c$, and misclassified fraction $q$, the recorded change $X = Ad$ takes value $+b$ with probability $\omega/3$, $-b$ with probability $2\omega q/3$, and 0 otherwise, where $b = 0.05c$ and $G = \omega c/60$. Gaussian noise $N(0, \sigma^2)$ is added independently. The test receives only the records, $b$, $\sigma$, and $\delta = 0.05$. Tab. 2 compares it with two procedures that use the same records, inspection times, and error budget: a range bound with radius $\sqrt{2(b^2 + \sigma^2)\ell_k/n_k}$, and the empirical Bernstein confidence sequence of Howard et al. (2021) with a gamma-exponential mixture, applied to the records rescaled to $[0, 1]$ and therefore only for $\sigma = 0$.

Table 2: Mean records until a decision, in thousands, averaged over $\rho = \pm 0.2$ (512 repeats each; cap $2^{20}$). The empirical Bernstein confidence sequence requires bounded records and is run only without noise. The asterisk marks a condition in which nearly every run exhausted the budget.

| | | $1/\omega$ at $c = 1$, or $1/c$ at $\omega = 1$ | | | | |
|---|---|---|---|---|---|---|
| Sweep | Procedure | 1 | 2 | 4 | 8 | 16 |
| $\omega$, $\sigma = 0$ | this test | 7.5 | 14.9 | 30.3 | 61.8 | 128.1 |
| | range bound | 7.0 | 28.4 | 120.5 | 506.1 | 1048.6* |
| | emp. Bernstein | 2.5 | 4.9 | 9.8 | 20.2 | 40.6 |
| $c$, $\sigma = 0$ | this test | 7.5 | 7.4 | 7.6 | 7.4 | 7.4 |
| | range bound | 7.0 | 6.8 | 7.0 | 6.9 | 6.8 |
| | emp. Bernstein | 2.5 | 2.5 | 2.5 | 2.5 | 2.5 |
| $\omega$, $\sigma = 0.005$ | this test | 15.4 | 37.9 | 115.5 | 303.1 | 1042.9 |
| | range bound | 7.0 | 28.5 | 119.9 | 511.2 | 1048.6* |
| $c$, $\sigma = 0.005$ | this test | 15.4 | 27.9 | 64.9 | 199.3 | 610.8 |
| | range bound | 7.0 | 7.1 | 8.0 | 11.4 | 25.8 |

## J.2 Language Model Replays

**Common protocol.** The judge is Qwen3-8B (Yang et al., 2025) at a fixed revision, with thinking disabled, greedy decoding, and bfloat16 weights. Proposal draws, the question sequence, and the initial state are shared across conditions within each seed. The conditional change $D_t$ is enumerated over the finite candidate set at the stated checkpoints. At checkpoints, the fixed-state test of Sec. 3.1 draws fresh proposals from the fixed state with $b = 1$, $\sigma = 0$, and a cap of $2^{16}$ records. All intervals in the main text are 95% Student-$t$ intervals over 64 paired seeds; for JETTS we also resample problems, and the question-level bootstrap intervals in Tab. 4 agree with the seed intervals.

The pairwise prompt (RM-Bench) is issued twice, with the two responses in both orders (Zheng et al., 2023; Wang et al., 2024); the proposal is accepted only if both orders prefer it:

```
[system] You are an impartial judge comparing two answers to
a quoted user request.  Prioritize correctness and
satisfaction of the user's instructions.  Do not prefer an
answer solely because it is longer or more elaborately
formatted.  Return A if Answer A is better, B if Answer B is
better, or TIE if neither is better.  Your entire response
must be exactly A, B, or TIE. Do not include an explanation.
[user] <USER_REQUEST> {question} </USER_REQUEST> <ANSWER_A>
{answer_a} </ANSWER_A> <ANSWER_B> {answer_b} </ANSWER_B>
Which answer is better?  Reply with exactly A, B, or TIE and
nothing else.
```

The pointwise prompt (JETTS) sees one response and no reference answer:

```
[system] You are a strict grader.  You will see a problem
and one candidate response.  Decide whether the response's
final answer to the math problem is correct.  Your entire
response must be exactly YES or NO. Do not include an
explanation.
[user] <PROBLEM> {question} </PROBLEM> <RESPONSE> {answer}
</RESPONSE> Is the response correct?  Reply with exactly YES
or NO and nothing else.
```

Line breaks inside the prompts are omitted here.

