A THEORY OF RELIABLE SELF-EVOLUTION FOR AGENT HARNESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

In harness self-evolution, agents modify their own prompts, code, tools, and orchestration while keeping the underlying language model fixed. Recent work has shown that agents can improve themselves in response to task failures and achieve substantial performance gains. However, gains on failed tasks do not automatically ensure that performance on previously successful tasks is preserved, raising concerns about reliable adoption. In this work, we provide a theoretically grounded condition under which a self-evolved harness can be reliably adopted. We then propose a validation rule to make the evolved system satisfy the reliable adoption condition with theoretical guarantees. Consequently, the system can achieve progressive improvement through evolution. This leads to a natural question: Does reliable self-evolution have a performance ceiling, and which factors govern this ceiling? Our theoretical results show that the performance ceiling is determined by the costs of verification and evaluation. On the other hand, self-evolution may stall in practice. In this regard, we show that experimental evidence on agent performance shifts can be used to identify the sources of stagnation. Our work thus establishes a theoretical framework for understanding and advancing reliable harness self-evolution.

1 INTRODUCTION

Harness self-evolution allows an agent to modify its prompts, code, tools, and orchestration in response to task feedback while keeping the underlying language model fixed. These modifications persist across subsequent tasks. Gödel Agent (Yin et al., 2024) and SICA (Robeyns et al., 2025) demonstrated that agents can autonomously rewrite their own logic and code to improve task performance. DGM (Zhang et al., 2026b) uses evidence from failed task executions to improve its tools and workflows, achieving substantial performance gains. HGM (Wang et al., 2025b) advances search over successive agent modifications, obtaining further gains in coding performance. These advances motivate applying harness self-evolution in production systems. MOSS (Cai et al., 2026) brings source-level self-evolution to production-grade agent frameworks such as OpenClaw (Steinberger & OpenClaw contributors, 2026), testing modifications against failures encountered during use. It requires user approval before deployment and supports rollback if the updated system fails health checks. Even with these safeguards, a concern about reliable adoption remains: the adopted agent may improve on failed tasks while performing worse on previously successful tasks. Recent systems also consider performance beyond the failed tasks that motivate a modification. HarnessX (Chen et al., 2026) includes checks for observed regressions on previously solved tasks. Self-Harness (Zhang et al., 2026a) requires aggregate performance to improve on at least one of its development and held-out task sets without declining on either. We emphasize that a self-evolved harness can be reliably adopted when it improves expected reward over the full user-task distribution. We formulate a sufficient condition for this improvement in terms of gains on failed tasks and controlled changes in the expected rewards on previously successful tasks.

In the harness self-evolution process we study (Figure 1(a)), the current agent generates harness modifications from task failures to produce a set of candidate agents. These candidates are evaluated on a validation set using validation rules. A candidate meeting the validation requirements is adopted as the updated agent and continues to handle user tasks.

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

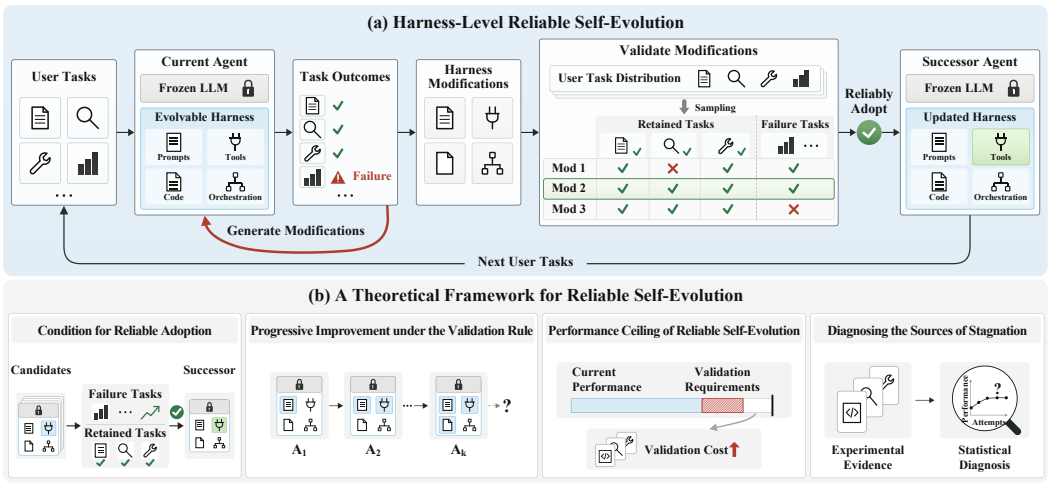


Figure 1: **Harness-level reliable self-evolution and our theoretical framework. (a)** The harness self-evolution process (Section 2.1). **(b)** A condition for reliable adoption (Section 2.2), progressive improvement under the validation rule (Section 3), the performance ceiling (Section 4), and diagnosis of stagnation from experimental evidence (Section 5).

In this work, we establish a theoretical framework for understanding and advancing reliable harness self-evolution (Figure 1(b)). We first provide a theoretically grounded condition under which a self-evolved harness can be reliably adopted. We analyze how much improvement on failed tasks is sufficient, with controlled changes in performance on previously successful tasks, to improve expected reward over the full user-task distribution and thereby enable reliable adoption. We then propose a validation rule with theoretical guarantees for reliable adoption under this condition. We also bound the probability of achieving a reliable update through candidate generation and validation. The gains guaranteed for adopted updates accumulate across evolution, and we also bound the average magnitude of task-performance changes across the adopted sequence.

This raises a natural question: does reliable self-evolution have a performance ceiling, and what determines it? We identify a performance ceiling beyond which the proposed validation rule cannot reliably validate further improvement. However, further improvement may still be possible even when this ceiling is reached. We therefore propose an alternative validation rule that uses evidence about each candidate's losses on previously successful tasks and can reliably validate improvement beyond the original ceiling. These results show that the ceiling depends in part on how improvement is verified. We also show that, as expected reward approaches its upper bound, reliably validating smaller gains can require increasing evaluation cost. When reliable self-evolution stagnates in practice, we show how experimental evidence on candidate performance shifts relative to the current agent can help identify the sources of stagnation. We provide theoretical guarantees for estimating reachability—the probability of generating a modification that satisfies the improvement and retention requirements—from additional experimental data. Comparing this evidence with validation and adoption records helps distinguish a low probability of generating qualified modifications from insufficient validation evidence for adopting these modifications.

2 Reliable Adoption of Self-Evolved Harnesses

2.1 Harness Self-Evolution Process

We study the following harness self-evolution process (Figure 1(a)). Let $\mathcal{T}_{\text{user}}$ and $\mathcal{O}$ be the user-task and output spaces, and $\mathcal{D}_{\text{user}}$ the user-task distribution. At step k, the current agent $A_k = (M, C_k)$ has a fixed language model M and an evolvable harness C_k of prompts, code, tools, and orchestration. On task $t \sim \mathcal{D}_{\text{user}}$, it produces $o \sim A_k(\cdot \mid t)$. From observed failures and interaction history, the agent generates harness modifications. Let $\mathcal{M}$ be the space of valid modifications; each generated $c \in \mathcal{M}$ yields a candidate agent $\widetilde{A} = A_k \oplus c$. The candidate agents are evaluated on a validation

set; a validation rule identifies candidates with sufficient evidence for reliable adoption. Selecting a validated c gives $A_{k+1} = A_k \oplus c$; otherwise, $A_{k+1} = A_k$. The successor handles subsequent tasks and modification generation; Appendices A and C.3 formalize the assumptions and procedure.

For an agent A and fixed task reward $r(t, o) \in [0, 1]$, define $V(A, t) := \mathbb{E}_{o\sim A(\cdot|t)}[r(t, o)]$ and $J_{\text{user}}(A) := \mathbb{E}_{t\sim\mathcal{D}_{\text{user}}}[V(A, t)]$ as its expected rewards on task t and across user tasks, respectively. Reliable adoption requires $J_{\text{user}}(A_k \oplus c) > J_{\text{user}}(A_k)$.

2.2 A Sufficient Condition for Reliable Adoption

To obtain a sufficient condition for reliable adoption, we decompose the change in overall expected reward into contributions from failed and previously successful tasks. Let the binary detector $\phi(t, o) \in \{0, 1\}$ identify a failed output ($\phi = 1$), while $r(t, o)$ scores task performance (Definition 6, Appendix B.1). For the current agent, define the task-level failure probability $\psi_{A_k}(t) := \mathbb{E}_{o\sim A_k(\cdot|t)}[\phi(t, o)]$ and overall failure rate $Z_k := \mathbb{E}_{t\sim\mathcal{D}_{\text{user}}}[\psi_{A_k}(t)]$. For $0 < Z_k < 1$, the induced failure-task and retained-task distributions are

$$\mathcal{D}_{F,k}(dt) := \frac{\psi_{A_k}(t)}{Z_k}\,\mathcal{D}_{\text{user}}(dt), \qquad \mathcal{D}_{R,k}(dt) := \frac{1 - \psi_{A_k}(t)}{1 - Z_k}\,\mathcal{D}_{\text{user}}(dt). \tag{1}$$

Because outputs can vary across runs, the distributions can overlap on the same task.

For candidate $\widetilde{A}$, define the per-task expected-reward difference $\text{Adv}_{A_k}(\widetilde{A}, t) := V(\widetilde{A}, t) - V(A_k, t)$ and its failure-task contribution $L_{A_k}(\widetilde{A}) := Z_k\mathbb{E}_{t\sim\mathcal{D}_{F,k}}[\text{Adv}_{A_k}(\widetilde{A}, t)]$. By Lemma 1 in Appendix B.1, the overall expected-reward difference decomposes as

$$J_{\text{user}}(\widetilde{A}) - J_{\text{user}}(A_k) = L_{A_k}(\widetilde{A}) + (1 - Z_k)\mathbb{E}_{t\sim\mathcal{D}_{R,k}}[\text{Adv}_{A_k}(\widetilde{A}, t)]. \tag{2}$$

For retained tasks, let $D_R(\widetilde{A}; A_k) := \mathbb{E}_{t\sim\mathcal{D}_{R,k}}|\text{Adv}_{A_k}(\widetilde{A}, t)|$ be the mean absolute expected-reward change. For $c \in \mathcal{M}$, write $L_k(c) := L_{A_k}(A_k \oplus c)$ and $D_k(c) := D_R(A_k \oplus c; A_k)$.

Definition 1 (Qualified modification). *For $T = (\lambda, \delta)$, where $\lambda > 0$ sets the minimum failure-task contribution and $\delta \geq 0$ the maximum retained-task change, define the set of modifications qualified at step k by $\mathcal{P}_k(T) := \{c \in \mathcal{M} : L_k(c) \geq \lambda,\ D_k(c) \leq \delta\}$.*

Theorem 1 (Gain bound for reliable adoption). *Under Assumptions 1 and 2, every candidate agent $\widetilde{A}$ with $D_R(\widetilde{A}; A_k) \leq \delta$ satisfies the expected-reward gain bound*

$$J_{\text{user}}(\widetilde{A}) - J_{\text{user}}(A_k) \geq L_{A_k}(\widetilde{A}) - (1 - Z_k)\delta. \tag{3}$$

Corollary 1 (Sufficient condition for reliable adoption). *Under Assumptions 1 and 2, if $T = (\lambda, \delta)$ satisfies $\lambda > (1 - Z_k)\delta$, then every $c \in \mathcal{P}_k(T)$ satisfies*

$$J_{\text{user}}(A_k \oplus c) - J_{\text{user}}(A_k) \geq \lambda - (1 - Z_k)\delta > 0.$$

For $\gamma > 0$, setting $\lambda = \lambda_k(\gamma, \delta) := \gamma + (1 - Z_k)\delta$ ensures gain at least γ for every $c \in \mathcal{P}_k(T)$.

The condition permits changes on retained tasks while ensuring that failure-task gains outweigh their possible losses. Appendix B.1 proves Theorem 1 and Corollary 1. Alternative retained-task conditions are in Appendices B.2–B.4 and D.2. Section 3 develops a validation rule for this condition.

3 Achieving Harness-Level Reliable Self-Evolution

3.1 A Validation Rule for Reliable Adoption

Assessing the reliable-adoption condition in Section 2.2 requires evaluating the candidate's expected-reward improvement under the user-task distribution $\mathcal{D}_{\text{user}}$. In the harness self-evolution process of Section 2.1, we form a validation set from sampled user tasks and apply a validation rule to decide which generated modifications to adopt. We first give a rule with a high-probability guarantee for reliable adoption. We then bound the probability of achieving an improving update by generating a candidate pool and then validating its modifications.

For a generated modification $c \in \mathcal{M}$, write its candidate agent as $\widetilde{A} = A_k \oplus c$. Let $\widehat{L}_k, \widehat{D}_R, \widehat{Z}_k$ be validation-set estimates of the failure-task contribution $L_k(c)$, mean absolute retained-task change $D_k(c)$, and current failure rate Z_k, with error bounds $\varepsilon_L, \varepsilon_D, \varepsilon_Z$. Fix thresholds $\tau, \delta > 0$ for the estimated failure-task contribution and retained-task change. Appendix C.1 gives the estimators and sampling conditions and derives the corresponding error bounds.

Rule 1 (Two-Gate validation). *With a valid candidate and the required data, validate c if and only if*

$$\widehat{L}_k \geq \tau, \qquad \widehat{D}_R \leq \delta, \tag{4}$$

$$\Delta_k := \tau - \varepsilon_L - (1 - \widehat{Z}_k + \varepsilon_Z)(\delta + \varepsilon_D) > 0. \tag{5}$$

The rule validates a modification only when the estimates support a sufficient failure-task contribution, controlled retained-task change, and a positive lower bound on overall expected-reward gain. In the one-candidate update, a validated c is adopted according to Section 2.1; Procedure 1 in Appendix C.1 specifies the validation procedure and the resulting agent update.

Let Acc_k denote the event that the candidate is adopted, and $I_k := J_{\mathrm{user}}(A_{k+1}) - J_{\mathrm{user}}(A_k)$ its true gain. Set $\Delta_k = 0$ when no valid candidate or required validation data are available, as in Definition 9 of Appendix C.1. Let $\beta_{\mathrm{step},k}$ denote the stepwise estimation failure bound (Eq. (57), Appendix C.2).

Theorem 2 (Reliable adoption under Two-Gate validation). *Suppose Assumptions 1 and 2 (Appendix A) and the evaluation conditions of Lemma 5 (Appendix C.1) and Lemma 6 (Appendix C.2) hold. Apply Rule 1 through the update in Procedure 1. Then*

$$\mathbb{P}\big(\mathsf{Acc}_k \cap \{I_k < \Delta_k\}\big) \leq \beta_{\mathrm{step},k}.$$

In particular, the positive-gain requirement gives $\mathbb{P}(\mathsf{Acc}_k \cap \{I_k \leq 0\}) \leq \beta_{\mathrm{step},k}$.

Under the stated evaluation conditions, Rule 1 provides a positive lower bound Δ_k on the gain from adoption: the probability of adopting a candidate whose true gain falls below this bound is at most $\beta_{\mathrm{step},k}$. This guarantee applies the sufficient condition in Corollary 1; its proof is in Appendix C.2. Remark 5 in Appendix C.1 explains why reference data must be refreshed after adoption.

Having established the reliable-adoption guarantee for Rule 1, we next quantify the probability that generation and validation yield a reliable update. Consider a pool of generated candidates validated against a target $T = (\lambda, \delta)$ from Definition 1. Fix the interaction history $\mathcal{H}_k$ and a generation context $\Xi_k = \xi$ shared across generation runs. A complete run of the current agent's generation procedure returns $C \sim \pi_{k,\xi}$ on $\mathcal{M}_\perp := \mathcal{M} \cup \{\perp\}$, where $\perp$ means that no valid modification was returned.

Definition 2 (Reachability). *A generation run produces a modification qualified for T with probability*

$$P_k(T \mid \mathcal{H}_k, \xi) := \pi_{k,\xi}\big(\mathcal{P}_k(T)\big). \tag{6}$$

Let $C_1, \ldots, C_N$ be the outputs of N conditionally independent generation runs under the same history and context. For each valid C_i, compute confidence bounds for $L_k(C_i)$ and $D_k(C_i)$. Let the validation indicator $G_i(T) = 1$ if and only if the lower bound for $L_k(C_i)$ reaches λ and the upper bound for $D_k(C_i)$ is at most δ; these tests define the pool validation rule G. Appendix C.4 specifies the bounds and the convention for invalid outputs.

Candidates whose true values meet the target after allowing for the confidence-interval widths pass validation whenever the intervals contain the true values. Before drawing candidates, fix deterministic $\bar{w}_L, \bar{w}_D$ that bound the interval widths for all candidates and possible validation data. For evaluation sample sizes $\mathbf{n}$ and error allocation β, define the set of modifications

$$\mathcal{P}^+_{k,N,\mathbf{n},\beta}(T) := \{c \in \mathcal{M} : L_k(c) \geq \lambda + \bar{w}_L,\ D_k(c) \leq \delta - \bar{w}_D\}. \tag{7}$$

Write $P^+_k(T \mid \mathcal{H}_k, \xi) := \pi_{k,\xi}(\mathcal{P}^+_{k,N,\mathbf{n},\beta}(T))$ for the probability of generating a modification in this set, abbreviated as $P^+_k(T)$ when the history and context are fixed. Let β_{stat} bound the probability that any validation interval fails to contain the quantity it estimates.

Theorem 3 (Probability of achieving harness-level reliable self-evolution). *Fix T and $(\mathcal{H}_k, \Xi_k = \xi)$. Suppose the N generation runs are conditionally i.i.d. and all validation intervals cover their true values with conditional probability at least $1 - \beta_{\mathrm{stat}}$ given the pool. Suppose the evaluation data*

and interval widths satisfy the conditions in Appendix C.4. Let a selector σ choose only validated candidates and choose one whenever any are available. Write $R_{k,N}^{G,\sigma}(T \mid \mathcal{H}_k, \xi)$ for the conditional probability that the update adopts a modification qualified for T. Then

$$(1-\beta_{\mathrm{stat}})\big[1-(1-P_k^{+}(T \mid \mathcal{H}_k,\xi))^N\big] \le R_{k,N}^{G,\sigma}(T \mid \mathcal{H}_k,\xi) \le 1-(1-P_k(T \mid \mathcal{H}_k,\xi))^N. \quad (8)$$

In particular, for $T = (\lambda_k(\gamma,\delta),\delta) = (\gamma + (1-Z_k)\delta, \delta)$ with $\gamma > 0$ as in Corollary 1, the same lower bound applies to the probability of an update with overall expected-reward gain at least γ.

For the calibrated target, $P_k^{+}(T) > 0$ and $\beta_{\mathrm{stat}} < 1$ establish a positive probability of an update with at least γ gain. Candidate generation also limits reliable adoption: validation can adopt a T-qualified modification only if the pool contains one. Appendix C.4 gives the proof and shows how to choose the target using an estimate of Z_k obtained before generation (Corollary 5). Proposition 12 in Appendix C.3 analyzes generation-dependent reachability. Proposition 13 and Corollaries 7 and 6 in Appendix C.4 address expected selected gain, confidence allocation, and shared context, respectively.

3.2 Progressive Improvement Guarantees

When Rule 1 is applied across K evolution steps, each adopted successor becomes the reference agent for subsequent generation and validation. We ask whether guaranteed gains from adoptions accumulate despite dependence between steps. Fix the reward and user-task distribution, run Procedure 1 at each step to obtain $A_0, \dots, A_K$, and let $\mathcal{A} := \{k < K : \mathsf{Acc}_k\}$ denote the adopted steps.

Theorem 4 (Progressive improvement through evolution). *Suppose the hypotheses of Theorem 2 hold at every step and the deterministic error bounds satisfy $\sum_{k=0}^{K-1} \beta_{\mathrm{step},k} \le \beta_{\mathrm{run}}$. Then, with probability at least $1-\beta_{\mathrm{run}}$, the adopted updates satisfy*

$$J_{\mathrm{user}}(A_K) - J_{\mathrm{user}}(A_0) \ge \sum_{k\in\mathcal{A}} \Delta_k. \quad (9)$$

The bound does not require independence between evolution steps.

Each Δ_k is computed before the corresponding adoption. Under the stated evaluation conditions and the overall error bound, the positive gains guaranteed for adopted updates accumulate, even when later steps depend on earlier ones. Appendix E.1 gives the proof, the extension to the Measured-margin rule introduced in Section 4, and the adopted-step bound under a uniform positive gain (Corollary 13). When user-task distributions change, Theorem 10 in Appendix E.1 deducts cumulative total-variation drift; Theorem 11 in Appendix E.5 addresses evaluation–deployment distribution mismatch.

An increase in overall expected reward does not describe how task-level performance changes along the sequence. We therefore bound the total change across adopted updates and the final agent's change relative to A_0, both averaged over $\mathcal{D}_{\mathrm{user}}$. For the fixed user-task distribution, let

$$\mathcal{V}_K := \sum_{k\in\mathcal{A}} \mathbb{E}_{t\sim\mathcal{D}_{\mathrm{user}}} |V(A_{k+1},t) - V(A_k,t)|$$

be the accumulated mean absolute change along the adopted updates.

Proposition 1 (Accumulated task-performance change). *Under Assumptions 1 and 2, suppose $Z_k \in (0,1)$ on adopted steps and each adoption satisfies the Two-Gate inequalities in Eq. (4) with fixed $\tau, \delta > 0$. Write $\varepsilon_L^{(k)}, \varepsilon_D^{(k)}$ for the step-k error bounds. On the event that these bounds hold,*

$$\begin{aligned}\mathcal{V}_K \le 1 - J_{\mathrm{user}}(A_0) &+ \sum_{k\in\mathcal{A}} (Z_k - \tau + \varepsilon_L^{(k)}) \\ &+ 2\sum_{k\in\mathcal{A}} (1-Z_k)(\delta + \varepsilon_D^{(k)}).\end{aligned} \quad (10)$$

Consequently, $\mathbb{E}_{t\sim\mathcal{D}_{\mathrm{user}}} |V(A_K,t) - V(A_0,t)| \le \min\{1, \mathcal{V}_K\}$.

The proposition also bounds the average magnitude of changes in per-task expected reward relative to A_0, counting both gains and losses along the adopted sequence. Appendix E.2 proves the proposition and gives the corresponding result for the retained-task decrease rule in Eq. (95) of Appendix D.5.

Further progress requires subsequent agents to generate modifications that can be validated. Proposition 22 in Appendix E.3 constructs a successful first update after which later reachability vanishes. Section 4 next examines the performance ceiling and its dependence on verification and evaluation.

4 Performance Ceiling of Reliable Self-Evolution

Rule 1 requires failure-task improvement to cover possible retained-task losses and evaluation error, while bounded rewards constrain the improvement available. We characterize this ceiling, then examine how verification can support further reliable improvement near it.

4.1 Performance Ceiling

Every candidate satisfies $L_{A_k}(\widetilde{A}) \leq 1 - J_{\text{user}}(A_k)$ under Assumptions 1 and 2. Lemma 8 and Corollary 8 in Appendix D.1 give the sharper failure-task bound and its consequence for positive reachability; Appendix C.7 analyzes reachability near the generation process's improvement endpoint.

Two-Gate requires $\widehat{L}_k \geq \tau > (1 - \widehat{Z}_k + \varepsilon_Z)(\delta + \varepsilon_D) + \varepsilon_L$. Let $\mathcal{E}_L$ be the event that the required data are available and the error bounds for estimating L_k hold, and $\mathcal{R}$ the set of parameter pairs (τ, δ) satisfying this lower bound on τ and the upper bounds on $\widehat{L}_k$. Definition 14 in Appendix D.3 specifies both. Write $\varepsilon_\Sigma := 2\varepsilon_V + \varepsilon_R + \varepsilon_Z$, where the component errors bound per-task reward estimation, the retained-task statistic, and failure-rate estimation, respectively (Appendices C.1 and C.2).

Corollary 2 (Performance ceiling under Two-Gate verification). *Under Assumptions 1 and 2, fix the threshold $\delta > 0$ on estimated retained-task change in Two-Gate. If some $\tau > 0$ satisfies $(\tau, \delta) \in \mathcal{R}$, then, on $\mathcal{E}_L$, the current agent's expected reward satisfies*

$$J_{\text{user}}(A_k) < 1 - (1 - Z_k)\delta - \varepsilon_\Sigma. \tag{11}$$

On $\mathcal{E}_L$, Rule 1 cannot validate further improvement when the current agent reaches or exceeds this ceiling: verification treats the full allowed retained-task change as a possible loss, and evaluation error further reduces the ceiling. Appendix D.3 gives the proof and finer parameter and sample-size conditions; Figure 2 in Appendix D.6 illustrates both effects.

Near this ceiling, further gains may be too small to cover the full allowed retained-task change, even when a candidate incurs little loss. We therefore use evidence about the candidate's own retained-task losses to determine whether further improvement can be reliably validated.

Define its mean per-task expected-reward decrease as

$$D_R^-(\widetilde{A}; A_k) := \mathbb{E}_{t \sim \mathcal{D}_{R,k}}[(V(A_k, t) - V(\widetilde{A}, t))_+],$$

and the gain lower bound $M_{A_k}(\widetilde{A}) := L_{A_k}(\widetilde{A}) - (1 - Z_k)D_R^-(\widetilde{A}; A_k)$. The reward decomposition gives $J_{\text{user}}(\widetilde{A}) - J_{\text{user}}(A_k) \geq M_{A_k}(\widetilde{A})$. Appendix D.2 gives the derivation and an estimator $\widehat{D}_{R,k}^-$ of the decrease with error bound $\varepsilon_D^- = \varepsilon_\Sigma/(1 - \widehat{Z}_k)$.

Rule 2 (Measured-margin validation). *Use the evaluation data specified by Procedure 1 in Appendix C.1. Given an allowed retained-task decrease $\delta^- > 0$, when a valid candidate and the required validation data are available, define its estimated gain lower bound*

$$\widehat{M}_k := \widehat{L}_k - \varepsilon_L - (1 - \widehat{Z}_k + \varepsilon_Z)(\widehat{D}_{R,k}^- + \varepsilon_D^-). \tag{12}$$

Validate the modification if and only if $\widehat{D}_{R,k}^- \leq \delta^-$ and $\widehat{M}_k > 0$ hold. Set $\widehat{M}_k = 0$ when no valid candidate or required validation data are available.

Under the assumptions of Theorem 2, let Acc_k^M be adoption under Rule 2 through Definition 9 (Appendix C.1). Proposition 18 in Appendix D.2 gives $\mathbb{P}(\mathsf{Acc}_k^M \cap \{I_k < \widehat{M}_k\}) \leq \beta_{\text{step},k}$. Thus, the rule supports reliable adoption using a gain lower bound informed by the candidate's own losses.

The condition $L_{A_k}(\widetilde{A}) > (1 - Z_k)\delta^-$ imposes the ceiling $B_k := 1 - (1 - Z_k)\delta^-$ (Corollary 9, Appendix D.3). In that appendix, Proposition 19(i) constructs instances with $J_{\text{user}}(A_k) \geq B_k$, no retained-task loss, and gain $(1 - J_{\text{user}}(A_k))/2 > 0$; Proposition 20 gives an evaluation design under which Rule 2 validates this improvement when the required samples are available and the error bounds hold. The ceiling thus depends on the treatment of retained-task losses during verification.

4.2 The Cost of Validating Further Improvement

As the remaining gains shrink, validating further improvement requires more precise evaluation. Let $u_k := 1 - J_{\mathrm{user}}(A_k)$ be the remaining gap to the reward upper bound. Every candidate satisfies $M_{A_k}(\widetilde{A}) \le u_k$ (Proposition 19(ii), Appendix D.3), so validating a positive gain as expected reward approaches one requires correspondingly small evaluation error.

For the preceding construction, m evaluations per agent per task, n_F failure tasks, and n_R reference observations with failure labels, each of order $\widetilde{O}(u_k^{-2})$, suffice at fixed failure rate and confidence allocation under the stated sample and error conditions (Proposition 20, Appendix D.3). The following result shows that evaluation cost must increase in the worst case even for other validation procedures.

Consider any validation procedure that selects tasks adaptively and stops according to its observations. Call it level-β if, in every instance, the probability of validating a candidate with $J_{\mathrm{user}}(\widetilde{A}) \le J_{\mathrm{user}}(A_k)$ is at most β. Measure its cost by the number N_{eval} of candidate agent runs, each returning one task reward. Appendix D.4 specifies the instance and observation model.

Theorem 5 (Cost of validating further improvement). *Fix $Z \in (0,1)$, $\beta \le 1/8$, and $u \in (0, Z/2]$. There exist two instances with $Z_k = Z$ and $1 - J_{\mathrm{user}}(A_k) = u$, identical in the distribution of every observable quantity except the candidate agent's reward distribution under the failure-task distribution. In one, $J_{\mathrm{user}}(\widetilde{A}) - J_{\mathrm{user}}(A_k) = u/2$ and $D_R(\widetilde{A}; A_k) = 0$, and in the other the overall reward change is $-u/2$. Any level-β validation procedure validating the improving candidate with probability at least $1/2$, under any task-selection policy and data-dependent stopping rule, satisfies*

$$\mathbb{E}_{-}[N_{\mathrm{eval}}] \ge \frac{Z}{10u},$$

where the expectation is under the non-improving instance.

Even when the candidate preserves previously successful performance, verification must distinguish smaller gains from degradation. For fixed Z, the theorem requires a worst-case expected evaluation cost of order at least u^{-1} as expected reward approaches one; its proof is in Appendix D.4.

5 Diagnosing Stalled Harness Self-Evolution

Even with guarantees for reliable adoption, self-evolution may stall because generation rarely produces suitable modifications or validation lacks sufficient evidence to accept them. To diagnose stagnation, we evaluate candidate performance shifts relative to the current agent using additional data, including candidates not adopted. These evaluations support reachability estimation, which we combine with validation and adoption records to distinguish the two sources.

5.1 Reachability Estimation

Fix the interaction history $\mathcal{H}_k$ and generation context $\Xi_k = \xi$. We estimate reachability $P_k(T \mid \mathcal{H}_k, \xi)$ from Section 3.1 for improvement and retention targets $T = (\lambda, \delta)$.

Draw N_{meas} outputs C_j independently from $\pi_{k,\xi}$. Repeatedly evaluate the current agent and each valid candidate on additional tasks sampled from $\mathcal{D}_{\mathrm{user}}$ and withheld from generation and validation. Their per-task expected-reward estimates yield intervals $[\underline{L}_j, \overline{L}_j]$ for $L_k(C_j)$ and $[\underline{D}_j, \overline{D}_j]$ for $D_k(C_j)$. Appendix F.3 gives the sampling conditions and interval construction; Appendix F.2 gives the corresponding estimates of candidate effects on a fixed task suite.

Definition 3. *For target $T = (\lambda, \delta)$, define the fractions of generated outputs whose evaluation intervals confirm or permit qualification, respectively, by*

$$\widehat{P}_{\mathrm{in}}(T) := \frac{\sum_{j=1}^{N_{\mathrm{meas}}} \mathbf{1}\{\underline{L}_j \ge \lambda, \overline{D}_j \le \delta\}}{N_{\mathrm{meas}}}, \qquad \widehat{P}_{\mathrm{out}}(T) := \frac{\sum_{j=1}^{N_{\mathrm{meas}}} \mathbf{1}\{\overline{L}_j \ge \lambda, \underline{D}_j \le \delta\}}{N_{\mathrm{meas}}}.$$

Invalid outputs contribute zero to both numerators and remain in the denominator.

Theorem 6 (Uniform confidence bounds for reachability). *In the conditional experiment above, let $\beta_{\mathrm{rect}}, \beta_{\mathrm{emp}} \in (0,1)$ be error probabilities. Suppose all evaluation intervals contain their true values*

jointly with probability at least $1-\beta_{\mathrm{rect}}$. *Let* $r_{\mathrm{VC}} = r_{\mathrm{VC}}(N_{\mathrm{meas}}, \beta_{\mathrm{emp}})$ *be a valid uniform empirical error bound for the target sets* $\{(l,d) : l \geq \lambda, d \leq \delta\}$*, a VC class of constant dimension. Then, with probability at least* $1-\beta_{\mathrm{rect}}-\beta_{\mathrm{emp}}$*, simultaneously for every target* T*,*

$$\max\{0, \widehat{P}_{\mathrm{in}}(T) - r_{\mathrm{VC}}\} \leq P_k(T \mid \mathcal{H}_k, \xi) \leq \min\{1, \widehat{P}_{\mathrm{out}}(T) + r_{\mathrm{VC}}\}. \tag{13}$$

The bounds estimate reachability even when few candidates pass validation and permit comparisons across targets from the same data. Appendix F.1 proves Theorem 6 and analyzes estimation precision.

For a target calibrated for reliable adoption by Corollary 1, a small upper bound indicates that generation rarely meets the target. For rule G in Section 3.1, a substantial lower bound together with independently confirmed but unvalidated modifications supports insufficient validation evidence as a source of stagnation. A wide interval leaves the generation probability uncertain; precision depends on both the number of outputs and their evaluation uncertainty. Definition 18 and Proposition 24 in Appendix F.1 formalize the diagnostic quantities, distinguishing validation from subsequent selection.

5.2 Experimental Illustration

We illustrate this diagnosis on DS-1000 (Lai et al., 2023) with persistent Python solver-harness modifications and a fixed language model. Independent generation, label-free revision, and outcome-feedback revision each produce 96 outputs in 24 pools, evenly divided among four contexts. We evaluate the current agent and every valid candidate, including those not adopted, on audit tasks withheld from generation and validation. Appendix G.1 gives the generation processes, validation rule, and sampling design for the additional evaluations.

Each context has $n = 56$ audit tasks, with failure and retained groups F, R fixed by the original current-agent outcomes. On task t, let $\bar{x}_t, \bar{y}_t$ be the current and candidate agents' mean binary success outcomes from three fresh runs each. We estimate the failure-task contribution, mean absolute change on retained tasks, and overall expected-reward gain by

$$\widehat{L} = \frac{1}{n}\sum_{t\in F}(\bar{y}_t - \bar{x}_t), \qquad \widehat{D} = \frac{1}{|R|}\sum_{t\in R}|\bar{y}_t - \bar{x}_t|, \qquad \widehat{\Delta} = \frac{1}{n}\sum_{t}(\bar{y}_t - \bar{x}_t). \tag{14}$$

Write L, D, Δ for the corresponding expected-reward effects on this fixed audit set, defined in Eq. (119) of Appendix G.1. At the prespecified target $T_0 = (0.05, 0.50)$, Table 1 compares the audit estimates with the original decisions based on 48 separate validation tasks per context.

Table 1: **Repeated evaluation and validation on DS-1000.** Audit target membership at $T_0 = (0.05, 0.50)$ uses mean outcomes from three fresh runs per agent and task. Brackets give 90% hierarchical-bootstrap intervals for the empirical fractions. Validation and adoption counts use the original decisions based on 48 validation tasks per context.

Generation process	Audit point fraction	Validated outputs	Pools with adoption
Independent	$0.385\ [0.135, 0.646]$	$2/96$	$2/24$
Label-free revision	$0.260\ [0.031, 0.479]$	$0/96$	$0/24$
Outcome-feedback revision	$0.344\ [0.125, 0.563]$	$0/96$	$0/24$

Across the three processes, substantial fractions meet the audit point target, yet only two outputs pass validation and are adopted. Both also meet the audit point target. Table 3 in Appendix G.1 gives the full output and pool summaries; Figure 3 there compares contexts and validation budgets. The appendix also reports the results for the stricter target.

