WHICH SELF-IMPROVEMENTS SHOULD WE TRUST? RELIABLE SELF-IMPROVEMENT WHEN AGENTS REUSE THEIR BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

As recursive self-improvement (RSI) rapidly advances, reliable evaluation becomes critical for guiding adaptive search. RSI typically relies on finite evaluation resources, such as fixed benchmarks, to determine which modifications are retained and what is proposed next. However, when these finite resources are repeatedly reused, new candidates are proposed based on feedback from the same evaluation set, so the search trajectory can adaptively overfit and empirical improvement may not reflect genuine population improvement on the underlying task distribution. Some existing methods account for multiple comparisons but assume that candidates are chosen independently of the evaluation set, and therefore do not control this adaptive dependence. To address this, we propose REUSE (**R**isk-controlled **E**valuation **U**nder **S**equential **E**volution), a certified evaluation and promotion framework that allows a fixed evaluation set to support repeated adaptive decisions while providing statistical guarantees. Specifically, for a user-specified error level α, with probability at least $1-\alpha$, every promoted modification is a genuine population improvement on the underlying task distribution. REUSE achieves this by strictly limiting the evaluation feedback returned to the search process and accounting for possible promotion histories within the error budget. We develop a detailed statistical theory for RSI evaluation in this setting, including simultaneous error control, valid lower bounds on cumulative improvement, and a characterization of the fundamental limits of adaptive evaluation reuse. In live self-improvement experiments, REUSE commits substantially fewer false promotions than evaluation frameworks from current RSI systems and error-controlled baselines, reducing the proportion of promotions that are false from up to 20.7% to 0%, while achieving final true population performance comparable to the best baselines.

1 INTRODUCTION

Recent progress in recursive self-improvement (RSI) has significantly advanced the capabilities of self-improving AI systems. For instance, agents now rewrite their own tools (Zhang et al., 2026; Robeyns et al., 2025), evolve production-ready algorithms (Novikov et al., 2025), and automate complex workflows like model alignment and post-training (Chen et al., 2026b; Rank et al., 2026). As agents become more capable at proposing and executing changes (Zelikman et al., 2024; Hu et al., 2025; Lu et al., 2024; Wang et al., 2026b), evaluation becomes increasingly critical (Manheim & Garrabrant, 2018; Skalse et al., 2022; Gao et al., 2023; Guo et al., 2026). The evaluation feedback dictates which candidate modifications are retained, expanded, or used to generate future proposals. Ideally, this feedback would favor only changes that improve performance on the underlying task distribution (Hastie et al., 2009; Recht et al., 2019). In practice, current RSI systems employ diverse search strategies but typically obtain this feedback by repeatedly querying a finite set of fixed benchmarks (Wang et al., 2026c).

This creates a fundamental statistical challenge. When RSI systems rely on a finite set of benchmarks to guide adaptive search, the search trajectory becomes increasingly tuned to the evaluation signal (Shawn, 2026), and empirical improvement may reflect adaptation to these benchmark tasks rather than genuine improvement (Wang et al., 2026c; Chen et al., 2026b). False promotions in this setting arise from three main sources.

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.

- **Within-round selection.** Selecting the best of several candidates within a round favors candidates whose advantage on the evaluation set is overstated by chance (Cawley & Talbot, 2010; Efron, 2011; Xu et al., 2026).
- **Repeated testing.** Testing across many rounds increases the probability that at least one test makes an error (Dunn, 1961; Holm, 1979).
- **Adaptive dependence.** Later candidates are adapted to earlier evaluation outcomes (Dwork et al., 2015b; Hardt & Ullman, 2014; Blum & Hardt, 2015).

The first two would arise even if every round used fresh evaluation data, and standard multiple comparison corrections handle them by accounting for the number of tests (Dunn, 1961; Holm, 1979; Bretz et al., 2009; Tian & Ramdas, 2021). The third source arises from benchmark reuse. When later candidates are generated based on earlier outcomes on the evaluation set and then tested on that same set, the candidates themselves become functions of the evaluation set. This adaptive dependence violates the assumption of standard statistical tests that candidates are chosen independently of the evaluation set (Kriegeskorte et al., 2009; Berk et al., 2013; Dwork et al., 2015a). Consequently, a candidate may be promoted because it overfits the specific evaluation set, rather than because it genuinely improves on the underlying task distribution. Recent work attempts to bring statistical error control into self-improvement loops and iterative evaluation (Shawn, 2026; Wu et al., 2025; Sengupta, 2026; Guo et al., 2026; Bertran et al., 2026). However, none of these methods account for all three sources of false promotion simultaneously. In particular, their statistical guarantees rely on each candidate being independent of the data on which it is evaluated, and restricting feedback alone does not account for the information revealed by the promotion decisions themselves. Because they do not address the adaptive dependence created by reusing the same evaluation set, they cannot guarantee that every promoted modification is a genuine population improvement.

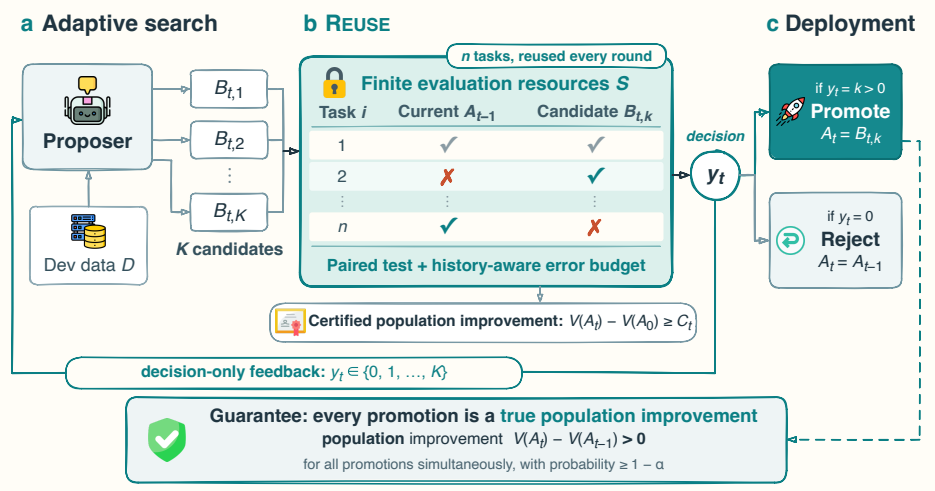


Figure 1: Overview of REUSE. (a) At round t, the proposer generates K candidate modifications using development data and earlier promotion decisions. (b) REUSE compares each candidate with the current system A_{t-1} on a fixed evaluation set S of n tasks that is reused in every round. (c) Only the promotion decision y_t is returned to the proposer. With probability at least $1 - \alpha$, every promotion is a true improvement in population performance, $V(A_t) > V(A_{t-1})$, simultaneously over the entire run, and the cumulative improvement $V(A_t) - V(A_0)$ is certified to be at least C_t.

In this paper, we propose REUSE (**R**isk-controlled **E**valuation **U**nder **S**equential **E**volution), a certified evaluation and promotion framework for adaptive self-improvement that controls all three sources jointly. REUSE allows a fixed evaluation set to be reused throughout an adaptive search while preserving valid evidence of population improvement. In each round, candidate modifications are evaluated against the current system, but the search process receives only the promotion decision rather than evaluation scores or task-level outcomes (Figure 1). Later candidates may still adapt to earlier decisions, but they can depend on the evaluation set only through this decision history. For each possible history, the candidates are determined without reference to the evaluation set, so the

corresponding comparison is a valid test. REUSE therefore turns the dependence on the evaluation set into a multiple comparison problem over all possible decision histories, and controls it together with the first two sources within a global error budget. As a result, for a user-specified level α, with probability at least $1 - \alpha$, every promoted modification over the entire self-improvement process is a genuine population improvement on the underlying task distribution.