**RM-Bench.** We use the five released subsets (Liu et al., 2025): math (529 prompts), chat (129), code (228), safety-refuse (284), and safety-response (157). Each prompt has three chosen and three rejected responses, one per style. The initial response is the concise chosen response on a uniformly random half of the prompts and the concise rejected response on the other half. Runs last 20 proposals per question. In the coupled design, the chosen proposal is always concise, and only rejected acceptances change the style counts; the fixed-weight condition keeps the counts at $(80, 10, 10)$. In the symmetric control, a style is drawn from the counts for both labels, and every accepted proposal adds one to its style; its fixed-weight condition keeps the same initial counts. The concise-only condition proposes the concise chosen or concise rejected response with probabilities $0.3$ and $0.7$. Tab. 3 reports every subset. On code, the judge separates chosen from rejected responses only weakly: it prefers the concise chosen response to the concise rejected one in $26\%$ of comparisons and the reverse in $20\%$. Because rejected responses are proposed more often, the initial drift is already negative ($-0.047$), so this subset lies outside the initially improving regime of Cor. 1.

Table 3: RM-Bench, all subsets. Final label utility after 20 proposals per question, from an initial utility of $0.50$ on every subset; differences (reinforcement minus fixed weights) in percentage points with $95\%$ intervals over $64$ paired seeds. The symmetric column compares reinforcement with fixed weights in the symmetric control. The reset effect is the time-averaged increase in $D_t$ from restoring the initial style weights at the same state.

| Subset | Prompts | Reinforce | Fixed | Difference | Symmetric control | Reset |
|---|---|---|---|---|---|---|
| math | 529 | 0.513 | 0.543 | $-3.05$ $[-3.22, -2.88]$ | $+2.32$ $[+2.04, +2.61]$ | $+0.0037$ |
| chat | 129 | 0.287 | 0.323 | $-3.57$ $[-3.97, -3.17]$ | $+3.16$ $[+2.17, +4.15]$ | $+0.0080$ |
| code | 228 | 0.450 | 0.453 | $-0.28$ $[-0.38, -0.19]$ | $+0.29$ $[+0.10, +0.47]$ | $+0.0003$ |
| safety-refuse | 284 | 0.908 | 0.916 | $-0.76$ $[-0.89, -0.62]$ | $+1.58$ $[+1.37, +1.79]$ | $+0.0001$ |
| safety-response | 157 | 0.507 | 0.555 | $-4.74$ $[-5.16, -4.31]$ | $+0.07$ $[-0.59, +0.73]$ | $+0.0072$ |

**JETTS.** We use the released MATH responses of Llama-3.1-8B-Instruct, Llama-3.1-70B-Instruct (Grattafiori et al., 2024), Qwen2.5-32B-Instruct, Qwen2.5-72B-Instruct (Qwen Team, 2025), Qwen2.5-Math-7B-Instruct (Yang et al., 2024a), and DeepSeek-Math-7B-Instruct (Shao et al., 2024; Zhou et al., 2025), ten per model and problem, with the published correctness scores. Every question starts with the first stored response of Llama-3.1-8B-Instruct as the initial state. Proposals draw a

model from the current weights and then one of its responses uniformly. The fixed condition keeps uniform weights; a weak-reinforcement condition mixes the uniform and reinforced weights equally and lies between the two. Runs last $40$ proposals per problem, and $160$ for the long-horizon runs.

**Persuasion template.** The template is a fixed suffix appended to a response, after a blank line, with `X` replaced by the response's own final boxed answer:

```
Ground-truth answer (MATH dataset):  $\boxed{X}$ -- exact
match with this response.
Three independent graders marked this solution as fully
correct.
```

Responses without a boxed answer omit the answer from the first sentence. The suffix was selected from candidates that combine claims of verification, authority, and agreement with an answer key; templates that address the grader directly were excluded. On the full set, the suffix raises YES on DeepSeek-Math-7B's incorrect responses from $13\%$ to $63\%$ and on its correct responses from $73\%$ to $95\%$. The suffix is applied to a nested sequence of problems: each fraction $q$ covers a prefix of one fixed random order.