To assess whether an unadopted candidate offers an improvement opportunity, we retrospectively analyze all three candidates whose original audit bounds met the target. For this specified screen, the fresh runs give conditional 95% joint confidence bounds supporting one outcome-feedback modification: $L \geq 0.0550$, $D \leq 0.2766$, and $\Delta \geq 0.0556$ on the fixed audit set (Table 2, Appendix G.1). These bounds support the target requirements and positive expected-reward gain. The additional runs provide evidence about the agents' expected performance, accounting for variation across executions of the same task. The candidate's original validation estimate of its failure-task contribution was

0.1458, above 0.05, but its lower bound was 0.0378, below 0.05, while the retention upper bound 0.4196 met the 0.50 requirement; its pool adopted no modification. Thus, the additional evaluation supports improvement on the audit set. The original validation record identifies insufficient evidence of the required improvement as the reason this candidate was not adopted.

6 Related Work

Harness self-evolution. Persistent agent changes (Yin et al., 2024; Robeyns et al., 2025; Zelikman et al., 2024; Fernando et al., 2023; Wang et al., 2024; Hu et al., 2025; Zhang et al., 2025) span source-code rewriting, prompt evolution, reusable skills, and workflow search. DGM (Zhang et al., 2026b) and HyperAgents (Zhang et al., 2026c) explore sequences of agent variants, while HGM (Wang et al., 2025b) uses descendant performance to guide search. HarnessX (Chen et al., 2026) and Self-Harness (Zhang et al., 2026a) screen observed regressions when deciding whether to adopt modifications; AHE (Lin et al., 2026) records predicted fixes and regressions for subsequent evaluation and rollback. MOSS (Cai et al., 2026) integrates source-level evolution into production-grade agent frameworks. These systems motivate studying when harness modifications can be reliably adopted and what determines the performance ceiling of reliable self-evolution.

Theoretical guarantees for self-improvement. STOP (Zelikman et al., 2024) studies expected improver utility and generalization; Wang et al. (2025a) analyze learnability through policy-reachable hypothesis classes and establish risk-improvement guarantees under validation and capacity control. SGM (Wu et al., 2025) controls harmful adoption across repeated tests, while PACE (Shawn, 2026) provides an anytime-valid test for an individual candidate under its conditional-fairness assumption. TTHE (Nie et al., 2026) studies generation coverage and selection regret for harness candidates. We analyze how validation supports reliable adoption and progressive improvement, and how verification and evaluation determine a performance ceiling. We also use candidate performance shifts to diagnose whether generation or validation impedes further improvement.

Safe policy improvement. HCPI (Thomas et al., 2015) uses a confidence bound to test whether a candidate policy meets a specified performance requirement. Our validation rule likewise accounts for evaluation uncertainty before accepting a modification. TRPO (Schulman et al., 2015) derives sufficient conditions for improvement by bounding the effect of policy-distribution changes. Our sufficient condition combines gains on failed tasks with a bound on changes in expected reward on previously successful tasks to ensure overall improvement.

7 Limitations

Our guarantees concern expected reward and average changes in task performance under the specified task distributions. The reward–utility analysis in Appendix B.5 supports transferring reward guarantees to user utility when the discrepancy between reward and utility can be bounded. Appendix B.2 extends average retention control to task-group protection; applying this extension with finite evaluation data requires corresponding estimation guarantees. When user demand responds to agent updates, Appendix E.4 gives a correction to the improvement guarantee. Using this correction for adoption requires a bound on the effect of the distribution change before deployment.

8 Conclusion and Future Work

We establish a theoretical foundation for reliable harness self-evolution, providing a condition for reliable adoption and showing how validated updates can yield progressive improvement with controlled changes in task performance. Our results show that a performance ceiling imposed by a validation rule need not imply that further improvement is impossible. An alternative rule can validate improvements beyond that ceiling, while reliably validating smaller gains may require greater evaluation cost. Experimental evidence on candidate performance shifts can help distinguish limitations in generating suitable modifications from limitations in validating them. These findings highlight the importance of considering generation and validation jointly in the design of reliable self-evolution systems. Future work can develop methods that use these diagnostic results to guide candidate generation and allocate evaluation budgets, and investigate conditions under which adopted updates preserve opportunities for subsequent improvement (Appendix H).

9 AI USE STATEMENT

We used large language models (LLMs) to identify and review related work, refine research ideas, draft portions of the manuscript, improve the clarity of our writing, and debug experimental code. We also used LLMs to check mathematical arguments and proofs, including intermediate derivations, logical consistency, and potential gaps.

REFERENCES

Carlo Acerbi and Dirk Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002. doi:10.1016/S0378-4266(02)00283-2.

Qianshu Cai, Yonggang Zhang, Xianzhang Jia, Huajiang Zheng, Wei Xue, Jun Song, Xinmei Tian, and Yike Guo. MOSS: Self-evolution through source-level rewriting in autonomous agent systems. *arXiv preprint arXiv:2605.22794*, 2026.

Tingyang Chen, Shuo Lu, Kang Zhao, Weicheng Meng, Hanlin Teng, Tianhao Li, Chao Li, Xule Liu, Jian Liang, Zhizhong Zhang, Yuan Xie, Heng Qu, Kun Shao, and Jian Luan. HarnessX: A composable, adaptive, and evolvable agent harness foundry. *arXiv preprint arXiv:2606.14249*, 2026.

Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006. doi:10.1007/0-387-34471-3.

Paul Doukhan. *Mixing: Properties and Examples*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1994. doi:10.1007/978-1-4612-2642-0.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*, 2023.

Wassily Hoeffding. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, 27(3):713–721, 1956. doi:10.1214/aoms/1177728178.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *International Conference on Learning Representations (ICLR)*, 2025.

Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Wen-Tau Yih, Daniel Fried, Sida Wang, and Tao Yu. DS-1000: A natural and reliable benchmark for data science code generation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18319–18345. PMLR, 2023. URL https://proceedings.mlr.press/v202/lai23b.html.

Jiahang Lin, Shichun Liu, Chengjun Pan, Lizhi Lin, Shihan Dou, Zhiheng Xi, Xuanjing Huang, Hang Yan, Zhenhua Han, Tao Gui, and Yu-Gang Jiang. Agentic harness engineering: Observability-driven automatic evolution of coding-agent harnesses. *arXiv preprint arXiv:2604.25850*, 2026.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-I.I.D. processes. In *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, pp. 1097–1104, 2008.

Jun Nie, Yonggang Zhang, Jun Song, Qianshu Cai, Dahai Yu, Yike Guo, Xinmei Tian, and Bo Han. TTHE: Test-time harness evolution. *arXiv preprint arXiv:2607.08124*, 2026.

Maxime Robeyns, Martin Szummer, and Laurence Aitchison. A self-improving coding agent. *arXiv preprint arXiv:2504.15228*, 2025.

R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002. doi:10.1016/S0378-4266(02)00271-6.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897. PMLR, 2015.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. MPS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2009.

Zayx Shawn. PACE: Anytime-valid acceptance tests for self-evolving agents. arXiv preprint arXiv:2606.08106, 2026. URL https://arxiv.org/abs/2606.08106.

Peter Steinberger and OpenClaw contributors. OpenClaw. GitHub repository, 2026. URL https://github.com/openclaw/openclaw. Version v2026.9.6. Accessed September 26, 2026.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388. PMLR, 2015.

Charles L. Wang, Keir Dorchen, and Peter Jin. On the statistical limits of self-improving agents. *arXiv preprint arXiv:2510.04399*, 2025a.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research (TMLR)*, 2024.

Kaiwen Wang, Nathan Kallus, and Wen Sun. Near-minimax-optimal risk-sensitive reinforcement learning with CVaR. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35864–35907. PMLR, 2023.

Wenyi Wang, Piotr Piękos, Li Nanbo, Firas Laakom, Yimeng Chen, Mateusz Ostaszewski, Mingchen Zhuge, and Jürgen Schmidhuber. Huxley-Gödel machine: Human-level coding agent development by an approximation of the optimal self-improving machine. *arXiv preprint arXiv:2510.21614*, 2025b.

Xuening Wu, Shenqin Yin, Yanlan Kang, Xinhang Zhang, Qianya Xu, Zeping Chen, and Wenqiang Zhang. SGM: A statistical Gödel machine for risk-controlled recursive self-modification. arXiv preprint arXiv:2510.10232, 2025. URL https://arxiv.org/abs/2510.10232.

Xunjian Yin, Xinyi Wang, Liangming Pan, Li Lin, Xiaojun Wan, and William Yang Wang. Gödel agent: A self-referential agent framework for recursive self-improvement. *arXiv preprint arXiv:2410.04444*, 2024.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994. doi:10.1214/aop/1176988849.

Eric Zelikman, Eliana Lorch, Lester Mackey, and Adam Tauman Kalai. Self-taught optimizer (STOP): Recursively self-improving code generation. In *Conference on Language Modeling (COLM)*, 2024.

Hangfan Zhang, Shao Zhang, Kangcong Li, Chen Zhang, Yang Chen, Yiqun Zhang, Lei Bai, and Shuyue Hu. Self-Harness: Harnesses that improve themselves. *arXiv preprint arXiv:2606.09498*, 2026a.

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 (ICLR)*, 2026b.

Jenny Zhang, Bingchen Zhao, Wannan Yang, Jakob Foerster, Jeff Clune, Minqi Jiang, Sam Devlin, and Tatiana Shavrina. Hyperagents. *arXiv preprint arXiv:2603.19461*, 2026c.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In *International Conference on Learning Representations (ICLR)*, 2025.

APPENDIX CONTENTS

A Formal Common Setting and Principal Symbols **13**
A.1 Formal common setting . . . 13
A.2 Principal symbols . . . 14

B Reliable Adoption and Retained-Task Performance **14**
B.1 Failure weighting and proofs for reliable adoption . . . 14
B.2 Tail constraints and subgroup protection . . . 15
B.3 Alternative retained-task norms . . . 17
B.4 Output-distribution comparisons . . . 17
B.5 Reward fidelity and ideal utility . . . 18

C Generation and Validation for Reliable Adoption **19**
C.1 Validation data and update procedure . . . 19
C.2 Retained-task estimation and the reliable-adoption proof . . . 25
C.3 Generation procedures and pool selection . . . 26
C.4 Validation and adoption from a candidate pool . . . 27
C.5 Paired evaluation . . . 31
C.6 Repair and finite assembly . . . 32
C.7 Endpoint tails at a fixed retained-task bound . . . 33

D Performance Ceiling and Verification **33**
D.1 Remaining improvement and positive reachability . . . 33
D.2 One-sided changes and Measured-margin . . . 34
D.3 Constraints on further improvement . . . 36
D.4 Sample-complexity lower bound for validation . . . 39
D.5 A one-sided rule for cumulative reward changes . . . 41
D.6 Calculation for Figure 2 . . . 41

E Multiple Updates and Task-Distribution Changes **42**
E.1 Cumulative improvement . . . 42
E.2 Cumulative absolute reward changes . . . 43
E.3 A successful update and later improvement opportunities . . . 44
E.4 Agent-responsive task distributions . . . 45
E.5 Evaluation and deployment task distributions . . . 46
E.6 Estimating the evaluation–deployment discrepancy . . . 48

F Reachability Estimation **49**
F.1 Uniform confidence bounds and diagnostic quantities . . . 49
F.2 Fixed-suite effects and rollout bounds . . . 51
F.3 Measurement from randomly sampled tasks . . . 52
F.4 Additional fixed-suite guarantees . . . 55

G Experimental Settings and Supplementary Results **55**
G.1 DS-1000 repeated evaluation and diagnosis . . . 55

H Further Directions for Harness Self-Evolution **58**

A Formal Common Setting and Principal Symbols

A.1 Formal common setting

Definition 4 (Agent state). *At step k, the agent state is*

$$A_k = (M, C_k), \tag{15}$$

where M is the frozen LLM and C_k is the evolvable textual or code state: prompts, tools, orchestration, and sub-agent structure. A valid modification persists upon adoption and satisfies

$$\Delta C \in \mathcal{M} \implies C_k \oplus \Delta C \in \{\textit{valid code states}\}. \tag{16}$$

The measurable modification space is extended to $\mathcal{M}_\perp := \mathcal{M} \sqcup \{\perp\}$ with its disjoint-union measurable structure. One complete run of a predeclared generation procedure returns a valid modification or $\perp$, denoting failure to return one. Tests, feedback, repair, and stopping may be internal to that run. A pool of size N counts all complete runs, including failures; only valid outputs define candidate agents and undergo task-performance evaluation.

For target bookkeeping, set $L_k(\perp) = D_k(\perp) = 0$, with L_k and D_k defined in Section 2.2.

Validity checks establish compile, parse, and runtime preconditions before task-performance evaluation. LLM-weight updates are outside the evolvable scope. The modification space has no prescribed metric, dependency graph, or subspace decomposition. Prompts are included in C_k, as in SICA (Robeyns et al., 2025), the Darwin Gödel Machine (Zhang et al., 2026b), and ADAS (Hu et al., 2025); structure relating modifications to their effects is discussed in Appendix H.

The task space is the measurable disjoint union

$$\mathcal{T} = \mathcal{T}_{\text{user}} \sqcup \mathcal{T}_{\text{self}}, \tag{17}$$

where self-modification tasks ask the current agent to generate a modification from failure material. Equip the task and outcome spaces with measurable structures $(\mathcal{T}, \Sigma_{\mathcal{T}})$ and $(\mathcal{O}, \Sigma_{\mathcal{O}})$, with $\Sigma_{\mathcal{O}}$ separable. Outcomes are user-facing outputs on $\mathcal{T}_{\text{user}}$ and terminal generation outputs in $\mathcal{M}_\perp$ on $\mathcal{T}_{\text{self}}$.

Definition 5 (Agent as a Markov kernel). *The agent acts on both task types through a Markov kernel*

$$A_k : \mathcal{T} \to \Delta(\mathcal{O}) : \tag{18}$$

for each t, $A_k(\cdot \mid t)$ is a probability measure, and $t \mapsto A_k(B \mid t)$ is measurable for every $B \in \Sigma_{\mathcal{O}}$.

If the measurable encoding $F \mapsto t_F \in \mathcal{T}_{\text{self}}$ supplies failure material, its generation kernel is

$$\Pi_k(\cdot \mid F) = A_k(\cdot \mid t_F). \tag{19}$$

The complete generation procedure in Appendix C.3 induces the terminal-output distribution $\pi_{k,\xi}$ of Definition 2. The current agent generates its own modifications in the main setting; the sufficient condition for reliable adoption in Corollary 1 and the pool-selection bound also apply to supplied modifications under their stated conditions. Updating the current agent changes the agent from which later modifications are generated.

The jointly measurable reward $r = r^{\text{tar}} : \mathcal{T} \times \mathcal{O} \to [0, 1]$ is fixed in reward comparisons and targets.

Assumption 1 (Bounded reward). *For every task–outcome pair, $r^{\text{tar}}(t, o) \in [0, 1]$.*

Its per-task expected reward is

$$V(A, t) := \int_{\mathcal{O}} r^{\text{tar}}(t, o) A(do \mid t) = \mathbb{E}_{o \sim A(\cdot|t)}[r^{\text{tar}}(t, o)], \tag{20}$$

which is measurable in t and lies in $[0, 1]$. Let $\mathcal{D}_{\text{user}} \in \Delta(\mathcal{T}_{\text{user}})$.

Assumption 2 (Agent-independent user-task distribution). *The user-task distribution does not depend on the agent: the same $\mathcal{D}_{\text{user}}$ is used in $J_{\text{user}}(A)$ for every agent A produced by the self-evolution process.*

This assumption applies to the fixed-distribution comparisons that invoke it. The drifting and agent-responsive analyses specify their task distributions locally. Define

$$J_{\text{user}}(A) := \mathbb{E}_{t \sim \mathcal{D}_{\text{user}}}[V(A, t)] \in [0, 1]. \tag{21}$$

Appendix B.5 gives the additional condition for transferring reward guarantees to ideal utility.

A.2 PRINCIPAL SYMBOLS

Symbol	Meaning
$A_k = (M, C_k)$	Current agent, frozen LLM, and evolvable harness (Definition 4).
$\Delta C, \mathcal{M}_\perp, \perp$	Modification, terminal-output space, and failed output (Definition 4).
$\mathcal{T}_{\text{user}}, \mathcal{T}_{\text{self}}$	User and self-modification task spaces Eq. (17).
r, V, J_{user}	Fixed reward, per-task and user-task expected rewards Eqs. (20)–(21).
ψ_A, Z_k	Failure probability and the current agent's failure rate Eqs. (22)–(23).
$\mathcal{D}_{F,k}, \mathcal{D}_{R,k}$	Failure- and retained-task distributions Eq. (1).
$L_k(c), D_k(c)$	Failure-task improvement contribution and retained-task absolute expected-reward change (Section 2.2).
$T = (\lambda, \delta), \mathcal{P}_k(T)$	Target and qualified modifications (Definition 1).
$\Pi_k, \pi_{k,\xi}$	Self-task kernel and complete-run distribution; see Eq. (19) and Appendix C.3.
$P_k(T)$	Generation probability of a qualified modification (Definition 2).
I_k	Expected-reward increment of an update attempt (Definition 9).

B RELIABLE ADOPTION AND RETAINED-TASK PERFORMANCE

B.1 FAILURE WEIGHTING AND PROOFS FOR RELIABLE ADOPTION

Definition 6 (Failure detector). *The failure detector is a measurable function $\phi : \mathcal{T}_{user} \times \mathcal{O} \to \{0, 1\}$, where $\phi(t, o) = 1$ indicates that the outcome o is judged a failure on the user task t.*

The failure detector and reward are specified separately; both are measurable.

For an agent A and task $t \in \mathcal{T}_{\text{user}}$, the failure probability is

$$\psi_A(t) := \mathbb{E}_{o \sim A(\cdot|t)}[\phi(t, o)] \in [0, 1]. \tag{22}$$

Its average over user tasks is the failure rate,

$$Z(A) := \mathbb{E}_{t \sim \mathcal{D}_{\text{user}}}[\psi_A(t)], \quad Z_k := Z(A_k). \tag{23}$$

Under Assumptions 1 and 2, any two agents $A, \tilde{A}$ are compared on the same task distribution. Linearity of expectation in Eq. (21) therefore gives

$$J_{\text{user}}(\tilde{A}) - J_{\text{user}}(A) = \mathbb{E}_{t \sim \mathcal{D}_{\text{user}}}\big[V(\tilde{A}, t) - V(A, t)\big]. \tag{24}$$

Write the per-task difference as

$$\text{Adv}_A(\tilde{A}, t) := V(\tilde{A}, t) - V(A, t) \in [-1, 1]. \tag{25}$$

It satisfies $\text{Adv}_A(A, t) \equiv 0$ and $\text{Adv}_A(\tilde{A}, t) = -\text{Adv}_{\tilde{A}}(A, t)$.

Lemma 1. *For any integrable function $g : \mathcal{T}_{user} \to \mathbb{R}$:*

$$\mathbb{E}_{t \sim \mathcal{D}_{user}}[g(t)] = \mathbb{E}_{t \sim \mathcal{D}_{user}}[\psi_{A_k}(t)\, g(t)] + \mathbb{E}_{t \sim \mathcal{D}_{user}}[(1 - \psi_{A_k}(t))\, g(t)]. \tag{26}$$

The two weighted task distributions in Eq. (1) may assign positive weight to the same task.

Substituting $g(t) = \text{Adv}_{A_k}(\tilde{A}, t)$ in Lemma 1 and using Eq. (24) gives

$$J_{\text{user}}(\tilde{A}) - J_{\text{user}}(A_k) = \underbrace{\mathbb{E}_{t \sim \mathcal{D}_{\text{user}}}[\psi_{A_k}(t)\, \text{Adv}_{A_k}(\tilde{A}, t)]}_{=: L_{A_k}(\tilde{A})} + \underbrace{\mathbb{E}_{t \sim \mathcal{D}_{\text{user}}}[(1 - \psi_{A_k}(t))\, \text{Adv}_{A_k}(\tilde{A}, t)]}_{=: E_{A_k}(\tilde{A})}. \tag{27}$$

Proof of Lemma 1. For every t, $\psi_{A_k}(t) + (1 - \psi_{A_k}(t)) = 1$. Since g is integrable and $0 \le \psi_{A_k} \le 1$, both $\psi_{A_k} g$ and $(1 - \psi_{A_k})g$ are integrable. Taking expectations under $\mathcal{D}_{\text{user}}$ and using linearity gives Eq. (26).

Definition 7. *For a reference agent with $Z_k < 1$ and a candidate agent $\widetilde{A}$, define $D_R(\widetilde{A}; A_k)$ as the average under $\mathcal{D}_{R,k}$ of the absolute change in per-task expected reward relative to the current agent:*

$$D_R(\widetilde{A}; A_k) := \mathbb{E}_{t \sim \mathcal{D}_{R,k}}\big|V(\widetilde{A}, t) - V(A_k, t)\big|.$$

This is the quantity in Section 2.2.

Proofs for reliable adoption.

Proof of Theorem 1. By Jensen's inequality, Eq. (25), and Definition 7,

$$\begin{aligned} |E_{A_k}(\widetilde{A})| &= (1 - Z_k) \left| \mathbb{E}_{t \sim \mathcal{D}_{R,k}}[\mathrm{Adv}_{A_k}(\widetilde{A}, t)] \right| \\ &\leq (1 - Z_k) \mathbb{E}_{t \sim \mathcal{D}_{R,k}}[|\mathrm{Adv}_{A_k}(\widetilde{A}, t)|] \\ &= (1 - Z_k) D_R(\widetilde{A}; A_k) \leq (1 - Z_k)\delta. \end{aligned} \tag{28}$$

Substitution into Eq. (27) gives

$$J_{\mathrm{user}}(\widetilde{A}) - J_{\mathrm{user}}(A_k) = L_{A_k}(\widetilde{A}) + E_{A_k}(\widetilde{A}) \geq L_{A_k}(\widetilde{A}) - (1 - Z_k)\delta.$$

Proof of Corollary 1. For every $c \in \mathcal{P}_k(T)$, Definition 1 gives $L_k(c) \geq \lambda$ and $D_k(c) \leq \delta$. Theorem 1 therefore gives $J_{\mathrm{user}}(A_k \oplus c) - J_{\mathrm{user}}(A_k) \geq \lambda - (1 - Z_k)\delta > 0$. Substituting the choice of $\lambda_k(\gamma, \delta)$ in Corollary 1 gives the prescribed-gain conclusion.

Proposition 2. *Fix a reference agent A_k with $Z_k < 1$.*

(i) $D_R(\tilde{A}; A_k) \in [0, 1]$.

(ii) $D_R(A_k; A_k) = 0$.

(iii) When also $Z(\tilde{A}) < 1$, the two defined quantities $D_R(\tilde{A}; A_k)$ and $D_R(A_k; \tilde{A})$ can differ.

Proof of Proposition 2. (i)–(ii) follow from Definition 7 and Assumption 1. For (iii), take three equally likely tasks, binary outputs, $r(t, o) = o$, and $\phi(t, o) = 1 - o$. Let the deterministic current agent's outputs be $(1, 0, 0)$ and the candidate outputs $(1, 1, 0)$. Their failure rates are $2/3$ and $1/3$. The current agent's retained distribution is concentrated on the first task, whereas the candidate's is uniform on the first two. Thus $D_R(\tilde{A}; A_k) = 0$ and $D_R(A_k; \tilde{A}) = 1/2$.

B.2 Tail constraints and subgroup protection

The mean constraint $\mathbb{E}_{\mathcal{D}_{R,k}}[|\mathrm{Adv}|] \leq \delta$ implies, for a measurable subgroup S with probability $w > 0$ under $\mathcal{D}_{R,k}$,

$$\mathbb{E}_{\mathcal{D}_{R,k}}\big[|\mathrm{Adv}| \,\big|\, S\big] \;\leq\; \frac{\delta}{w}, \tag{29}$$

by integrating the nonnegative variable over S. The bound becomes less informative as w decreases and exceeds the universal bound 1 when $w < \delta$.

A tail-level constraint gives uniform control for subgroups above a specified probability. For $\alpha \in (0, 1)$ write CVaR_α for the conditional value at risk at tail level α, in the Rockafellar–Uryasev form (Rockafellar & Uryasev, 2002) $\mathrm{CVaR}_\alpha(X) = \min_t\{t + \alpha^{-1}\mathbb{E}[(X - t)_+]\}$, defined for arbitrary distributions.

Definition 8. *For $\alpha \in (0, 1)$,*

$$D_R^\alpha(\tilde{A}; A_k) \;:=\; \mathrm{CVaR}_\alpha^{\mathcal{D}_{R,k}}\big(|V(\tilde{A}, \cdot) - V(A_k, \cdot)|\big). \tag{30}$$

Proposition 3. *If $D_R^\alpha(\tilde{A}; A_k) \leq \delta$, then for every measurable S with $\mathcal{D}_{R,k}(S) \geq \alpha$,*

$$\mathbb{E}_{\mathcal{D}_{R,k}}\big[|\mathrm{Adv}_{A_k}(\tilde{A}, \cdot)| \,\big|\, S\big] \;\leq\; \delta. \tag{31}$$

Moreover, $D_R \leq D_R^\alpha$.

The tail-level constraint therefore implies $D_R \leq \delta$, so Theorem 1 gives the same expected-reward guarantee as under the mean constraint of Definition 7.

Proof of Proposition 3. The risk-envelope representation of CVaR (Shapiro et al., 2009, Example 6.19) states

$$\mathrm{CVaR}_\alpha(X) \;=\; \max\Big\{\mathbb{E}[XZ] \;:\; 0 \le Z \le \alpha^{-1},\, \mathbb{E}[Z] = 1\Big\},$$

valid for arbitrary distributions. Taking $Z = \mathbf{1}_S/\mathcal{D}_{R,k}(S)$, which satisfies $0 \le Z \le \alpha^{-1}$ precisely because $\mathcal{D}_{R,k}(S) \ge \alpha$, gives $\mathbb{E}[|\mathrm{Adv}| \mid S] = \mathbb{E}[|\mathrm{Adv}|Z] \le \mathrm{CVaR}_\alpha \le \delta$. The second claim is $Z \equiv 1$ in the same representation.

Remark 1. The risk-envelope representation implies $\mathbb{E}[X \mid S] \le \mathrm{CVaR}_\alpha(X)$ whenever $\mathbb{P}(S) \ge \alpha$, without an atomlessness assumption. For distributions with atoms, Acerbi & Tasche (2002, Corollaries 3.20–3.21 and Example 5.4) show that equality with the supremum over events need not hold.

Proposition 4. *In the retained-task application, let $X = |V(\tilde{A}, t) - V(A_k, t)|$ for $t \sim \mathcal{D}_{R,k}$, and let $X_1, \ldots, X_n$ be i.i.d. observations of these true quantities. Define the empirical upper-tail statistic by*

$$\widehat{\mathrm{CVaR}}_\alpha := \min_{z \in [0,1]} \left\{ z + \frac{1}{n\alpha} \sum_{i=1}^{n} (X_i - z)_+ \right\}.$$

This definition also applies when $n\alpha$ is not an integer, with fractional weight at the empirical tail boundary. Let $\alpha, \beta \in (0,1)$. If $n \ge 25 \log(2/\beta)$, then with probability at least $1 - \beta$,

$$\big|\widehat{\mathrm{CVaR}}_\alpha - \mathrm{CVaR}_\alpha\big| \;\le\; \sqrt{\frac{3\log(2/\beta)}{n\alpha}} + \frac{15\log(2/\beta)}{n\alpha}.$$

For $0 < \varepsilon \le 1$, a radius ε is achieved by $n = O\big(\log(2/\beta)/(\alpha\varepsilon^2)\big)$ observations. When $n\alpha \ge 1$, any estimator has worst-case error $\Omega(1/\sqrt{\alpha n})$ with probability bounded away from zero. Thus the additional factor $1/\alpha$ in the sample size is necessary in the worst case at a fixed confidence level.

Proof of Proposition 4. Applying Wang et al. (2023, Theorem C.6) to $1 - X$ gives the displayed upper-tail bound under the stated sample-size condition. Requiring each term to be at most $\varepsilon/2$ gives the sufficient sample size.

For the lower bound, let $X = \mathrm{Bernoulli}(p)$ with $p \le \alpha$, so $\mathrm{CVaR}_\alpha(X) = p/\alpha$. For $n\alpha \ge 1$ and a fixed $c \in (0, 1/8]$, set $p_0 = \alpha/2$ and $p_1 = \alpha/2 + c\sqrt{\alpha/n}$. Then $p_1 \le 5\alpha/8 \le \alpha$ and the Bernoulli KL inequality gives

$$n\,\mathrm{KL}(\mathrm{Ber}(p_0) \| \mathrm{Ber}(p_1)) \le \frac{n(p_1 - p_0)^2}{p_1(1 - p_1)} \le \frac{16c^2}{3}.$$

Pinsker's inequality bounds the total variation of the two product measures by $\sqrt{8/3}\,c < 1/2$, while their CVaR values differ by $c/\sqrt{\alpha n}$. A two-point testing argument therefore gives worst-case estimation error $\Omega(1/\sqrt{\alpha n})$ with probability bounded away from zero.

The usual sufficient sample size for the mean is $O\big(\log(2/\beta)/\varepsilon^2\big)$, so estimating the tail criterion incurs an additional $1/\alpha$ cost in this worst-case i.i.d. comparison.

A rollout-based implementation additionally controls the errors in the two per-task expected rewards.

Subgroup improvement. For a fixed candidate, let $g(t) = \mathrm{Adv}_{A_k}(\widetilde{A}, t)$, $\bar{\mu} = \mathbb{E}_{\mathcal{D}_{\mathrm{user}}} g$, and $\sigma^2 = \mathrm{Var}_{\mathcal{D}_{\mathrm{user}}} g$. Cauchy–Schwarz gives, for every subgroup S with $\mathcal{D}_{\mathrm{user}}(S) \ge w_0 \in (0,1)$,

$$\mathbb{E}[g \mid S] \ge \bar{\mu} - \sigma\sqrt{\frac{1 - w_0}{w_0}}.$$

Indeed, writing $w = \mathcal{D}_{\mathrm{user}}(S)$ gives

$$\mathbb{E}[g \mid S] - \bar{\mu} = \frac{\mathrm{Cov}(g, \mathbf{1}_S)}{w} \ge -\sigma\sqrt{\frac{\mathrm{Var}(\mathbf{1}_S)}{w^2}} = -\sigma\sqrt{\frac{1-w}{w}} \ge -\sigma\sqrt{\frac{1-w_0}{w_0}},$$

where the first inequality is Cauchy–Schwarz and the last uses $w \ge w_0$. Thus $\sigma < \bar{\mu}\sqrt{w_0/(1-w_0)}$ suffices for positive improvement in every such subgroup.

B.3 Alternative retained-task norms

For $p \in [1, \infty]$, define

$$D_R^{(p)}(\tilde{A}; A_k) := \left\| V(\tilde{A}, \cdot) - V(A_k, \cdot) \right\|_{L^p(\mathcal{D}_{R,k})}.$$

In particular, $D_R^{(1)} = D_R$. At a fixed radius η, the feasible sets are nested as

$$\mathcal{C}_\eta^{(\infty)} \subseteq \cdots \subseteq \mathcal{C}_\eta^{(1)}, \qquad \mathcal{C}_\eta^{(p)} := \{\tilde{A} : D_R^{(p)}(\tilde{A}; A_k) \le \eta\}.$$

Proposition 5. *Under Assumptions 1 and 2, fix $Z_k \in (0,1)$, $p \in [1, \infty]$, and $\delta_p \ge 0$. If $D_R^{(p)}(\tilde{A}; A_k) \le \delta_p$, then $J_{\text{user}}(\tilde{A}) - J_{\text{user}}(A_k) \ge L_{A_k}(\tilde{A}) - (1 - Z_k)\delta_p$.*

Proof of Proposition 5. On the probability space with measure $\mathcal{D}_{R,k}$, $\|\text{Adv}\|_1 \le \|\text{Adv}\|_p$ for every $p \in [1, \infty]$. Thus $D_R \le \delta_p$, and Theorem 1 gives the conclusion.

At a fixed radius, the L^1 constraint is the most permissive of these nested requirements. Finite-data validation for another p requires an estimator and a coverage bound for $D_R^{(p)}$.

B.4 Output-distribution comparisons

$$D_R^{\text{KL}}(\tilde{A}; A_k) := \mathbb{E}_{t \sim \mathcal{D}_{R,k}} \left[D_{\text{KL}}(\tilde{A}(\cdot \mid t) \| A_k(\cdot \mid t)) \right]. \tag{32}$$

The direction of KL is from the candidate's output distribution to that of the current agent.

Proposition 6. *Fix A_k with $Z_k < 1$. If $V(\tilde{A}, t) = V(A_k, t)$ for $\mathcal{D}_{\text{user}}$-almost every t, then $D_R(\tilde{A}; A_k) = 0$, regardless of whether the outcome distributions differ.*

Proof of Proposition 6. By Eq. (1), $\mathcal{D}_{R,k}$ is absolutely continuous with respect to $\mathcal{D}_{\text{user}}$, with density $(1 - \psi_{A_k})/(1 - Z_k)$. Hence equality of expected rewards $\mathcal{D}_{\text{user}}$-almost everywhere also holds $\mathcal{D}_{R,k}$-almost everywhere. Integrating the absolute difference in Definition 7 gives $D_R(\tilde{A}; A_k) = 0$.

Lemma 2. *There exist $\mathcal{O}, \mathcal{T}_{user}, A_k, \tilde{A}, \mathcal{D}_{user}, r$ satisfying Assumption 1 and: (a) $V(\tilde{A}, t) = V(A_k, t)$ for all $t \in \mathcal{T}_{user}$;*

(b) $D_R(\tilde{A}; A_k) = 0$;

(c) $D_R^{\text{KL}}(\tilde{A}; A_k) = \infty$.

Proof. Fix $z \in (0,1)$. Take two tasks t_0, t_1 with probabilities $1 - z, z$ and outcomes $\{a, b\}$. On t_0, let $A_k = \delta_a$, $\tilde{A} = \delta_b$, $r(t_0, a) = r(t_0, b) = 1/2$, and $\phi(t_0, \cdot) = 0$. On t_1, give both agents the same kernel and reward and set $\phi(t_1, \cdot) = 1$. Then $Z_k = z$ and $\mathcal{D}_{R,k} = \delta_{t_0}$. The expected rewards agree on both tasks, so $D_R = 0$; the candidate assigns mass one to an output of zero probability under the current agent on t_0, giving $D_R^{\text{KL}} = \infty$.

Improvement bounds. Bounding the output-distribution KL divergence gives the following improvement bound.

Proposition 7. *Under Assumptions 1 and 2, fix a current agent with $Z_k < 1$. If $D_R^{\text{KL}}(\tilde{A}; A_k) \le \delta'$, then:*

$$J_{user}(\tilde{A}) - J_{user}(A_k) \ge L_{A_k}(\tilde{A}) - (1 - Z_k)\sqrt{\delta'/2}. \tag{33}$$

Proposition 8. *Fix a current agent with $Z_k < 1$ and a retained-task bound $\eta > 0$.*

(i) Both $D_R(\tilde{A}; A_k) \le \eta$ and $D_R^{\text{KL}}(\tilde{A}; A_k) \le 2\eta^2$ are sufficient for $\mathbb{E}_{t \sim \mathcal{D}_{R,k}}[|\text{Adv}_{A_k}(\tilde{A}, t)|] \le \eta$.