Our main contributions are:

- We propose REUSE, a certified evaluation and promotion framework for adaptive self-improvement that jointly addresses the statistical challenges of within-round selection, repeated testing, and adaptive dependence under finite evaluation resources. For a user-specified error level α, with probability at least $1 - \alpha$, every promoted modification is a genuine population improvement on the underlying task distribution.
- We develop statistical theory for evaluation under benchmark reuse, including simultaneous error control over the entire process, valid lower bounds on cumulative improvement, the minimum detectable improvement, and the fundamental limits of adaptive evaluation reuse.
- We evaluate REUSE in live self-improvement experiments against evaluation frameworks from current RSI systems and error-controlled baselines. REUSE commits substantially fewer false promotions than competing methods, reducing the proportion of promotions that are false from up to 20.7% to 0%, while achieving final true population performance within 0.04 percentage points of the best baselines, demonstrating that its rigorous statistical guarantees do not compromise overall progress.

2 METHODOLOGY AND THEORETICAL GUARANTEES

In this section, we formalize the adaptive self-improvement setting, present REUSE, and establish its statistical guarantees. More comprehensive theoretical results and complete proofs are given in Section A.

2.1 SETTING AND OBJECTIVE

A system $A \in \mathcal{A}$ denotes the complete agent or AI system evaluated at a given round. A modification may change its prompt, code, tools, training procedure, or model parameters. Let Q denote the underlying population distribution on which system performance is measured. For an evaluation draw $W \sim Q$, the score of system A is $s(A, W) \in [0, 1]$. Any randomness used in the evaluation is included in W. The population performance of A is

$$V(A) = \mathbb{E}_{W \sim Q}[s(A, W)], \qquad \Delta(B, A) = V(B) - V(A),$$

where $\Delta(B, A)$ is the population improvement of candidate B over system A.

The self-improvement process starts from A_0. At round t, the current system is A_{t-1} and the proposer generates K candidates $B_{t,1}, \ldots, B_{t,K}$. The promotion decision is $y_t \in \{0, 1, \ldots, K\}$. If $y_t = k > 0$, candidate $B_{t,k}$ becomes the new system. If $y_t = 0$, all candidates are rejected and the incumbent is retained. We write $p_t = \sum_{s=1}^{t} \mathbf{1}\{y_s \neq 0\}$ for the number of promotions through round t.

Promotion decisions are based on a fixed evaluation set

$$S = (W_1, \ldots, W_n), \qquad W_i \overset{\text{i.i.d.}}{\sim} Q,$$

drawn before the process begins and repeatedly reused across rounds. For any pair of systems (B, A), define the paired differences

$$Z_i(B, A) = s(B, W_i) - s(A, W_i),$$

which are i.i.d. with mean $\Delta(B, A)$. The proposer receives information from S only through previous promotion decisions. It may otherwise adapt using its own randomness R, development data D, and those decisions. We formalize this setting with two assumptions.

Assumption 1 (Independent evaluation set)**.** The evaluation set S is independent of the tuple (R, D, A_0) consisting of the proposer's randomness, the development data, and the initial system.

Assumption 2 (Decision-only feedback)**.** Write $\xi = (R, D, A_0)$. For every round $t \geq 1$ and candidate index $k \leq K$, there is a map $f_{t,k}$ such that $B_{t,k} = f_{t,k}(\xi, y_{1:t-1})$. Moreover, the map $(\xi, w) \mapsto s(A_0, w)$ and, for every $t \geq 1$, $k \leq K$, and fixed history $h \in \{0, 1, \ldots, K\}^{t-1}$, the map $(\xi, w) \mapsto s(f_{t,k}(\xi, h), w)$ are jointly measurable.

Assumption 1 requires that the evaluation set S be drawn independently of everything that shapes the search from the start, which is exactly what a benchmark is meant to be, a fresh sample from the target task distribution that plays no role in development. Assumption 2 requires that S influence the search only through the promotion decisions, while the proposer is otherwise unrestricted and can freely use any language model or development pipeline.

Together, the two assumptions ensure that candidate generation depends on S only through the decision history. Once ξ is fixed, every possible decision history determines the incumbent and the next candidates without reference to S, and by Assumption 1 S remains an i.i.d. sample from Q. Every comparison determined in this way is therefore a valid test, and each comparison actually made during the search is one of them. This turns the adaptive dependence created by benchmark reuse into a multiple comparison problem over all possible decision histories, which the allocation in Section 2.2 explicitly controls. The measurability condition is technical and holds for essentially any practical proposer, for example whenever systems are prompts, programs, or configurations.

For a practical improvement threshold $\gamma \geq 0$, a promotion is false if $\Delta(A_t, A_{t-1}) \leq \gamma$. Our goal is to control all promotions simultaneously:

$$\mathbb{P}(\Delta(A_t, A_{t-1}) > \gamma \text{ for every } t \text{ with } y_t \neq 0) \geq 1 - \alpha. \tag{1}$$

2.2 REUSE

To achieve the guarantee in (1), REUSE tests each candidate against the current system using paired evaluation and allocates a single error budget across all candidate comparisons that may arise during the adaptive search. At each round, the proposer generates candidates using development data and previous promotion decisions, and receives only the evaluation framework's promotion decision. Figure 1 illustrates this process. We first consider pass/fail benchmarks with $\gamma = 0$, where $s(A, W) \in \{0, 1\}$. The same framework extends naturally to bounded scores $s(A, W) \in [0, 1]$, where the detailed construction is given in Section A.1.

For a fixed candidate B and incumbent A, let n_+ be the number of evaluation tasks on which B succeeds and A fails, let n_- be the number on which A succeeds and B fails, and let $M = n_+ + n_-$ be the number of tasks on which the two systems have different outcomes. We test whether the candidate genuinely improves on the incumbent:

$$H_0 : \Delta(B, A) \leq 0 \qquad \text{versus} \qquad H_1 : \Delta(B, A) > 0. \tag{2}$$

Under H_0, conditional on a task where the two systems disagree, the probability that B wins is at most $1/2$. Conditional on M, this gives the exact one-sided paired sign-test p-value

$$\pi(n_+, M) = \mathbb{P}\big(\mathrm{Bin}(M, \tfrac{1}{2}) \geq n_+\big), \qquad \pi(0, 0) = 1. \tag{3}$$

A small value of $\pi(n_+, M)$ provides evidence in favor of $\Delta(B, A) > 0$, and its finite-sample validity is established in Lemma A2.

A single hypothesis test controls the error for one fixed candidate–incumbent comparison. In the self-improvement loop, however, many comparisons are made on the same evaluation set, and later candidates depend on earlier promotion decisions. To control false promotions over the entire run, REUSE allocates the global error budget $\alpha \in (0, 1)$ across all comparisons that can arise under all possible decision histories.