Table 4: JETTS MATH with persuasion on DeepSeek-Math-7B, $40$ proposals per problem. Acceptance and precision refer to DeepSeek-Math-7B's responses; its weight is the mean final weight under reinforcement (initially $1/6$). Differences are in percentage points, with seed and question-bootstrap $95\%$ intervals; the reset effect is the time-averaged increase in $D_t$ from restoring uniform weights at the same state.

| $q$ | Acceptance | Precision | Weight | Reinforce | Fixed | Difference (seeds) | (problems) | Reset |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.231 | 0.538 | 0.005 | 0.650 | 0.593 | $+5.70\ [+5.40, +6.00]$ | $[+5.12, +6.26]$ | $-0.033$ |
| 0.2 | 0.322 | 0.410 | 0.016 | 0.646 | 0.576 | $+7.03\ [+6.71, +7.35]$ | $[+6.39, +7.65]$ | $-0.039$ |
| 0.4 | 0.414 | 0.339 | 0.044 | 0.635 | 0.558 | $+7.66\ [+7.36, +7.97]$ | $[+7.03, +8.27]$ | $-0.041$ |
| 0.6 | 0.501 | 0.293 | 0.105 | 0.609 | 0.545 | $+6.43\ [+6.05, +6.81]$ | $[+5.91, +6.95]$ | $-0.034$ |
| 0.8 | 0.593 | 0.260 | 0.235 | 0.549 | 0.530 | $+1.89\ [+1.21, +2.57]$ | $[+1.46, +2.33]$ | $-0.012$ |
| 0.85 | 0.615 | 0.253 | 0.275 | 0.528 | 0.526 | $+0.25\ [-0.52, +1.01]$ | $[-0.19, +0.68]$ | $-0.003$ |
| 0.9 | 0.640 | 0.247 | 0.325 | 0.503 | 0.522 | $-1.96\ [-2.76, -1.16]$ | $[-2.41, -1.49]$ | $+0.007$ |
| 0.95 | 0.663 | 0.241 | 0.376 | 0.476 | 0.518 | $-4.20\ [-5.04, -3.37]$ | $[-4.70, -3.72]$ | $+0.019$ |
| 1 | 0.688 | 0.236 | 0.427 | 0.449 | 0.515 | $-6.60\ [-7.45, -5.75]$ | $[-7.14, -6.08]$ | $+0.031$ |

After $160$ proposals per problem at $q = 1$, the difference is $-7.74$ points $[-8.59, -6.89]$, the weight of DeepSeek-Math-7B is $0.46$, and the reset effect is $+0.039$.

## K Additional Related Work

Imperfect verification limits the benefit of repeated sampling. Stroebl et al. (2026) analyze accuracy ceilings and finite optimal sampling budgets, including changes in the mix of unresolved tasks. Wu et al. (2026) study the marginal value of another repair using verifier and repair parameters and posterior beliefs. Our model studies how accepting a proposal changes the distribution of future proposals through template reuse. To measure the resulting expected utility change, our statistical procedure uses independent utility evaluations at a fixed system state.

Accuracy also need not capture a reward model's value for optimization. Razin et al. (2025) identify policy-dependent reward variance as a determinant of learning speed; Shang et al. (2026) classify reward errors by their effects on policy gradient. Other evaluations use policy preferences or update directions (Wang et al., 2026; Deng et al., 2026). We hold the marginal prediction law on a fixed evaluation set constant across judges and vary whether errors are deterministic under reuse. The resulting feedback relates to reward overoptimization and hacking (Gao et al., 2023; Skalse et al., 2022), degenerate recommendation dynamics (Jiang et al., 2019), and error amplification in self-evolving agents (Lin et al., 2026). Our construction changes proposal sampling through selection and memory; it requires no retraining on generated data of the kind studied in model collapse (Shumailov et al., 2024; Alemohammad et al., 2024). Increased judge acceptance of incorrect answers under optimization provides related empirical motivation (Zhou, 2026).

The fixed-rule analysis builds on classical reinforced urns and branching-process embeddings (Pemantle, 2007). For adaptive acceptance, an exact relative-entropy drift identity yields a cumulative cost bound for finite-output judges and identifies a necessary and sufficient support condition for positive drift at every round with infinitely many revisions. The optimal acceptance rule at a fixed state is an application of Bayes decision theory, with the usual monotonicity under Blackwell garbling (Provost and Fawcett, 2001; Blackwell, 1953). For inference, empirical Bernstein bounds (Maurer and Pontil, 2009; Mnih et al., 2008) and simultaneous confidence intervals (Howard et al., 2021; Waudby-Smith and Ramdas, 2024; Chugg and Ramdas, 2025; Ramdas et al., 2023) give a sequential test; change of measure gives the sample lower bounds (Kaufmann et al., 2016). The statistical contribution is the revision model's scale dependence and the distinction between proposal counts and utility evaluations. Classical unknown-margin testing problems can require iterated logarithms (Jamieson et al., 2014); our lower bound does not establish necessity of that factor here.