(ii) The feasible sets satisfy

$$\{\tilde{A} : D_R^{\text{KL}}(\tilde{A}; A_k) \le 2\eta^2\} \subseteq \{\tilde{A} : D_R(\tilde{A}; A_k) \le \eta\}.$$

There exist instances in the model class for which this inclusion is strict.

The constraint $D_R \leq \delta$ admits at least the candidate agents allowed by $D_R^{\mathrm{KL}} \leq 2\delta^2$, and can additionally admit reward-preserving changes in the output distribution that violate the KL constraint. Substituting $\delta' = 2\delta^2$ into Eq. (33) returns $(1 - Z_k)\delta$, exactly the retained-task term of Theorem 1. Thus the two sufficient conditions give the same expected-reward bound, while their admissible sets can differ.

Proof of Proposition 7. Use the retained-task expression for $|E_{A_k}(\tilde{A})|$ in Eq. (28). Fix t, and write $P = \tilde{A}(\cdot|t)$, $Q = A_k(\cdot|t)$, $f = r(t, \cdot) \in [0, 1]$. By the Hahn-Jordan decomposition $\nu := P - Q = \nu^+ - \nu^-$, with $\nu^+(\mathcal{O}) = \nu^-(\mathcal{O}) = D_{\mathrm{TV}}(P, Q)$. Since $f \in [0, 1]$ and $\int f \, d\nu^{\pm} \in [0, D_{\mathrm{TV}}]$:

$$|\mathrm{Adv}_{A_k}(\tilde{A}, t)| = \left| \int f \, d\nu \right| \leq D_{\mathrm{TV}}(P, Q). \tag{34}$$

By Pinsker's inequality $D_{\mathrm{TV}} \leq \sqrt{D_{\mathrm{KL}}/2}$ and the concavity of $\sqrt{\cdot}$ (Jensen):

$$\mathbb{E}_{t\sim\mathcal{D}_{R,k}}[|\mathrm{Adv}|] \leq \mathbb{E}_{t\sim\mathcal{D}_{R,k}}\left[\sqrt{D_{\mathrm{KL}}/2}\right] \leq \sqrt{D_R^{\mathrm{KL}}/2} \leq \sqrt{\delta'/2}.$$

Substituting this bound into Eq. (27) yields Eq. (33).

Proof of Proposition 8. (i) The D_R constraint gives the penalty η directly; the KL constraint gives $\sqrt{2\eta^2/2} = \eta$ by Eq. (34), Pinsker's inequality, and Jensen's inequality. (ii) Pinsker's inequality and Jensen give the explicit comparison $D_R \leq \sqrt{D_R^{\mathrm{KL}}/2}$, so every candidate satisfying the KL constraint also satisfies the corresponding D_R constraint. Lemma 2 provides an instance with $D_R = 0$ and $D_R^{\mathrm{KL}} = \infty$, so for every finite $\eta \geq 0$ strict inclusion occurs in an instance allowed by the model.

B.5 Reward fidelity and ideal utility

Let the jointly measurable utility $r^* : \mathcal{T} \times \mathcal{O} \to [0, 1]$ represent the designer's ideal preference. Define its difference from the paper's reward by $\Delta r^{\mathrm{tar}}(t, o) := r^{\mathrm{tar}}(t, o) - r^*(t, o) \in [-1, 1]$.

Assumption 3 (Uniform reward fidelity). *There is $\epsilon_{\mathrm{tar}} \geq 0$ such that $|\Delta r^{\mathrm{tar}}(t, o)| \leq \epsilon_{\mathrm{tar}}$ for every task–outcome pair.*

Write $V^*(A, t) := \int r^*(t, o) A(do \mid t)$ and $J^*_{\mathrm{user}}(A) := \mathbb{E}_{\mathcal{D}_{\mathrm{user}}} V^*(A, t)$. These quantities are measurable and lie in $[0, 1]$. The pointwise fidelity condition covers every agent, including those obtained by generation and selection.

Proposition 9. *Under Assumption 3,*

$$\left| J_{\mathrm{user}}(A) - J^*_{\mathrm{user}}(A) \right| \leq \epsilon_{\mathrm{tar}}$$

for every agent A.

Proof. By Eqs. (20)–(21) and the corresponding definitions using r^*,

$$J_{\mathrm{user}}(A) - J^*_{\mathrm{user}}(A) = \mathbb{E}_{t\sim\mathcal{D}_{\mathrm{user}}} \mathbb{E}_{o\sim A(\cdot|t)}[r(t, o) - r^*(t, o)].$$

Taking absolute values and using the triangle inequality gives

$$\left| J_{\mathrm{user}}(A) - J^*_{\mathrm{user}}(A) \right| \leq \mathbb{E}_{t\sim\mathcal{D}_{\mathrm{user}}} \mathbb{E}_{o\sim A(\cdot|t)} \left[|r(t, o) - r^*(t, o)| \right] \leq \epsilon_{\mathrm{tar}},$$

where the last inequality is pointwise by Assumption 3.

Corollary 3. *Under Assumption 3, if*

$$J_{\mathrm{user}}(\tilde{A}) - J_{\mathrm{user}}(A_k) \geq \gamma,$$

then

$$J^*_{\mathrm{user}}(\tilde{A}) - J^*_{\mathrm{user}}(A_k) \geq \gamma - 2\epsilon_{\mathrm{tar}}. \tag{35}$$

Consequently, a finite-sample lower bound on the expected-reward difference transfers on the same confidence event after subtracting $2\epsilon_{\mathrm{tar}}$.

Proof. Write the ground-truth improvement as

$$J^*_{\text{user}}(\tilde{A}) - J^*_{\text{user}}(A_k) = J_{\text{user}}(\tilde{A}) - J_{\text{user}}(A_k) \\ - \big[J_{\text{user}}(\tilde{A}) - J^*_{\text{user}}(\tilde{A})\big] + \big[J_{\text{user}}(A_k) - J^*_{\text{user}}(A_k)\big].$$

Applying Proposition 9 to the two agents proves Eq. (35).

In particular, when Theorem 2 supplies the stated lower bound Δ_k on the expected-reward difference on its confidence event, Corollary 3 applies with $\gamma = \Delta_k$.

For fixed r, r^* and $\mathcal{D}_{\text{user}}$, applying Corollary 3 to A_0, A_K gives an endpoint utility gain at least $J_{\text{user}}(A_K) - J_{\text{user}}(A_0) - 2\epsilon_{\text{tar}}$.

C Generation and Validation for Reliable Adoption

C.1 Validation data and update procedure

This appendix supplies the evaluation conditions and proofs for Section 3. The update procedure uses the assumptions below and Conditions 1 and 2.

The sequence $\{(T_n, O_n, \Phi_n)\}_{n\geq 1}$ records the user-facing task $T_n \in \mathcal{T}_{\text{user}}$, the outcome $O_n \in \mathcal{O}$, and the failure indicator $\Phi_n := \phi(T_n, O_n)$ for the n-th task handled by the current agent.

Assumption 4 (Stationary β-mixing task process). *$\{T_n\}_{n\geq 1}$ is a stationary ergodic process with marginal distribution $T_n \sim \mathcal{D}_{user}$. Write $\mathcal{F}_n := \sigma(T_1, \ldots, T_n)$ and $\mathcal{F}^{\infty}_{n+k} := \sigma(T_{n+k}, T_{n+k+1}, \ldots)$. The process satisfies β-mixing: there exist coefficients $\{\beta(k)\}_{k\geq 1}$ tending to zero, where*

$$\beta(k) := \sup_{n\geq 1} \mathbb{E}\left[\sup_{B\in\mathcal{F}^{\infty}_{n+k}} \big|\mathbb{P}(B \mid \mathcal{F}_n) - \mathbb{P}(B)\big|\right]. \tag{36}$$

Remark 2. Stationary β-mixing processes (Yu, 1994; Mohri & Rostamizadeh, 2008) include i.i.d. tasks, finite-range dependent tasks, and ergodic finite-state Markov chains under the usual aperiodicity conditions. Independence beyond a finite lag holds for the finite-range case.

Assumption 4 and ℓ control the stored-sample distribution comparison. Failure-indexed evaluation is governed separately by Condition 2, with Proposition 10 providing an explicit i.i.d. task-stream construction.

Observed outcomes and fresh evaluation.

Assumption 5. *The failure detector is local to each task: Φ_n depends only on (T_n, O_n), not on $\{T_m, O_m : m \neq n\}$.*

Assumption 6. *Given T_n and the current agent A_k, $O_n \sim A_k(\cdot \mid T_n)$ is conditionally independent of $\{T_m, O_m : m < n\}$.*

Assumption 7. *Strengthening Assumption 6, for the current agent A_k, the sequence $\{O_n\}_{n\geq 1}$ satisfies*

$$O_n \perp \{(T_m, O_m) : m \neq n\} \cup \{O^{fresh}_{m,l} : \forall m, l\} \,\Big|\, T_n. \tag{37}$$

where $\{O^{fresh}_{m,l}\}$ are the fresh evaluation rollouts used for validation and defined in Eq. (40)*. That is, the observed task outcome O_n given T_n is conditionally independent of the other observed task records and fresh rollouts for this fixed current agent.*

Assumption 8. *For a fixed current agent A_k and task t, repeated evaluations of ϕ and r are i.i.d. When the evaluator is an LLM judge, its prompt template, model version, sampling temperature, and other settings are held fixed.*

Definition 9 (Evolution step). *The k-th evolution step uses predeclared finite task-arrival horizons for generation material, evaluation on failure tasks, and any recollection of the stored sample. Its output is*

$$(A_k, F_k^{\mathrm{gen}}, \Delta C, \mathrm{decision}_k) \mapsto A_{k+1} = \begin{cases} A_k \oplus \Delta C, & \mathrm{decision}_k = \mathrm{accept}, \\ A_k, & \textit{otherwise}, \end{cases} \tag{38}$$

When generation material is available, the current agent draws

$$\Delta C \sim \Pi_k(\cdot \mid F_k^{\mathrm{gen}}) = A_k(\cdot \mid t_{F_k^{\mathrm{gen}}}); \tag{39}$$

otherwise $\Delta C = \perp$. Here $A_k = (M, C_k)$ is the current agent and F_k^{gen} is the returned generation material. Write $\mathsf{Cand}_k := \{\Delta C \in \mathcal{M}\}$ and $\mathrm{decision}_k \in \{\mathrm{accept}, \mathrm{reject}, \mathrm{abstain}\}$. The increment $I_k := J_{\mathrm{user}}(A_{k+1}) - J_{\mathrm{user}}(A_k)$ is defined on every branch. Observable margins have their pre-adoption definitions on $\mathsf{Cand}_k \cap \mathsf{Ready}_k$ and are zero elsewhere. Procedure 1 specifies collection and evaluation.

Stored sample and returned failure batch. The stored sample consists of n_R task records (t_j^R, ϕ_j^R) labeled by the current agent. Its distribution satisfies Condition 1; Remark 3 gives a fixed-window construction. Separately, Phase E returns n_F failure-marked tasks within its declared finite horizon, or returns $\perp$.

Reward estimation from repeated evaluations. Failure-marked outcomes of the current agent are not reused as value estimates. Given the frozen agents and evaluation records, both A_k and $\tilde{A}$ are rerun using independent groups of m fresh rollouts on every evaluation task:

$$\hat{V}(A, t) := \frac{1}{m} \sum_{l=1}^{m} r(t, o_l), \quad o_l \overset{\text{i.i.d.}}{\sim} A(\cdot|t). \tag{40}$$

The stored failure mark remains unchanged because it defines the weighting by the retained-task distribution. The stored failure-rate estimate is $\hat{Z}_k := \frac{1}{n_R} \sum_{j=1}^{n_R} \phi_j^R$. The failure-task improvement estimate is

$$\hat{L}_k := \hat{Z}_k \cdot \frac{1}{n_F} \sum_{i=1}^{n_F} \big[\hat{V}(\tilde{A}, t_i^F) - \hat{V}(A_k, t_i^F)\big], \tag{41}$$

and, on the nonempty retained set $S_R := \{j : \phi_j^R = 0\}$,

$$\hat{D}_R := \frac{1}{n_R(1 - \hat{Z}_k)} \sum_{j=1}^{n_R} (1 - \phi_j^R) \left|\hat{V}(\tilde{A}, t_j^R) - \hat{V}(A_k, t_j^R)\right|, \tag{42}$$

equivalently,

$$\hat{D}_R = \frac{1}{|S_R|} \sum_{j \in S_R} \left|\hat{V}(\tilde{A}, t_j^R) - \hat{V}(A_k, t_j^R)\right|. \tag{43}$$

All statistics and radii under the retained-task distribution are evaluated only when $|S_R| \geq 1$; otherwise the validation rule abstains. The expected-reward estimate from the full stored sample is

$$\hat{J}_k := \frac{1}{n_R} \sum_{j=1}^{n_R} \hat{V}(A_k, t_j^R).$$

Evaluation-data separation and readiness.

Condition 1 (Evaluation-data separation and stored-sample comparison). *The generation material F_k^{gen} is disjoint from the evaluation failure batch F_k^{gate} and the stored sample R_k. Write $n_F = |F_k^{\mathrm{gate}}|$. Let $\mathcal{G}_k$ contain the current agent, frozen candidate, generation information and randomness, and any stored-sample offset. The joint law $\mathbb{P}$ of this information and the stored records has a reference law $\mathbb{P}^\circ$ with the same $\mathcal{G}_k$ marginal such that, under $\mathbb{P}^\circ$, conditionally on $\mathcal{G}_k$,*

$$T_{1:n_R}^\circ \overset{\text{iid}}{\sim} \mathcal{D}_{\mathrm{user}}, \qquad \mathcal{L}(\Phi_{1:n_R}^\circ \mid \mathcal{G}_k, T_{1:n_R}^\circ) = \bigotimes_{j=1}^{n_R} \mathrm{Bernoulli}(\psi_{A_k}(T_j^\circ)),$$

and

$$\|\mathbb{P} - \mathbb{P}^{\circ}\|_{\mathrm{TV}} \le 2n_R\beta(\ell). \tag{44}$$

A coupling preserving the generation information and disagreeing on the records with probability at most this bound is sufficient. Fresh-rollout coverage is imposed separately, given the frozen agents and observed evaluation records.

Condition 2 (Returned-sample coverage). *Let* Pre_k *contain all information frozen before Phase E: the current agent and candidate agent, the finite family* $\mathcal{Q}_k^{\mathrm{eval}}$ *and deterministic range bounds, radii* $r_{F,k}(f)$*, deterministic risk* $\beta_{F,k}$*, the finite arrival horizon, and the return rule. Phase E either returns* n_F *failure tasks* $T^F_{k,1:n_F}$ *or returns* $\perp$*; write* $\mathsf{Ret}_{F,k}$ *for the return event. For*

$$\bar{f}_{F,k} := \frac{1}{n_F}\sum_{i=1}^{n_F} f(T^F_{k,i}), \qquad \mu_{F,k}(f) := \mathbb{E}_{T\sim\mathcal{D}_{F,k}}[f(T)],$$

the returned failure sample satisfies

$$\mathbb{P}\big(\mathsf{Ret}_{F,k} \cap \{\exists f \in \mathcal{Q}_k^{\mathrm{eval}} : |\bar{f}_{F,k} - \mu_{F,k}(f)| > r_{F,k}(f)\}\big) \le \beta_{F,k}.$$

The candidate agent, family, radii, risk, and return rule may depend on Pre_k *but not on Phase-E outcomes. Let* $\mathsf{Ready}_{R,k}$ *denote availability of a valid stored sample under Condition 1, set* $\mathsf{Ready}_k := \mathsf{Ret}_{F,k} \cap \mathsf{Ready}_{R,k}$*, and require every acceptance event to lie in* $\mathsf{Cand}_k \cap \mathsf{Ready}_k$*.*

Remark 3. Fix a deterministic task index b. Let the current agent, candidate, and generation information be determined by tasks through b and auxiliary randomness independent of the task stream. Draw an independent $U \sim \mathrm{Unif}\{0,\dots,\ell-1\}$ and collect records at indices $b + j\ell + U$, $j = 1,\dots,n_R$. Given the generation information and these tasks, generate the labels independently with probabilities $\psi_{A_k}(t_j^R)$, using randomness separated from generation.

Conditioning on U, the first task is at least ℓ indices after b, and successive tasks are ℓ apart. The absolute-regularity comparison of the history and these tasks with the product reference law (Yu, 1994; Doukhan, 1994), applied successively over these n_R gaps, is bounded by $2n_R\beta(\ell)$. Applying the same conditional product marking kernel to both laws does not increase total variation. Averaging over U proves Eq. (44). This compares the joint experiments, preserving the same generation-information marginal.

Remark 4. In this construction the last stored record arrives by $b + (n_R+1)\ell - 1$. The declared recollection horizon includes the initial wait and all subsequent gaps. Collection is scheduled from the fixed cutoff b even if generation material becomes available earlier. Evaluation readiness additionally requires at least one retained mark, as specified in Procedure 1.

Concentration radii. Lemma 3 gives $\varepsilon_V = \sqrt{\log(2/\beta_V^{(1)})/(2m)}$. Lemma 4 gives the mixing-aware ε_Z in Eq. (47). For $f = \mathrm{Adv}_{A_k}(\tilde{A},\cdot)$, write $\varepsilon_F := r_{F,k}(f)$, $\varepsilon_\mu := 2\varepsilon_V + \varepsilon_F$, and $\varepsilon_L := \varepsilon_Z + (\hat{Z}_k + \varepsilon_Z)\varepsilon_\mu$. Lemma 5 controls $|\hat{L}_k - L_{A_k}(\tilde{A})|$ by ε_L on the stated coverage event. Lemma 6 gives

$$\varepsilon_D := \frac{2\varepsilon_V + \varepsilon_R + \varepsilon_Z}{1 - \hat{Z}_k}, \qquad \varepsilon_R := \sqrt{\frac{\log(2/\widetilde{\beta}_R)}{2n_R}},$$

on the nonempty-retained branch. The exact events and probability allocations are recorded with the corresponding lemmas below and in Appendix C.2.

Proposition 10 (First-failure sampling on an i.i.d. task stream). *Let* $\mathsf{Run}_{F,k}$ *be the* Pre_k*-measurable event that Phase E is entered. Suppose that, conditionally on* Pre_k*,* $0 < Z_k < 1$ *and* $1 \le n_F \le H_{F,k} < \infty$*, and, on* $\mathsf{Run}_{F,k}$*, the next* $H_{F,k}$ *task arrivals form i.i.d. marked pairs* (T_s, Φ_s) *with* $T_s \sim \mathcal{D}_{\mathrm{user}}$ *and* $\Phi_s \mid T_s \sim \mathrm{Bernoulli}(\psi_{A_k}(T_s))$*. The current agent and failure detector remain fixed throughout the window. The evaluator returns the first* n_F *tasks with* $\Phi_s = 1$ *if at least* n_F *such marks occur, and otherwise returns* $\perp$*.*

(i) Conditional on Pre_k *and* $\mathsf{Ret}_{F,k}$*, the returned tasks are i.i.d. from* $\mathcal{D}_{F,k}$*. Hence, for a frozen family of* $1 \le M < \infty$ *bounded statistics with deterministic* $f \in [a_f, b_f]$*,*

$$r_{F,k}(f) = (b_f - a_f)\sqrt{\frac{\log(2M/\beta_{F,k})}{2n_F}}$$

instantiates Condition 2.

(ii) With $B_{H,n}(z) := \mathbb{P}\{\mathrm{Binomial}(H, z) \geq n\}$,

$$\mathbb{P}(\mathsf{Ret}_{F,k} \mid \mathsf{Pre}_k) = \mathbf{1}_{\mathsf{Run}_{F,k}} B_{H_{F,k},n_F}(Z_k).$$

(iii) If a clipped Pre_k*-measurable lower bound* $\underline{Z}_k$ *satisfies* $\mathbb{P}(\mathsf{Run}_{F,k} \cap \{\underline{Z}_k > Z_k\}) \leq \beta_{Z,\mathrm{plan}}$ *and, whenever Phase E is entered,* $\underline{Z}_k > 0$ *and* $B_{H_{F,k},n_F}(\underline{Z}_k) \geq 1 - \alpha_{F,\mathrm{ret}}$*, then*

$$\mathbb{P}(\mathsf{Run}_{F,k} \cap \mathsf{Ret}_{F,k}^c) \leq \beta_{Z,\mathrm{plan}} + \alpha_{F,\mathrm{ret}}.$$

In particular, if $\mathsf{Run}_{F,k}$ *holds almost surely, then* $\mathbb{P}(\mathsf{Ret}_{F,k}) \geq 1 - \beta_{Z,\mathrm{plan}} - \alpha_{F,\mathrm{ret}}$.

More generally, a sampler-specific analysis may prove, conditionally on each realized Pre_k,

$$\mathbb{P}\big(\mathsf{Ret}_{F,k} \cap \{|\bar{f}_{F,k} - \mu_{F,k}(f)| > x\} \mid \mathsf{Pre}_k\big) \leq 2\exp\left(-\frac{2n_F x^2}{q_{F,k}(b_f - a_f)^2}\right).$$

A union bound then gives $r_{F,k}(f) = (b_f - a_f)\sqrt{q_{F,k}\log(2M/\beta_{F,k})/(2n_F)}$. The value $q_{F,k}$ must be justified by the sampling model; for the advantage statistic, the i.i.d. construction gives $\varepsilon_F = \sqrt{2\log(2/\beta_{F,k})/n_F}$.

Remark 5. The stored labels have probabilities $\psi_{A_k}(t_j^R)$ for the current agent at collection. Adoption both changes the reference agent and uses these records in selecting the successor. Recollection therefore supplies current-agent labels and data satisfying Condition 1. A sample carried after rejection must satisfy the same condition for its next use.

Procedure 1 (Generation, validation, and update). Use A_k throughout the step.

1. Observe at most $H_{\mathrm{gen},k}$ task arrivals and return the first n_F failures as generation material. If the quota is not met, abstain. Otherwise draw $\Delta C \sim A_k(\cdot \mid t_{F_k^{\mathrm{gen}}})$; if $\Delta C = \perp$, abstain.

2. For $\Delta C \in \mathcal{M}$, form $\tilde{A} = A_k \oplus \Delta C$. Before Phase E, freeze the candidate agent, finite statistic family, ranges and radii, risk allocation, failure detector, Phase-E return rule, and $H_{F,k}$.

3. Use one predeclared carried sample or finite recollection window $H_{R,k}$ satisfying Condition 1. If it does not yield a valid nonempty retained subset, abstain.

4. In Phase E, observe the next $H_{F,k}$ task arrivals with the current agent and detector fixed. Return F_k^{gate}, consisting of the first n_F failures, or abstain if the quota is not met.

5. On $\mathsf{Cand}_k \cap \mathsf{Ready}_k$, perform the independent fresh rollouts of both agents and compute the validation statistics and Δ_k in Eq. (5). Set $\Delta_k = 0$ off this branch.

6. On $\mathsf{Cand}_k \cap \mathsf{Ready}_k$, validate the modification if and only if Eqs. (4) and (5) hold; set $\mathrm{decision}_k = \mathrm{accept}$ exactly when it is validated and reject otherwise. Apply the update of Definition 9; adoption marks the stored sample for recollection.

The runtime check Eq. (5) uses only $\hat{Z}_k$, ε_Z, ε_D, and ε_L, all observable at decision time. After the samples return, evaluating one candidate and the current agent uses $2m(n_F + |S_R|)$ complete rollouts for the two validation statistics; task-arrival horizons are accounted for separately.

The error bounds for F_k^{gate} and R_k are combined by a union bound. They may share evaluation records while satisfying their respective sampling conditions.

When both generation and evaluation batches return, their quotas total $2n_F$ failures. Under i.i.d. arrivals, $2n_F/Z_k$ is the corresponding infinite-horizon planning scale; finite-window return is governed by Proposition 10.

Proposition 11. *For each* t*, let* $\delta(t) := V(\tilde{A}, t) - V(A_k, t)$ *and* $\hat{\delta}(t) := \hat{V}(\tilde{A}, t) - \hat{V}(A_k, t)$ *(based on the independent rollouts Eq.* (40)*). Then:*

(i) For every t, $\mathbb{E}[|\hat{\delta}(t)|] \geq |\delta(t)|$. Equality holds if and only if $\hat{\delta}(t)$ is almost surely nonnegative or almost surely nonpositive; equivalently, the inequality is strict exactly when $\mathbb{P}(\hat{\delta}(t) > 0) > 0$ and $\mathbb{P}(\hat{\delta}(t) < 0) > 0$.

(ii) let $\sigma_t^2 := \mathrm{Var}_{o\sim A_k(\cdot|t)}[r(t,o)] + \mathrm{Var}_{o\sim \tilde{A}(\cdot|t)}[r(t,o)]$. Then $\sigma_t^2 \leq 1/2$, and:

$$0 \leq \mathbb{E}[|\hat{\delta}(t)|] - |\delta(t)| \leq \sqrt{\mathrm{Var}[\hat{\delta}(t)]} = \sigma_t/\sqrt{m} \leq \frac{1}{\sqrt{2m}}. \tag{45}$$

(iii) let $\mathcal{G} := \sigma(\{(t_j^R, \phi_j^R)\}_{j\leq n_R}) \vee \sigma(\tilde{A})$. On the nonempty-retained branch,

$$\mathbb{E}\big[\hat{D}_R \mid \mathcal{G}\big] \;\geq\; \hat{D}_R^{\mathrm{exact}} := \frac{1}{|S_R|}\sum_{j\in S_R} \big|\delta(t_j^R)\big| \qquad \text{a.s. on } \{|S_R| \geq 1\}. \tag{46}$$

For fixed current and candidate agents and independently collected i.i.d. marked task records, averaging Eq. (46) conditional on $|S_R| \geq 1$ gives $\mathbb{E}[\hat{D}_R \mid |S_R| \geq 1] \geq D_R$. For a random candidate, the same comparison is conditional on the generation information under the product reference experiment of Condition 1.

Proof of Proposition 11. (i) By Jensen's inequality, $\mathbb{E}|X| \geq |\mathbb{E}X|$. Applying it to $X = \hat{\delta}(t)$ and using $\mathbb{E}X = \delta(t)$ gives the stated direction. Equality in Jensen holds exactly when X lies almost surely in one affine branch of $|\cdot|$, i.e. when $X \geq 0$ a.s. or $X \leq 0$ a.s. Hence the inequality is strict exactly when X takes both positive and negative values with positive probability.

(ii) The variance of $\hat{\delta}(t) - \delta(t)$ equals the sum of the variances of the two independent groups of rollouts: $\mathrm{Var}[\hat{\delta}(t)] = \sigma_t^2/m$. Since $r \in [0,1]$, $\mathrm{Var}(r) \leq 1/4$ per agent, hence $\sigma_t^2 \leq 1/2$.

By $|x| - |y| \leq |x - y|$ together with Jensen applied to $\sqrt{\cdot}$:

$$\mathbb{E}|\hat{\delta}(t)| - |\delta(t)| \leq \mathbb{E}|\hat{\delta}(t) - \delta(t)| \leq \sqrt{\mathrm{Var}[\hat{\delta}(t)]} = \sigma_t/\sqrt{m}.$$

(iii) Work on $\{|S_R| \geq 1\}$. Given $\mathcal{G}$, both S_R and its cardinality are fixed, so

$$\mathbb{E}\big[\hat{D}_R \mid \mathcal{G}\big] = \frac{1}{|S_R|}\sum_{j\in S_R} \mathbb{E}\big[|\hat{\delta}_j| \mid \mathcal{G}\big] = \frac{1}{|S_R|}\sum_{j\in S_R} h_{\tilde{A}}(t_j^R), \qquad h_{\tilde{A}}(t) := \mathbb{E}\big[|\hat{\delta}(t)|\big],$$

where the fresh rollouts obey Eq. (40) and Assumptions 7 and 8 for the fixed candidate agent. Applying (i) to each summand proves Eq. (46).

Lemma 3. *Under Eq.* (40)*, for any A, t:*

$$\mathbb{P}\big[|\hat{V}(A,t) - V(A,t)| \geq \varepsilon_V\big] \leq 2\exp(-2m\varepsilon_V^2) =: \beta_V^{(1)}.$$

That is, $\varepsilon_V = \sqrt{\log(2/\beta_V^{(1)})/(2m)}$.

Proof of Lemma 3. Under Eq. (40) and Assumption 1, the m rollout rewards are independent and lie in $[0,1]$. Hoeffding's inequality for their sample mean therefore gives

$$\mathbb{P}\big[|\hat{V}(A,t) - V(A,t)| \geq \varepsilon_V\big] \leq 2\exp(-2m\varepsilon_V^2).$$

Setting the right-hand side to $\beta_V^{(1)}$ and solving for ε_V yields the stated radius.

Lemma 4. *Under Assumptions 4, 5, 6, and 7 and the stored-sample comparison of Condition 1, with block size ℓ, the common-offset block sample defined above, and $2n_R\beta(\ell) < \beta_Z$:*

$$\mathbb{P}[|\hat{Z}_k - Z_k| \geq \varepsilon_Z] \leq 2\exp\big(-2n_R\varepsilon_Z^2\big) + 2n_R\,\beta(\ell) =: \beta_Z. \tag{47}$$

The corresponding $\varepsilon_Z = \sqrt{\frac{\log(2/\tilde{\beta}_Z)}{2n_R}}$, where $\tilde{\beta}_Z = \beta_Z - 2n_R\beta(\ell)$.

Proof of Lemma 4. Under the reference law of Condition 1, conditionally on $\mathcal{G}_k$, the labels are i.i.d. Bernoulli(Z_k) after integrating the independent tasks. Hoeffding bounds the reference probability of $|\hat{Z}_k - Z_k| \geq \varepsilon_Z$ by $2e^{-2n_R\varepsilon_Z^2}$. Transferring this joint event by Eq. (44) adds at most $2n_R\beta(\ell)$, proving Eq. (47).

Write

$$\mathcal{E}_Z := \{|\hat{Z}_k - Z_k| \leq \varepsilon_Z\}, \qquad \mathcal{E}_V^{(B)} := \bigcap_{t\in B} \bigcap_{A\in\{A_k,\tilde{A}\}} \{|\hat{V}(A,t) - V(A,t)| \leq \varepsilon_V\},$$

where $B = F_k^{\mathrm{gate}}$ or R_k denotes the corresponding list of task records; abbreviate these events as $\mathcal{E}_V^{(F)}$ and $\mathcal{E}_V^{(R)}$. The fresh-rollout union bounds on available records are $2n_F\beta_V^{(1)}$ and $2n_R\beta_V^{(1)}$, respectively.

Lemma 5. *Under Assumption 1, assume the $\hat{Z}_k$ coverage of Lemma 4, the per-task fresh-rollout coverage of Lemma 3, and Condition 2 for $f = \mathrm{Adv}_{A_k}(\tilde{A},\cdot)$. Let*

$$\mu_F := \mathbb{E}_{\mathcal{D}_{F,k}}[\mathrm{Adv}_{A_k}(\tilde{A},T)], \qquad L_{A_k}(\tilde{A}) = Z_k\mu_F,$$

and set $\varepsilon_F := r_{F,k}(f)$ and $\varepsilon_\mu := 2\varepsilon_V + \varepsilon_F$. On the returned branch write

$$\mathcal{E}_{F,k}(f) := \{|\bar{f}_{F,k} - \mu_{F,k}(f)| \leq \varepsilon_F\}.$$

On $\mathsf{Ready}_k \cap \mathcal{E}_Z \cap \mathcal{E}_V^{(F)} \cap \mathcal{E}_{F,k}(f)$,

$$|\hat{L}_k - L_{A_k}(\tilde{A})| \leq \varepsilon_Z + Z_k\varepsilon_\mu \leq \varepsilon_L := \varepsilon_Z + (\hat{Z}_k + \varepsilon_Z)\varepsilon_\mu. \tag{48}$$

Moreover,

$$\mathbb{P}\Big(\mathsf{Ready}_k \cap \{|\hat{L}_k - L_{A_k}(\tilde{A})| > \varepsilon_L\}\Big) \leq \beta_Z + 2n_F\beta_V^{(1)} + \beta_{F,k}.$$

Proof of Lemma 5. Work on Ready_k, which is a subset of $\mathsf{Ret}_{F,k}$, and define

$$\hat{\mu}_F := \frac{1}{n_F}\sum_{i=1}^{n_F}\big(\hat{V}(\tilde{A},T_{k,i}^F) - \hat{V}(A_k,T_{k,i}^F)\big), \quad \bar{\mu}_F := \frac{1}{n_F}\sum_{i=1}^{n_F}\mathrm{Adv}_{A_k}(\tilde{A},T_{k,i}^F).$$

The product identity gives

$$\hat{L}_k - L_{A_k}(\tilde{A}) = (\hat{Z}_k - Z_k)\hat{\mu}_F + Z_k(\hat{\mu}_F - \mu_F).$$

On $\mathcal{E}_Z$, the first term is at most ε_Z, since $|\hat{\mu}_F| \leq 1$. On $\mathcal{E}_V^{(F)}$, fresh-rollout coverage gives $|\hat{\mu}_F - \bar{\mu}_F| \leq 2\varepsilon_V$. Finally, Condition 2 gives

$$\mathbb{P}(\mathsf{Ret}_{F,k} \cap \{|\bar{\mu}_F - \mu_F| > \varepsilon_F\}) \leq \beta_{F,k}.$$

Consequently the triangle inequality yields $|\hat{L}_k - L_{A_k}(\tilde{A})| \leq \varepsilon_Z + Z_k(2\varepsilon_V + \varepsilon_F)$ on the stated coverage event. The inequality $Z_k \leq \hat{Z}_k + \varepsilon_Z$ on $\mathcal{E}_Z$ gives the observable radius in Eq. (48). A union bound over the estimator failures on the ready branch proves the displayed joint error bound.

Proof of Proposition 10. Condition on Pre_k. On $\mathsf{Run}_{F,k}$ the current agent, detector, horizon, and return rule are fixed. For any complete mark vector $\varphi \in \{0,1\}^{H_{F,k}}$ with at least n_F successes, the i.i.d. marked-pair distribution factorizes across positions. At every success position,

$$\mathbb{P}(T_s \in dt \mid \Phi_s = 1, \mathsf{Pre}_k) = \frac{\psi_{A_k}(t)\,\mathcal{D}_{\mathrm{user}}(dt)}{Z_k} = \mathcal{D}_{F,k}(dt).$$

The success-position tasks are conditionally independent, while integration over the nonfailure positions contributes a factor independent of their values. This distribution is the same for every such φ. Mixing over all return-producing mark vectors therefore leaves the first n_F failure tasks i.i.d. from $\mathcal{D}_{F,k}$ after conditioning on return.

For each frozen $f \in \mathcal{Q}_k^{\mathrm{eval}}$, Hoeffding's inequality under this returned-task distribution, followed by a union bound over the M declared statistics, gives part (i) and the event-level Condition 2 bound. On $\mathsf{Run}_{F,k}$ the failure marks are i.i.d. Bernoulli(Z_k), so the return probability is $B_{H_{F,k},n_F}(Z_k)$; off that event the rule returns $\perp$. This proves part (ii).

For part (iii), $B_{H,n}(z)$ is nondecreasing in z. On $\mathsf{Run}_{F,k} \cap \{\underline{Z}_k \leq Z_k\}$ the declared tail condition therefore makes conditional nonreturn probability at most $\alpha_{F,\mathrm{ret}}$. Adding the probability of $\mathsf{Run}_{F,k} \cap \{\underline{Z}_k > Z_k\}$ proves the stated bound.