Suppose p promotions have occurred before round t. The p promotion rounds can be chosen in $\binom{t-1}{p}$ ways, and each promotion can select one of K candidates. Hence there are $\binom{t-1}{p} K^p$ possible decision histories with p previous promotions. Under Assumption 2, once (R, D, A_0) are fixed, each such history determines the incumbent and the K candidates at round t. Let $(w_t)_{t \geq 1}$ and $(v_p)_{p \geq 0}$ be nonnegative weights satisfying $\sum_{t \geq 1} w_t \leq 1$ and $\sum_{p \geq 0} v_p \leq 1$. These weights distribute the overall error budget across the round index and the number of previous promotions. At round t after

Algorithm 1 REUSE at round t

$p \leftarrow p_{t-1}$ and $\delta \leftarrow \delta_{t,p}$
for $k = 1, \dots, K$ **do**
  Compute $(n_{+,k}, n_{-,k})$ and $\bar{Z}_k$ for $B_{t,k}$ against A_{t-1}
  $\text{pass}_k \leftarrow [\pi(n_{+,k}, n_{+,k} + n_{-,k}) \le \rho\delta]$
end for
if at least one candidate passes **then**
  $k^\star \leftarrow \arg\max_{k:\text{pass}_k} \bar{Z}_k$
  Compute a lower confidence bound ℓ for $\Delta(B_{t,k^\star}, A_{t-1})$ at error level $(1-\rho)\delta$
  $y_t \leftarrow k^\star$
  $C_t \leftarrow C_{t-1} + \max\{\ell, 0\}$
else
  $y_t \leftarrow 0$, $C_t \leftarrow C_{t-1}$
end if

p promotions, each candidate comparison is assigned the error budget

$$\delta_{t,p} = \frac{\alpha w_t v_p}{\binom{t-1}{p} K^{p+1}}. \tag{4}$$

This explicit allocation addresses all three sources of false promotion mentioned in Section 1 simultaneously. The weights over rounds t control repeated testing, the additional factor K in the denominator accounts for within-round selection, and the $\binom{t-1}{p}K^p$ term handles adaptive dependence by covering all possible histories. Throughout the paper we use $w_t = 1/[t(t+1)]$ and $v_p = 1/[(p+1)(p+2)]$.

To also quantify cumulative improvement, we divide $\delta_{t,p}$ into two parts. For a fixed $\rho \in (0,1)$, a fraction $\rho\delta_{t,p}$ is used for the sign test, and the remaining $(1-\rho)\delta_{t,p}$ is used to construct a lower confidence bound ℓ on the candidate's population improvemetn (Lemma A3). When a candidate is promoted, REUSE adds $\max\{\ell, 0\}$ to a running certificate C_t, initialized at $C_0 = 0$. The certificate is kept for auditing and is not returned to the proposer during the search.

For candidate k at round t, define its empirical improvement over the incumbent as

$$\bar{Z}_k = \frac{1}{n}\sum_{i=1}^{n} Z_i(B_{t,k}, A_{t-1}).$$

If multiple candidates pass the promotion test, REUSE promotes the one with the largest $\bar{Z}_k$. The complete procedure is given in Algorithm 1.

Conditional on (R, D, A_0), each fixed-history comparison has probability at most $\delta_{t,p}$ of either an erroneous promotion decision or an invalid lower confidence bound. A union bound over all rounds, all possible histories, and all candidate comparisons gives

$$\sum_{t\ge 1}\sum_{p=0}^{t-1}\binom{t-1}{p}K^p \cdot K\delta_{t,p} = \alpha\sum_{t\ge 1} w_t \sum_{p=0}^{t-1} v_p \le \alpha.$$

This gives the following main result.

Theorem 1 (Simultaneous validity and cumulative progress)**.** *Consider pass/fail evaluation, with $s(A, W) \in \{0, 1\}$ and $\gamma = 0$. Suppose Assumptions 1 and 2 hold, and* REUSE *is run as in Algorithm 1 with the allocation* (4)*. Then, for every adaptive proposer satisfying these assumptions, with probability at least $1-\alpha$, $\Delta(A_t, A_{t-1}) > 0$ for every t with $y_t \neq 0$. On the same event, the running certificate satisfies $V(A_t) - V(A_0) \ge C_t$ for every $t \ge 1$.*

The cumulative bound follows by summing the valid stepwise contributions $\max\{\ell, 0\}$ over promotions, since the corresponding population improvement telescope to $V(A_t) - V(A_0)$.

Development data may also be used to select a single candidate before S is queried. If only that candidate is tested on S, Theorem 1 applies with $K = 1$ in (4), since there is no within-round candidate selection on the evaluation set.

We now establish how large a true improvement must be for reliable detection. We let $\mathcal{L}_{t,p} = \ln(2/\delta_{t,p})$ and let $d = \mathbb{P}(Z_i(B, A) \neq 0)$ denote the probability that the candidate and incumbent

have different outcomes on a random task. The following result gives the corresponding minimum detectable population improvement.

Proposition 1 (Minimum detectable improvement)**.** *Consider pass/fail evaluation with $\gamma = 0$ and $\rho = 1/2$ at a fixed decision history node. Suppose the candidate and incumbent pairs are fixed independently of S and each has disagreement probability $d \in (0, 1]$. Define*

$$\Delta_{\min}(n, t, p, K, d) = \sqrt{\frac{2d\,\mathcal{L}_{t,p}}{n}}.$$

Under the regularity conditions in Section A.6, for any fixed $\eta \in (0, 1)$, a candidate with $\Delta(B, A) \geq (1 + \eta)\Delta_{\min}$ passes the promotion test with probability tending to one, whereas if all K candidates satisfy $\Delta(B_{t,k}, A_{t-1}) \leq (1 - \eta)\Delta_{\min}$, the probability of any promotion tends to zero.

The quantity $\Delta_{\min}$ represents the smallest population improvement that we can reliably detect. The formula shows that a larger evaluation set n in the denominator decreases this detection threshold. At the same time the term $\mathcal{L}_{t,p}$ in the numerator depends on the per-round error level $\delta_{t,p}$. To maintain our global guarantee we must allocate smaller error budgets $\delta_{t,p}$ as the round count t and promotion count p increase. This decrease in $\delta_{t,p}$ causes $\mathcal{L}_{t,p}$ to grow which makes larger true improvements necessary for reliable detection at later stages. This behavior directly reflects our statistical intuition. By demanding stronger evidence as the search proceeds we prevent random noise from accumulating into false promotions over a long trajectory.

This dependence on the adaptive history is not solely an artifact of our particular allocation. In Section A.7 we show that for a class of threshold rules a comparable first order growth in the promotion threshold is necessary to prevent an adaptive proposer from eventually inducing a false promotion.

3 Experiments

In this section, we evaluate REUSE against existing methods in live self-improvement loops, where an LLM-based proposer generates modifications and a finite evaluation set is repeatedly reused. Our experiments address two key questions: First, how often do current RSI promotion schemes accept false improvements? Second, can REUSE prevent these false promotions without sacrificing true population improvement?

3.1 Experimental setup

We consider the agent tasked with machine learning classification. Specifically, we use the Covertype dataset (Blackard, 1998)[1] for the binary classification of cover type 2 against the remaining six classes. A fixed random split yields a training set D_{tr} of 50,000 examples for fitting candidates, a development set D_{dev} of 20,000 examples for proposer feedback, and an evaluation pool $\mathcal{E}$ of 60,000 examples, from which each run draws an evaluation set S of size $n \in \{2{,}000, 10{,}000\}$ that is reused throughout the search. Together, D_{tr} and D_{dev} form D of Section 2.1. Since the underlying task distribution is unknown, we measure population improvement on a disjoint held-out set H of 50,000 examples. Each run lasts $T = 200$ rounds. At every round, Qwen2.5-7B-Instruct (Yang et al., 2024) proposes up to $K = 8$ modifications of a scikit-learn pipeline (Pedregosa et al., 2011) based on the prompt templates detailed in Section B.2. Each modification changes at most two components among the model family, feature subset, pre-processing, and hyperparameters. All methods start from the same gradient-boosting configuration, share the proposer, candidate space, and training data, and see the same evaluation set within a seed. We run 30 seeds at each evaluation-set size, and Section B gives the full protocol.