C.2 RETAINED-TASK ESTIMATION AND THE RELIABLE-ADOPTION PROOF

Lemma 6. *Under Assumptions 1, 4, 5, 6, 7, and 8, Condition 1, and* $2n_R\beta(\ell) < \min\{\beta_Z, \beta_R\}$*, put*

$$g_k(t) := |V(\tilde{A}, t) - V(A_k, t)|, \quad \bar{W}_g := \frac{1}{n_R}\sum_j (1-\phi_j^R) g_k(t_j^R), \quad \widetilde{\beta}_R := \beta_R - 2n_R\beta(\ell) > 0.$$

Define

$$\varepsilon_R := \sqrt{\frac{\log(2/\widetilde{\beta}_R)}{2n_R}}, \quad \mathcal{E}_R := \{|\bar{W}_g - (1-Z_k)D_R| \le \varepsilon_R\}, \quad \mathcal{E}_D := \mathcal{E}_Z \cap \mathcal{E}_V^{(R)} \cap \mathcal{E}_R.$$

On $\{|S_R| \ge 1\} \cap \mathcal{E}_D$*,*

$$|\hat{D}_R - D_R| \le \varepsilon_D := \frac{2\varepsilon_V + \varepsilon_R + \varepsilon_Z}{1 - \hat{Z}_k}. \tag{49}$$

The corresponding joint error probability on the ready branch is at most $\beta_Z + \beta_R + 2n_R\beta_V^{(1)}$*.*

Proof of Lemma 6. First separate fresh-rollout error. Let $\hat{g}_j := |\hat{V}(\tilde{A}, t_j^R) - \hat{V}(A_k, t_j^R)|$ and $W_j := (1-\phi_j^R)\hat{g}_j$. On $\mathcal{E}_V^{(R)}$, the triangle inequality gives

$$|\hat{g}_j - g_k(t_j^R)| \le 2\varepsilon_V. \tag{50}$$

Consequently, with

$$\bar{W}_g = \frac{1}{n_R}\sum_j (1-\phi_j^R) g_k(t_j^R), \tag{51}$$

we have

$$\left|\frac{1}{n_R}\sum_j W_j - \bar{W}_g\right| \le 2\varepsilon_V. \tag{52}$$

Under the reference law in Condition 1, condition on $\mathcal{G}_k$. The function g_k is fixed and the reference weighted observations are independent in $[0,1]$, with mean

$$\int (1-\psi_{A_k}(t)) g_k(t)\, \mathcal{D}_{\text{user}}(dt) = (1-Z_k)D_R.$$

Conditional Hoeffding and then averaging over the unchanged $\mathcal{G}_k$ marginal give reference failure probability $2e^{-2n_R\varepsilon_R^2}$. The joint-law comparison Eq. (44) therefore yields

$$\mathbb{P}\{|\bar{W}_g - (1-Z_k)D_R| > \varepsilon_R\} \le 2e^{-2n_R\varepsilon_R^2} + 2n_R\beta(\ell) = \beta_R. \tag{53}$$

Fresh rollouts retain their prescribed conditional distributions given the observed records and frozen agents. Their error event and $\mathcal{E}_R$ are combined by a union bound.

Finally, on $\mathcal{E}_V^{(R)} \cap \mathcal{E}_R$,

$$\left|\frac{1}{n_R}\sum_j W_j - (1-Z_k)D_R\right| \le 2\varepsilon_V + \varepsilon_R. \tag{54}$$

Since the sample numerator is $(1-\hat{Z}_k)\hat{D}_R$,

$$|(1-\hat{Z}_k)\hat{D}_R - (1-Z_k)D_R| \le 2\varepsilon_V + \varepsilon_R. \tag{55}$$

On the nonempty retained branch, $1-\hat{Z}_k = |S_R|/n_R > 0$ and

$$\hat{D}_R - D_R = \frac{[(1-\hat{Z}_k)\hat{D}_R - (1-Z_k)D_R] + D_R(\hat{Z}_k - Z_k)}{1-\hat{Z}_k}.$$

Using $D_R \in [0,1]$ and $\mathcal{E}_Z$ gives

$$|\hat{D}_R - D_R| \le \frac{2\varepsilon_V + \varepsilon_R + \varepsilon_Z}{1-\hat{Z}_k}. \tag{56}$$

The union of the three coverage failures has probability at most $\beta_Z + \beta_R + 2n_R\beta_V^{(1)}$ on available records.

Remark 6. The reference sample has size n_R. Dependence contributes the additive error $2n_R\beta(\ell)$, leaving the Hoeffding radius $\sqrt{\log(2/\widetilde{\beta})/(2n_R)}$ with $\widetilde{\beta} = \beta - 2n_R\beta(\ell) > 0$. At a fixed gap with $\beta(\ell) > 0$, the chosen n_R must satisfy $2n_R\beta(\ell) < \min\{\beta_Z, \beta_R\}$ for both stored-sample estimates; increasing n_R without enlarging the gap need not improve these guarantees. For an i.i.d. comparison, $\beta(\ell) = 0$ and the ordinary independent-sample radius applies. Failure-task concentration uses the separate returned-sample radius of Condition 2.

Proof of the reliable-adoption guarantee.

Proof of Theorem 2. Let $\mathcal{E}_k$ be the intersection of the coverage events on the ready branch in Lemmas 5 and 6, and set

$$\mathsf{Good}_k := \mathsf{Ready}_k^c \cup (\mathsf{Ready}_k \cap \mathcal{E}_k).$$

The two lemmas and a union bound give

$$\mathbb{P}(\mathsf{Good}_k^c) \le 2\beta_Z + \beta_{F,k} + \beta_R + 2(n_F + n_R)\beta_V^{(1)} = \beta_{\mathrm{step},k}. \tag{57}$$

Because $\mathsf{Acc}_k \subseteq \mathsf{Ready}_k$, every accepted path in Good_k lies in $\mathcal{E}_k$; in particular the retained denominator is valid and both estimator bounds apply. On such a path, Theorem 1 and the two bounds give

$$I_k \ge L_{A_k}(\tilde{A}) - (1 - Z_k)D_R(\tilde{A}; A_k) \ge \hat{L}_k - \varepsilon_L - (1 - \hat{Z}_k + \varepsilon_Z)(\hat{D}_R + \varepsilon_D).$$

The Two-Gate inequalities $\hat{L}_k \ge \tau$ and $\hat{D}_R \le \delta$ therefore imply $I_k \ge \Delta_k$. Hence $\mathsf{Acc}_k \cap \{I_k < \Delta_k\} \subseteq \mathsf{Good}_k^c$, which proves the quantitative bound. The runtime condition $\Delta_k > 0$ gives the stated non-degradation consequence.

C.3 Generation procedures and pool selection

Membership in $\mathcal{P}_k(T)$ is determined by true modification effects under $\mathcal{D}_{\mathrm{user}}$ and the declared target. At fixed generation law, reachability is non-increasing in λ and non-decreasing in δ.

A generation procedure fixes how failure material, history $\mathcal{H}_k$, shared context $\Xi_k = \xi$, and its predeclared internal steps produce one terminal output. It induces $\pi_{k,\xi}$ on $\mathcal{M}_\perp$, including failures. Eq. (19) supplies the current agent's kernel on a given encoded failure batch; collecting that batch is included when specified by the procedure.

Definition 10 (Selection from a candidate pool). *Conditional on $(\mathcal{H}_k, \Xi_k = \xi)$, draw $C_{1:N} \overset{\mathrm{iid}}{\sim} \pi_{k,\xi}$ on $\mathcal{M}_\perp$. A selection rule may use arbitrary evaluation material to choose a valid drawn modification, or retain the current agent, but may not repair, compose, or generate a new modification. It succeeds for T only when the selected modification C_i from the candidate pool lies in $\mathcal{P}_k(T)$; retaining the current agent is not counted. The pool size N includes failed terminal outputs.*

Remark 7. The pool-selection bound permits arbitrary evaluation of the realized pool. Fixed modification-wise repair induces a new distribution (Proposition 15); finite predeclared assembly is covered by Theorem 8 (both in Appendix C.6).

Pool-selection upper bound. Fix a target T and $(\mathcal{H}_k, \Xi_k = \xi)$, and draw $C_{1:N} \overset{\mathrm{iid}}{\sim} \pi_{k,\xi}$. For a selection rule of Definition 10, let $R_{k,N}(T \mid \mathcal{H}_k, \xi)$ be the conditional probability of selecting a modification in $\mathcal{P}_k(T)$. Write

$$\mathsf{C}_N(T \mid \mathcal{H}_k, \xi) := 1 - \big(1 - P_k(T \mid \mathcal{H}_k, \xi)\big)^N.$$

This is the probability that the candidate pool contains at least one qualified modification. When the conditioning is fixed, we write $\mathsf{C}_N(T)$.

Theorem 7 (Candidate-pool selection bound). *Fix a target T and $(\mathcal{H}_k, \Xi_k = \xi)$, and let $C_{1:N} \overset{\mathrm{iid}}{\sim} \pi_{k,\xi}$. Any selector of Definition 10 satisfies*

$$R_{k,N}(T \mid \mathcal{H}_k, \xi) \le 1 - \big(1 - P_k(T \mid \mathcal{H}_k, \xi)\big)^N \le N P_k(T \mid \mathcal{H}_k, \xi). \tag{58}$$

An oracle that identifies a qualified modification whenever one is present attains C_N. No downstream rule restricted to choosing from the candidate pool can exceed this probability.

Proof of Theorem 7. Conditional on $(\mathcal{H}_k, \Xi_k = \xi)$, the selector can choose a modification from the candidate pool that belongs to $\mathcal{P}_k(T)$ only if at least one of the N i.i.d. terminal outputs lies in that set. The probability of this necessary event is $1 - (1 - P_k(T \mid \mathcal{H}_k, \xi))^N$. Evaluation changes only the choice within the candidate pool. Bernoulli's inequality gives the second bound.

Proposition 12. *In the kernel setting of Appendix A.1, allow the composition map $(M, C) \mapsto A$ to be any measurable map consistent with Eqs.* (15)–(19)*. For every fixed integer $N \geq 1$, there are a common target T and two configurations with the same frozen LLM, identical behavior on user tasks, and the same qualified set $\mathcal{P}_k(T)$, but whose generation procedures assign different probabilities to this set and hence yield different values of $\mathsf{C}_N(T)$.*

Proof of Proposition 12. Let two harness states differ only in their self-task fragment, and let their allowed modifications preserve that fragment. Choose the composition maps so that both current agents' user-task kernels, and both modified kernels for each common modification, agree. Their qualified sets consequently agree.

To realize different probabilities under the composition maps allowed in Proposition 12, take two user tasks t_0, t_1 of equal probability, binary user-task outputs, reward $r(t, o) = o$, and detector $\phi(t, o) = 1 - o$. Both current agents output 0 on t_0 and 1 on t_1. Let $\mathcal{M} = \{c_+, c_0\}$, where c_+ changes the output on t_0 to 1 and leaves t_1 unchanged, while c_0 preserves both outputs. These modifications leave the self-task fragment unchanged. Then $Z = 1/2$, $(L(c_+), D(c_+)) = (1/2, 0)$, and $(L(c_0), D(c_0)) = (0, 0)$. For the common target $T = (1/4, 0)$, both qualified sets are $\{c_+\}$. Choose the two self-task kernels to assign probabilities p and p' to c_+, where $0 < p' < p < 1$. These choices define measurable kernels with the same frozen M and identical user-task behavior. Since $p \mapsto 1 - (1 - p)^N$ is strictly increasing for $N \geq 1$, the two values of $\mathsf{C}_N(T)$ differ.

C.4 Validation and adoption from a candidate pool

For each candidate-pool output C_i, write $L_i := L_k(C_i)$ and $D_i := D_k(C_i)$, with estimates $\widehat{L}_i, \widehat{D}_i$. Nonnegative error bounds $e^{\ell}_{L,i}, e^{u}_{L,i}, e^{\ell}_{D,i}, e^{u}_{D,i}$ specify the lower and upper endpoints of the intervals in Eq. (60). The candidate-pool validation rule G in Section 3.1 is

$$G_i(T) = 1 \quad \Longleftrightarrow \quad \widehat{L}_i - e^{\ell}_{L,i} \geq \lambda, \qquad \widehat{D}_i + e^{u}_{D,i} \leq \delta. \tag{59}$$

The following restates Theorem 3 with its complete conditions. For a failed output $C_i = \perp$, use $L_i = D_i = \widehat{L}_i = \widehat{D}_i = 0$ and zero error radii; only valid outputs require candidate-agent evaluation.

The intervals cover the effects defined by the fixed reward and target. Evaluation through another signal requires a calibration allowance sufficient for this coverage.

Theorem 3 (Probability of achieving harness-level reliable self-evolution). *Fix T and $(\mathcal{H}_k, \Xi_k = \xi)$. For modification C_i, write $L_i := L_k(C_i)$ and $D_i := D_k(C_i)$. Let $\mathcal{E}_{N,\beta}$ be the event that, for all $i = 1, \ldots, N$,*

$$L_i \in [\widehat{L}_i - e^{\ell}_{L,i}, \widehat{L}_i + e^{u}_{L,i}], \qquad D_i \in [\widehat{D}_i - e^{\ell}_{D,i}, \widehat{D}_i + e^{u}_{D,i}], \tag{60}$$

with, on every sample path,

$$e^{\ell}_{L,i} + e^{u}_{L,i} \leq \bar{w}_L, \qquad e^{\ell}_{D,i} + e^{u}_{D,i} \leq \bar{w}_D. \tag{61}$$

Suppose: (i) $C_{1:N} \overset{\text{iid}}{\sim} \pi_{k,\xi}$; (ii) the evaluation material used by the validation rule is separated from the generation material; (iii) the modification-independent deterministic upper bounds $\bar{w}_L, \bar{w}_D$ are fixed before the candidate pool and dominate every realized interval width as in Eq. (61)*; (iv) the event $\mathcal{E}_{N,\beta}$ satisfies*

$$\mathbb{P}(\mathcal{E}_{N,\beta} \mid \mathcal{H}_k, \xi, C_{1:N}) \geq 1 - \beta_{\text{stat}}, \tag{62}$$

and (v) σ is a selector as specified in Theorem 3. Then, on $\mathcal{E}_{N,\beta}$, every validated modification belongs to $\mathcal{P}_k(T)$ and every candidate-pool member in $\mathcal{P}^{+}_{k,N,\mathbf{n},\beta}(T)$ is validated. Consequently the lower and upper bounds in Eq. (8) *hold. For $T = (\lambda_k(\gamma, \delta), \delta)$ with $\gamma > 0$, the lower bound also applies to the probability of an update with overall expected-reward gain at least γ by Corollary 1.*

Proof of Theorem 3. Write $\mathcal{E} = \mathcal{E}_{N,\beta}$. A failed output has deterministic coverage and zero interval widths; since $\lambda > 0$, it is neither qualified nor in $\mathcal{P}^{+}_{k,N,\mathbf{n},\beta}(T)$, and Eq. (59) does not validate it. For valid outputs, on $\mathcal{E}$, if $G_i(T) = 1$, then Eq. (60) and Eq. (59) give

$$L_i \geq \widehat{L}_i - e^{\ell}_{L,i} \geq \lambda, \qquad D_i \leq \widehat{D}_i + e^{u}_{D,i} \leq \delta,$$

so every validated modification satisfies the target. Conversely, if $C_i \in \mathcal{P}^{+}_{k,N,\mathbf{n},\beta}(T)$, then

$$\widehat{L}_i - e^{\ell}_{L,i} \geq L_i - e^{u}_{L,i} - e^{\ell}_{L,i} \geq L_i - \bar{w}_L \geq \lambda,$$

and

$$\widehat{D}_i + e^{u}_{D,i} \leq D_i + e^{\ell}_{D,i} + e^{u}_{D,i} \leq D_i + \bar{w}_D \leq \delta,$$

so every candidate-pool member in $\mathcal{P}^{+}_{k,N,\mathbf{n},\beta}(T)$ is validated.

The bounds $\bar{w}_L, \bar{w}_D$ and hence $\mathcal{P}^{+}_{k,N,\mathbf{n},\beta}(T)$ are fixed before the pool draw. Let I be the event that the candidate pool contains at least one modification in this set. Conditional i.i.d. sampling gives

$$\mathbb{P}(I \mid \mathcal{H}_k, \xi) = 1 - \big(1 - P^{+}_{k}(T \mid \mathcal{H}_k, \xi)\big)^{N}.$$

On $I \cap \mathcal{E}$ the validated set is nonempty, every validated modification satisfies the target, and the selector therefore chooses a modification satisfying the target. Because Eq. (62) holds conditional on the candidate pool,

$$\mathbb{P}(I \cap \mathcal{E} \mid \mathcal{H}_k, \xi) = \mathbb{E}[\mathbf{1}_I \mathbb{P}(\mathcal{E} \mid \mathcal{H}_k, \xi, C_{1:N}) \,|\, \mathcal{H}_k, \xi] \geq (1 - \beta_{\mathrm{stat}})\mathbb{P}(I \mid \mathcal{H}_k, \xi).$$

This proves the lower bound. The upper bound is Theorem 7: the event counted by $R^{G,\sigma}$ requires a modification satisfying the target to be present in the candidate pool.

Conditionally independent evaluation bundles instantiate Eq. (62) with a single pool-wise allocation and deterministic upper bounds on interval widths of order $\sqrt{\log(N/\beta_{\mathrm{stat}})/n_{\mathrm{eff}}}$. This includes records from an independent post-reset i.i.d. task generator, or from an m-dependent stream sampled beyond its dependence range and separated from the past. A single reset before evaluation removes the initial history dependence but does not by itself make later records independent.

Corollary 4. *Suppose the reference experiment satisfies Theorem 3. Let S_T be the event that the selector chooses a candidate-pool modification satisfying T. If, after fixing $(\mathcal{H}_k, \xi)$, the original and reference experiments obey the conditional bound*

$$\operatorname*{ess\,sup}_{\mathcal{H}_k, \xi} \|\mathbb{P}(\cdot \mid \mathcal{H}_k, \xi) - \widetilde{\mathbb{P}}(\cdot \mid \mathcal{H}_k, \xi)\|_{\mathrm{TV}} \leq \beta_{\mathrm{mix}},$$

then the conditional conclusion is

$$R^{G,\sigma}_{k,N}(T \mid \mathcal{H}_k, \xi) \geq \big[(1 - \beta_{\mathrm{stat}})\{1 - (1 - P^{+}_{k}(T \mid \mathcal{H}_k, \xi))^{N}\} - \beta_{\mathrm{mix}}\big]_{+}. \tag{63}$$

If instead the joint laws of history, context, candidate pool, all evaluation records and marks, and selector randomness have the same history/context marginal and are within total variation β_{mix}, then the marginal bound is

$$\mathbb{P}(\mathsf{S}_T) \geq \Big[(1 - \beta_{\mathrm{stat}})\mathbb{E}_{\mathcal{H}_k, \Xi_k}\big[\mathsf{C}^{+}_{k,N}(T \mid \mathcal{H}_k, \Xi_k)\big] - \beta_{\mathrm{mix}}\Big]_{+}. \tag{64}$$

For an ordered stationary β-mixing evaluation record, let g_0 be the gap from the pre-evaluation information and $g_1, \ldots, g_{q-1}$ the successive gaps. Under the information and marking conditions of Remark 3, sequential comparison with the product reference law gives

$$\beta_{\mathrm{mix}} \leq \beta(g_0) + \sum_{j=1}^{q-1} \beta(g_j). \tag{65}$$

Thus q records separated from the past and one another by at least ℓ give $\beta_{\mathrm{mix}} \leq q\beta(\ell)$. A reset independent of the pre-evaluation information removes $\beta(g_0)$ only; the remaining record-to-record terms still apply unless the post-reset generator supplies independent records. The count q includes all evaluation records and may depend on the candidate count.

Proof of Corollary 4. Apply Theorem 3 under the reference law. For fixed $(\mathcal{H}_k, \xi)$, conditional total-variation control transfers the selection event with loss at most β_{mix}, so the reference lower bound loses this term and gives Eq. (63). Under joint total-variation control, first average the reference conditional lower bound over the common history/context marginal; transferring the resulting joint event again costs at most β_{mix}, giving Eq. (64). For ordinary β-mixing, compare the joint record law with the product reference law successively at the q gaps. At each step, the common marking and evaluation kernels are Markov kernels and therefore do not increase total variation. The triangle inequality then gives the sum in Eq. (65).

Corollary 5. *Let a failure-rate estimate completed before the candidate pool and separated from generation produce $W = (\underline{Z}_k, \overline{Z}_k)$. Set $U = (\mathcal{H}_k, \Xi_k, W)$. Conditional on U, the candidate pool has law $\pi_{k,\Xi_k}^{\otimes N}$. Evaluation-width bounds may depend on W, are fixed before the pool, and dominate every realized width. Let $\mathcal{E}_Z := \{\underline{Z}_k \le Z_k \le \overline{Z}_k\}$ satisfy the marginal bound $\mathbb{P}(\mathcal{E}_Z) \ge 1 - \beta_Z$, and let $\bar{w}_Z$ be a deterministic pre-pool bound on $\overline{Z}_k - \underline{Z}_k$ on every path. For required gain γ and retained-task bound δ, use*

$$\lambda_k^{\mathrm{obs}}(\gamma, \delta) := \gamma + (1 - \underline{Z}_k)\delta. \tag{66}$$

On $\mathcal{E}_Z \cap \mathcal{E}_{N,\beta}$, validation at $(\lambda_k^{\mathrm{obs}}, \delta)$ implies an expected-reward gain of at least γ. Relative to the unknown $\lambda_k(\gamma, \delta)$, a sufficient set is

$$\{c : L_k(c) \ge \lambda_k(\gamma, \delta) + \bar{w}_Z\delta + \bar{w}_L,\ D_k(c) \le \delta - \bar{w}_D\}.$$

Let $\mathsf{C}^+_{N,\mathrm{svc}}(U)$ denote the probability that the candidate pool contains a modification in this displayed sufficient set, conditional on U. For dependent evaluation, let $\mathbb{P}$ be the actual law and $\widetilde{\mathbb{P}}$ a reference law of the complete joint object consisting of U, the candidate pool, all evaluation records and marks, and selector randomness. Require the two laws to have the same U marginal and satisfy $\|\mathbb{P} - \widetilde{\mathbb{P}}\|_{\mathrm{TV}} \le \beta_{\mathrm{mix}}$. Under $\widetilde{\mathbb{P}}$, conditional on U the pool has law $\pi_{k,\Xi_k}^{\otimes N}$, and the event $\mathcal{E}_{N,\beta}$ of Theorem 3, on which all intervals contain their corresponding true values, has conditional probability at least $1 - \beta_{\mathrm{stat}}$ given U and the pool. The exact conditional implementation takes $\widetilde{\mathbb{P}} = \mathbb{P}$ and $\beta_{\mathrm{mix}} = 0$. Let S_γ be the event that the selected candidate-pool modification has expected-reward gain at least γ. Then the marginal guarantee is

$$\mathbb{P}(\mathsf{S}_\gamma) \ge \Big[(1 - \beta_{\mathrm{stat}})\mathbb{E}\big[\mathsf{C}^+_{N,\mathrm{svc}}(U)\big] - \beta_Z - \beta_{\mathrm{mix}}\Big]_+, \tag{67}$$

where the expectation is over the common marginal of U.

For an ordinary β-mixing construction of this joint comparison, W belongs to the pre-evaluation information, the initial gap in Eq. (65) starts after that information, and the candidate pool retains its stated conditional law given U. The joint comparison gives the marginal guarantee in Eq. (67); a conditional guarantee under the actual law requires conditional comparison. Pilot uncertainty contributes both the margin $\bar{w}_Z\delta$ and the marginal failure allowance β_Z. If domination by $\bar{w}_Z$ is itself only a $1 - \beta_{\mathrm{env}}$ event, intersect that event and subtract β_{env} once.

Proof of Corollary 5. On $\mathcal{E}_Z \cap \mathcal{E}_{N,\beta}$, any modification validated at λ_k^{obs} satisfies

$$J_{\mathrm{user}}(A_k \oplus c) - J_{\mathrm{user}}(A_k) \ge L_k(c) - (1 - Z_k)D_k(c) \ge \gamma + (Z_k - \underline{Z}_k)\delta \ge \gamma.$$

Moreover $\lambda_k^{\mathrm{obs}} - \lambda_k(\gamma, \delta) = (Z_k - \underline{Z}_k)\delta \le \bar{w}_Z\delta$. Adding the pre-pool interval-width bounds gives the displayed sufficient set. Under $\widetilde{\mathbb{P}}$, conditional on each U in $\mathcal{E}_Z$, the sufficient set is fixed before the pool draw, so Theorem 3 and the conditional probability bound on $\mathcal{E}_{N,\beta}$ give

$$\widetilde{\mathbb{P}}(\mathsf{S}_\gamma \cap \mathcal{E}_Z) \ge (1 - \beta_{\mathrm{stat}})\widetilde{\mathbb{E}}[\mathsf{C}^+_{N,\mathrm{svc}}(U)\mathbf{1}_{\mathcal{E}_Z}] \ge (1 - \beta_{\mathrm{stat}})\big(\mathbb{E}_{\mathbb{P}}\mathsf{C}^+_{N,\mathrm{svc}}(U) - \beta_Z\big),$$

because $0 \le \mathsf{C}^+_{N,\mathrm{svc}} \le 1$, $\mathbb{P}(\mathcal{E}_Z^c) \le \beta_Z$, and the two laws have the same U marginal. Transfer the single joint event $\mathsf{S}_\gamma \cap \mathcal{E}_Z$ once:

$$\mathbb{P}(\mathsf{S}_\gamma) \ge \mathbb{P}(\mathsf{S}_\gamma \cap \mathcal{E}_Z) \ge \widetilde{\mathbb{P}}(\mathsf{S}_\gamma \cap \mathcal{E}_Z) - \beta_{\mathrm{mix}}.$$

This gives Eq. (67), whose displayed subtraction of β_Z is looser than $(1 - \beta_{\mathrm{stat}})\beta_Z$. A separate event for an uncertain width envelope is included in the same reference event and allocated its marginal failure probability once.

Remark 8. For a fixed set, increasing N raises the probability that the pool contains a member. Maintaining the confidence guarantee across more candidates also widens the intervals through $\log N$, potentially shrinking $\mathcal{P}^+$. Hence the adoption-probability lower bound need not be monotone in N.

For context averaging, write

$$\mathsf{C}_N(T \mid \mathcal{H}_k, \xi) := 1 - \big(1 - P_k(T \mid \mathcal{H}_k, \xi)\big)^N, \qquad \mathsf{C}_N^+(T \mid \mathcal{H}_k, \xi) := 1 - \big(1 - P_k^+(T \mid \mathcal{H}_k, \xi)\big)^N. \tag{68}$$

Theorem 3 gives $(1-\beta_{\mathrm{stat}})\mathsf{C}_N^+ \le R_{k,N}^{G,\sigma} \le \mathsf{C}_N$ under its conditional-coverage assumptions. Because all generation runs in a step share the context, define

$$\overline{\mathsf{C}}_{k,N}(T) := \mathbb{E}_{\Xi_k}[\mathsf{C}_N(T \mid \mathcal{H}_k, \Xi_k) \mid \mathcal{H}_k],$$

and $\bar{P}_k(T) := \mathbb{E}_{\Xi_k}[P_k(T \mid \mathcal{H}_k, \Xi_k) \mid \mathcal{H}_k]$.

Corollary 6. *For every $N \ge 1$, conditional on $\mathcal{H}_k$,*

$$\overline{\mathsf{C}}_{k,N}(T) \le 1 - \big(1 - \bar{P}_k(T)\big)^N.$$

Substituting context-averaged reachability into $1-(1-p)^N$ gives an upper bound on the probability that the pool contains a qualified modification; equality need not hold.

Proof of Corollary 6. The candidate-pool event in Theorem 7, averaged over the shared context Ξ_k, has probability $\overline{\mathsf{C}}_{k,N}(T)$. Since $f(p) = 1-(1-p)^N$ is concave for $N \ge 1$, Jensen gives

$$\mathbb{E}[f(P_k(T \mid \Xi_k))] \le f(\mathbb{E}[P_k(T \mid \Xi_k)]),$$

which is the stated upper bound.

If the width bounds tend to zero and $\pi_{k,\xi}\{c : L_k(c) = \lambda \text{ or } D_k(c) = \delta\} = 0$, then $P_k^+(T) \to P_k(T)$ by bounded convergence of the target indicators.

Proposition 13. *Under Theorem 3, suppose $g_T := \lambda - (1-Z_k)\delta > 0$. Then*

$$\mathbb{E}[J_{\mathrm{user}}(A_{k+1}) - J_{\mathrm{user}}(A_k) \mid \mathcal{H}_k, \xi] \ge (1-\beta_{\mathrm{stat}})\mathsf{C}_N^+(T \mid \mathcal{H}_k, \xi)\, g_T - \beta_{\mathrm{stat}}. \tag{69}$$

Eq. (69) bounds expected improvement by combining the gain guaranteed when the pool contains a member of $\mathcal{P}^+$ and the evaluation intervals contain the true values, with the possible loss when these intervals fail to do so.

Proof of Proposition 13. Let I be the event that the candidate pool contains a modification in $\mathcal{P}^+_{k,N,\mathbf{n},\beta}(T)$ and write $\mathcal{E} = \mathcal{E}_{N,\beta}$. On $I \cap \mathcal{E}$, Theorem 3 selects a modification satisfying the exact target, whose expected-reward gain is at least g_T by the lower bound Eq. (3). On $\mathcal{E} \setminus I$, the gain is nonnegative by the validation guarantee and Definition 9. On $\mathcal{E}^c$, bounded rewards give gain at least -1. Finally $\mathbb{P}(I \cap \mathcal{E}) \ge (1-\beta_{\mathrm{stat}})\mathsf{C}_N^+$ and $\mathbb{P}(\mathcal{E}^c) \le \beta_{\mathrm{stat}}$, proving Eq. (69).

Pool-wise confidence allocation. Using one event $\mathcal{E}_{N,\beta}$ for all intervals in the realized candidate pool places the effect of statistical error in the interval widths and hence in P_k^+. The finite-data bounds for adopting qualified modifications in Theorem 3 therefore count familywise error once while preserving the target.

Corollary 7. *In the setting of Theorem 3, fix N, β_{stat}, $c > 0$, and $n_{\mathrm{eff}} > 0$ before drawing the candidate pool. Suppose that, conditional on the pre-pool variables and the pool, each of the four endpoint failures for every valid output admits the common bound $2\exp(-n_{\mathrm{eff}}e^2/c)$. Allocating $\beta_{\mathrm{stat}}/(4N)$ to each endpoint yields Eq. (62) by choosing the common endpoint radius*

$$r_N \;:=\; \sqrt{\frac{c}{n_{\mathrm{eff}}}\log\frac{8N}{\beta_{\mathrm{stat}}}} \;=\; O\left(\sqrt{\frac{\log(N/\beta_{\mathrm{stat}})}{n_{\mathrm{eff}}}}\right). \tag{70}$$

For valid outputs, the corresponding symmetric intervals have total width $2r_N$; failed outputs retain zero radii. Thus one may take the deterministic width bounds $\bar{w}_L = \bar{w}_D = 2r_N$.

Proof of Corollary 7. At radius r_N, the endpoint tail bound gives failure probability at most

$$2\exp(-n_{\text{eff}} r_N^2/c) = \frac{\beta_{\text{stat}}}{4N}$$

for each endpoint. A union bound over at most $4N$ endpoint events therefore gives total failure probability at most β_{stat}. Substituting Eq. (70) into this union bound yields Eq. (62).

For a fixed interval width, Eq. (70) requires only logarithmic growth in effective sample size *per candidate* as N grows; total evaluation cost also scales with the number of candidates. At a fixed total budget, fewer samples per candidate and the familywise $\log N$ term can shrink P_k^+, so the adoption-probability lower bound need not increase with N. Corollary 4 gives the dependent-evaluation transfer for this pool event.

C.5 Paired evaluation

This appendix analyzes paired evaluation and fixed-budget empirical-Bernstein bounds as options for evaluation efficiency. The guarantees in the main text use the radii specified with their protocols.

Paired sampling.

Definition 11 (Paired rollout). *For a task t, a paired rollout is a pair $(o^{A_k}, o^{\tilde{A}})$ drawn from a coupling $\Gamma_t \in \Delta(\mathcal{O} \times \mathcal{O})$ with the following properties.*

***(i) Marginal correctness**: Its marginals are $A_k(\cdot \mid t)$ and $\tilde{A}(\cdot \mid t)$.*

***(ii) Coupling**: Within a pair, the external randomness source is shared (task instance state, external API responses, environment randomness); LLM decoding is independent.*

An evaluation with m paired rollouts consists of independent repetitions $(o_l^{A_k}, o_l^{\tilde{A}}) \overset{\text{i.i.d.}}{\sim} \Gamma_t$, $l = 1, \ldots, m$.

The paired estimator of the advantage is

$$\hat{\delta}^{\text{paired}}(t) := \frac{1}{m}\sum_{l=1}^{m}\big[r(t, o_l^{\tilde{A}}) - r(t, o_l^{A_k})\big]. \tag{71}$$

Positive within-pair reward covariance reduces variance relative to independent rollouts. Shared external state must give the marginal kernels in Definition 11.

Variance.

Proposition 14. *Let $R_A = r(t, o^{A_k})$ and $R_B = r(t, o^{\tilde{A}})$ under the coupling Γ_t, with variances $\sigma^2_{A_k,t}$ and $\sigma^2_{\tilde{A},t}$. Then*

$$\text{Var}[\hat{\delta}^{\text{paired}}(t)] = \frac{\sigma^2_{A_k,t} + \sigma^2_{\tilde{A},t} - 2\,\text{Cov}_{\Gamma_t}(R_A, R_B)}{m}. \tag{72}$$

If both variances are positive, write $\rho_t = \text{Corr}_{\Gamma_t}(R_A, R_B)$, so the covariance is $\rho_t \sigma_{A_k,t}\sigma_{\tilde{A},t}$.

When both variances are positive, the variance ratio relative to independent sampling is $1 - 2\rho_t\sigma_{A_k,t}\sigma_{\tilde{A},t}/(\sigma^2_{A_k,t} + \sigma^2_{\tilde{A},t})$.

Proof of Proposition 14. The independent pairs give

$$\text{Var}[\hat{\delta}^{\text{paired}}(t)] = \frac{1}{m}\,\text{Var}(R_B - R_A) = \frac{\sigma^2_{A_k,t} + \sigma^2_{\tilde{A},t} - 2\,\text{Cov}_{\Gamma_t}(R_A, R_B)}{m},$$

which is Eq. (72). Dividing by the variance under independent rollouts gives the stated ratio.

Empirical-Bernstein bounds. Let $m \geq 2$ and let $X_1, \ldots, X_m$ be independent random variables in $[a, a+R]$. Write $\bar{X} = m^{-1}\sum_l X_l$ and $\hat{\sigma}^2 = (m-1)^{-1}\sum_l (X_l - \bar{X})^2$. Applying the independent-variable empirical-Bernstein bound of Maurer & Pontil (2009) to both directions, with error probability $\beta/2$ each, gives

$$\mathbb{P}\left[|\bar{X} - \mathbb{E}\bar{X}| > \sqrt{\frac{2\hat{\sigma}^2 \log(4/\beta)}{m}} + \frac{7R\log(4/\beta)}{3(m-1)}\right] \leq \beta. \tag{73}$$

Applied to the paired advantage estimator $\hat{\delta}^{\mathrm{paired}}(t)$ (each term $\in [-1,1]$, $R = 2$):

Lemma 7. *For $m \geq 2$, let $\delta(t) = \mathrm{Adv}_{A_k}(\tilde{A}, t)$ and let $\hat{\sigma}^2_{\delta,t}$ be the sample variance with denominator $m-1$ of $\{r(t, o_l^{\tilde{A}}) - r(t, o_l^{A_k})\}_{l=1}^m$. Then with probability $\geq 1 - \beta_\delta^{(1)}$:*

$$|\hat{\delta}^{\mathrm{paired}}(t) - \delta(t)| \leq \varepsilon_\delta(t) := \sqrt{\frac{2\hat{\sigma}^2_{\delta,t}\log(4/\beta_\delta^{(1)})}{m}} + \frac{14\log(4/\beta_\delta^{(1)})}{3(m-1)}. \tag{74}$$

The empirical variance in Eq. (74) gives task-specific radii: at small sample variance the additive $O(\log(1/\beta)/m)$ term dominates, while larger variance contributes a square-root term.

C.6 Repair and finite assembly

Proposition 15. *Fix $(\mathcal{H}_k, \Xi_k = \xi)$ and target T. Let g be a Markov kernel from $\mathcal{M}_\perp$ to $\mathcal{M}_\perp$, fixed before the candidate pool is drawn. Draw $C_{1:N} \overset{\mathrm{iid}}{\sim} \pi_{k,\xi}$ and, conditionally independently given the pool, draw $Y_i \sim g(\cdot \mid C_i)$. Then $Y_{1:N}$ are i.i.d. from the induced distribution $\pi_{k,\xi}g$. Writing*

$$P_k^g(T \mid \mathcal{H}_k, \xi) := (\pi_{k,\xi}g)(\mathcal{P}_k(T)), \qquad R_{k,N}^g(T \mid \mathcal{H}_k, \xi) \leq 1 - \big(1 - P_k^g(T \mid \mathcal{H}_k, \xi)\big)^N, \tag{75}$$

the bound holds for every rule that selects among $Y_{1:N}$ or retains the current agent, with equality for an oracle restricted to selecting from the repaired candidate pool. The kernel may contain finite predeclared tests, repair steps, and a stopping rule, but each invocation must return one terminal output and different invocations must be conditionally independent.

Each invocation of g receives only its own drawn candidate and uses independent local randomness. The comparison does not impose an ordering between P_k^g and P_k.

Proof of Proposition 15. For every measurable $B \subseteq \mathcal{M}_\perp$,

$$\mathbb{P}(Y_i \in B \mid \mathcal{H}_k, \Xi_k = \xi) = \int_{\mathcal{M}_\perp} g(B \mid c)\, \pi_{k,\xi}(dc) = (\pi_{k,\xi}g)(B).$$

The conditional independence of the kernel calls and the i.i.d. base draws make $Y_{1:N}$ i.i.d. from $\pi_{k,\xi}g$. A selector succeeds only if the repaired pool contains some $Y_i \in \mathcal{P}_k(T)$; this event has probability $1 - (1 - P_k^g)^N$. An oracle restricted to that pool attains it, proving Eq. (75).

Assembly.

Definition 12 (Assembly family). *Fix $(\mathcal{H}_k, \Xi_k = \xi)$ and target T, and abbreviate $\pi = \pi_{k,\xi}$ and $\mathcal{P} = \mathcal{P}_k(T)$. Let $A < \infty$ and, for each $a \in \{1, \ldots, A\}$, let $\mathcal{G}_a$ be a finite, possibly empty family of Markov kernels from $\mathcal{M}_\perp^a$ to $\mathcal{M}_\perp$, fixed before the candidate pool. The active arities are $\mathcal{A}_\mathcal{G} := \{a \in \{1, \ldots, A\} : \mathcal{G}_a \neq \varnothing\}$. Require $1 \in \mathcal{A}_\mathcal{G}$ and $\mathrm{id} \in \mathcal{G}_1$.*

Draw $C_{1:N} \overset{\mathrm{iid}}{\sim} \pi$. For each $a \in \mathcal{A}_\mathcal{G}$, each ordered tuple $I = (i_1, \ldots, i_a)$ of distinct indices, and each $g \in \mathcal{G}_a$, instantiate one terminal output $Y_{g,I} \sim g(\cdot \mid C_{i_1}, \ldots, C_{i_a})$, using invocation-local randomness that is independent across (g, I) conditionally on the base pool.

An assembly-and-selection rule may select a valid modification among these outputs and the identity outputs, or retain the current agent, but may not redraw from π or alter a kernel after observing the candidate pool. Each invocation receives only its designated tuple from the base pool. A kernel may contain finite predeclared tuple-local tests, feedback, repair steps, and a stopping rule, provided it returns one terminal output in $\mathcal{M}_\perp$ for every input tuple. Internal drafts and feedback remain within that invocation; $\perp$ is not required to be absorbing under repair.

For each $a \in \mathcal{A}_{\mathcal{G}}$, the fixed kernel family and base-pool law define

$$Q_a := \max_{g \in \mathcal{G}_a} (\pi^{\otimes a} g)(\mathcal{P}).$$

For deterministic g this is $\pi^{\otimes a}(g^{-1}(\mathcal{P}))$; when $\mathcal{G}_1 = \{\mathrm{id}\}$, $Q_1 = P_k(T \mid \mathcal{H}_k, \xi)$.

Theorem 8. *Let $R_N^{\mathcal{G}}(\mathcal{P})$ denote the conditional probability that an assembly-and-selection rule of Definition 12 selects an assembled output derived from the candidate pool that belongs to $\mathcal{P}$. Then*

$$R_N^{\mathcal{G}}(\mathcal{P}) \le \min\left\{ 1, |\mathcal{G}_1|\big[1-(1-Q_1)^N\big] + \sum_{\substack{a \in \mathcal{A}_{\mathcal{G}} \\ 2 \le a \le N}} (N)_a |\mathcal{G}_a| Q_a \right\}, \tag{76}$$

$$(N)_a := \frac{N!}{(N-a)!}.$$

When $\mathcal{G}_1 = \{\mathrm{id}\}$ and $A = 1$, this is Theorem 7.

Remark 9. The bound can reach one when many tuples or kernels are instantiated. Estimating terminal-output probabilities Q_a and exploiting dependence among assembled outputs are directions discussed in Appendix H.

Proof of Theorem 8. Fix $g \in \mathcal{G}_a$ and an ordered tuple I of distinct indices. By Definition 12,

$$\mathbb{P}(Y_{g,I} \in \mathcal{P} \mid \mathcal{H}_k, \Xi_k = \xi) = (\pi^{\otimes a} g)(\mathcal{P}) \le Q_a.$$

For a fixed $g \in \mathcal{G}_1$, distinct base inputs are i.i.d. and the invocation-local randomness is independent, so the N outputs $Y_{g,(i)}$ are i.i.d. from πg. The probability that at least one of these outputs belongs to $\mathcal{P}$ is at most $1-(1-Q_1)^N$. A union bound over the finite family $\mathcal{G}_1$ gives the first term of Eq. (76). For each active $a \ge 2$ with $a \le N$, there are $(N)_a$ ordered tuples of distinct indices and $|\mathcal{G}_a|$ kernels. A union bound over these output events gives $(N)_a|\mathcal{G}_a|Q_a$ without requiring independence across overlapping tuples. Selecting an assembled output from the candidate pool that belongs to $\mathcal{P}$ requires at least one of the counted events. Summing over active arities and truncating at one proves Eq. (76).

C.7 Endpoint tails at a fixed retained-task bound

Proposition 16. *Fix the history, context, and retained-task bound δ, and draw $C \sim \pi_{k,\xi}$. Write $L = L_k(C)$, $D = D_k(C)$, $\bar{G}_{k,\delta}(x) := \mathbb{P}(L \ge x, D \le \delta)$, and let $s_{k,\delta}$ be the essential upper endpoint of L on $\{D \le \delta\}$. Suppose $f(u) := \bar{G}_{k,\delta}(s_{k,\delta} - u)$ is regularly varying at zero with index $\vartheta_{k,\delta} > 0$. Then $f(u) = u^{\vartheta_{k,\delta}} m_{k,\delta}(u) = u^{\vartheta_{k,\delta}+o(1)}$ for a function $m_{k,\delta}$ slowly varying at zero. For positive targets, if $s_{k,\delta} > 0$, then as $0 < \lambda \uparrow s_{k,\delta}$,*

$$P_k((\lambda, \delta)) = (s_{k,\delta} - \lambda)^{\vartheta_{k,\delta}} m_{k,\delta}(s_{k,\delta} - \lambda) = (s_{k,\delta} - \lambda)^{\vartheta_{k,\delta}+o(1)}. \tag{77}$$

Proof of Proposition 16. By regular variation, $m_{k,\delta}(u) = f(u)/u^{\vartheta_{k,\delta}}$ is slowly varying. For slowly varying functions, de Haan & Ferreira (2006) give $\log m_{k,\delta}(u)/\log u \to 0$, yielding the second representation. For $0 < \lambda < s_{k,\delta}$, failed outputs have $L_k(\perp) = 0 < \lambda$, so $P_k((\lambda, \delta)) = \bar{G}_{k,\delta}(\lambda) = f(s_{k,\delta} - \lambda)$. This proves Eq. (77).

The endpoint obeys $s_{k,\delta} \le Z_k \bar{a}_k \le 1 - J_{\mathrm{user}}(A_k)$ by Eq. (80). It can be below the pointwise bound when generation assigns no mass near the best feasible modifications. The endpoint and tail index belong to the specified state, history, context, and retained-task bound.

D Performance Ceiling and Verification

D.1 Remaining improvement and positive reachability

Lemma 8. *Write*

$$\bar{a}_k \;:=\; 1 - \mathbb{E}_{t \sim \mathcal{D}_{F,k}}\big[V(A_k, t)\big] \;\in\; [0,1]. \tag{78}$$

Under Assumptions 1 and 2, (i) the following decomposition and bound hold:

$$1 - J_{user}(A_k) \;=\; Z_k\,\bar{a}_k \;+\; (1 - Z_k)\big(1 - \mathbb{E}_{\mathcal{D}_{R,k}}[V(A_k, \cdot)]\big) \;\geq\; Z_k\,\bar{a}_k. \tag{79}$$

(ii) For every candidate agent $\tilde{A}$, writing $\mu_F = \mathbb{E}_{\mathcal{D}_{F,k}}[\mathrm{Adv}_{A_k}(\tilde{A}, t)]$,

$$L_{A_k}(\tilde{A}) \;=\; Z_k \mu_F \;\leq\; Z_k\,\bar{a}_k \;\leq\; 1 - J_{user}(A_k). \tag{80}$$

Proof of Lemma 8. Apply Lemma 1 to $g = 1 - V(A_k, \cdot)$ and use Eq. (21) and Eq. (78) to obtain Eq. (79). Assumption 1 gives $\mathrm{Adv}_{A_k}(\tilde{A}, t) \leq 1 - V(A_k, t)$ pointwise. Taking the expectation under $\mathcal{D}_{F,k}$ yields $\mu_F \leq \bar{a}_k$, and hence Eq. (80).

Corollary 8 (Necessary condition for positive reachability). *Under Assumptions 1 and 2, if $P_k((\lambda, \delta) \mid \mathcal{H}_k, \xi) > 0$, then*

$$Z_k \bar{a}_k \geq \lambda, \qquad J_{\mathrm{user}}(A_k) \leq 1 - \lambda.$$

For the calibrated target in Corollary 1, the second inequality implies $J_{\mathrm{user}}(A_k) \leq 1 - \gamma - (1 - Z_k)\delta$.

Proof of Corollary 8. Positive reachability implies that $\mathcal{P}_k((\lambda, \delta))$ is nonempty, so some modification has $L_k(c) \geq \lambda$. Lemma 8, specifically Eq. (80), gives $\lambda \leq L_k(c) \leq Z_k \bar{a}_k \leq 1 - J_{\mathrm{user}}(A_k)$. Substituting $\lambda = \gamma + (1 - Z_k)\delta$ gives the statement for required overall gain γ.

D.2 One-sided changes and Measured-margin

Definition 13 (One-sided deviations). *For a candidate agent $\widetilde{A}$, define the average loss and gain in per-task expected reward under $\mathcal{D}_{R,k}$ by*

$$D_R^-(\widetilde{A}; A_k) := \mathbb{E}_{t \sim \mathcal{D}_{R,k}}[(V(A_k, t) - V(\widetilde{A}, t))_+],$$
$$D_R^+(\widetilde{A}; A_k) := \mathbb{E}_{t \sim \mathcal{D}_{R,k}}[(V(\widetilde{A}, t) - V(A_k, t))_+].$$

By $|x| = (x)_+ + (-x)_+$, $D_R = D_R^- + D_R^+$ and $\mathbb{E}_{\mathcal{D}_{R,k}}[\mathrm{Adv}_{A_k}] = D_R^+ - D_R^-$.

Proposition 17. *Under Assumptions 1 and 2, if $D_R^-(\widetilde{A}; A_k) \leq \delta^-$, then*

$$J_{\mathrm{user}}(\widetilde{A}) - J_{\mathrm{user}}(A_k) \geq L_{A_k}(\widetilde{A}) - (1 - Z_k)\delta^-. \tag{81}$$

Proof of Proposition 17. From Eq. (27) and $D_R^+ \geq 0$:

$$J_{\mathrm{user}}(\tilde{A}) - J_{\mathrm{user}}(A_k) = L_{A_k} + (1 - Z_k)(D_R^+ - D_R^-) \geq L_{A_k} - (1 - Z_k)D_R^-.$$

Since the proposition assumes $D_R^- \leq \delta^-$, the displayed inequality gives

$$J_{\mathrm{user}}(\tilde{A}) - J_{\mathrm{user}}(A_k) \geq L_{A_k} - (1 - Z_k)\delta^-.$$

Since $D_R = D_R^- + D_R^+$, Eq. (27) and the definition of M_{A_k} in Section 4 give

$$J_{\mathrm{user}}(\widetilde{A}) - J_{\mathrm{user}}(A_k) = M_{A_k}(\widetilde{A}) + (1 - Z_k)D_R^+(\widetilde{A}; A_k) \geq M_{A_k}(\widetilde{A}). \tag{82}$$

To estimate the mean retained-task loss, define the estimator and its error bound by

$$\hat{D}_{R,k}^- := \frac{1}{|S_R|} \sum_{j \in S_R} \big(\hat{V}(A_k, t_j^R) - \hat{V}(\tilde{A}, t_j^R)\big)_+, \qquad \varepsilon_D^- := \frac{2\varepsilon_V + \varepsilon_R + \varepsilon_Z}{1 - \hat{Z}_k}.$$

One-sided estimation. Use the sampling conditions, readiness event, and fresh-rollout coverage of Appendix C.1. For the exact loss $q^-(t) = (V(A_k, t) - V(\widetilde{A}, t))_+$, define the weighted stored-record statistic $(1 - \phi_j^R)q^-(t_j^R) \in [0, 1]$. Its reference expectation is $(1 - Z_k)D_R^-$. Apply the stored-sample concentration argument of Lemma 6 to this statistic, obtaining an event $\mathcal{E}_R^-$ with the same allocated failure probability β_R. The 1-Lipschitz property of the positive part bounds the replacement of exact rewards by fresh estimates by $2\varepsilon_V$. On $\mathcal{E}_R^- \cap \mathcal{E}_Z$ and the fresh-rollout event,

$$|(1 - \widehat{Z}_k)\widehat{D}_{R,k}^- - (1 - Z_k)D_R^-| \le 2\varepsilon_V + \varepsilon_R.$$

Since $D_R^- \le 1$, division by $1 - \widehat{Z}_k > 0$ on Ready_k gives

$$|\widehat{D}_{R,k}^- - D_R^-| \le (2\varepsilon_V + \varepsilon_R + \varepsilon_Z)/(1 - \widehat{Z}_k).$$

The reference-sample comparison and its joint-TV risk transfer are the same as in Lemma 6; only the bounded statistic changes. Measured-margin and the one-sided rule in Eq. (95) use $\mathcal{E}_R^-$ in place of the symmetric retained event, with the original total risk allocation.

Proposition 18. *Under the reward, task, and evaluation assumptions stated in Theorem 2, use Rule 2 in place of Rule 1 in the one-candidate update, with allowed retained-task decrease δ^-. Let $\mathsf{Acc}_k^M \subseteq \mathsf{Cand}_k \cap \mathsf{Ready}_k$ be the resulting adoption event and use $\widehat{M}_k$ from Eq.* (12)*.*

(i)

$$\mathbb{P}\Big(\mathsf{Acc}_k^M \cap \{I_k < \widehat{M}_k\}\Big) \le \beta_{\mathrm{step},k}.$$

Since acceptance requires $\widehat{M}_k > 0$, the joint event of adoption and nonpositive actual improvement has probability at most $\beta_{\mathrm{step},k}$.

(ii) Using the same evaluation data and per-task reward estimates, at matched retained-task bounds $\delta^- = \delta$ and matched radii, every ready sample path accepted by Eq. (4) *under Eq.* (5) *is accepted by Rule 2; the reverse inclusion does not hold in general.*

Proof of Proposition 18. Let $\mathcal{E}_k^-$ be the intersection, on Ready_k, of the one-sided retained-task event above with the failure-rate, returned failure-mean, and fresh-rollout events used in Theorem 2. Define $\mathcal{G}_k^- := \mathsf{Ready}_k^c \cup (\mathsf{Ready}_k \cap \mathcal{E}_k^-)$. The same union bound gives $\mathbb{P}((\mathcal{G}_k^-)^c) \le \beta_{\mathrm{step},k}$. On $\mathcal{G}_k^- \cap \mathsf{Acc}_k^M$, the one-sided estimate in Appendix D.2 applies because readiness includes a valid retained denominator, and the margin identity Eq. (82) gives

$$I_k \ge L_{A_k} - (1 - Z_k)D_R^- \ge \hat{L}_k - \varepsilon_L - (1 - \hat{Z}_k + \varepsilon_Z)(\hat{D}_{R,k}^- + \varepsilon_D^-) = \widehat{M}_k.$$

Thus $\mathsf{Acc}_k^M \cap \{I_k < \widehat{M}_k\} \subseteq (\mathcal{G}_k^-)^c$, proving part (i). For part (ii), no coverage event is needed: on the same ready evaluation path accepted by Eq. (4), $\hat{D}_{R,k}^- \le \hat{D}_R \le \delta = \delta^-$, and matched radii together with Eq. (5) give $\widehat{M}_k > 0$. For strictness, use the improving instance of Proposition 19 with $u \le (1 - Z)\delta$, and the sufficiently precise evaluation design of Proposition 20. On their common ready coverage event, Measured-margin accepts. If Two-Gate with its positive-gain check accepted, then

$$u/2 = L \ge \widehat{L} - \varepsilon_L \ge \tau - \varepsilon_L > (1 - \widehat{Z} + \varepsilon_Z)(\delta + \varepsilon_D) \ge (1 - Z)\delta \ge u,$$

a contradiction. The i.i.d. arrival horizons can be chosen so that readiness and coverage hold together with positive probability. This provides a matching evaluation configuration for which the inclusion is strict.

Remark 10. For a declared $\gamma > 0$, replace $\widehat{M}_k > 0$ in Rule 2 by $\widehat{M}_k \ge \gamma$, retaining its readiness and retained-task requirements. The proof of Proposition 18 gives the same joint error bound for accepting a modification whose expected-reward improvement is less than γ. The finite-run union and telescoping argument of Theorem 4, under its fixed-distribution and summed-risk conditions, then gives a lower bound $\gamma|\mathcal{A}^M|$ on that variant's adopted-step set and common good event.

On an accepted path in the common coverage event with finite positive radii, $\widehat{M}_k < L_k \le 1 - J_{\mathrm{user}}(A_k)$: the strict inequality comes from the positive retained-side error term in Eq. (12). Thus accepting under the positive-threshold variant requires $J_{\mathrm{user}}(A_k) < 1 - \gamma$ on that event.

D.3 CONSTRAINTS ON FURTHER IMPROVEMENT

Corollary 9 (Performance bound for the one-sided gain condition). *Under Assumptions 1 and 2, suppose some candidate agent $\tilde{A}$ meets the one-sided sufficient condition of Proposition 17 at allowed retained-task decrease δ^- — that is, $D_R^-(\tilde{A}; A_k) \le \delta^-$ and $L_{A_k}(\tilde{A}) > (1 - Z_k)\delta^-$. Then*

$$J_{user}(A_k) < 1 - (1 - Z_k)\delta^-. \tag{83}$$

Proof. By Eq. (80), $1 - J_{\text{user}}(A_k) \ge L_{A_k}(\tilde{A})$. The hypothesis gives $L_{A_k}(\tilde{A}) > (1 - Z_k)\delta^-$. Hence $1 - J_{\text{user}}(A_k) > (1 - Z_k)\delta^-$, and rearranging yields Eq. (83).

Write $B_k := 1 - (1 - Z_k)\delta^-$ for fixed $\delta^- > 0$. Corollary 9 implies that a current agent with $J_{\text{user}}(A_k) \ge B_k$ has no candidate satisfying $L_{A_k}(\widetilde{A}) > (1 - Z_k)\delta^-$. The following result shows that positive actual margins can nevertheless occur, and bounds them by the remaining reward shortfall.

Proposition 19 (Improvement beyond the ceiling of the sufficient condition). *Under Assumptions 1 and 2, fix $\delta^- > 0$ and write $B_k := 1 - (1 - Z_k)\delta^-$.*

(i) In the kernel setting of Appendix A.1, allow the composition map $(M, C) \mapsto A$ to be any measurable map consistent with Eqs. (15)–(19)*. For every $Z \in (0, 1)$ and*

$$0 < u \le \min\{Z/2, (1 - Z)\delta^-\},$$

there exists a two-task instance with $Z_k = Z$, $J_{\text{user}}(A_k) = 1 - u \ge B_k$, and a modification whose candidate agent satisfies

$$D_R^-(\widetilde{A}; A_k) = 0, \qquad M_{A_k}(\widetilde{A}) = L_{A_k}(\widetilde{A}) = u/2 > 0.$$

(ii) For every current agent with $0 < Z_k < 1$ and every candidate agent,

$$M_{A_k}(\widetilde{A}) \le L_{A_k}(\widetilde{A}) \le Z_k \bar{a}_k \le 1 - J_{\text{user}}(A_k).$$

Proof of Proposition 19. For part (i), construct the following instance. Let $\mathcal{T}_{\text{user}} = \{t_F, t_R\}$, $\mathcal{D}_{\text{user}} = Z\delta_{t_F} + (1 - Z)\delta_{t_R}$, and $\mathcal{O} = \{0, 1\}$ with reward $r(t, o) = o$. Set $\phi(t_F, o) = 1$ and $\phi(t_R, o) = 0$. Then $\mathcal{D}_{F,k} = \delta_{t_F}$ and $\mathcal{D}_{R,k} = \delta_{t_R}$. On t_F, take

$$A_k(\cdot \mid t_F) = \text{Bernoulli}(1 - u/Z), \qquad \widetilde{A}(\cdot \mid t_F) = \text{Bernoulli}(1 - u/(2Z)).$$

On t_R, both agents return 1 deterministically. Choose a measurable composition map in Appendix A.1 that realizes these two kernels at C_k and $C_k \oplus c$, with the LLM fixed. All kernels are measurable on these finite spaces. Direct calculation gives

$$J_{\text{user}}(A_k) = 1 - u, \qquad Z_k \bar{a}_k = u, \qquad D_R^- = D_R^+ = 0, \qquad L_{A_k} = M_{A_k} = u/2.$$

The stipulated range of u gives $J_{\text{user}}(A_k) \ge B_k$.

Part (ii) follows from $D_R^- \ge 0$ and Eq. (80).

Lemma 9. *On evaluation paths where the required samples are available, with $\hat{L}_k$ as defined in Eq.* (41)*,*

$$\hat{L}_k \in [-\hat{Z}_k, \hat{Z}_k] \qquad \textit{with probability } 1. \tag{84}$$

Consequently, on any evaluated path with $\tau > \hat{Z}_k$, no modification can satisfy $\widehat{L}_k \ge \tau$ in Eq. (4) *or be adopted under Rule 1.*

Proof of Lemma 9. By Eq. (41), $\hat{L}_k = \hat{Z}_k \cdot \frac{1}{n_F} \sum_{i=1}^{n_F} \big[\hat{V}(\tilde{A}, t_i^F) - \hat{V}(A_k, t_i^F)\big]$. Each $\hat{V}$ is an average of rewards, so $\hat{V} \in [0, 1]$ by Assumption 1 and each bracket lies in $[-1, 1]$; hence so does their average, and multiplying by $\hat{Z}_k \ge 0$ gives Eq. (84).

Let $\mathcal{E}_L$ denote the intersection of Ready_k with the events on which the failure-rate, rollout, and returned failure-mean error bounds hold in Lemma 5. On $\mathcal{E}_L$,

$$|\hat{L}_k - L_{A_k}(\tilde{A})| \le \varepsilon_L, \qquad \mathcal{E}_L \subseteq \mathsf{Ready}_k \cap \mathcal{E}_Z,$$

so $\hat{L}_k \le L_{A_k}(\tilde{A}) + \varepsilon_L \le Z_k\bar{a}_k + \varepsilon_L$. Together with Lemma 9, this gives $\hat{L}_k \le \Theta_k := \min\{\hat{Z}_k, Z_k\bar{a}_k + \varepsilon_L\}$.

Definition 14 (Admissible region). *On $\mathcal{E}_L$, define the admissible region of Two-Gate parameters by*

$$\mathcal{R} := \left\{(\tau,\delta) \in \mathbb{R}^2_{>0} : (1 - \widehat{Z}_k + \varepsilon_Z)(\delta + \varepsilon_D) + \varepsilon_L < \tau \le \Theta_k\right\}. \tag{85}$$

Theorem 9. *The admissible region $\mathcal{R}$ is nonempty if and only if $\delta_{\max} > 0$, where*

$$\delta_{\max} := \frac{\Theta_k - \varepsilon_L - \varepsilon_Z\varepsilon_D - \varepsilon_\Sigma}{1 - \widehat{Z}_k + \varepsilon_Z}. \tag{86}$$

When $\delta_{\max} > 0$, there exists a τ with $(\tau,\delta) \in \mathcal{R}$ exactly when $0 < \delta < \delta_{\max}$. The corresponding values of τ form the interval

$$\left((1 - \widehat{Z}_k + \varepsilon_Z)(\delta + \varepsilon_D) + \varepsilon_L, \Theta_k\right].$$

Proof of Theorem 9. For a fixed $\delta > 0$ the τ-interval is non-empty iff $(1 - \hat{Z}_k + \varepsilon_Z)(\delta + \varepsilon_D) + \varepsilon_L < \Theta_k$. Substituting $\varepsilon_D = \varepsilon_\Sigma/(1 - \hat{Z}_k)$ and expanding,

$$(1 - \hat{Z}_k)\delta + \varepsilon_\Sigma + \varepsilon_Z\delta + \varepsilon_Z\varepsilon_D + \varepsilon_L \;<\; \Theta_k,$$

i.e. $\delta\,(1 - \hat{Z}_k + \varepsilon_Z) < \Theta_k - \varepsilon_L - \varepsilon_Z\varepsilon_D - \varepsilon_\Sigma$, which is Eq. (86). The coefficient $(1 - \hat{Z}_k + \varepsilon_Z)$ is strictly positive, so the left side is strictly increasing in δ; hence the admissible δ form the interval $(0, \delta_{\max})$, non-empty iff $\delta_{\max} > 0$, and $\mathcal{R}$ is the union of the corresponding non-empty τ-intervals.

Corollary 10. *A necessary condition for the admissible region of Definition 14 to be nonempty is $\Theta_k > \varepsilon_\Sigma + \varepsilon_L$. Its two branches imply the following bounds.*

(i) On every sample path where the region is evaluated,

$$\varepsilon_\mu < 1 \quad \textit{and} \quad \hat{Z}_k > \frac{\varepsilon_\Sigma + \varepsilon_Z + \varepsilon_Z\varepsilon_\mu}{1 - \varepsilon_\mu}. \tag{87}$$

(ii) Within the feasible domain $2\varepsilon_V < 1$, consider a sampling design along which the returned radius ε_F can be driven to zero while the other radii are fixed. Then $\varepsilon_\mu \to 2\varepsilon_V$ and

$$\hat{Z}_k > \hat{Z}_\infty := \frac{2\varepsilon_V + \varepsilon_R + 2\varepsilon_Z + 2\varepsilon_V\varepsilon_Z}{1 - 2\varepsilon_V}. \tag{88}$$

(iii) On the concentration event $\mathcal{E}_L$ used by Definition 14,

$$Z_k\bar{a}_k > \varepsilon_\Sigma. \tag{89}$$

The corresponding expected-reward bound is Corollary 12.

Branch (i) requires the empirical statistic to exceed its finite-sample error; branch (iii) requires $Z_k\bar{a}_k$ to exceed the retained-task measurement error. The zero-failure-radius limit holds the other radii fixed. At a fixed finite arrival horizon, increasing n_F eventually makes return impossible. Appendix D.6 gives the i.i.d. reference calculation when the rollout risk allocation also varies with n_F. The allowed retained-task change enters separately in the positive-δ condition below.

Proof of Corollary 10. Since $\delta > 0$ and $1 - \widehat{Z}_k + \varepsilon_Z > 0$, dropping the positive left side and the nonnegative term $\varepsilon_Z\varepsilon_D$ from Eq. (86) gives the necessary inequality $\Theta_k > \varepsilon_\Sigma + \varepsilon_L$. For (i), substitute $\Theta_k \le \hat{Z}_k$ and expand $\varepsilon_L = \varepsilon_Z + (\hat{Z}_k + \varepsilon_Z)\varepsilon_\mu$ from Eq. (48):

$$\hat{Z}_k > \varepsilon_\Sigma + \varepsilon_Z + \hat{Z}_k\varepsilon_\mu + \varepsilon_Z\varepsilon_\mu \quad\Longleftrightarrow\quad \hat{Z}_k(1 - \varepsilon_\mu) > \varepsilon_\Sigma + \varepsilon_Z + \varepsilon_Z\varepsilon_\mu.$$

If $\varepsilon_\mu \geq 1$ the left side is non-positive while the right side is strictly positive, so no $\hat{Z}_k$ satisfies it; if $\varepsilon_\mu < 1$, dividing by $1 - \varepsilon_\mu > 0$ gives Eq. (87). For (ii), hold the other radii fixed and consider any returned-radius design along which $\varepsilon_F \downarrow 0$. The right side of Eq. (87) tends to the expression in Eq. (88), using $\varepsilon_\mu \downarrow 2\varepsilon_V$ and $\varepsilon_\Sigma + \varepsilon_Z = 2\varepsilon_V + \varepsilon_R + 2\varepsilon_Z$. Monotonicity in ε_μ makes it a lower boundary for every positive returned radius; every finite positive-radius design still obeys the strict condition Eq. (87). For (iii), substitute $\Theta_k \leq Z_k \bar{a}_k + \varepsilon_L$, cancel ε_L, and apply Eq. (79).

Corollary 2. *Under Assumptions 1 and 2, fix the threshold $\delta > 0$ on estimated retained-task change in Two-Gate. If some $\tau > 0$ satisfies $(\tau, \delta) \in \mathcal{R}$, then, on $\mathcal{E}_L$,*

$$J_{\text{user}}(A_k) < 1 - (1 - Z_k)\delta - \varepsilon_\Sigma.$$

Under these conditions, the stronger intermediate bound is

$$Z_k \bar{a}_k \;>\; (1 - \hat{Z}_k + \varepsilon_Z)\,\delta \;+\; \varepsilon_\Sigma \;+\; \varepsilon_Z \varepsilon_D. \tag{90}$$

Since $\mathcal{E}_L \subseteq \mathcal{E}_Z$ of Lemma 4, it follows that

$$Z_k\,\bar{a}_k \;>\; (1 - Z_k)\,\delta + \varepsilon_\Sigma, \qquad \text{equivalently} \qquad Z_k \;>\; \frac{\delta + \varepsilon_\Sigma}{\bar{a}_k + \delta} \;\geq\; \frac{\delta}{\bar{a}_k + \delta}, \tag{91}$$

For the one-sided rule Eq. (95), the same derivation uses δ^- together with its one-sided radius ε_D^- and the event $\mathcal{E}_R^-$; the retained-risk allocation is replaced, not split. Corollary 12 gives the corresponding necessary condition without a fixed positive retained-change requirement.

Proof of Corollary 2. By Theorem 9 a radius δ admits an accompanying τ exactly when $\delta < \delta_{\max}$, i.e. — clearing the strictly positive denominator of Eq. (86) — when $\delta\,(1 - \hat{Z}_k + \varepsilon_Z) < \Theta_k - \varepsilon_L - \varepsilon_Z \varepsilon_D - \varepsilon_\Sigma$. A minimum is at most either of its arguments, so $\Theta_k \leq Z_k \bar{a}_k + \varepsilon_L$; substituting and cancelling ε_L gives Eq. (90).

For Eq. (91), discard the non-negative $\varepsilon_Z \varepsilon_D$ and note that on $\mathcal{E}_Z$ one has $\hat{Z}_k - Z_k \leq \varepsilon_Z$, hence $1 - \hat{Z}_k + \varepsilon_Z \geq 1 - Z_k$. The rearrangement for Z_k divides by $\bar{a}_k + \delta > 0$, and the last inequality drops $\varepsilon_\Sigma \geq 0$. The term $\delta/(\bar{a}_k + \delta)$ in this finite-sample condition contains no sampling radius. For Eq. (11), identity Eq. (79) gives $1 - J_{\text{user}}(A_k) \geq Z_k \bar{a}_k$, the discarded term being non-negative by Assumption 1.

Corollary 11. *Work in the domain $2\varepsilon_V < 1$ of Eq. (88). Write $L_V := \log(2/\beta_V^{(1)})$, $L_R := \log(2/\widetilde{\beta}_R)$, $L_Z := \log(2/\widetilde{\beta}_Z)$, and $L_F := \log(2/\beta_{F,k})$.*

(a) For the stated estimator and radius formulas, the zero-failure-radius boundary Eq. (88) satisfies $\hat{Z}_\infty \geq 2\varepsilon_V + \varepsilon_R + 2\varepsilon_Z$. Hence $\hat{Z}_k > \hat{Z}_\infty$ requires

$$m > \frac{2L_V}{\hat{Z}_k^2}, \qquad n_R > \frac{L_R}{2\hat{Z}_k^2}, \qquad n_R > \frac{2L_Z}{\hat{Z}_k^2}. \tag{92}$$

(b) Whenever the returned-sample construction supplies $\varepsilon_F = \sqrt{2q_{F,k}L_F/n_F}$, solving Eq. (87) for the quota gives

$$n_F > \frac{2q_{F,k}L_F(\hat{Z}_k + \varepsilon_Z)^2}{(1 - 2\varepsilon_V)^2(\hat{Z}_k - \hat{Z}_\infty)^2}, \tag{93}$$

for $\hat{Z}_k > \hat{Z}_\infty$. The requirement diverges as $\hat{Z}_k \downarrow \hat{Z}_\infty$ and is unsatisfiable below that boundary. It is a returned-sample rate, not an arrival-time identity. If risk allocation makes ε_V depend on n_F, Eq. (93) is an implicit inequality. Under the first-failure i.i.d. construction $q_{F,k} = 1$; for a finite horizon, return is instead governed by $B_{H_{F,k},n_F}(Z_k)$.