We compare REUSE with several evaluation frameworks that reflect how current RSI systems use their benchmarks. Empirical best-of-K edits the incumbent and promotes the candidate with the largest positive gain on S (Chen et al., 2026b). SICA-style follows the archive-based search of SICA (Robeyns et al., 2025), DGM-style follows the parent selection of DGM (Zhang et al., 2026), and Elite and Niche-elite use population-based search inspired by AlphaEvolve (Novikov et al., 2025). We also evaluate baselines that add statistical control to the same loop. PACE-style applies a one-sided

[1] https://archive.ics.uci.edu/dataset/31/covertype

Table 1: Live self-improvement on Covertype over 30 seeds at each evaluation-set size, with all methods run on the same seeds. Pop. Impr. is the final population improvement of the deployed system over starting system A_0 in percentage points (pp), estimated on the held-out set H and reported as mean $\pm$ standard error. Prom. is the mean number of promotions per run, and False is the number of promotions with non-positive population improvement, summed over the 30 runs. Opt. Gap is the final improvement measured on S minus the population improvement measured on H, which measures how much the reused evaluation set overstates the final improvement. Select., Repeat., and Adapt. indicate whether a rule controls errors from within-round selection, repeated testing, and adaptive dependence, the three sources of false promotion discussed in Section 1.

				$n = 2{,}000$				$n = 10{,}000$			
Method	Select.	Repeat.	Adapt.	Pop. Impr. ↑	Prom.	False ↓	Opt. Gap ↓	Pop. Impr. ↑	Prom.	False ↓	Opt. Gap ↓
Empirical best-of-K	✗	✗	✗	7.04 ± 0.08	13.0	75	0.43 ± 0.11	7.19 ± 0.06	14.2	71	0.15 ± 0.06
SICA-style	✗	✗	✗	6.36 ± 0.18	12.3	72	0.34 ± 0.10	6.56 ± 0.16	12.5	62	0.14 ± 0.05
DGM-style	✗	✗	✗	3.89 ± 0.17	6.4	34	0.06 ± 0.11	4.03 ± 0.18	7.2	18	0.01 ± 0.05
Elite	✗	✗	✗	6.72 ± 0.14	14.3	89	0.47 ± 0.12	6.85 ± 0.12	13.7	59	0.16 ± 0.05
Niche-elite	✗	✗	✗	6.46 ± 0.15	12.3	71	0.40 ± 0.13	6.37 ± 0.15	12.4	44	0.13 ± 0.05
PACE-style	✗	✗	✗	6.45 ± 0.19	6.4	4	0.31 ± 0.10	7.23 ± 0.02	8.1	3	0.10 ± 0.05
Bonferroni-style	✓	✓	✗	2.35 ± 0.37	1.3	0	0.21 ± 0.08	6.90 ± 0.12	6.5	0	0.10 ± 0.05
REUSE (ours)	✓	✓	✓	7.02 ± 0.26	3.2	**0**	0.05 ± 0.12	7.19 ± 0.13	5.9	**0**	0.06 ± 0.07

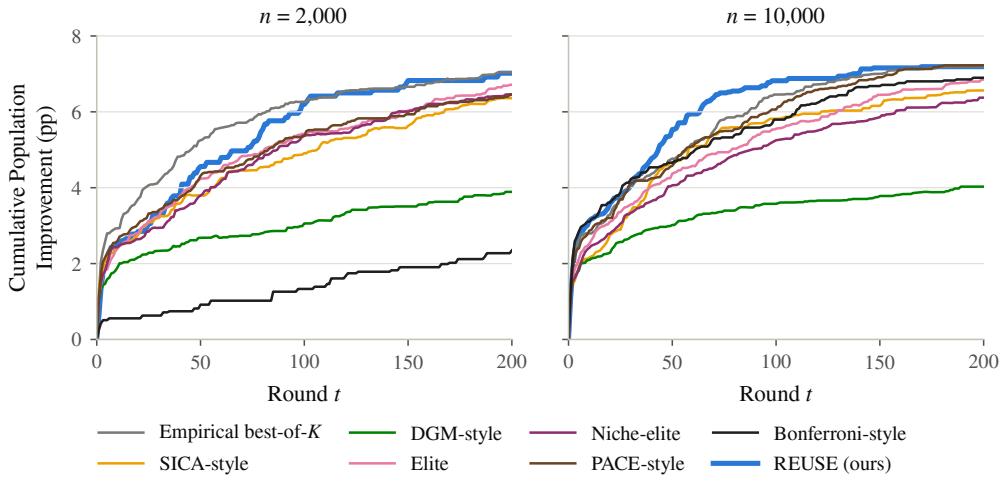


Figure 2: Population improvement of the deployed system over starting system A_0 across rounds, estimated on H and averaged over 30 seeds, in percentage points (pp). The left panel uses $n = 2{,}000$ and the right panel uses $n = 10{,}000$.

McNemar test (Dietterich, 1998) to each candidate at level $\alpha = 0.05$, a fixed-sample analogue of the per-decision error control in PACE (Shawn, 2026). Bonferroni-style applies the same test at level $0.05/(TK)$, splitting the error budget evenly over all potential comparisons of a run, as in the union-bound allocation of SGM (Wu et al., 2025).

To measure true improvement, we estimate the population improvement of each promoted modification by its paired improvement on the held-out set H, which no method ever accesses, and call a promotion false when this improvement is at most zero. For each method we report the final population improvement of the deployed system over the starting system A_0, the mean number of promotions per run, and the total number of false promotions across runs. We also report the optimism gap of the final system, defined as its improvement over A_0 measured on S minus its population improvement, which measures how much the reused evaluation set overstates the final improvement.

3.2 RESULTS

Table 1 reports the final population improvement over A_0 for each method alongside the number of promotions, the count of false promotions, and the optimism gap (Opt. Gap) of the final system. The results highlight how existing evaluation methods fail and how REUSE guarantees reliability

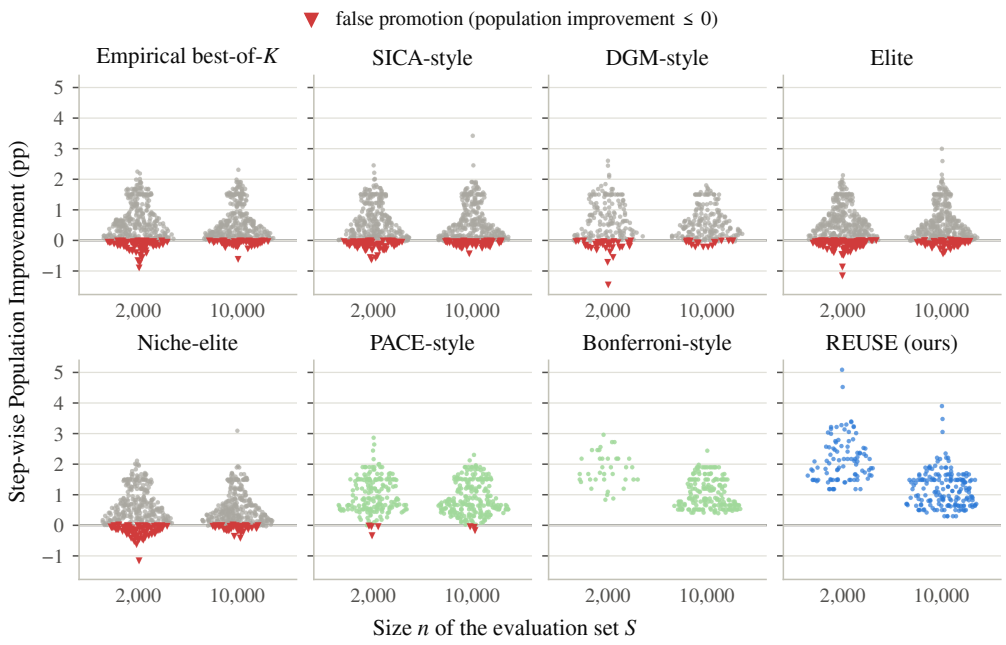


Figure 3: Population improvement of every promoted modification over the system it replaces, estimated on H and pooled over 30 seeds, in percentage points (pp). Each point is one promotion, and red triangles mark false promotions, whose population improvement is at most zero. Within each panel, the left column uses $n = 2{,}000$ and the right column uses $n = 10{,}000$.

without sacrificing true performance. For instance, at $n = 2{,}000$, Empirical best-of-K makes 75 false promotions, that is, nearly 20 percent of its accepted updates yield non-positive population improvement. Even when the evaluation set expands to $n = 10{,}000$, this framework still commits 71 false promotions. Although Empirical best-of-K achieves high final population improvement, its gains only match those of REUSE. Other benchmark-driven frameworks like SICA-style also commit numerous false promotions while yielding substantially lower population improvements.