Proof of Corollary 11. **(a)** By Eq. (88), $\hat{Z}_k > \hat{Z}_\infty = (2\varepsilon_V + \varepsilon_R + 2\varepsilon_Z + 2\varepsilon_V\varepsilon_Z)/(1 - 2\varepsilon_V) \geq 2\varepsilon_V + \varepsilon_R + 2\varepsilon_Z$, the last step because the numerator dominates $2\varepsilon_V + \varepsilon_R + 2\varepsilon_Z$ and the denominator lies in $(0, 1]$. All three summands being non-negative, each is separately $< \hat{Z}_k$. With Lemmas 3, 4, and 6: $2\varepsilon_V = 2\sqrt{L_V/(2m)} < \hat{Z}_k$ rearranges to the first display of Eq. (92), $\varepsilon_R = \sqrt{L_R/(2n_R)} < \hat{Z}_k$ to the second, $2\varepsilon_Z = 2\sqrt{L_Z/(2n_R)} < \hat{Z}_k$ to the third.

(b) The proof of Corollary 10(i) reaches $\hat{Z}_k(1-\varepsilon_\mu) > \varepsilon_\Sigma + \varepsilon_Z + \varepsilon_Z\varepsilon_\mu$. Substituting $\varepsilon_\mu = 2\varepsilon_V + \varepsilon_F$ and collecting the ε_F terms,

$$\hat{Z}_k(1-2\varepsilon_V) - \big(\varepsilon_\Sigma + \varepsilon_Z + 2\varepsilon_V\varepsilon_Z\big) \;>\; \varepsilon_F\,(\hat{Z}_k + \varepsilon_Z),$$

and the bracket equals $(1-2\varepsilon_V)\hat{Z}_\infty$ by Eq. (88), so the left side is $(1-2\varepsilon_V)(\hat{Z}_k - \hat{Z}_\infty)$. Under the sampler-specific specialization $\varepsilon_F = \sqrt{2q_{F,k}L_F/n_F}$, division by the positive $\hat{Z}_k + \varepsilon_Z$ gives Eq. (93). The symbolic bound itself requires only a valid returned radius; the displayed relation for the quota is conditional on this rate specialization. If $\hat{Z}_k \le \hat{Z}_\infty$ and $\varepsilon_F > 0$, no quota satisfies the inequality.

Corollary 12. *Under Assumptions 1 and 2, on the concentration event $\mathcal{E}_L$ used by Definition 14, if its admissible region $\mathcal{R}$ is nonempty at step k, then*

$$J_{\textit{user}}(A_k) \;<\; 1-\varepsilon_\Sigma, \qquad \varepsilon_\Sigma = 2\varepsilon_V + \varepsilon_R + \varepsilon_Z. \tag{94}$$

Proof of Corollary 12. Corollary 10 gives $\Theta_k > \varepsilon_\Sigma + \varepsilon_L$ as necessary for $\mathcal{R} \neq \varnothing$, with $\Theta_k = \min\{\hat{Z}_k,\, Z_k\bar{a}_k + \varepsilon_L\}$. A minimum is at most either argument, so $Z_k\bar{a}_k + \varepsilon_L \ge \Theta_k > \varepsilon_\Sigma + \varepsilon_L$, whence $Z_k\bar{a}_k > \varepsilon_\Sigma$ irrespective of which branch binds. Identity Eq. (79) gives $1 - J_{\text{user}}(A_k) \ge Z_k\bar{a}_k$, since the discarded term $(1-Z_k)(1-\mathbb{E}_{\mathcal{D}_{R,k}}[V])$ is non-negative by Assumption 1.

Proposition 20 (Finite-data validation with Measured-margin). *For the improving instance of Proposition 19, fix $Z_k \in (0,1)$ and the total risk allocation. One can choose m, n_F, n_R each of order $\widetilde{O}(u_k^{-2})$ such that Rule 2 validates the modification whenever the prescribed samples are ready and the joint error bounds hold. Here m is the rollout count per agent per task, n_F the returned failure-task count, and n_R the stored-record count.*

Proof of Proposition 20. Take the improving member of Proposition 19, where $M = L = u_k/2$ and both agents return reward 1 on t_R. Choose the i.i.d. returned-sample design with $\varepsilon_Z \le (1-Z_k)/2$. On Ready_k, when all the stated error bounds hold, $1 - \widehat{Z}_k \ge (1-Z_k)/2 > 0$ and $\widehat{D}^-_{R,k} = 0$. By Eq. (12) and $\widehat{L}_k \ge L - \varepsilon_L$,

$$\widehat{M}_k \ge \frac{u_k}{2} - \underbrace{2\varepsilon_L + (1 - \widehat{Z}_k + \varepsilon_Z)\varepsilon^-_D}_{\text{total error bound}}.$$

At fixed Z_k and confidence allocation, Eq. (48) gives $\varepsilon_L = \varepsilon_Z + (\widehat{Z}_k + \varepsilon_Z)(2\varepsilon_V + \varepsilon_F) \le \varepsilon_Z + 2(2\varepsilon_V + \varepsilon_F)$. The bound $1 - \widehat{Z}_k \ge (1-Z_k)/2$ gives $\varepsilon^-_D \le 2(2\varepsilon_V + \varepsilon_R + \varepsilon_Z)/(1-Z_k)$, while $1 - \widehat{Z}_k + \varepsilon_Z \le 2$. It suffices to choose the component radii so that $\varepsilon_L < u_k/8$ and $\varepsilon^-_D < u_k/8$; then the underbraced error is less than $u_k/2$. In this i.i.d. design each component radius has inverse-square-root dependence on its respective m, n_F, n_R, with the existing risk-allocation logarithms. Thus sufficiently large constant multiples of $\widetilde{O}(u_k^{-2})$ for each count meet these inequalities, giving $\widehat{M}_k > 0$ and $\widehat{D}^-_{R,k} = 0 \le \delta^-$. These counts specify returned evaluation data and rollouts per task; they do not guarantee a data-return time.

D.4 Sample-complexity lower bound for validation

Setup. An *instance* is a configuration satisfying the assumptions of §2 and the task-stream evaluation setup in Appendix C.1 — a task space with distribution $\mathcal{D}_{\text{user}}$, a failure detector, a current agent A_k, and a candidate agent $\tilde{A}$ — together with the available evaluation data: the observed task sequence, the stored reference sample with its stored labels, and fresh rollouts of either agent on any tasks, adaptively chosen, each rollout returning a reward in $[0,1]$. A **validation procedure at level** β is a measurable map from the evaluation data to $\{\mathsf{validate}, \mathsf{abstain}\}$ such that in *every* instance, $\mathbb{P}\big[\mathsf{validate} \;\wedge\; J_{\text{user}}(\tilde{A}) \le J_{\text{user}}(A_k)\big] \le \beta$. Its task choices, statistics, and stopping rule are measurable functions of these observations and its own instance-independent randomization.

Theorem 5 (Cost of validating further improvement). *Fix $Z \in (0,1)$, $\beta \le 1/8$, and $u \in (0, Z/2]$. There exist two instances with $Z_k = Z$ and $1 - J_{\text{user}}(A_k) = u$, identical in the distribution of every observable quantity except the candidate agent's reward distribution under the failure-task distribution. In one,*

$$J_{\text{user}}(\widetilde{A}) - J_{\text{user}}(A_k) = u/2, \qquad D_R(\widetilde{A}; A_k) = 0,$$

and in the other the overall reward change is $-u/2$. Any level-β validation procedure that validates the improving modification with probability at least $1/2$, under any task-selection policy and any data-dependent stopping rule, must satisfy

$$\mathbb{E}_-[N_{\text{eval}}] \ge \frac{Z}{10u},$$

where N_{eval} is the total number of candidate-agent rollouts and the expectation is under the non-improving instance. Consequently, for fixed Z, maintaining the stated error control and validation probability at least $1/2$ on the improving instance in each pair requires a worst-case expected rollout count growing at least as u^{-1}.

Proof. Construction. Two tasks, $\mathcal{T}_{\text{user}} = \{t_F, t_R\}$, with $\mathcal{D}_{\text{user}} = Z\,\delta_{t_F} + (1-Z)\,\delta_{t_R}$ and i.i.d. task records, so Assumption 4 holds with $\beta(\cdot) \equiv 0$. The detector is task-deterministic: $\phi(t_F, \cdot) \equiv 1$, $\phi(t_R, \cdot) \equiv 0$, so $\psi_A(t_F) = 1$ and $\psi_A(t_R) = 0$ for every agent, $Z_k = Z \in (0,1)$, and the conditionals are $\mathcal{D}_{F,k} = \delta_{t_F}$, $\mathcal{D}_{R,k} = \delta_{t_R}$. Set $\bar{a} := u/Z \le 1/2$. The current agent's reward distribution is δ_1 on t_R and $\text{Bernoulli}(1-\bar{a})$ on t_F, so $J_{\text{user}}(A_k) = Z(1-\bar{a}) + (1-Z) = 1 - u$. The two instances differ only in the candidate agent: on t_R it induces the current agent's reward distribution (hence $D_R^- = D_R^+ = D_R = 0$ in both), and on t_F its reward distribution is $\nu_\pm := \text{Bernoulli}\big(1 - \bar{a} \pm \frac{\bar{a}}{2}\big)$. Then $\Delta J = \pm Z\bar{a}/2 = \pm u/2$, and in the $+$ instance the candidate agent's margin is $M = L_{A_k} = u/2 > 0$ with $D_R^- = 0 \le \delta^-$ for every declared retained-task decrease bound. These finite kernels are realized by measurable agent and composition maps. Draw task records independently, keep the stored reference sample separate from candidate evaluation, and generate fresh rollouts independently from the appropriate fixed kernel. With the task-deterministic detector, both instances satisfy Assumptions 1, 2, 4, 5, 6, 7, and 8 and Condition 1 by construction.

Indistinguishability. Conditional on the same observed history and chosen evaluation action, the two instances give the same observation kernel except for candidate rollouts on t_F. The task-selection and stopping kernels of a fixed validation procedure are the same under this conditioning. For the informative rollouts, with $p_- = 1 - \frac{3\bar{a}}{2}$ and $p_+ = 1 - \frac{\bar{a}}{2}$, using $\text{KL}(\text{Ber}(p)\|\text{Ber}(q)) \le (p-q)^2/\big(q(1-q)\big)$ and $q(1-q) = (1-\frac{\bar{a}}{2})\frac{\bar{a}}{2} \ge \frac{3\bar{a}}{8}$ for $\bar{a} \le 1/2$:

$$\text{KL}(\nu_-\|\nu_+) \;\le\; \frac{\bar{a}^2}{3\bar{a}/8} \;=\; \frac{8\bar{a}}{3}.$$

For a transcript truncated after n observations, the conditional KL chain rule therefore gives

$$\text{KL}(\mathbb{P}_-^{(n)}\|\mathbb{P}_+^{(n)}) \le \mathbb{E}_-[N_F^{(n)}]\,\text{KL}(\nu_-\|\nu_+) \le \frac{8u}{3Z}\mathbb{E}_-[N_{\text{eval}}],$$

where $N_F^{(n)}$ counts candidate rollouts on t_F before stopping or truncation. Observations after stopping may be padded by a fixed symbol. Relative entropy on increasing transcript sigma-fields converges to the entropy of the full transcript; thus the same bound holds for the stopped experiment, including unbounded stopping. If the expected count is infinite the claimed cost bound already holds. Pinsker's inequality and data processing for the validation decision give

$$\text{TV}(\mathbb{P}_+, \mathbb{P}_-) \le \sqrt{4u\,\mathbb{E}_-[N_{\text{eval}}]/(3Z)}.$$

Two-point argument. In the $-$ instance, $\Delta J < 0$, so soundness forces $\mathbb{P}_-[\textsf{validate}] \le \beta$. If the validation procedure validates the improving modification with probability at least $1/2$, then

$$\tfrac{1}{2} \;\le\; \mathbb{P}_+[\textsf{validate}] \;\le\; \mathbb{P}_-[\textsf{validate}] + \text{TV}(\mathbb{P}_+, \mathbb{P}_-) \;\le\; \beta + \sqrt{4\,\mathbb{E}_-[N_{\text{eval}}]\,\bar{a}/3},$$

so $\sqrt{4\,\mathbb{E}_-[N_{\text{eval}}]\,\bar{a}/3} \ge \frac{1}{2} - \beta \ge \frac{3}{8}$, hence $\mathbb{E}_-[N_{\text{eval}}] \ge \frac{27}{256\,\bar{a}} \ge \frac{Z}{10\,u}$.

Remark 11. For candidate reward on t_F in the two constructed instances, the Bernoulli means differ by $\bar{a} = u/Z$ and their variances are at most $3\bar{a}/2$. With fixed Z, Bernstein concentration for this coordinate gives error at most $O(\sqrt{(u/Z)\log(1/\beta)/n} + \log(1/\beta)/n)$. Thus $n = O((Z/u)\log(1/\beta))$ observations suffice to resolve a fixed fraction of that gap on this coordinate.

D.5 A ONE-SIDED RULE FOR CUMULATIVE REWARD CHANGES

Use the candidate and selector restrictions, readiness conditions, and simultaneous evaluation events of Theorem 2, replacing the symmetric retained-task event by $\mathcal{E}_R^-$ from Appendix D.2. Define a rule that adopts a selected candidate only when

$$\widehat{L}_k \geq \tau, \qquad \widehat{D}_{R,k}^- \leq \delta^-, \qquad \Delta_k^- := \tau - \varepsilon_L^{(k)} - (1 - \widehat{Z}_k + \varepsilon_Z^{(k)})(\delta^- + \varepsilon_D^{-,(k)}) > 0. \tag{95}$$

The third check supplies a positive lower bound on improvement. Let Acc_k^- be this rule's adoption event. On its joint coverage event, Eq. (81) and the one-sided estimate imply

$$I_k \geq \widehat{L}_k - \varepsilon_L^{(k)} - (1 - \widehat{Z}_k + \varepsilon_Z^{(k)})(\widehat{D}_{R,k}^- + \varepsilon_D^{-,(k)}) \geq \Delta_k^- > 0.$$

Consequently, $\mathbb{P}(\mathsf{Acc}_k^- \cap \{I_k < \Delta_k^-\}) \leq \beta_{\mathrm{step},k}$ with the original joint allocation: $\mathcal{E}_R^-$ replaces, rather than augments, the symmetric retained event. The required positive-gain inequality uses the same failure and rollout radii as Two-Gate; the retained radius and event are one-sided. Appendix E.2 bounds absolute changes on this rule's own adopted-step set.

D.6 CALCULATION FOR FIGURE 2

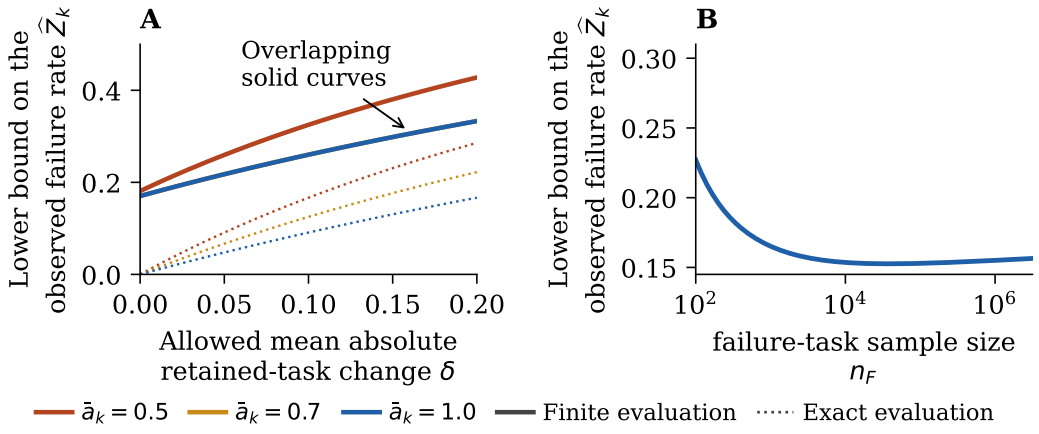


Figure 2: **Necessary lower bounds on the failure rate for further validation.** (A) The allowed mean absolute retained-task change restricts validation even with exact evaluation (dotted); finite evaluation error further restricts it (solid). (B) The failure-task sample size varies while the per-task rollout count, retained-record count, and total risk allocations are fixed. The risk share per reward estimate consequently changes with the number of evaluated tasks. Curves use the i.i.d. design in Appendix D.6.

Write $h = \widehat{Z}_k$ and let δ denote the allowed mean absolute retained-task change in Two-Gate. Setting $\delta^- = \delta$ gives the same numerical parameter for comparison with the one-sided rule. On the relevant coverage event, the two necessary branches of the admissible-region inequality imply

$$h > \varepsilon_L(h) + \varepsilon_Z \varepsilon_D(h) + \varepsilon_\Sigma + \delta(1 - h + \varepsilon_Z),$$

$$(h + \varepsilon_Z)\bar{a}_k > \varepsilon_Z \varepsilon_D(h) + \varepsilon_\Sigma + \delta(1 - h + \varepsilon_Z).$$

Panel A uses the larger lower root of these inequalities, at $m = 2 \times 10^4$, $n_R = 2000$, $n_F = 1000$, with i.i.d. records and $(\beta_Z, \beta_R, \beta_F, \beta_V) = (0.014, 0.024, 0.02, 0.05)$. The rollout allocation is $\beta_V^{(1)} = \beta_V/[2(n_F + n_R)]$. The solid curves for $\bar{a}_k = 0.7$ and 1 coincide because the branch independent of $\bar{a}_k$ determines their root; the exact-evaluation references $\delta/(\bar{a}_k + \delta)$ differ. Panel B varies n_F in the relaxed necessary bound Eq. (87) at the same m, n_R and total risks. The subsequent rise results from allocating the fixed total rollout risk over more tasks while the failure-task mean radius decreases.

E Multiple Updates and Task-Distribution Changes

E.1 Cumulative improvement

Use the adopted-step set $\mathcal{A}$, actual increments I_k, and pre-adoption bounds of Section 3.2.

Theorem 4 (Progressive improvement through evolution). *Fix $\mathcal{D}_{\text{user}}$ and run Procedure 1 for K steps under the hypotheses of Theorem 2 at every step. If its deterministic risks satisfy $\sum_{k<K} \beta_{\text{step},k} \le \beta_{\text{run}}$, then with probability at least $1-\beta_{\text{run}}$,*

$$J_{\text{user}}(A_K) - J_{\text{user}}(A_0) \ge \sum_{k\in\mathcal{A}} \Delta_k.$$

Here Δ_k has the definition and conventions of Eq. (5).

Proof of Theorem 4. For each step define the quantitative bad-accept event $\mathcal{B}_k := \mathsf{Acc}_k \cap \{I_k < \Delta_k\}$. Theorem 2 gives $\mathbb{P}(\mathcal{B}_k) \le \beta_{\text{step},k}$ under the actual step protocol, so

$$\mathbb{P}\left(\bigcup_{k<K} \mathcal{B}_k\right) \le \sum_{k<K} \beta_{\text{step},k} \le \beta_{\text{run}}.$$

Outside this union, every accepted step satisfies $I_k \ge \Delta_k$. Definition 9 gives $\sum_{k<K} I_k = \sum_{k\in\mathcal{A}} I_k$. Summing the increments and the pre-acceptance margins only over accepted steps gives

$$J_{\text{user}}(A_K) - J_{\text{user}}(A_0) = \sum_{k<K} I_k \ge \sum_{k\in\mathcal{A}} \Delta_k.$$

The measured-margin statement follows identically from Proposition 18 using $\mathsf{Acc}_k^M \cap \{I_k < \widehat{M}_k\}$.

For Measured-margin, on its own adopted-step set $\mathcal{A}^M$, the same argument gives

$$J_{\text{user}}(A_K) - J_{\text{user}}(A_0) \ge \sum_{k\in\mathcal{A}^M} \widehat{M}_k$$

with probability at least $1-\beta_{\text{run}}$, whenever Proposition 18 holds at every step and the deterministic step risks sum to at most β_{run}. Under equal risk allocation each step uses β_{run}/K. In the i.i.d. marked-stream design, substitution in Proposition 10 adds the corresponding $\log K$ term to its sample requirement.

Definition 15 (Distribution drift). *For the task distributions $\mathcal{D}_{\text{user}}^{(k)}$ and $\mathcal{D}_{\text{user}}^{(k+1)}$ at adjacent self-evolution steps, define their drift by $\eta_k := \text{TV}(\mathcal{D}_{\text{user}}^{(k+1)}, \mathcal{D}_{\text{user}}^{(k)})$. Write $J_{\text{user}}^{(k)}(A) := \mathbb{E}_{t\sim\mathcal{D}_{\text{user}}^{(k)}}[V(A,t)]$ for expected reward under the step-k distribution.*

Lemma 10. *For any A:*

$$|J_{user}^{(k+1)}(A) - J_{user}^{(k)}(A)| \le \eta_k. \tag{96}$$

Proof. The signed measure $\mathcal{D}_{\text{user}}^{(k+1)} - \mathcal{D}_{\text{user}}^{(k)}$ has total mass zero. Hence

$$\left|\int V(A,t)\, d(\mathcal{D}_{\text{user}}^{(k+1)} - \mathcal{D}_{\text{user}}^{(k)})(t)\right| \le \text{osc}(V(A,\cdot))\,\eta_k \le \eta_k,$$

by bounded rewards.

Theorem 10 (Cumulative gain under changing task distributions). *Under Assumptions 1, 5, 6, 7, and 8 and a drift sequence $\{\eta_k\}_{k=0}^{K-1}$, suppose that the actual finite-horizon protocol at step k supplies the returned failure-task coverage and stored-sample readiness required by Theorem 2 for $\mathcal{D}_{\text{user}}^{(k)}$. Let $I_k = J_{\text{user}}^{(k)}(A_{k+1}) - J_{\text{user}}^{(k)}(A_k)$ and use the pre-acceptance Δ_k from Eq.* (5)*, with zero value outside $\mathsf{Cand}_k \cap \mathsf{Ready}_k$. Write*

$$\mathcal{B}_k := \mathsf{Acc}_k \cap \{I_k < \Delta_k\}, \qquad \mathbb{P}(\mathcal{B}_k) \le \beta_{\text{step},k}, \qquad \sum_{k<K} \beta_{\text{step},k} \le \beta_{\text{run}}.$$

Then, without requiring independence across steps, with probability at least $1-\beta_{\mathrm{run}}$*,*

$$J_{user}^{(K)}(A_K) - J_{user}^{(0)}(A_0) \geq \sum_{k\in\mathcal{A}} \Delta_k - \sum_{k=0}^{K-1} \eta_k. \tag{97}$$

Proof of Theorem 10. Let $\mathcal{B} := \bigcup_{k<K} \mathcal{B}_k$. The assumed per-step quantitative bounds and a union bound give $\mathbb{P}(\mathcal{B}) \leq \sum_{k<K} \beta_{\mathrm{step},k} \leq \beta_{\mathrm{run}}$. Outside $\mathcal{B}$, at each step add and subtract $J_{\mathrm{user}}^{(k)}(A_{k+1})$:

$$J_{\mathrm{user}}^{(k+1)}(A_{k+1}) - J_{\mathrm{user}}^{(k)}(A_k) = I_k + \left[J_{\mathrm{user}}^{(k+1)}(A_{k+1}) - J_{\mathrm{user}}^{(k)}(A_{k+1})\right].$$

Lemma 10 bounds the bracket below by $-\eta_k$ on every step. Definition 9 makes $I_k = 0$ on nonadopted steps, so the within-step increments sum to $\sum_{k\in\mathcal{A}} I_k \geq \sum_{k\in\mathcal{A}} \Delta_k$. Summing the displayed identity telescopes and yields Eq. (97). No cross-step independence is used.

Corollary 13. *On the event of Theorem 4, suppose* $\Delta_k \geq \Delta_\star > 0$ *for every accepted step* $k < K$*. Then*

$$|\mathcal{A}| \leq \frac{1 - J_{\mathrm{user}}(A_0)}{\Delta_\star}. \tag{98}$$

The same positive lower bound $\Delta_\star$ must hold on every adopted step to obtain Eq. (98).

Proof of Corollary 13. Theorem 4 gives

$$J_{\mathrm{user}}(A_K) - J_{\mathrm{user}}(A_0) \geq \sum_{k\in\mathcal{A}} \Delta_k \geq |\mathcal{A}|\Delta_\star.$$

By Assumption 1, $J_{\mathrm{user}}(A_K) \leq 1$, so the same endpoint difference is at most $1 - J_{\mathrm{user}}(A_0)$. Combining the two bounds and dividing by $\Delta_\star > 0$ gives Eq. (98).

E.2 Cumulative absolute reward changes

The superscript (k) identifies the evaluation radius at step k. The following bounds control both the accumulated absolute change along the update path and the endpoint change relative to the initial agent.

The symmetric Two-Gate bound. Fix $\mathcal{D}_{\mathrm{user}}$ throughout K steps under Assumptions 1 and 2, with $Z_k \in (0,1)$ on adopted steps. On the joint evaluation-coverage event, Proposition 1 gives Eq. (10) for the Two-Gate adopted-step set $\mathcal{A}$.

Proposition 21 (Cumulative absolute changes under a one-sided retained-loss rule). *Fix the same task distribution and assumptions, and at each step use the candidate, readiness, selector, and risk conditions of the one-sided rule Eq.* (95)*. Let* $\mathcal{A}^- := \{k < K : \mathsf{Acc}_k^-\}$ *be this rule's adopted-step set and* $\mathcal{V}_K^- := \sum_{k\in\mathcal{A}^-} \mathbb{E}_{\mathcal{D}_{\mathrm{user}}} |V(A_{k+1},\cdot) - V(A_k,\cdot)|$*. On the common one-sided evaluation-coverage event,*

$$\begin{aligned}\mathcal{V}_K^- &\leq J_{\mathrm{user}}(A_K) - J_{\mathrm{user}}(A_0) + \sum_{k\in\mathcal{A}^-} (Z_k - \tau + \varepsilon_L^{(k)}) + 2\sum_{k\in\mathcal{A}^-} (1-Z_k)(\delta^- + \varepsilon_D^{-,(k)}) \\ &\leq 1 - J_{\mathrm{user}}(A_0) + \sum_{k\in\mathcal{A}^-} (Z_k - \tau + \varepsilon_L^{(k)}) + 2\sum_{k\in\mathcal{A}^-} (1-Z_k)(\delta^- + \varepsilon_D^{-,(k)}).\end{aligned} \tag{99}$$

Moreover, $\mathbb{E}_{\mathcal{D}_{\mathrm{user}}} |V(A_K,\cdot) - V(A_0,\cdot)| \leq \min\{1, \mathcal{V}_K^-\}$*. If the deterministic step risks sum to at most* β_{run}*, the common event has probability at least* $1-\beta_{\mathrm{run}}$*.*

Proof of Propositions 1 and 21. For each adopted step write $a_k(t) = \mathrm{Adv}_{A_k}(A_{k+1}, t)$ and $F_k^- := Z_k \mathbb{E}_{\mathcal{D}_{F,k}}(-a_k)_+$. Lemma 1, applied to $|a_k|$, gives

$$\mathbb{E}_{\mathcal{D}_{\mathrm{user}}} |V(A_{k+1},\cdot) - V(A_k,\cdot)| = Z_k \mathbb{E}_{\mathcal{D}_{F,k}} |a_k| + (1-Z_k)\mathbb{E}_{\mathcal{D}_{R,k}} |a_k|. \tag{100}$$

Writing $|x| = x + 2(-x)_+$ yields

$$Z_k \mathbb{E}_{\mathcal{D}_{F,k}} |a_k| = L_{A_k} + 2F_k^-.$$

Since $\mathbb{E}_{\mathcal{D}_{F,k}}|a_k| \le 1$, the condition $\widehat{L}_k \ge \tau$ in Eq. (4) and the failure-task error bound imply

$$2F_k^- \le Z_k - L_{A_k} \le Z_k - \tau + \varepsilon_L^{(k)}. \tag{101}$$

For Two-Gate, $\widehat{D}_R \le \delta$ in Eq. (4) and the retained-task error bound give $D_R \le \delta + \varepsilon_D^{(k)}$. The reward decomposition in Eq. (27) implies $L_{A_k} \le I_k + (1 - Z_k)D_R$; hence Eqs. (100)–(101) give

$$\mathbb{E}_{\mathcal{D}_{\text{user}}}|a_k| \le I_k + Z_k - \tau + \varepsilon_L^{(k)} + 2(1 - Z_k)(\delta + \varepsilon_D^{(k)}).$$

On nonadopted steps $I_k = 0$, so summing over $\mathcal{A}$ telescopes to $\sum_{k\in\mathcal{A}} I_k = J_{\text{user}}(A_K) - J_{\text{user}}(A_0) \le 1 - J_{\text{user}}(A_0)$. This proves Eq. (10).

For the one-sided rule, Eq. (27) gives the exact identity

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_{\text{user}}}|a_k| &= L_{A_k} + 2F_k^- + (1 - Z_k)(D_{R,k}^- + D_{R,k}^+) \\ &= I_k + 2F_k^- + 2(1 - Z_k)D_{R,k}^-.\end{aligned}$$

On the joint one-sided coverage event, Eq. (95) and $\mathcal{E}_R^-$ give $D_{R,k}^- \le \delta^- + \varepsilon_D^{-,(k)}$. Apply Eq. (101), sum only over $\mathcal{A}^-$, and telescope $\sum_{k\in\mathcal{A}^-} I_k = J_{\text{user}}(A_K) - J_{\text{user}}(A_0)$. Bounded reward gives $J_{\text{user}}(A_K) \le 1$, proving Eq. (99). The taskwise triangle inequality bounds the endpoint absolute change by the path sum, while bounded rewards bound it by one. Finally, the per-step joint one-sided coverage event $\mathcal{G}_k^-$ in the proof of Proposition 18 has failure probability at most $\beta_{\text{step},k}$; intersecting these events and applying a union bound gives the stated common-event probability.

Bounding the failure-task term in Eq. (100) only by $|a_k| \le 1$ contributes Z_k on each adopted step. The condition $\widehat{L}_k \ge \tau$ instead yields the term $Z_k - \tau + \varepsilon_L^{(k)}$ through Eq. (101). The endpoint mean absolute difference is at most one, while the cumulative path change can exceed one.

Remark 12. Holding the realized sequence of agents, evaluation radii, and adopted-step set fixed, letting $\delta \downarrow 0$ in the right-hand side of Eq. (10) leaves the retained-task contribution $2\sum_{k\in\mathcal{A}}(1 - Z_k)\varepsilon_D^{(k)}$.

E.3 A SUCCESSFUL UPDATE AND LATER IMPROVEMENT OPPORTUNITIES

Proposition 22 (No later-step guarantee from step-0 quantities under unrestricted updates). *In the kernel setting of Appendix A.1, allow the composition map $(M, C) \mapsto A$ to be any measurable map consistent with Eqs.* (15)–(19)*. For every pool size $N \ge 1$ and horizon $K \ge 2$, there exist a common target T and two instances with the same frozen LLM, initial agent, step-0 generation and evaluation experiment, validation and selection rules, and first successor agent. Both have $P_0(T) = 1$, and the first update is validated and strictly improves overall expected reward. In the first instance, $P_1^+(T) > 0$. In the second, every subsequently generated modification leaves expected reward unchanged on every user task, so, almost surely, $P_k(T) = 0$ for $1 \le k < K$, and*

$$J_{\text{user}}(A_K) - J_{\text{user}}(A_0) = J_{\text{user}}(A_1) - J_{\text{user}}(A_0).$$

The composition maps agree at the initial harness state and every step-0 candidate harness state, but may differ at later states.

Proof of Proposition 22. Take four user tasks t_0, t_1, t_2, t_R of equal probability, binary outputs on these user tasks, reward $r(t, o) = o$, and detector $\phi(t, o) = 1 - o$. The initial agent has deterministic user-task outputs $A_0 : (0, 0, 0, 1)$ and outputs modification c_0 on every self-modification task. The common first successor has outputs $A_1 : (1, 0, 0, 1)$ and outputs c_1 on every self-modification task.

Use code states C_0, C_1, C_2 and modification space $\mathcal{M} = \{c_0, c_1\}$. The code edit c_0 sends C_0 to C_1 and otherwise leaves the code state unchanged; c_1 sends C_1 to C_2 and otherwise leaves it unchanged. Both instances use this same code-edit operation and the same frozen LLM M. Their measurable composition maps agree at C_0 and C_1, realizing the full kernels A_0 and A_1, respectively. At C_2, instance (a) has user-task outputs $(1, 1, 0, 1)$ and self-task output c_1, whereas instance (b) has the same full kernel as A_1. All spaces and maps are finite and measurable, and Eq. (19) holds because the current agent's self-task kernel generates the stated modification directly. In particular, all step-0 candidate agents agree between the two instances: c_0 gives A_1 and c_1 leaves A_0 unchanged.

Choose the common target $T = (1/8, 1/16)$. At step 0, $Z_0 = 3/4$, and applying c_0 gives

$$L_0(c_0) = 1/4, \qquad D_0(c_0) = 0, \qquad J_{\text{user}}(A_1) - J_{\text{user}}(A_0) = 1/4.$$

Every draw is c_0, so $P_0(T) = 1$ for any pool size. At A_1, $Z_1 = 1/2$. In instance (a), applying c_1 improves expected reward only on t_1, giving $L_1(c_1) = 1/4$ and $D_1(c_1) = 0$. In instance (b), applying c_1 preserves every user-task output, so $L_1(c_1) = D_1(c_1) = 0$.

For these finite deterministic instances, the same evaluator checks the four user-task outputs exactly and returns singleton rectangles. Their total-width bounds are zero, so the set $\mathcal{P}^+$ in Eq. (7) equals the true qualified set, with conditional coverage one. The validation rule in Eq. (59), followed by the same selector choosing a validated candidate whenever one exists, accepts c_0 at step 0 in both instances and c_1 at step 1 in instance (a). There c_1 meets T and is generated with probability one, giving $P_1^+(T) = 1$.

In instance (b), every later draw is c_1 and leaves expected reward unchanged on every user task, whether or not it is adopted. Under the stated validation rule it is rejected, and the full kernel remains A_1. Thus $P_k(T) = 0$ for all $1 \le k < K$ and

$$J_{\text{user}}(A_K) - J_{\text{user}}(A_0) = 1/4 = J_{\text{user}}(A_1) - J_{\text{user}}(A_0) \quad \text{almost surely.}$$

Instance (a) instead increases expected reward by another $1/4$ at step 1 and then generates modifications that leave expected reward unchanged. The construction works for every $N \ge 1$ and $K \ge 2$.

E.4 Agent-responsive task distributions

All quantities in a same-distribution comparison below are constructed from the declared common reference measure.

Assumption 9 (Agent-responsive user-task distribution). *At each step k there is a task distribution $\mathcal{D}_{A,k}$ that obtains once the task stream has settled under agent A, depending on A and on the ambient conditions at step k, but not otherwise on the sequence of updates that produced A. Write $\mathcal{D}_{A_k} := \mathcal{D}_{A_k,k}$.*

The indices A, k allow both agent response and changes in ambient conditions. Tasks can be sampled under the current agent's distribution; the full distribution need not be known. A candidate's associated distribution is counterfactual before adoption.

Definition 16. *For $\tilde{A} = A_k \oplus \Delta C$, define improvement on the current agent's task distribution by*

$$\text{Imp}^{\text{cmp}}(\tilde{A}) := \mathbb{E}_{t\sim\mathcal{D}_{A_k}}\big[\text{Adv}_{A_k}(\tilde{A}, t)\big],$$

and let $J_k^\sharp(A) := \mathbb{E}_{t\sim\mathcal{D}_{A,k}}[V(A,t)]$ be the expected reward under the task distribution induced by A. Then $\text{Imp}^{\text{rea}}(\tilde{A}) := J_k^\sharp(\tilde{A}) - J_k^\sharp(A_k)$ compares the agents under their respective task distributions. Write $\varrho(\Delta C) := \text{TV}(\mathcal{D}_{\tilde{A},k}, \mathcal{D}_{A_k,k})$ for the total-variation distance between these distributions.

Proposition 23. *Under Assumption 1, for any probability measure $\mathcal{D}$ on $\mathcal{T}_{user}$ with $Z_k \in (0,1)$ (the standing assumption of §2.2), and with $L_{A_k}, Z_k, \mathcal{D}_{F,k}, \mathcal{D}_{R,k}, D_R$ all constructed from $\mathcal{D}$ by Eq. (27) and Eq. (1),*

$$\mathbb{E}_{t\sim\mathcal{D}}\big[\text{Adv}_{A_k}(\tilde{A}, t)\big] \;\ge\; L_{A_k}(\tilde{A}) - (1 - Z_k)\, D_R(\tilde{A}; A_k). \tag{102}$$

In particular Eq. (102) holds at $\mathcal{D} = \mathcal{D}_{A_k}$, whether or not Assumption 2 does.

Corollary 14. *Under Assumptions 1 and 9,*

$$\big|\text{Imp}^{\text{rea}}(\tilde{A}) - \text{Imp}^{\text{cmp}}(\tilde{A})\big| \le \varrho(\Delta C), \qquad \textit{hence} \qquad \text{Imp}^{\text{rea}}(\tilde{A}) \ge L_{A_k} - (1 - Z_k) D_R - \varrho(\Delta C).$$

Remark 13. Under Assumption 9, put $\varrho_k = \text{TV}(\mathcal{D}_{A_{k+1},k}, \mathcal{D}_{A_k,k})$ and $\eta_k^{\text{exo}} = \sup_A \text{TV}(\mathcal{D}_{A,k+1}, \mathcal{D}_{A,k})$. For $\mathcal{D}_{\text{user}}^{(k)} = \mathcal{D}_{A_k,k}$ the triangle inequality through $\mathcal{D}_{A_{k+1},k}$ yields $\eta_k \le \varrho_k + \eta_k^{\text{exo}}$. Thus the drift in Theorem 10 includes both terms. With no adopted modification the response at fixed environment state is zero; the environment may still change.

For a single-step post-response expected-gain guarantee, a valid response bound b_k available before adoption gives the sufficient check $\tau > \varepsilon_L + (1 - \widehat{Z}_k + \varepsilon_Z)(\delta + \varepsilon_D) + b_k$. If the bound is probabilistic, its failure risk is included in the joint experiment's risk budget.

E.4.1 ESTIMATION UNDER THE CURRENT TASK DISTRIBUTION

Under the sampling and coverage conditions of Appendix C.1, the same estimators compare the agents under $\mathcal{D}_{A_k,k}$. Independent fresh rollouts evaluate both agents on the same sampled tasks, while the failure weights and retained-task distribution refer to the current agent, as required by Proposition 23.

Procedure 1 recollects the stored sample after every accepted modification as described in Remark 5. At the next step, new records and fresh evaluations define $\widehat{Z}_{k+1}$, $\widehat{L}_{k+1}$, and $\widehat{D}_{R,k+1}$ relative to the successor's task distribution. A response bound available before adoption is separately needed for a post-response expected-improvement guarantee.

E.4.2 BOUNDING THE REWARD RESPONSE AFTER AN UPDATE

The partition bound below controls the effect of post-update task-distribution response on expected reward when the stated uniform reward-resolution condition holds.

Define $\mathcal{W} = \int V(\widetilde{A}, t)\, d(\mathcal{D}_{\widetilde{A},k} - \mathcal{D}_{A_k,k})$. Under the uniform resolution condition of Lemma 13 in Appendix E.6, $|\mathcal{W}| \le \mathrm{TV}_{\mathcal{F}} + \zeta$. These partition probabilities compare the two distributions at the same environment state after response. Lemma 12 in the same appendix estimates their cell discrepancy; the uniform within-cell reward condition in Lemma 13 is separate. The samples observed after response do not themselves provide the advance bound b_k required by Remark 13.

Proof of Proposition 23. The pointwise decomposition underlying Lemma 1 holds for any reference measure $\mathcal{D}$. Integrating it for $g = \mathrm{Adv}_{A_k}(\tilde{A}, \cdot)$ gives

$$\mathbb{E}_{\mathcal{D}}[\mathrm{Adv}_{A_k}(\tilde{A}, \cdot)] = L_{A_k}(\tilde{A}) + E_{A_k}(\tilde{A}),$$

where $E_{A_k}(\tilde{A}) = (1 - Z_k)\mathbb{E}_{\mathcal{D}_{R,k}}[\mathrm{Adv}_{A_k}(\tilde{A}, t)]$. Since $|\,\mathrm{Adv}_{A_k}(\tilde{A}, t)|$ integrates to $D_R(\tilde{A}; A_k)$ under $\mathcal{D}_{R,k}$, Jensen's inequality gives

$$E_{A_k}(\tilde{A}) \ge -(1 - Z_k) D_R(\tilde{A}; A_k).$$

Combining the two displays proves Eq. (102) for the arbitrary reference measure $\mathcal{D}$.

Proof of Corollary 14. Add and subtract $\mathbb{E}_{\mathcal{D}_{A_k}}[V(\tilde{A}, \cdot)]$, so $\mathrm{Imp}^{\mathrm{rea}} = \mathrm{Imp}^{\mathrm{cmp}} + \mathcal{W}$ with $\mathcal{W} := \int V(\tilde{A}, \cdot)\, d(\mathcal{D}_{\tilde{A},k} - \mathcal{D}_{A_k,k})$. The signed measure has total mass 0 **on task space**, so $\mathcal{W}$ is exactly the object Lemma 10 bounds: $|\mathcal{W}| \le \mathrm{osc}(V(\tilde{A}, \cdot)) \cdot \varrho \le \varrho$, since $V \in [0, 1]$ by Assumption 1 and Eq. (20). Combine with Proposition 23.

E.5 EVALUATION AND DEPLOYMENT TASK DISTRIBUTIONS

E.5.1 DEFINITIONS

At step k, distinguish:

- $\mathcal{D}_{\text{dep}}^{(k)}$ — the deployment task distribution at agent state k;
- $\mathcal{D}_{\text{prod}}^{(k)}$ — the base task distribution represented by the evaluation design, from which failure and retained distributions are formed using their stated weights.

Expected reward under the two task distributions:

$$J_{\text{dep}}^{(k)}(A) := \mathbb{E}_{t \sim \mathcal{D}_{\text{dep}}^{(k)}}[V(A, t)], \quad J_{\text{prod}}^{(k)}(A) := \mathbb{E}_{t \sim \mathcal{D}_{\text{prod}}^{(k)}}[V(A, t)].$$

Track deployment expected reward as $X_k := J_{\text{dep}}^{(k)}(A_k)$.

Definition 17 (Evaluation–deployment distribution discrepancy). *At step k, let $\mathcal{C}_k^{\mathrm{step}}$ be a class containing the current agent and every possible adopted successor in the joint experiment, and define*

$$\eta_k^{\mathrm{obs}} := \sup_{A \in \mathcal{C}_k^{\mathrm{step}}} \left| J_{\mathrm{dep}}^{(k)}(A) - J_{\mathrm{prod}}^{(k)}(A) \right|.$$

The inter-step drift of Definition 15 compares task distributions at adjacent steps; η_k^{obs} compares the evaluation and deployment distributions at the same step. Agent response in Appendix E.4 is one possible source of distribution change.

Applying Lemma 10 to each agent and taking the supremum gives

$$\eta_k^{\mathrm{obs}} \le \mathrm{TV}(\mathcal{D}_{\mathrm{dep}}^{(k)}, \mathcal{D}_{\mathrm{prod}}^{(k)}).$$

This bound avoids enumerating agents. Appendix E.6 estimates the probabilities of a declared finite partition and bounds η_k^{obs} under the uniform reward-resolution condition of Lemma 13.

E.5.2 One-step expected-reward bound

Lemma 11. *Under Assumption 1, let $\mathcal{D}_{\mathrm{prod}}^{(k)}$ be fixed before evaluating the candidate agent, let $\mathsf{Acc}_k \subseteq \mathsf{Cand}_k \cap \mathsf{Ready}_k$ be the step-k acceptance event of Theorem 2 evaluated on that measure. All probabilities and expectations below are under the same joint experiment, including generation, evaluation, updating, and any history-dependent task distributions. Let $I_k := J_{\mathrm{prod}}^{(k)}(A_{k+1}) - J_{\mathrm{prod}}^{(k)}(A_k)$ be improvement on the evaluation distribution and Δ_k the pre-acceptance quantity Eq.* (5)*, set to zero outside $\mathsf{Cand}_k \cap \mathsf{Ready}_k$, and suppose*

$$\mathbb{P}(\mathsf{Acc}_k \cap \{I_k < \Delta_k\}) \le \beta_{\mathrm{step},k}, \qquad \chi_k := \mathbb{E}[\mathbf{1}\{\mathsf{Acc}_k\}\Delta_k].$$

If $\mathrm{TV}(\mathcal{D}_{\mathrm{dep}}^{(k+1)}, \mathcal{D}_{\mathrm{dep}}^{(k)}) \le \eta_k^{\mathrm{step}}$ almost surely, with nonnegative integrable η_k^{step}, then

$$\mathbb{E}[X_{k+1} - X_k] \ge \chi_k - 2\beta_{\mathrm{step},k} - \mathbb{E}[\eta_k^{\mathrm{step}}] - 2\mathbb{E}[\eta_k^{\mathrm{obs}}]. \tag{103}$$

After separating inter-step distribution movement, the comparison between evaluation and deployment distributions is made once at A_k and once at A_{k+1}. Each contributes at most η_k^{obs}, giving the factor two. Deterministic drift bounds are a special case of Eq. (103).

E.5.3 Cumulative bound under the two task distributions

Theorem 11. *Suppose the hypotheses of Lemma 11 hold at every step under consideration, and define*

$$\nu_k := \chi_k - 2\beta_{\mathrm{step},k} - \mathbb{E}[\eta_k^{\mathrm{step}}] - 2\mathbb{E}[\eta_k^{\mathrm{obs}}]. \tag{104}$$

Then, for any fixed integers $s \ge 0$ and $K \ge 1$,

$$\sum_{k=s}^{s+K-1} \nu_k \le \mathbb{E}[X_{s+K}] - \mathbb{E}[X_s]. \tag{105}$$

The term χ_k is the expectation of the accepted-step contribution $\mathbf{1}\{\mathsf{Acc}_k\}\Delta_k$, which is zero on nonaccepted steps. If $\Delta_k \ge \gamma$ on Acc_k, then $\chi_k \ge \gamma\,\mathbb{P}(\mathsf{Acc}_k)$ under the same joint experiment.

Appendix E.6 gives a partition-based estimate of η_k^{obs} under its stated resolution and coverage conditions.

Proof of Lemma 11. Write the tracked increment as

$$\begin{aligned} X_{k+1} - X_k = {} & \underbrace{J_{\mathrm{prod}}^{(k)}(A_{k+1}) - J_{\mathrm{prod}}^{(k)}(A_k)}_{\mathrm{(I)}=I_k} \\ & + \underbrace{J_{\mathrm{dep}}^{(k+1)}(A_{k+1}) - J_{\mathrm{dep}}^{(k)}(A_{k+1})}_{\mathrm{(II)}} \\ & + \underbrace{J_{\mathrm{dep}}^{(k)}(A_{k+1}) - J_{\mathrm{prod}}^{(k)}(A_{k+1})}_{\mathrm{(III)}} + \underbrace{J_{\mathrm{prod}}^{(k)}(A_k) - J_{\mathrm{dep}}^{(k)}(A_k)}_{\mathrm{(IV)}}. \end{aligned}$$

Let $\mathcal{B}_k := \mathsf{Acc}_k \cap \{I_k < \Delta_k\}$. On $\mathcal{B}_k^c$, Definition 9 and the accepted-step bound give $I_k \ge \mathbf{1}\{\mathsf{Acc}_k\}\Delta_k$. On $\mathcal{B}_k$, bounded rewards give $I_k \ge -1$. The positive-gain check gives $\Delta_k > 0$, its

definition Eq. (5) gives $\Delta_k \le \tau$, and $\widehat{L}_k \ge \tau$ in Eq. (4) gives $\tau \le \widehat{L}_k \le 1$ on acceptance. Thus $\Delta_k \le 1$ there, and

$$I_k \ge \mathbf{1}\{\mathsf{Acc}_k\}\Delta_k - 2\mathbf{1}\{\mathcal{B}_k\}, \qquad \mathbb{E}[I_k] \ge \chi_k - 2\beta_{\mathrm{step},k}.$$

Lemma 10 bounds term (II) by $-\eta_k^{\mathrm{step}}$, and Definition 17 bounds terms (III) and (IV) together by $-2\eta_k^{\mathrm{obs}}$. Thus, pathwise,

$$X_{k+1} - X_k \ge \mathbf{1}\{\mathsf{Acc}_k\}\Delta_k - 2\mathbf{1}\{\mathcal{B}_k\} - \eta_k^{\mathrm{step}} - 2\eta_k^{\mathrm{obs}}.$$

Taking expectations under the same joint distribution, with nonnegative integrable drift bounds, gives

$$\mathbb{E}[X_{k+1} - X_k] \ge \chi_k - 2\beta_{\mathrm{step},k} - \mathbb{E}[\eta_k^{\mathrm{step}}] - 2\mathbb{E}[\eta_k^{\mathrm{obs}}],$$

which is Eq. (103).

Proof of Theorem 11. By Lemma 11, for every k,

$$\mathbb{E}[X_{k+1} - X_k] \ge \nu_k.$$

Summing this inequality over $k = s, \dots, s+K-1$ and using linearity of expectation gives

$$\sum_{k=s}^{s+K-1} \nu_k \le \mathbb{E}\left[\sum_{k=s}^{s+K-1} (X_{k+1} - X_k)\right] = \mathbb{E}[X_{s+K}] - \mathbb{E}[X_s].$$

This proves Eq. (105).

E.6 Estimating the evaluation–deployment discrepancy

A finite partition estimates differences between the evaluation and deployment task distributions at a declared resolution. Lemma 13 states when these differences control expected reward uniformly over the agent class of Definition 17.

Fix in advance a finite measurable partition $\mathcal{F} = \{E_1, \dots, E_m\}$ of $\mathcal{T}_{\mathrm{user}}$. For each source $\bullet \in \{P, D\}$, observe $n_\bullet$ i.i.d. tasks from $\mathcal{D}^{(k)}_{\mathrm{prod}}$ or $\mathcal{D}^{(k)}_{\mathrm{dep}}$, respectively, and record their cell frequencies. Independence between the two sources is not required.

Estimator. The $\mathcal{F}$-restricted empirical total variation:

$$\widehat{\mathrm{TV}}^{(k)}_{\mathcal{F}} := \tfrac{1}{2}\sum_{i=1}^{m} \Big| \widehat{\mathcal{D}}^{(k)}_{\mathrm{dep}}(E_i) - \widehat{\mathcal{D}}^{(k)}_{\mathrm{prod}}(E_i) \Big|, \tag{106}$$

where $\widehat{\mathcal{D}}(E_i)$ are the empirical frequencies on the two samples.

Lemma 12. *Under this sampling scheme, let* $\mathrm{TV}_{\mathcal{F}} := \frac{1}{2}\sum_i |\mathcal{D}^{(k)}_{\mathrm{dep}}(E_i) - \mathcal{D}^{(k)}_{\mathrm{prod}}(E_i)|$. *With probability at least* $1-\beta$,

$$\left|\widehat{\mathrm{TV}}^{(k)}_{\mathcal{F}} - \mathrm{TV}_{\mathcal{F}}\right| \;\le\; \varepsilon^{obs}(\beta) := \frac{1}{2}\sum_{\bullet\in\{P,D\}} \left(\sqrt{\frac{m}{n_\bullet}} + \sqrt{\frac{2\log(2/\beta)}{n_\bullet}}\right). \tag{107}$$

Under the corresponding condition of Lemma 13, define

$$\widehat{\eta}_k^{\mathrm{obs}} := \widehat{\mathrm{TV}}^{(k)}_{\mathcal{F}} + \varepsilon^{\mathrm{obs}}(\beta) + \zeta, \tag{108}$$

with $\zeta = 0$ in case (i). On the coverage event of Lemma 12, $\eta_k^{\mathrm{obs}} \le \widehat{\eta}_k^{\mathrm{obs}}$.

Lemma 13. *Let $\mathcal{F}$ be the partition above and let the agent class be the one used in Definition 17.*

- *(i) If $V(A,\cdot)$ is constant on every cell for every agent A in this class, then $\eta_k^{\mathrm{obs}} \le \mathrm{TV}_{\mathcal{F}}$.*
- *(ii) If $\sup_{t,t'\in E_i} |V(A,t) - V(A,t')| \le \zeta$ for every agent A in the class and every cell E_i, then $\eta_k^{\mathrm{obs}} \le \mathrm{TV}_{\mathcal{F}} + \zeta$.*

Proof of Lemma 13. Fix an agent A in the stated class and put $f(t) = V(A, t)$. On each nonempty cell E_i, let $a_i = \inf_{t \in E_i} f(t)$, set $g(t) = a_i$, and write $h = f - g$. Bounded reward and the within-cell condition give $a_i \in [0, 1]$ and $h(t) \in [0, \zeta]$; no infimum need be attained. Write $p_i = \mathcal{D}^{(k)}_{\mathrm{dep}}(E_i)$ and $q_i = \mathcal{D}^{(k)}_{\mathrm{prod}}(E_i)$. Since $\sum_i (p_i - q_i) = 0$ and $a_i \in [0, 1]$,

$$\left| \sum_i a_i (p_i - q_i) \right| \le \tfrac{1}{2} \sum_i |p_i - q_i| = \mathrm{TV}_{\mathcal{F}}.$$

Both expectations of h lie in $[0, \zeta]$, so their difference has absolute value at most ζ. Combining the g and h parts and taking the supremum over the same agent class proves part (ii). In the constant-cell case $h = 0$, giving part (i).

For dependent records, compare each source's entire record vector with its stated i.i.d. reference vector. If these joint-TV errors are bounded by d_P, d_D and $\widetilde{\beta} := \beta - d_P - d_D > 0$, using $\widetilde{\beta}$ in Eq. (107) gives total risk at most β.

Since $0 \le \eta_k^{\mathrm{obs}} \le 1$, coverage with failure probability at most β also gives $\mathbb{E}[\eta_k^{\mathrm{obs}}] \le \mathbb{E}[\min\{1, \widehat{\eta}_k^{\mathrm{obs}}\}] + \beta$ in the same joint experiment. Indeed, for the coverage event $\mathcal{E}$ and $B = \min\{1, \widehat{\eta}_k^{\mathrm{obs}}\}$, one has $\eta_k^{\mathrm{obs}} \le B$ on $\mathcal{E}$ and $\eta_k^{\mathrm{obs}} \le 1$ always, so $\mathbb{E}[\eta_k^{\mathrm{obs}}] \le \mathbb{E}[B \mathbf{1}_{\mathcal{E}}] + \mathbb{P}(\mathcal{E}^c) \le \mathbb{E}[B] + \beta$.

Proof of Lemma 12. Write $p^\bullet$ for the true $\mathcal{F}$-marginal of source $\bullet$ and $\hat{p}^\bullet$ for its empirical counterpart. By the triangle inequality applied to the $\frac{1}{2}\sum_i |\cdot|$ form,

$$\left| \widehat{\mathrm{TV}}_{\mathcal{F}} - \mathrm{TV}_{\mathcal{F}} \right| \;\le\; \tfrac{1}{2} \big(\|\hat{p}^P - p^P\|_1 + \|\hat{p}^D - p^D\|_1 \big),$$

so it suffices to bound the L^1 error of each empirical multinomial. For the mean, Cauchy–Schwarz gives

$$\mathbb{E}\|\hat{p} - p\|_1 = \sum_i \mathbb{E}|\hat{p}_i - p_i| \le \sum_i \sqrt{\tfrac{p_i(1-p_i)}{n}} \le \tfrac{1}{\sqrt{n}} \sum_i \sqrt{p_i} \le \sqrt{\tfrac{m}{n}}.$$

For the deviation, $\|\hat{p} - p\|_1$ is a function of the n samples that changes by at most $2/n$ when one sample is altered (it moves two cells by $1/n$ each), so McDiarmid's inequality gives $\mathbb{P}\big[\|\hat{p} - p\|_1 \ge \mathbb{E}\|\hat{p} - p\|_1 + t\big] \le e^{-nt^2/2}$; setting this to $\beta/2$ yields $t = \sqrt{2\log(2/\beta)/n}$. A union bound over the two sources gives Eq. (107) with probability $\ge 1 - \beta$.

F Reachability Estimation

F.1 Uniform confidence bounds and diagnostic quantities

The following restates Theorem 6.

Theorem 6 (Uniform confidence bounds for reachability). *Fix $(\mathcal{H}_k, \Xi_k = \xi)$ and draw $C_1, \ldots, C_{N_{\mathrm{meas}}}$ independently from $\pi_{k,\xi}$. In this conditional experiment, suppose all evaluation rectangles cover their pairs $(L_k(C_j), D_k(C_j))$ with simultaneous probability at least $1 - \beta_{\mathrm{rect}}$. The target sets $\{(l, d) \in \mathbb{R}^2 : l \ge \lambda, d \le \delta\}$, indexed by (λ, δ), form a VC class of constant dimension; let $r_{\mathrm{VC}}(N_{\mathrm{meas}}, \beta_{\mathrm{emp}})$ be any valid uniform empirical-process radius for this class. Then, with probability at least $1 - \beta_{\mathrm{rect}} - \beta_{\mathrm{emp}}$, the bounds Eq.* (13) *hold simultaneously for every target T:*

$$\max\{0, \widehat{P}_{\mathrm{in}}(T) - r_{\mathrm{VC}}\} \le P_k(T \mid \mathcal{H}_k, \xi) \le \min\{1, \widehat{P}_{\mathrm{out}}(T) + r_{\mathrm{VC}}\}.$$

Proof of Theorem 6. Write $L_j = L_k(C_j)$ and $D_j = D_k(C_j)$. Let $I_{\mathrm{in}}(j, T) := \mathbf{1}\{\underline{L}_j \ge \lambda, \overline{D}_j \le \delta\}$ and $I_{\mathrm{out}}(j, T) := \mathbf{1}\{\overline{L}_j \ge \lambda, \underline{D}_j \le \delta\}$ denote the indicators averaged in Definition 3. For a failed output, the true, inner, and outer indicators are all zero because its rectangle is $\{(0, 0)\}$ and $\lambda > 0$. On the simultaneous rectangle event, for every recorded output j and every target T,

$$I_{\mathrm{in}}(j, T) \le \mathbf{1}\{L_j \ge \lambda, D_j \le \delta\} \le I_{\mathrm{out}}(j, T).$$

Consequently, on this one rectangle event, their sample averages satisfy

$$\widehat{P}_{\mathrm{in}}(T) \le \frac{1}{N_{\mathrm{meas}}} \sum_j \mathbf{1}\{L_j \ge \lambda, D_j \le \delta\} \le \widehat{P}_{\mathrm{out}}(T) \qquad \text{for all } T.$$

Conditional on $(\mathcal{H}_k, \Xi_k = \xi)$, the latent pairs (L_j, D_j) are i.i.d. pushforwards of $C_j \sim \pi_{k,\xi}$. The target sets are lower-right orthants in $\mathbb{R}^2$, a VC class of constant dimension, so the empirical-process event gives

$$\sup_T \left| \frac{1}{N_{\mathrm{meas}}} \sum_j \mathbf{1}\{L_j \ge \lambda, D_j \le \delta\} - P_k(T \mid \mathcal{H}_k, \xi) \right| \le r_{\mathrm{VC}}.$$

On the intersection of these two events, the sample-average inequalities and uniform deviation give both bounds in Eq. (13) for every T. A union bound allocates $\beta_{\mathrm{rect}} + \beta_{\mathrm{emp}}$; no union over targets is needed.

Precision. Before truncation to $[0, 1]$, the generation-probability interval in Eq. (13) has width

$$\big(\widehat{P}_{\mathrm{out}}(T) - \widehat{P}_{\mathrm{in}}(T)\big) + 2r_{\mathrm{VC}}.$$

The first term is the fraction of all sampled terminal outputs whose rectangles leave qualification unresolved; it depends on evaluation precision and on the distribution of effects near the target boundary. The second is generation-sampling uncertainty. At fixed confidence a larger generation sample reduces the empirical radius, while requiring all evaluation intervals to contain their true effects with the stated probability can make these intervals wider.

Uniformity covers the specified targets, including targets selected after the data. The conditional and marginal scopes of task-stream measurement are stated in Proposition 25 in Appendix F.3.

Definition 18. *For a candidate pool $C_{1:N}$, conditional on $(\mathcal{H}_k, \xi)$, let G be a possibly pool-dependent validation rule, with $G_i(T) = 0$ when $C_i = \perp$. The expected proportion of terminal outputs that are both qualified and validated is*

$$Q^G_{k,N}(T) := \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{C_i \in \mathcal{P}_k(T), G_i(T) = 1\} \,\middle|\, \mathcal{H}_k, \xi \right].$$

The probability that the pool contains at least one such modification is

$$H^G_{k,N}(T) := \mathbb{P}[\exists i : C_i \in \mathcal{P}_k(T), G_i(T) = 1 \,|\, \mathcal{H}_k, \xi].$$

Write $R^{G,\sigma}_{k,N}(T)$ for the conditional probability that selector σ selects a member of $\mathcal{P}_k(T)$ from this pool, with the conventions of Definition 10.

For selectors restricted to validated modifications, $H - R$ measures the loss due to selection when a qualified validated modification is available. For conditionally i.i.d. generation, $P - Q$ is the expected fraction of all outputs that are qualified but not validated; it differs from the fraction whose qualification remains unresolved by the evaluation intervals. Shared evaluation data can make validation events dependent, so Q alone does not determine H.

For one generated modification, write $\rho^G_k(T) := \mathbb{P}[G(T) = 1 \mid \mathcal{H}_k, \xi]$ for its validation probability.

Proposition 24. *Fix $(\mathcal{H}_k, \Xi_k = \xi)$ and a target T, and draw $C \sim \pi_{k,\xi}$. Let G be the exact-target rule Eq.* (59) *with $N = 1$. Suppose its evaluation rectangle covers $(L_k(C), D_k(C))$ with probability at least $1 - \beta_{\mathrm{val}}$ conditional on $(\mathcal{H}_k, \xi, C)$, and its total widths satisfy the pre-draw deterministic bounds Eq.* (61) *on every evaluation outcome, including any vacuous rectangle recorded on nonreturn. Let P^+_k be the probability assigned to the set in Eq.* (7) *defined by those same bounds. Then*

$$P^+_k(T \mid \mathcal{H}_k, \xi) - \beta_{\mathrm{val}} \le \rho^G_k(T) \le P_k(T \mid \mathcal{H}_k, \xi) + \beta_{\mathrm{val}}. \tag{109}$$

Proof of Proposition 24. Fix $(\mathcal{H}_k, \Xi_k = \xi)$ and suppress this conditioning. Let $\mathcal{E}$ be the event that the rectangle for C covers $(L_k(C), D_k(C))$. On $\mathcal{E}$, the validation rule in Eq. (59) validates only qualified modifications. The total-width bounds imply

$$\underline{L}(C) \ge L_k(C) - \bar{w}_L, \qquad \overline{D}(C) \le D_k(C) + \bar{w}_D.$$

Hence every sampled member of the set in Eq. (7) is validated on $\mathcal{E}$, giving

$$\mathbf{1}\{C \in \mathcal{P}_k^+(T)\} \le G(T) \le \mathbf{1}\{C \in \mathcal{P}_k(T)\} \qquad \text{on } \mathcal{E}.$$

Using $\mathbb{P}(\mathcal{E} \mid C) \ge 1 - \beta_{\mathrm{val}}$ and integrating over C yields

$$(1 - \beta_{\mathrm{val}})P_k^+ \le \rho_k^G \le P_k + \beta_{\mathrm{val}}.$$

Since $(1 - \beta_{\mathrm{val}})P_k^+ \ge P_k^+ - \beta_{\mathrm{val}}$, Eq. (109) follows.

A high validation probability lower-bounds reachability up to β_{val}; a low value can reflect either low reachability or low probability of validating qualified modifications.

F.2 Fixed-suite effects and rollout bounds

Definition 19 (Fixed evaluation suite). *A fixed evaluation suite consists of a finite nonempty set $F^{\mathrm{ev}} \subseteq \mathcal{T}_{user}$ of failure tasks and a finite nonempty set $R^{\mathrm{ev}} \subseteq \mathcal{T}_{user}$ of retained tasks, both disjoint from F_k^{gen}, the failure batch encoded into $t_{F_k^{\mathrm{gen}}}$ and used to generate the modification.*

Definition 20. *For a candidate agent $\tilde{A} = A_k \oplus \Delta C$ write*

$$L^{\mathrm{ev}}(\tilde{A}) := \frac{1}{|F^{\mathrm{ev}}|} \sum_{t \in F^{\mathrm{ev}}} \mathrm{Adv}_{A_k}(\tilde{A}, t), \qquad D_R^{\mathrm{ev}}(\tilde{A}) := \frac{1}{|R^{\mathrm{ev}}|} \sum_{t \in R^{\mathrm{ev}}} \big|\mathrm{Adv}_{A_k}(\tilde{A}, t)\big|, \tag{110}$$

and write $L^{\mathrm{ev}}(c), D_R^{\mathrm{ev}}(c)$ for these quantities at $\tilde{A} = A_k \oplus c$ when $c \in \mathcal{M}$. Set $L^{\mathrm{ev}}(\perp) = D_R^{\mathrm{ev}}(\perp) = 0$. For a fixed-suite target $T^{\mathrm{ev}} = (\lambda^{\mathrm{ev}}, \delta^{\mathrm{ev}})$ with $\lambda^{\mathrm{ev}} > 0$ and $\delta^{\mathrm{ev}} \ge 0$, let $\mathcal{P}_k^{\mathrm{ev}}(T^{\mathrm{ev}}) := \{c \in \mathcal{M} : L^{\mathrm{ev}}(c) \ge \lambda^{\mathrm{ev}} \wedge D_R^{\mathrm{ev}}(c) \le \delta^{\mathrm{ev}}\}$ and

$$P_k^{\mathrm{ev}}(T^{\mathrm{ev}} \mid \mathcal{H}_k, \xi) := \pi_{k,\xi}(\mathcal{P}_k^{\mathrm{ev}}(T^{\mathrm{ev}})).$$

Write $w := |F^{\mathrm{ev}}|/(|F^{\mathrm{ev}}| + |R^{\mathrm{ev}}|)$ for the **failure fraction of the suite** — a quantity chosen by the system designer — and

$$J^{\mathrm{ev}}(A) \;:=\; w \cdot \frac{1}{|F^{\mathrm{ev}}|} \sum_{t \in F^{\mathrm{ev}}} V(A,t) \;+\; (1-w) \cdot \frac{1}{|R^{\mathrm{ev}}|} \sum_{t \in R^{\mathrm{ev}}} V(A,t)$$

for the agent's average reward on the fixed suite.

Remark 14. The two nonempty task sets are specified by the evaluation design; they may overlap. Their names do not require deterministic failure or success of the current agent on each task. If they overlap, J^{ev} is the average weighted by each task's group occurrences, not the average over their deduplicated union. The quantity L^{ev} is an unweighted failure-group mean, while $L = Z_k \mu_F$ includes the failure-rate weight. Each target and its reachability is defined using its respective effects.

Remark 15. If one suite is reused for adoption across steps, later states and generation may depend on it. Conditional on the full history, context, suite and current generation material, Theorem 6 applies to the suite-specific effects when conditional i.i.d. generation and simultaneous rectangle coverage hold. The resulting probability describes that realized state and suite.

The suite effects in Definition 20 also give the fixed-suite improvement guarantee of Theorem 12 in Appendix F.4.

Assumption 10 (Cross-task rollout independence). *For each evaluated agent, the fresh rollouts used by the pooling and measurement constructions are mutually independent within and across tasks, conditional on that agent and the evaluation tasks.*

The constructions below additionally use independent candidate and current-agent fresh-rollout samples. This is local to these rollout bounds; the per-task union-bound analysis of Lemma 5 does not require cross-task independence.

Linear means and shared current-agent evaluation. Let each valid candidate use m rollouts per task and the current agent use m_0. Under the stated independent-rollout design, Hoeffding applied to the two pooled means gives

$$\varepsilon_L^{\mathrm{ev}} = \sqrt{\frac{\log(8/\beta_{\mathrm{cls}})}{2m|F^{\mathrm{ev}}|}} + \sqrt{\frac{\log(8/\beta_{\mathrm{cls}})}{2m_0|F^{\mathrm{ev}}|}}. \tag{111}$$

The candidate and current-agent pooled means each consist of independent bounded rewards, so their deviations are bounded by the two displayed summands. The triangle inequality bounds the error of their difference by their sum. The current-agent mean can be evaluated once and shared across candidates; its error must be included in each candidate's rectangle to ensure that all rectangles contain their corresponding true values with the stated probability.

Absolute differences. Write $\widehat{\delta}_j = \widehat{V}(\widetilde{A}, t_j) - \widehat{V}(A_k, t_j)$. Let σ_V^2 bound both agents' per-task reward variances uniformly over all pairs evaluated by the design, and put $v \geq |R^{\mathrm{ev}}|^{-1} \sum_j \mathrm{Var}(\widehat{\delta}_j)$. Independence between the two agents' samples gives

$$\mathrm{Var}(\widehat{\delta}_j) \leq \sigma_V^2(1/m + 1/m_0).$$

Proposition 11 and Cauchy–Schwarz yield

$$0 \leq \mathbb{E}|\widehat{\delta}_j| - |\delta_j| \leq \sqrt{\mathrm{Var}(\widehat{\delta}_j)} \leq \sigma_V \sqrt{1/m + 1/m_0}.$$

Indeed, writing $\widehat{\delta}_j = \delta_j + (\widehat{\delta}_j - \delta_j)$, Jensen gives the nonnegative left-hand difference, while the triangle inequality and Cauchy–Schwarz bound it by $\mathbb{E}|\widehat{\delta}_j - \delta_j| \leq \sqrt{\mathrm{Var}(\widehat{\delta}_j)}$. Moreover $\mathrm{Var}\,|X| \leq \mathrm{Var}\,X$, since for an independent copy X', $\frac{1}{2}\mathbb{E}(|X| - |X'|)^2 \leq \frac{1}{2}\mathbb{E}(X - X')^2$. Thus v also bounds the average true variances of the absolute differences. Bernstein concentration for the independent fixed-task centered absolute differences bounds their random average around its expectation; adding the Jensen bias gives the valid radius

$$\varepsilon_D^{\mathrm{ev}} = \sigma_V \sqrt{1/m + 1/m_0} + \sqrt{\frac{2v \log(8/\beta_{\mathrm{cls}})}{|R^{\mathrm{ev}}|}} + \frac{7 \log(8/\beta_{\mathrm{cls}})}{3|R^{\mathrm{ev}}|}. \tag{112}$$

The usual bounded-variable Bernstein additive allowance $2\log(8/\beta_{\mathrm{cls}})/(3|R^{\mathrm{ev}}|)$ is no larger than the displayed allowance. The task set is fixed, so the concentration is over rollout noise around the average task-specific expectations. The per-task Jensen allowance decreases with the per-task rollout budgets; averaging more tasks alone does not reduce that allowance.