The Opt. Gap column further clarifies these behaviors. Methods like Empirical best-of-K and PACE-style show large gaps because they adaptively overfit to the evaluation set. DGM-style exhibits a small gap but achieves very low population improvement, which reflects an inability to optimize rather than successful generalization. Statistical baselines also struggle to balance safety and progress. Bonferroni-style avoids false promotions but proves heavily conservative. It makes only 1.3 promotions per run at $n = 2{,}000$ and restricts its final population improvement to a mere 2.35 pp, well below the 7.02 pp achieved by REUSE. PACE-style improves performance but lacks formal guarantees and still commits false promotions at both evaluation set sizes. REUSE effectively solves this trade-off. It makes no false promotion while matching the strongest methods, achieving a population improvement of 7.02 pp at $n = 2{,}000$ and 7.19 pp at $n = 10{,}000$. Furthermore, its optimism gap stays within one standard error of zero at both sizes alongside high final true performance, proving that it eliminates adaptive overfitting without falling into the underfitting trap seen in DGM-style.

The trajectories across rounds in Figure 2 show that REUSE consistently stays near the top of the performance curves and establishes dominance when the sample size increases to $n = 10{,}000$. Figure 3 reveals the step-wise dynamics of individual promotions. Most competing methods display a dense cluster of points below zero, visibly marked by red triangles that represent false promotions. Bonferroni-style has no red triangles but produces very few points, confirming its overly conservative nature. In contrast, the entire cluster of points for REUSE sits noticeably higher, and every promotion is strictly positive.

4 Related Work

Recursive self-improvement and statistical control. The idea of a system that improves itself with guarantees dates back to the Gödel machine (Schmidhuber, 2007). Language models have turned this into practical search loops that rewrite prompts, agent scaffolds, algorithms, and entire pipelines (Zelikman et al., 2024; Fernando et al., 2024; Robeyns et al., 2025; Zhang et al., 2026; Novikov et al., 2025; Chen et al., 2026b;a). However, evaluating these systems by repeatedly querying fixed benchmarks often leads to false promotions and degraded deployment performance (Wang et al., 2026c; Shawn, 2026; Guo et al., 2026; Wang et al., 2026a). To address this, recent works gate self-modification with statistical tests, utilizing global error budgets (Wu et al., 2025; Sengupta, 2026), sequential testing (Shawn, 2026), or sealed audits (Guo et al., 2026; Mittal & Gupta, 2026). Yet, these approaches typically treat candidates as independent of the test data or restrict feedback without accounting for the information revealed by the promotion decisions themselves. REUSE overcomes this by combining decision-only feedback with an explicit error allocation over all possible decision histories, thereby controlling adaptive dependence on a reused evaluation set over the entire run.

Adaptive data analysis and holdout reuse. The risk of reusing a holdout set adaptively is a central topic in adaptive data analysis. Mechanisms based on differential privacy and description length allow a holdout set to answer adaptively chosen queries (Dwork et al., 2015b;a), and the Ladder mechanism maintains a leaderboard by releasing scores only upon significant improvement (Blum & Hardt, 2015). Information-theoretic arguments bound the bias of adaptive analysis (Russo & Zou, 2016; Xu & Raginsky, 2017; Bassily et al., 2016), though empirical benchmark reuse often shows less overfitting than worst-case theory predicts (Recht et al., 2019; Roelofs et al., 2019; Bertran et al., 2026). REUSE explicitly counts decision histories, turning the information bound into exact finite-sample tests for each promotion.

Multiple testing and sequential inference. Controlling the family-wise error over many hypotheses encompasses classical (Holm, 1979), weighted (Bretz et al., 2009), and sequential anytime-valid procedures (Lan & DeMets, 1983; Javanmard & Montanari, 2018; Howard et al., 2021; Ramdas et al., 2023; Waudby-Smith & Ramdas, 2024). Similarly, selective inference adjusts tests for data-driven selection biases (Berk et al., 2013; Fithian et al., 2014; Efron, 2011; Jamieson & Jain, 2018). However, these frameworks generally assume that hypotheses are fixed in advance, arrive with fresh data, or are selected by characterizable distributions. In a self-improvement loop, hypotheses are instead generated by an arbitrary proposer adapting to earlier outcomes on the same data. REUSE bridges this gap by allocating an error budget over all possible decision histories, ensuring each test is valid in finite samples (Clopper & Pearson, 1934; Maurer & Pontil, 2009).

5 Discussion

Self-improving systems increasingly decide what to keep by querying the same evaluation set over long trajectories. We showed that this reuse creates a source of error that multiple comparison corrections do not address, because later candidates depend on earlier outcomes on the evaluation set. REUSE controls this by returning only promotion decisions and allocating an error budget over all decision histories, which yields a simultaneous guarantee that every promoted modification is a true population improvement. In live self-improvement experiments, REUSE made much fewer false promotion than other frameworks without sacrificing true population performance. REUSE achieves this reliability while placing no restrictions on the underlying proposer or agent architecture. While our experiments demonstrate the framework on a machine learning classification task, extending it to larger agentic settings like coding and model post-training is a natural next step. The decision-only feedback currently provides strict statistical guarantees, but future work could accommodate richer signals including rounded scores, privatized summaries, or periodically refreshed evaluation sets by explicitly accounting for their information transcripts. Furthermore, expanding the framework to simultaneously control for secondary metrics like safety and cost offers a promising path toward generating the auditable and provable records of capability gains increasingly required by frontier AI safety frameworks (OpenAI, 2025).

REPRODUCIBILITY STATEMENT

Proofs of all theoretical results are given in Appendix Section A and the full experimental protocol and prompts in Appendix Section B. The supplementary material contains anonymized code for all experiments of Section 3.

AI USE STATEMENT

Large language models (Claude, Anthropic; ChatGPT, OpenAI) were used solely to assist with the writing and presentation of this paper. Based on the authors' draft, this assistance included refining wording for clarity and concision, improving paragraph organization, and adjusting the manuscript's formatting. The authors developed the research ideas and theoretical arguments, designed the experiments, and analyzed and interpreted the results. The authors reviewed each edit suggested by the models and verified that it did not alter any technical claim or reported numerical value. The authors take full responsibility for the final text, technical claims, and reported results.

ETHICS STATEMENT

Our experiments use the publicly available Covertype dataset (Blackard, 1998), which contains only cartographic variables and no personal or sensitive information. The task is a standard classification problem and involves no human subjects. We are not aware of ethical concerns arising from the data or the task, and we adhere to the ICLR Code of Ethics.

REFERENCES

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC)*, 2016.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

Martin Andres Bertran, Aaron Roth, and Zhiwei Steven Wu. What fits (into few tokens) doesn't overfit: Compression and generalization in ML research agents. arXiv preprint arXiv:2606.11045, 2026.

Jock A. Blackard. Covertype. UCI Machine Learning Repository, 1998.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 1006–1014, 2015.

Frank Bretz, Willi Maurer, Werner Brannath, and Martin Posch. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, 28(4):586–604, 2009.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

Mingguang Chen, Licheng Wang, and Bo Qu. Recursive self-improvement in AI: From bounded self-refinement to autonomous research loops. arXiv preprint arXiv:2607.07663, 2026a.

Yueh-Han Chen, Jiaxin Wen, and Jan Hendrik Kirchner. Automated researchers can mitigate well-characterized alignment failures. Anthropic Alignment Science, 2026b.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015b.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. arXiv preprint arXiv:1410.2597, 2014.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Diandian Guo, Cong Cao, Fangfang Yuan, Yingqi Wang, Yueshan Wang, and Dakui Wang. Self-authored verification is unreliable in heuristic self-improving agents. arXiv preprint arXiv:2607.24300, 2026.

Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 454–463, 2014.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edition, 2009.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

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

Kevin G. Jamieson and Lalit Jain. A bandit approach to sequential experimental design with false discovery control. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3660–3670, 2018.

Adel Javanmard and Andrea Montanari. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of Statistics*, 46(2):526–554, 2018.

Nikolaus Kriegeskorte, W. Kyle Simmons, Patrick S. F. Bellgowan, and Chris I. Baker. Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, 12(5): 535–540, 2009.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pp. 611–626, 2023.

K. K. Gordon Lan and David L. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 1983.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. arXiv preprint arXiv:2408.06292, 2024.

David Manheim and Scott Garrabrant. Categorizing variants of Goodhart's law. arXiv preprint arXiv:1803.04585, 2018.

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

Ayush Mittal and Dhruv Gupta. Signed compression progress on a sealed audit is goodhart-resistant. arXiv preprint arXiv:2606.11417, 2026.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, et al. AlphaEvolve: A coding agent for scientific and algorithmic discovery. arXiv preprint arXiv:2506.13131, 2025.

OpenAI. Our updated preparedness framework. OpenAI, 2025.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Ben Rank, Hardik Bhatnagar, Ameya Prabhu, Shira Eisenberg, Karina Nguyen, Matthias Bethge, and Maksym Andriushchenko. PostTrainBench: Can LLM agents automate LLM post-training? In *International Conference on Machine Learning (ICML)*, 2026.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 5389–5400, 2019.

Maxime Robeyns, Martin Szummer, and Laurence Aitchison. A self-improving coding agent. In *ICLR Workshop on Scaling Self-Improving Foundation Models without Human Supervision*, 2025.

Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1232–1240, 2016.

Jürgen Schmidhuber. Gödel machines: Fully self-referential optimal universal self-improvers. In *Artificial General Intelligence*, pp. 199–226. 2007.

Biswa Sengupta. Self-evolving agents with anytime-valid certificates. arXiv preprint arXiv:2607.00871, 2026.

Zayx Shawn. PACE: Anytime-valid acceptance tests for self-evolving agents. arXiv preprint arXiv:2606.08106, 2026.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jinjin Tian and Aaditya Ramdas. Online control of the familywise error rate. *Statistical Methods in Medical Research*, 30(4):976–993, 2021.

Charles L. Wang, Keir Dorchen, and Peter Jin. On the statistical limits of self-improving agents. *Transactions on Machine Learning Research*, 2026a.

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. In *International Conference on Learning Representations (ICLR)*, 2026b.

Yike Wang, Huaisheng Zhu, Zhengyu Hu, Yige Yuan, Zhengyu Chen, Shakti Senthil, Hannaneh Hajishirzi, Yulia Tsvetkov, Pradeep Dasigi, and Teng Xiao. Rethinking the evaluation of harness evolution for agents. In *COLM Workshop on Lifelong Agents: Learning, Aligning, and Evolving*, 2026c.

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

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

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Yang Xu, Jiefu Zhang, Haixiang Sun, Zihan Zhou, Tianyu Cao, and Vaneet Aggarwal. Towards reliable LLM evaluation: Correcting the winner's curse in adaptive benchmarking. arXiv preprint arXiv:2605.05973, 2026.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.

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.

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)*, 2026.

A Proofs and Additional Theory

A.1 A paired lower confidence bound for bounded scores

We first give a lower confidence bound for the mean of bounded paired differences. This result is used for non-binary scores and as one component of the certificate for pass/fail evaluation.

Let $z_1, \dots, z_n \in [-1, 1]$. Write

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i, \qquad \overline{z^2} = \frac{1}{n}\sum_{i=1}^{n} z_i^2.$$

Here $\mathrm{kl}(a \,\|\, b) = a\ln(a/b) + (1-a)\ln((1-a)/(1-b))$ is the Bernoulli relative entropy, with the usual continuous extensions at the endpoints. All logarithms are natural. For $L > 0$, define

$$\hat{s}^+(z; L) = \max\left\{q \in [\overline{z^2}, 1] : n\,\mathrm{kl}(\overline{z^2} \,\|\, q) \le L\right\}, \tag{5}$$

and

$$\varepsilon(v, L) = \frac{2L}{3n} + \sqrt{\left(\frac{2L}{3n}\right)^2 + \frac{2vL}{n}}. \tag{6}$$

We use the lower bound

$$\mathrm{lcb}(z; L) = \bar{z} - \varepsilon(\hat{s}^+(z; L), L). \tag{7}$$

The quantity $\hat{s}^+(z; L)$ is an upper confidence bound for $\mathbb{E}Z^2$, which in turn upper bounds $\mathrm{Var}(Z)$. The second term in (7) is a Bernstein deviation evaluated at this data-dependent variance proxy.

Lemma A1 (Paired lower confidence bound)**.** *Let $Z_1, \dots, Z_n$ be i.i.d. random variables in $[-1, 1]$ with mean μ. For $\delta \in (0, 1)$ and $L = \ln(2/\delta)$,*

$$\mathbb{P}(\mu < \mathrm{lcb}(Z_{1:n}; L)) \le \delta.$$

Proof. Let

$$m = \mathbb{E}Z_1^2, \qquad \sigma^2 = \mathrm{Var}(Z_1) \le m.$$

Since $Z_i^2 \in [0, 1]$ are i.i.d. with mean m, applying the Chernoff–Hoeffding bound to $1 - Z_i^2$ (Hoeffding, 1963) gives

$$\mathbb{P}\Big(\overline{Z^2} \le a\Big) \le \exp\{-n\mathrm{kl}(a \,\|\, m)\}, \qquad 0 \le a < m.$$

For $a = 0$ and $m < 1$, the same bound follows from $\mathbb{P}(Z_i^2 = 0) \le 1 - m$ and independence. If $m = 1$, then $Z_i^2 = 1$ almost surely, so the probability is zero for every $a < 1$, consistent with the convention $\exp(-\infty) = 0$. By the definition of $\hat{s}^+$,

$$\mathbb{P}(\hat{s}^+ < m) \le e^{-L} = \delta/2.$$

Next, the centered variables satisfy $Z_i - \mu \le 2$. Bernstein's inequality (Hoeffding, 1963), with range parameter $b = 2$, gives, for every $\epsilon > 0$,

$$\mathbb{P}(\bar{Z} - \mu \ge \epsilon) \le \exp\left\{-\frac{n\epsilon^2}{2\sigma^2 + \frac{4}{3}\epsilon}\right\}.$$

If $\sigma^2 = 0$, the left-hand side is zero and the inequality remains valid. The positive solution of

$$n\epsilon^2 = L\left(2\sigma^2 + \frac{4}{3}\epsilon\right)$$

is $\varepsilon(\sigma^2, L)$. Hence

$$\mathbb{P}\big(\bar{Z} - \mu \ge \varepsilon(\sigma^2, L)\big) \le e^{-L} = \delta/2.$$

The function $\varepsilon(v, L)$ is increasing in v. On the event $\{\hat{s}^+ \ge m\}$,

$$\varepsilon(\hat{s}^+, L) \ge \varepsilon(m, L) \ge \varepsilon(\sigma^2, L).$$

Therefore

$$\{\mu < \mathrm{lcb}(Z_{1:n}; L)\} \subseteq \{\hat{s}^+ < m\} \cup \left\{\bar{Z} - \mu > \varepsilon(\sigma^2, L)\right\}.$$

The two probabilities sum to at most δ. □

A.2 VALIDITY OF THE EXACT PAIRED SIGN TEST

For pass/fail evaluation,

$$Z_i \in \{-1, 0, +1\}.$$

Let

$$q_+ = \mathbb{P}(Z_i = +1), \qquad q_- = \mathbb{P}(Z_i = -1), \qquad d = q_+ + q_-.$$

The population gain is

$$\mu = \mathbb{E}Z_i = q_+ - q_-.$$

Recall that n_+ and n_- denote the number of candidate wins and losses and

$$M = n_+ + n_-.$$

The paired sign-test p-value is

$$\pi(n_+, M) = \mathbb{P}\big(\mathrm{Bin}(M, \tfrac{1}{2}) \geq n_+\big), \qquad \pi(0, 0) = 1.$$

Lemma A2 (Validity of the exact paired test)**.** *Let $Z_1, \ldots, Z_n$ be i.i.d. in $\{-1, 0, +1\}$ with mean $\mu \leq 0$. For any $\delta' \in (0, 1)$,*

$$\mathbb{P}(\pi(n_+, M) \leq \delta') \leq \delta'.$$

The same inequality holds conditionally on every value of M.