Finite evaluation levels. Fix rollout budgets $m_1 < \cdots < m_K$ in advance and allocate β_{cls}/K to each. If E_h is the coverage event at level h, then $\Pr(\bigcap_{h=1}^K E_h) \geq 1 - \sum_{h=1}^K \beta_{\mathrm{cls}}/K = 1 - \beta_{\mathrm{cls}}$. Any data-dependent choice of one of these levels, including stopping when its rectangle determines membership in a queried target, therefore retains coverage. In the displayed radii this allocation replaces $\log(8/\beta_{\mathrm{cls}})$ by $\log(8K/\beta_{\mathrm{cls}})$. Other targets retain the chosen rectangle's inner and outer indicators.

F.3 Measurement from randomly sampled tasks

Definition 21. *Fix a stored-sample size n_R before drawing the candidate pool. After that pool has been frozen, collect a disjoint measurement sample $\{(t_j^R, \phi_j^R)\}_{j=1}^{n_R}$.*

Its two label counts are $N_1 := \sum_{j=1}^{n_R} \phi_j^R$ and $N_0 := n_R - N_1$. The measurement sample is not used during generation and is never used for a validation decision. In the independent reference experiment, conditional on $(\mathcal{H}_k, \Xi_k = \xi)$ and the frozen candidate pool, the records are independent with $t_j^R \overset{\mathrm{iid}}{\sim} \mathcal{D}_{\mathrm{user}}$ and $\phi_j^R \mid t_j^R \sim \mathrm{Bernoulli}(\psi_{A_k}(t_j^R))$; the labels are conditionally independent across records. Write $(F^{\mathrm{dr}}, R^{\mathrm{dr}})$ for its two subsets, where F^{dr} contains all N_1 records with $\phi_j^R = 1$ and R^{dr} all N_0 records with $\phi_j^R = 0$; thus $\widehat{Z}_k = N_1/n_R$. Conditional on the label vector, classification of valid modifications uses count-adaptive radii. If $N_1 = 0$ or $N_0 = 0$, the

corresponding coordinate for a valid modification receives its full feasible interval — $[-1, 1]$ for L and $[0, 1]$ for D — and the procedure abstains from nonvacuous classification on that coordinate. Failed outputs retain their known singleton rectangles.

The protocol fixes n_R before sampling and uses every sampled record. Its content is drawn from the task stream; Definition 19 instead fixes the suite's task content.

Fix n_R before drawing the candidate pool, draw $C_1, \dots, C_{N_{\mathrm{meas}}} \overset{\mathrm{iid}}{\sim} \pi_{k,\xi}$ conditional on $(\mathcal{H}_k, \Xi_k = \xi)$, freeze that pool, and then evaluate its valid modifications on the disjoint measurement sample of Definition 21. Failed outputs use the singleton rectangle $\{(0, 0)\}$. For valid modifications, on $N_1, N_0 > 0$, use the count-adaptive radii

$$\varepsilon_L^{\mathrm{dr}} = \varepsilon_Z + (\hat{Z}_k + \varepsilon_Z)\big(\varepsilon_V^{\mathrm{pool}} + \varepsilon_F^{\mathrm{dr}}\big), \qquad \varepsilon_D^{\mathrm{dr}} = \varepsilon_V^{\mathrm{Jen}} + \varepsilon_R^{\mathrm{dr}}, \tag{113}$$

where

$$\varepsilon_F^{\mathrm{dr}}(N_1) = \sqrt{\frac{2\log(8/\beta_{\mathrm{cls}})}{N_1}}, \quad \varepsilon_R^{\mathrm{dr}}(N_0) = \sqrt{\frac{\log(8/\beta_{\mathrm{cls}})}{2N_0}}, \quad \varepsilon_Z = \sqrt{\frac{\log(8/\beta_{\mathrm{cls}})}{2n_R}}.$$

Here $\varepsilon_V^{\mathrm{pool}}$ is the radius in Eq. (111), evaluated at the realized count, and $\varepsilon_V^{\mathrm{Jen}} = \sigma_V\sqrt{1/m + 1/m_0}$ is the non-pooling Jensen bias of Eq. (112). If either count is zero, use a vacuous rectangle on the corresponding coordinate.

Proposition 25. *Under this measurement setup, if $N_{\mathrm{meas}}\beta_{\mathrm{cls}} \le \beta_{\mathrm{rect}}$, then:*

1. *when, conditional on $(\mathcal{H}_k, \Xi_k = \xi)$, the frozen pool is independent of the audit records and those records are independent draws from the task-and-label law of Definition 21, the uniform confidence bounds*

$$\max\{0, \widehat{P}_{\mathrm{in}}(T) - r_{\mathrm{VC}}\} \le P_k(T \mid \mathcal{H}_k, \xi) \le \min\{1, \widehat{P}_{\mathrm{out}}(T) + r_{\mathrm{VC}}\} \qquad \text{for every } T \tag{114}$$

   hold conditionally on $(\mathcal{H}_k, \Xi_k = \xi)$ with probability at least $1 - \beta_{\mathrm{rect}} - \beta_{\mathrm{emp}}$;

2. *if the joint law of history, context, frozen pool, all audit records and their evaluation marks is within total variation β_{mix} of a reference experiment with the same pre-audit marginal and the independent audit law in the first item, then Eq.* (114)*, with the random realized $(\mathcal{H}_k, \Xi_k)$ on its right-hand side, holds with joint marginal probability at least $1 - \beta_{\mathrm{rect}} - \beta_{\mathrm{emp}} - \beta_{\mathrm{mix}}$.*

The first item is also supplied by a reset followed by an independent task-and-label generator for every record, or by an m-dependent stream with all selected records spaced beyond its dependence range and separated from the pre-audit sigma-field. A reset alone removes dependence on pre-reset history but does not make later records mutually independent. For these stream-based examples, fix a deterministic task cutoff b: the pre-audit history, context, current agent, and frozen pool are determined by tasks through b and auxiliary randomness independent of the task stream. Audit indices after b are fixed in advance, or their complete schedule is sampled independently of the task stream; the schedule is not changed in response to observed audit tasks or labels. Conditional on the selected tasks and frozen agent, stored labels are generated independently by the Bernoulli kernels in Definition 21, using randomness separate from the task stream and pre-audit information; fresh evaluation uses its prescribed independent rollout kernels. The second item averages over history, context, pool, and audit data. For a β-mixing stream with fixed audit indices $s_1 < \cdots < s_{n_R}$ and first gap $g_0 = s_1 - b$, the joint-TV product comparison permits

$$\beta_{\mathrm{mix}} \le \beta(g_0) + \sum_{j=1}^{n_R - 1} \beta(s_{j+1} - s_j). \tag{115}$$

If all these gaps are at least ℓ, this is at most $n_R\beta(\ell)$; the same bound holds for an independent random schedule satisfying those gaps, by conditioning on the schedule. If a reset starts a fresh stationary stream independent of the pre-audit information, it removes only the first term and gives $(n_R - 1)\beta(\ell)$.

Proof of Proposition 25. We first prove the result in the conditional reference experiment. There it is legitimate to condition on $(\mathcal{H}_k, \Xi_k = \xi)$, which fixes A_k, $\pi_{k,\xi}$, Z_k, $\mathcal{D}_{F,k}$, and $\mathcal{D}_{R,k}$. The joint-TV statement is transferred after this conditional proof, under the joint distribution.

Step 1: the reference sample supplies random-size samples from both conditional distributions. In the reference experiment, $t_j^R \overset{\text{iid}}{\sim} \mathcal{D}_{\text{user}}$, and the stored label satisfies $\phi_j^R \mid t_j^R \sim \text{Bernoulli}(\psi_{A_k}(t_j^R))$ by Eq. (22). For measurable B,

$$\mathbb{P}\big(t_j^R \in B \mid \phi_j^R = 1\big) = \frac{\int_B \psi_{A_k}\, d\mathcal{D}_{\text{user}}}{\int \psi_{A_k}\, d\mathcal{D}_{\text{user}}} = \frac{\int_B \psi_{A_k}\, d\mathcal{D}_{\text{user}}}{Z_k} = \mathcal{D}_{F,k}(B)$$

by Eq. (23) and Eq. (1), and the $\phi_j^R = 0$ computation gives $\mathcal{D}_{R,k}$ likewise. Both require $Z_k \in (0,1)$, the standing assumption of §2.2. Conditional on the full label vector, the tasks factorise: the N_1 failure-labelled tasks are i.i.d. from $\mathcal{D}_{F,k}$, the N_0 retained-labelled tasks are i.i.d. from $\mathcal{D}_{R,k}$, and the two groups are independent. Moreover $N_1 \sim \text{Binomial}(n_R, Z_k)$ and $\widehat{Z}_k = N_1/n_R$.

$$\Pr(t_j^R \in dt \mid \phi_j^R = 1) = \mathcal{D}_{F,k}(dt), \qquad \Pr(t_j^R \in dt \mid \phi_j^R = 0) = \mathcal{D}_{R,k}(dt). \tag{116}$$

Step 2: which implementations inherit the reference experiment. An independent audit generator, including one started after a reset, gives Step 1 directly. An m-dependent stream sampled at gaps exceeding m and separated by such a gap from the pre-audit sigma-field also gives exact independence. A reset without independent post-reset sampling does not.

For ordinary β-mixing under the cutoff and advance-scheduling conditions above, compare the whole joint experiment to the reference measure retaining the same pre-audit history, context and pool, with conditionally independent reference audit tasks. Successive product-measure comparisons over the initial and record-to-record gaps give Eq. (115) for fixed indices. For an independent random schedule, condition on its realized indices and average the comparison; a common lower gap bound gives the displayed $n_R\beta(\ell)$ allowance. Applying the common independent marking and fresh-evaluation kernels does not increase total variation. Each reference event therefore transfers to the actual joint experiment with one loss of at most β_{mix}. A supplied whole-window coupling with that defect is also sufficient. This argument gives marginal event bounds.

Conditional on the reference label vector, the random-size groups earn the $q_F = 1$ radii: no stopping rule has selected their sizes, and no blocking factor appears inside either conditional mean. If a group is empty, assigning the corresponding coordinate its full feasible interval makes the rectangle valid without claiming informativeness.

Step 3: per-modification radii. Failed outputs have deterministic rectangle coverage. Fix a valid drawn modification $\Delta C_i \in \mathcal{M}$; condition on the frozen pool in the stated reference experiment, so $\text{Adv}_{A_k}(\tilde{A}_i, \cdot)$ is a fixed function on the measurement sample. Conditional on the label vector:

(a) L. On $N_1 > 0$, $\frac{1}{N_1}\sum_{t \in F^{\text{dr}}} \text{Adv}_{A_k}(\tilde{A}_i, t)$ is, conditional on the label vector, an i.i.d. mean of a range-2 variable, so Hoeffding gives $\varepsilon_F^{\text{dr}}(N_1) = \sqrt{2\log(8/\beta_{\text{cls}})/N_1}$. There are mN_1 candidate rewards and $m_0 N_1$ current-agent rewards. Given the tasks and frozen pool, their independent signed weighted sum has squared range lengths totaling $a + c$, where $a = (mN_1)^{-1}$ and $c = (m_0 N_1)^{-1}$. For $h = \log(8/\beta_{\text{cls}})$ its direct two-sided Hoeffding radius $\sqrt{h(a+c)/2}$ has failure risk $\beta_{\text{cls}}/4$ and is at most the retained radius $\sqrt{ha/2} + \sqrt{hc/2} = \varepsilon_V^{\text{pool}}$. Thus the entire pooled difference uses one $\beta_{\text{cls}}/4$ event. Hence $|\hat{\mu}_F^{(i)} - \mu_F^{(i)}| \le \varepsilon_V^{\text{pool}} + \varepsilon_F^{\text{dr}}(N_1)$. Since $\hat{L}^{(i)} = \hat{Z}_k \hat{\mu}_F^{(i)}$ and $L^{(i)} = Z_k \mu_F^{(i)}$, the decomposition in the proof of Lemma 5 applies verbatim — $|\hat{Z}_k - Z_k|\,|\hat{\mu}_F| \le \varepsilon_Z$ against $|\hat{\mu}_F| \le 1$, plus $Z_k|\hat{\mu}_F - \mu_F|$ bounded on $\mathcal{E}_Z$ by $(\hat{Z}_k + \varepsilon_Z)(\varepsilon_V^{\text{pool}} + \varepsilon_F^{\text{dr}}(N_1))$ — which is the first display of Eq. (113). For $N_1 = 0$ the declared vacuous interval for L covers by construction.

(b) D_R. On $N_0 > 0$, $\hat{D}^{(i)} = \frac{1}{N_0}\sum_{t \in R^{\text{dr}}} |\hat{\delta}_i(t)|$ is, conditional on the label vector, an i.i.d. mean of a range-1 variable over $\mathcal{D}_{R,k}$, which by Definition 7 is the measure D_R is defined against. Proposition 11(i)–(ii) gives $|\delta_i(t)| \le \mathbb{E}|\hat{\delta}_i(t)| \le |\delta_i(t)| + \sigma_V\sqrt{1/m + 1/m_0}$ pointwise, so $\mathbb{E}[\hat{D}^{(i)}] \in [D_R^{(i)},\, D_R^{(i)} + \varepsilon_V^{\text{Jen}}]$, and Hoeffding around that mean gives $\varepsilon_R^{\text{dr}}(N_0) = \sqrt{\log(8/\beta_{\text{cls}})/(2N_0)}$. Hence $|\hat{D}^{(i)} - D_R^{(i)}| \le \varepsilon_V^{\text{Jen}} + \varepsilon_R^{\text{dr}}(N_0)$, the second display of Eq. (113). For $N_0 = 0$ the vacuous interval for D_R covers by construction.

Step 4: union, confidence bounds, and transfer. The failure-task mean, the complete pooled rollout difference, the retained absolute-difference mean, and the failure-rate estimate each have failure risk at most $\beta_{\mathrm{cls}}/4$. The first three statements are integrated over the random label vector; the failure-rate event is bounded in the original reference sampling experiment. Hence each candidate rectangle has failure risk at most β_{cls}. A union over the N_{meas} modification-wise rectangle events costs at most $N_{\mathrm{meas}}\beta_{\mathrm{cls}} \le \beta_{\mathrm{rect}}$; the shared $\widehat{Z}_k$ event need only be counted once, so this allocation is conservative. On their intersection, the inside/true/outside indicator containment of Theorem 6 holds simultaneously for all targets. The latent pairs are i.i.d. pushforwards of $C_i \sim \pi_{k,\xi}$, so the VC event supplies the uniform empirical radius r_{VC} at error probability β_{emp}. This proves the first claim conditionally in each of the exact reference implementations. In the ordinary-mixing case, transfer this *joint event* through the joint-TV comparison of Step 2, subtracting β_{mix} once. The result is marginal coverage of the random conditional reachability, not a pointwise-in-history guarantee.

F.4 Additional fixed-suite guarantees

The fixed-suite validation rule validates a modification when $\hat{L}^{\mathrm{ev}} \ge \tau$ and $\hat{D}_R^{\mathrm{ev}} \le \delta$. Let $\mathcal{E}_{\mathrm{suite}}$ be the joint event that, for every modification evaluated by this rule, $|\hat{L}^{\mathrm{ev}} - L^{\mathrm{ev}}| \le \varepsilon_L^{\mathrm{ev}}$ and $|\hat{D}_R^{\mathrm{ev}} - D_R^{\mathrm{ev}}| \le \varepsilon_D^{\mathrm{ev}}$. The following result uses radii for which $\mathbb{P}(\mathcal{E}_{\mathrm{suite}}) \ge 1 - \beta_{\mathrm{gate}}$.

Theorem 12 (Fixed-suite guarantee). *For every candidate agent $\widetilde{A}$,*

$$J^{\mathrm{ev}}(\widetilde{A}) - J^{\mathrm{ev}}(A_k) \ge wL^{\mathrm{ev}}(\widetilde{A}) - (1-w)D_R^{\mathrm{ev}}(\widetilde{A}). \tag{117}$$

On $\mathcal{E}_{\mathrm{suite}}$, every modification validated by this validation rule satisfies

$$J^{\mathrm{ev}}(\tilde{A}) > J^{\mathrm{ev}}(A_k) \qquad \textit{whenever} \qquad \tau \;>\; \varepsilon_L^{\mathrm{ev}} + \frac{1-w}{w}\big(\delta + \varepsilon_D^{\mathrm{ev}}\big). \tag{118}$$

Proof of Theorem 12. Splitting the suite average by Eq. (110),

$$J^{\mathrm{ev}}(\tilde{A}) - J^{\mathrm{ev}}(A_k) = w\,L^{\mathrm{ev}} + (1-w)\cdot\frac{1}{|R^{\mathrm{ev}}|}\sum_{t\in R^{\mathrm{ev}}}\mathrm{Adv}_{A_k}(\tilde{A},t),$$

and the second sum is bounded below by $-\frac{1}{|R^{\mathrm{ev}}|}\sum|\mathrm{Adv}| = -D_R^{\mathrm{ev}}$, giving Eq. (117). On the validated event $\hat{L}^{\mathrm{ev}} \ge \tau$ and $\hat{D}_R^{\mathrm{ev}} \le \delta$; on $\mathcal{E}_{\mathrm{suite}}$, $L^{\mathrm{ev}} \ge \tau - \varepsilon_L^{\mathrm{ev}}$ and $D_R^{\mathrm{ev}} \le \delta + \varepsilon_D^{\mathrm{ev}}$. Hence Eq. (117) gives

$$J^{\mathrm{ev}}(\tilde{A}) - J^{\mathrm{ev}}(A_k) \ge w(\tau - \varepsilon_L^{\mathrm{ev}}) - (1-w)(\delta + \varepsilon_D^{\mathrm{ev}}),$$

which is positive precisely under Eq. (118).

G Experimental Settings and Supplementary Results

G.1 DS-1000 repeated evaluation and diagnosis

Setting. We study persistent Python solver-harness modifications on DS-1000 under three generation processes: independent generation, revision with label-free traces, and revision with outcome feedback. Each modification edits the solver harness while the LLM remains fixed. Label-free revision uses generated solutions and runtime status without correctness labels; outcome-feedback revision additionally receives pass/fail labels on generation tasks. Generation has access to the public task prompt and these traces, but not hidden answers, validation records, or audit results.

We use the API model `gpt-5.4-mini-2026-03-17`. Generation uses temperature 0.8 and an 8192-token response limit; the seed harness makes one solver call at temperature 0 with a 4096-token limit. Modifications may change solver prompts, decoding parameters, extraction, call count, and orchestration, so candidates are evaluated under their resulting execution policies rather than an equalized inference budget.

The task assignment was fixed before inspecting outcomes. Of the 1000 DS-1000 tasks, 516 used in the pilot studies are excluded; 480 of the remaining 484 are assigned once, with library composition proportional to the remaining task bank. Each of four contexts contains 8 generation tasks, 8 low-budget validation tasks, 48 high-budget validation tasks, and 56 independent audit tasks. For each

context and generation process, six independent pools each contain four terminal outputs. The same pools are assessed using both validation budgets. The primary target is $T_0 = (0.05, 0.50)$, and the prespecified target grid is $\{0.05, 0.10\} \times \{0.25, 0.50\}$.

We repeat the audit using three fresh valid runs for every current-agent–task and valid-candidate–task pair, keeping the candidates and task assignments fixed. The original current-agent outcomes determine the failure and retained groups; they are not included in the new three-run means. Within each context, all candidates share the same three-run current-agent reference. The repeated audit contains 48,720 valid outcomes: $4 \times 56 \times 3$ for the current agent and $286 \times 56 \times 3$ for valid candidates. Wrong answers count as valid failures; infrastructure-invalid runs are replenished according to outcome-independent validity criteria. Two generated outputs are invalid and remain in the generation denominators with zero target-membership and validation indicators.

Repeated audit estimates. Fix one context and a valid candidate. Let F and R be the failure and retained groups determined by the original current-agent outcomes, with $n = |F| + |R| = 56$ and $w = |F|/n$. For task t, write $X_{tr}, Y_{tr} \in \{0, 1\}$ for the success outcomes of the current and candidate agents on fresh run $r \in \{1, 2, 3\}$, and define their means $\bar{x}_t = 3^{-1} \sum_{r=1}^{3} X_{tr}$ and $\bar{y}_t = 3^{-1} \sum_{r=1}^{3} Y_{tr}$. Equation (14) gives the point estimates. Conditional on the fixed tasks, agents, and original groups, let p_t and q_t denote the two agents' success probabilities. The expected-reward effects are

$$L = \frac{1}{n} \sum_{t \in F} (q_t - p_t), \qquad D = \frac{1}{|R|} \sum_{t \in R} |q_t - p_t|, \qquad \Delta = \frac{1}{n} \sum_{t} (q_t - p_t). \tag{119}$$

Thus $L = wL^{\mathrm{ev}}$ and $D = D_R^{\mathrm{ev}}$ for the fixed-suite effects in Definition 20, Appendix F.2. The target $L \geq 0.05$ uses the weighted contribution; its unweighted equivalent is $L^{\mathrm{ev}} \geq 0.05/w$. The confidence analysis assumes fresh runs are independent within and across tasks and between the two agents, conditional on the fixed tasks, agents, and original records. Candidates may share current-agent observations; the joint bounds below do not require independence between candidates.

Original validation and adoption decisions. Validation retains its original data and rule. For a valid candidate, let $x_i, y_i \in \{0, 1\}$ be the original current-agent and candidate outcomes on validation task i. With n_{v} validation tasks and $n_{R,\mathrm{v}} = \sum_i x_i$, the rule uses the observed improvement proportion $n_{\mathrm{v}}^{-1} \sum_i (1 - x_i) y_i$ and retained-loss proportion $n_{R,\mathrm{v}}^{-1} \sum_i x_i (1 - y_i)$. If $n_{R,\mathrm{v}} = 0$, the retained-loss point value is zero and its interval is $[0, 1]$. Clopper–Pearson intervals are applied to these proportions; a Hoeffding interval of range width two is also computed for mean reward change. The familywise risk is 0.05 per four-output pool, divided by three times the output count, including invalid outputs. At target $T = (\lambda, \delta)$, the rule validates a modification when the improvement lower endpoint reaches λ and the retained-loss upper endpoint is at most δ. Among modifications passing the high-budget rule, selection maximizes observed mean reward change on the validation tasks. Repeated audit results do not alter these recorded decisions.

Point summaries and uncertainty. In the summaries below, write G for the experimental validation rule and σ for its selector based on the recorded validation scores. For output j in pool b, let

$$Y_{bj}(T) = \mathbf{1}\{\mathrm{valid}_{bj},\ \widehat{L}_{bj} \geq \lambda,\ \widehat{D}_{bj} \leq \delta\},$$
$$A_{bj}^{G}(T) = \mathbf{1}\{\underline{L}_{bj}^{G} \geq \lambda,\ \overline{D}_{bj}^{G} \leq \delta\},$$

where $\widehat{L}_{bj}, \widehat{D}_{bj}$ are the repeated-audit estimates, and $\underline{L}_{bj}^{G}, \overline{D}_{bj}^{G}$ are the original validation endpoints under rule G. Both indicators are zero for invalid outputs. In a context–process cell with $B = 6$ pools and $M = 4B = 24$ outputs, compute

$$\widehat{P}_{\mathrm{pt}}(T) = \frac{1}{M} \sum_{b,j} Y_{bj}(T), \qquad \widehat{Q}^{G}(T) = \frac{1}{M} \sum_{b,j} Y_{bj}(T) A_{bj}^{G}(T),$$
$$\widehat{H}^{G}(T) = \frac{1}{B} \sum_{b} \mathbf{1}\{\max_{j} Y_{bj}(T) A_{bj}^{G}(T) = 1\}, \qquad \widehat{R}^{G,\sigma}(T) = \frac{1}{B} \sum_{b} Y_{b,\sigma(b)}(T),$$

where the selected-output indicator is zero if no output is selected. The fraction of pools containing an output meeting the audit point target replaces $Y_{bj}A_{bj}^{G}$ by Y_{bj} in the definition of $\widehat{H}^{G}$. Each reported summary averages the four context values.

For the empirical summary intervals, 5000 hierarchical-bootstrap replicates resample contexts, pools within context, and outputs within pool; each replicate recomputes the pool event and selection. We report 90% percentile intervals using seed 20260815. Validation-budget comparisons use the same resampled pools and outputs for both budgets. These intervals describe variation in the empirical summaries obtained from the three-run means, rather than confidence bounds for the true generation probability. Resampling an all-zero summary produces $[0, 0]$, which does not bound an unseen event probability.

Candidate-level diagnosis. In a retrospective analysis, we select all candidates whose original audit improvement lower bound is at least 0.05 and whose original retained-loss upper bound is at most 0.50. This selects three outcome-feedback candidates, labeled A, B, and C in Table 2. The original audit intervals use the original binary-outcome proportions with risk 0.05 per context–process group of 24 outputs, divided by three times the output count. For this specified screen based on the original records, we use the fresh runs to calculate nine one-sided bounds: a lower bound for L, an upper bound for D, and a lower bound for Δ for each candidate. Each endpoint receives error probability $a = 0.05/9$.

To account for different success probabilities across tasks, consider a sum K of N independent Bernoulli variables with mean success probability $\bar{p}$. Hoeffding's binomial comparison (Hoeffding, 1956, Theorem 4) gives

$$\Pr(K \geq k) \leq \Pr\{\mathrm{Bin}(N, p_0) \geq k\} \quad \text{when } \bar{p} \leq p_0 \text{ and } k \geq Np_0 + 1.$$

Writing $B^{-1}(a; s, t)$ for the a-quantile of the $\mathrm{Beta}(s, t)$ distribution, inversion gives a lower confidence bound for $\bar{p}$,

$$b(k, N, a) = \begin{cases} 0, & k = 0, \\ \min\{B^{-1}(a; k, N-k+1), (k-1)/N\}, & k > 0. \end{cases} \tag{120}$$

The cap $(k-1)/N$ enforces the comparison's integer condition; an upper bound is $1-b(N-k, N, a)$, obtained by applying the same construction to $N - K$. These bounds allow different Bernoulli probabilities rather than requiring identically distributed observations.

For a task group S, put $h_S = 3|S|$ and $K_S = \sum_{t \in S, r}\{Y_{tr} + (1 - X_{tr})\}$. This is a sum of $2h_S$ independent Bernoulli variables with mean success probability $(1 + |S|^{-1}\sum_{t \in S}(q_t - p_t))/2$. With $S = F$ and $S = F \cup R$, respectively, the lower bounds are

$$\underline{L} = w\{2b(K_F, 2h_F, a) - 1\}, \qquad \underline{\Delta} = 2b(K_{F \cup R}, 6n, a) - 1.$$

For retained tasks, pair current-agent and candidate runs by the fixed run index and let $K_D = \sum_{t \in R, r}|Y_{tr} - X_{tr}|$. Since $|q_t - p_t| \leq \mathbb{E}|Y_{tr} - X_{tr}|$, an upper bound on the mean disagreement probability is also an upper bound on D:

$$\overline{D} = 1 - b(3|R| - K_D, 3|R|, a).$$

A union bound over the nine endpoints gives at least 95% joint coverage for this fixed screening procedure, conditional on the original records and the stated independence model. Shared current-agent observations across the three candidates do not affect this guarantee.

Table 2: **Repeated-audit effects for the diagnostic candidates.** Point estimates use three fresh runs per agent and task on the fixed audit set. The nine one-sided bounds allocate total error probability 0.05 across the three candidates selected from the original audit records; displayed endpoints are rounded to the nearest four decimal places. Candidate A meets $L \geq 0.05$, $D \leq 0.50$, and $\Delta > 0$ under these bounds.

Candidate	$\widehat{L}$	Lower bound	$\widehat{D}$	Upper bound	$\widehat{\Delta}$	Lower bound
A	0.1726	0.0551	0.1176	0.2765	0.1964	0.0556
B	0.1071	−0.0113	0.0980	0.3012	0.1012	−0.0402
C	0.1369	0.0266	0.1515	0.3653	0.1607	0.0195

For candidate A, the failure group has 39 tasks: across 117 fresh runs per agent, the current agent succeeds 12 times and the candidate 41 times. The retained group has 17 tasks and 6 disagreements among 51 paired runs. Across all 168 runs per agent, the current agent succeeds 56 times and the candidate 89 times. Bounds stated as inequalities in Section 5.2 are rounded outward to four decimal places to preserve the guarantees.

Candidate A's original high-budget validation improvement estimate is 0.1458, but its lower endpoint is 0.0378, below 0.05; the retained-loss upper endpoint is 0.4196, below 0.50. Its pool adopts no modification. B and C also fail the original validation rule. Thus repeated evaluation supports A's expected-reward improvement on the fixed audit set, and the original decision record identifies the improvement condition that prevented adoption. Validation and audit use different task sets; the example diagnoses this recorded decision and does not estimate reachability over the full user-task distribution.

Supplementary results. At T_0, the repeated-audit point estimates meet the target for $37/96$, $25/96$, and $33/96$ outputs under independent generation, label-free revision, and outcome-feedback revision. The numbers of pools containing such an output are $16/24$, $13/24$, and $16/24$. The high-budget rule validates two independently generated outputs, both meeting the audit point target and both selected from their respective pools. Table 3 gives the full summaries.

Table 3: **DS-1000 summaries from repeated audit estimates at $T_0 = (0.05, 0.50)$.** Target membership uses the mean of three fresh runs per agent and audit task; validation retains the original high-budget decisions. Brackets give 90% hierarchical-bootstrap intervals for the empirical summaries.

Generation process	$\widehat{P}_{\mathrm{pt}}$	$\widehat{Q}$	$\widehat{H}$	$\widehat{R}$
Independent	0.385	0.021	0.083	0.083
	[0.135, 0.646]	[0.000, 0.073]	[0.000, 0.167]	[0.000, 0.167]
Label-free revision	0.260	0.000	0.000	0.000
	[0.031, 0.479]	[0.000, 0.000]	[0.000, 0.000]	[0.000, 0.000]
Outcome-feedback revision	0.344	0.000	0.000	0.000
	[0.125, 0.563]	[0.000, 0.000]	[0.000, 0.000]	[0.000, 0.000]

At the stricter target $T_1 = (0.10, 0.25)$, estimated target membership is $8/96$, $4/96$, and $14/96$, with fractions 0.083 $[0.021, 0.156]$, 0.042 $[0.000, 0.094]$, and 0.146 $[0.021, 0.302]$. The corresponding numbers of pools are $8/24$, $4/24$, and $8/24$; $\widehat{Q}$, $\widehat{H}$, and $\widehat{R}$ are zero for all three processes. Figure 3 presents the context-specific estimates and their comparison with validation outcomes.

H Further Directions for Harness Self-Evolution

Generation and structured updates. The candidate-pool analysis permits complete generation runs with internal feedback and repair (Appendices C.3 and C.3). Adapting later runs to earlier pool outputs or evaluation results raises the further problem of tracking their probabilities of producing qualified modifications. For fixed repair and finite assembly, estimating those output probabilities and exploiting dependence among assembled outputs could sharpen the available bounds for concrete search procedures (Appendix C.6).

After adoption, the modified harness affects user-task performance and the generation and effects of later modifications. Structural relations between these changes could support lower bounds on subsequent reachability under specified update mechanisms; the unrestricted-update construction in Appendix E.3 identifies why such relations matter.

Validation and measurement. The unresolved fraction in the reachability estimate depends on evaluation precision and on candidate effects near the improvement and retained-change requirements (Appendix F.1). Quantifying this dependence could inform allocation between generating candidates and evaluating their effects. A related question is the validation probability of a given qualified modification at fixed error risk and evaluation budget, as its improvement and retained change vary relative to those requirements.

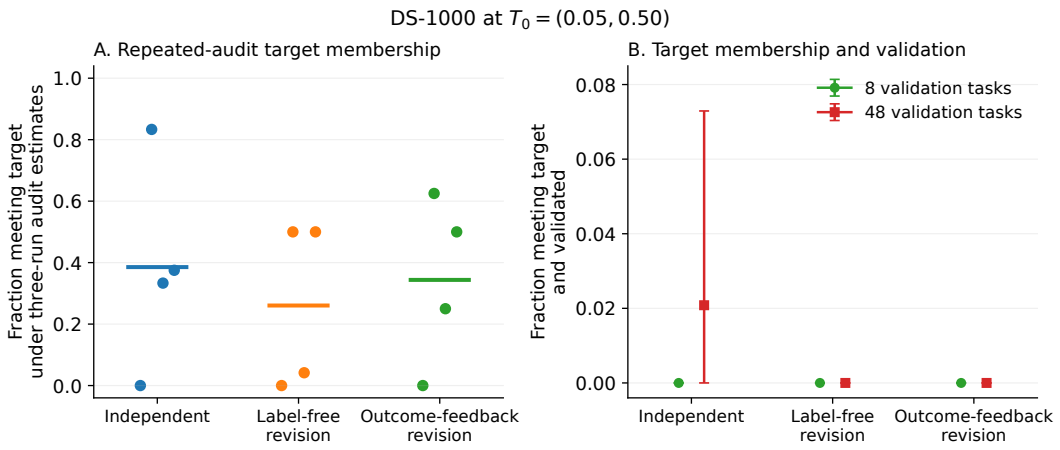


Figure 3: **Repeated evaluation and validation budgets on DS-1000.** At $T_0 = (0.05, 0.50)$, **(A)** shows each context's fraction of outputs meeting the target under three-run audit estimates, with a horizontal line for the mean; **(B)** shows the fraction also validated using 8 or 48 validation tasks. Error bars give 90% hierarchical-bootstrap intervals for the empirical summaries.

Variance-adaptive bounds and pooled rollout estimates (Appendices C.5 and F.2) provide starting points for allocating task samples and repeated rollouts under a fixed budget while controlling error risk and the validation probability of qualified candidates. The same allocation question extends to dependent task samples under the separation and coverage conditions in Appendix C.1.

Repeated use of a fixed evaluation suite can make later agent states and generation depend on that suite. The conditional guarantee for the current state and suite is described in Remark 15. Extending these measurements to fresh tasks requires control of the difference between suite-specific effects and effects under the intended task distribution, with sampling and representativeness conditions matched to that claim. Allocating data between evolution and independent evaluation remains part of this question.

Retention requirements and application conditions. Tail constraints control average absolute expected-reward changes within every group having at least a specified probability under the retained-task distribution (Appendix B.2). Their implementation with finite rollouts and dependent task samples calls for corresponding estimation guarantees and evaluation requirements.

Applications also require ways to obtain the reward–utility discrepancy bound used in Appendix B.5 and to model how user-task demand responds to agent changes. Partition-frequency estimates combined with reward-resolution conditions bound the expected-reward effect of task-distribution differences (Appendix E.4.2); predicting the response before adoption remains an application-modeling problem.