Proof. If $d = 0$, then $M = 0$ almost surely and $\pi(0, 0) = 1$. The test never rejects.

Assume $d > 0$ and define

$$q = \frac{q_+}{d} = \mathbb{P}(Z_i = +1 \mid Z_i \neq 0).$$

By independence, $(n_+, n_-, n - M)$ has the multinomial distribution with cell probabilities $(q_+, q_-, 1 - d)$. For $0 \leq k \leq m \leq n$, its probability mass factors as

$$\begin{aligned}\mathbb{P}(M = m, n_+ = k) &= \binom{n}{m}\binom{m}{k} q_+^k q_-^{m-k}(1 - d)^{n-m} \\ &= \mathbb{P}(M = m)\binom{m}{k} q^k (1 - q)^{m-k}.\end{aligned}$$

Thus, for every m with $\mathbb{P}(M = m) > 0$, $n_+ \mid M = m \sim \mathrm{Bin}(m, q)$. The null hypothesis $\mu \leq 0$ is equivalent to $q \leq 1/2$.

Fix such an m. Under the null, $\mathrm{Bin}(m, q)$ is stochastically dominated by $\mathrm{Bin}(m, 1/2)$, so the rejection probability of the upper-tail sign test is maximized over $q \leq 1/2$ at $q = 1/2$. Indeed, if there is a rejection threshold, let k_0 be the smallest integer in $\{0, \ldots, m\}$ with $\pi(k_0, m) \leq \delta'$. Monotonicity of $\pi(k, m)$ in k then gives

$$\begin{aligned}\mathbb{P}(\pi(n_+, M) \leq \delta' \mid M = m) &= \mathbb{P}(n_+ \geq k_0 \mid M = m) \\ &\leq \mathbb{P}\big(\mathrm{Bin}(m, \tfrac{1}{2}) \geq k_0\big) = \pi(k_0, m) \leq \delta'.\end{aligned}$$

If there is no rejection threshold, the rejection probability is zero. For $m = 0$ it is also zero. Averaging over M proves the unconditional claim. □

A.3 A LOWER CONFIDENCE BOUND FOR PASS/FAIL EVALUATION

For $d > 0$, define

$$q = \frac{q_+}{d}.$$

Then

$$\mu = q_+ - q_- = (2q - 1)d, \qquad \mathbb{E}Z_i^2 = d. \tag{8}$$

For k successes in m Bernoulli trials, let $\underline{\theta}(k; m, a)$ denote the one-sided Clopper–Pearson lower confidence limit at error level $a \in (0, 1)$ (Clopper & Pearson, 1934). For $1 \leq k \leq m$, it is the a-quantile of $\mathrm{Beta}(k, m - k + 1)$, and we set $\underline{\theta}(0; m, a) = 0$.

Given an error level $\delta_c \in (0,1)$, define

$$q_{\mathrm{lo}} = \underline{\theta}(n_+; M, \delta_c/4), \qquad d_{\mathrm{lo}} = \underline{\theta}(M; n, \delta_c/4),$$

and

$$\ell^{\mathrm{sgn}} = \begin{cases} (2q_{\mathrm{lo}} - 1)d_{\mathrm{lo}}, & q_{\mathrm{lo}} > 1/2, \\ -\infty, & q_{\mathrm{lo}} \le 1/2. \end{cases} \tag{9}$$

We combine this exact bound with the bounded-score lower confidence bound from Lemma A1:

$$\ell^\star(Z_{1:n}; \delta_c) = \max\left\{ \mathrm{lcb}\left(Z_{1:n}; \ln\frac{4}{\delta_c}\right), \ell^{\mathrm{sgn}} \right\}. \tag{10}$$

Lemma A3 (Lower confidence bound for pass/fail evaluation)**.** *For i.i.d. $Z_1, \ldots, Z_n \in \{-1, 0, +1\}$ with mean μ,*

$$\mathbb{P}(\mu < \ell^{\mathrm{sgn}}) \le \frac{\delta_c}{2},$$

and

$$\mathbb{P}(\mu < \ell^\star(Z_{1:n}; \delta_c)) \le \delta_c.$$

Proof. If $d = 0$, then $M = n_+ = 0$ almost surely, so $q_{\mathrm{lo}} = 0$ and $\ell^{\mathrm{sgn}} = -\infty$. The first claim is immediate, and the second follows from Lemma A1. Assume henceforth that $d > 0$.

By the conditional-count calculation in the proof of Lemma A2, the conditional law of n_+ given $M = m$ is $\mathrm{Bin}(m, q)$ whenever $\mathbb{P}(M = m) > 0$. Applying the one-sided Clopper–Pearson coverage property conditionally and then averaging over M gives

$$\mathbb{P}(q < q_{\mathrm{lo}}) \le \delta_c/4.$$

Separately, the indicators $\mathbf{1}\{Z_i \neq 0\}$ are i.i.d. Bernoulli variables with success probability d, so $M \sim \mathrm{Bin}(n, d)$. Applying the same coverage property directly to this count gives

$$\mathbb{P}(d < d_{\mathrm{lo}}) \le \delta_c/4.$$

Therefore, with probability at least $1 - \delta_c/2$,

$$q \ge q_{\mathrm{lo}}, \qquad d \ge d_{\mathrm{lo}}.$$

On this event, when $q_{\mathrm{lo}} > 1/2$,

$$\mu = (2q - 1)d \ge (2q_{\mathrm{lo}} - 1)d_{\mathrm{lo}} = \ell^{\mathrm{sgn}}.$$

When $q_{\mathrm{lo}} \le 1/2$, the bound equals $-\infty$ and the inequality holds automatically. Hence

$$\mathbb{P}(\mu < \ell^{\mathrm{sgn}}) \le \delta_c/2.$$

By Lemma A1,

$$\mathbb{P}\left(\mu < \mathrm{lcb}\left(Z_{1:n}; \ln\frac{4}{\delta_c}\right)\right) \le \delta_c/2.$$

A union bound over the two lower confidence bounds gives

$$\mathbb{P}(\mu < \ell^\star(Z_{1:n}; \delta_c)) \le \delta_c.$$

□

Remark A1 (Running cumulative bound)**.** The sign test may establish that $\mu > 0$ even when the numerical lower confidence bound $\ell^\star$ is non-positive. For the pass/fail test with $\gamma = 0$, REUSE therefore adds

$$\max\{\ell^\star, 0\}$$

to the running bound C_t. On the event of Theorem 1, both

$$\Delta(A_t, A_{t-1}) > 0$$

and

$$\Delta(A_t, A_{t-1}) \ge \ell^\star$$

hold. Hence

$$\Delta(A_t, A_{t-1}) \ge \max\{\ell^\star, 0\}.$$

This keeps C_t nondecreasing while preserving its lower-bound interpretation. For the bounded-score extension, promotion requires $\ell > \gamma \ge 0$, and the contribution is ℓ itself.

A.4 PROOF OF THE MAIN GUARANTEE

We prove Theorem 1 for general allocation weights (w_t) and (v_p) satisfying

$$\sum_{t\geq 1} w_t \leq 1, \qquad \sum_{p\geq 0} v_p \leq 1.$$

For pass/fail evaluation with $\gamma = 0$, let $\rho \in (0,1)$ denote the fraction of $\delta_{t,p}$ used for the promotion test; the remaining fraction is used for the lower confidence bound. The proof applies to every such ρ. If $\delta_{t,p} = 0$, all candidates are rejected and the running bound is unchanged. Such comparisons can be omitted from the argument below.

Proof of Theorem 1. Condition on the search procedure's randomness R, development data D, and starting system A_0. By Assumption 1, S remains i.i.d. from Q. All probabilities in this proof are conditional on (R, D, A_0).

For round t, let

$$y = y_{1:t-1}$$

denote a possible decision history. By Assumption 2, every such history determines an incumbent

$$a(y)$$

and candidates

$$b_1(y), \ldots, b_K(y)$$

that do not otherwise depend on S. Here y is a fixed hypothetical history: we do not condition on the event that the evaluation set produces that history.

Let $\mathcal{Y}_{t-1,p}$ be the set of decision histories of length $t-1$ containing exactly p promotions. Then

$$|\mathcal{Y}_{t-1,p}| = \binom{t-1}{p} K^p. \tag{11}$$

The factor $\binom{t-1}{p}$ selects the promotion rounds and the factor K^p selects the promoted candidate at each such round.

Fix t, p, $y \in \mathcal{Y}_{t-1,p}$, and candidate k. Write

$$\mu_{y,k} = \Delta(b_k(y), a(y)).$$

In what follows, $Z_{1:n}$, n_+, and M refer to this fixed pair $(b_k(y), a(y))$. For pass/fail evaluation with $\gamma = 0$, define the decision failure event

$$E^{\mathrm{dec}}_{t,y,k} = \{\mu_{y,k} \leq 0\} \cap \{\pi(n_+, M) \leq \rho\delta_{t,p}\},$$

and the confidence-bound failure event

$$E^{\mathrm{cert}}_{t,y,k} = \{\mu_{y,k} < \ell^\star(Z_{1:n}; (1-\rho)\delta_{t,p})\}.$$

By Lemma A2,

$$\mathbb{P}(E^{\mathrm{dec}}_{t,y,k}) \leq \rho\delta_{t,p},$$

and by Lemma A3,

$$\mathbb{P}(E^{\mathrm{cert}}_{t,y,k}) \leq (1-\rho)\delta_{t,p}.$$

Set $E_{t,y,k} = E^{\mathrm{dec}}_{t,y,k} \cup E^{\mathrm{cert}}_{t,y,k}$. Then

$$\mathbb{P}(E_{t,y,k}) \leq \delta_{t,p}.$$

For the bounded-score extension, define instead

$$E_{t,y,k} = \{\mu_{y,k} < \mathrm{lcb}(Z_{1:n}; \mathcal{L}_{t,p})\}.$$

By Lemma A1,

$$\mathbb{P}(E_{t,y,k}) \leq \delta_{t,p}.$$

Now sum over every round, promotion count, decision history, and candidate. Using (4) and (11),

$$\mathbb{P}\left(\bigcup_{t,p,y,k} E_{t,y,k}\right) \le \sum_{t\ge 1}\sum_{p=0}^{t-1}\binom{t-1}{p}K^p \cdot K \cdot \delta_{t,p} \tag{12}$$

$$= \alpha \sum_{t\ge 1} w_t \sum_{p=0}^{t-1} v_p \tag{13}$$

$$\le \alpha. \tag{14}$$

Therefore, with probability at least $1-\alpha$, every candidate comparison that can arise under every possible decision history is valid simultaneously.

Consider the realized sequence. If $y_t = k \neq 0$, the promoted candidate passed the promotion test. On the event above, with $\gamma = 0$ for the sign-test route and any $\gamma \ge 0$ for the bounded-score route,

$$\Delta(A_t, A_{t-1}) > \gamma,$$

which proves the first claim.

The same event also gives

$$\Delta(A_t, A_{t-1}) \ge \max\{\ell_t, \gamma\}$$

at every promoted round, where ℓ_t denotes the lower bound used for the promoted comparison. Summing over promoted rounds gives

$$V(A_t) - V(A_0) = \sum_{s\le t: y_s \neq 0} \Delta(A_s, A_{s-1}) \tag{15}$$

$$\ge \sum_{s\le t: y_s \neq 0} \max\{\ell_s, \gamma\} \tag{16}$$

$$= C_t. \tag{17}$$

This proves the cumulative lower bound for each fixed $\xi = (R, D, A_0)$. By the joint measurability in Assumption 2, the failure events used above, for each fixed hypothetical history, are measurable as functions of (ξ, S). Their union over rounds and histories is a countable union of measurable sets. Since S and ξ are independent by Assumption 1, integrating the section-wise probability bound with respect to the law of ξ gives the unconditional statement.

□

A.5 A DIRECT LOWER BOUND AGAINST THE STARTING SYSTEM

The running bound C_t adds one lower confidence bound for each promoted modification. We can also compare the current system A_t directly with the starting system A_0. This produces a second lower bound on total progress from the same stored evaluations.

Let $\alpha_c \in (0,1)$ be a separate error level. For a decision history of length t containing p promotions, define

$$\delta^c_{t,p} = \frac{\alpha_c w_t v_p}{\binom{t}{p}K^p}, \qquad \mathcal{L}^c_{t,p} = \ln\frac{2}{\delta^c_{t,p}}. \tag{18}$$

When $p_t \ge 1$, define

$$D_t = \ell^\star\left(Z_{1:n}(A_t, A_0); \delta^c_{t,p_t}\right), \tag{19}$$

for pass/fail outcomes. For general bounded scores, use $D_t = \mathrm{lcb}(Z_{1:n}(A_t, A_0); \mathcal{L}^c_{t,p_t})$ from Lemma A1 instead. If $p_t \ge 1$ and $\delta^c_{t,p_t} = 0$, set $D_t = -\infty$. When $p_t = 0$, set

$$D_t = 0,$$

since $A_t = A_0$ exactly.

Proposition A1 (Direct lower bound on total progress)**.** *Under Assumptions 1 and 2, with probability at least* $1 - \alpha_c$*,*

$$V(A_t) - V(A_0) \geq D_t \qquad \textit{for all } t \geq 1.$$

Under development-set screening, the same result holds with the factor K^p *removed from* (18)*. In the pass/fail setting with* $\gamma = 0$*, under the conditions of Theorem 1, we additionally have*

$$V(A_t) - V(A_0) \geq \max\{C_t, D_t\}$$

for all t *with probability at least* $1 - \alpha - \alpha_c$*.*

Proof. Condition on (R, D, A_0) as in the proof of Theorem 1. Probabilities below are conditional on this tuple.

A complete decision history

$$y = y_{1:t}$$

determines the current system $a_t(y)$. If the history contains exactly p promotions, then there are

$$\binom{t}{p} K^p$$

such histories.

For each history with $p \geq 1$, define

$$F_{t,y} = \left\{ \Delta(a_t(y), A_0) < \ell^\star \left(Z_{1:n}(a_t(y), A_0); \delta^c_{t,p} \right) \right\}.$$

For each fixed hypothetical history y, the pair $(a_t(y), A_0)$ is fixed given (R, D, A_0) and does not otherwise depend on S. We do not condition on observing y. Thus

$$\mathbb{P}(F_{t,y}) \leq \delta^c_{t,p}.$$

For bounded scores, replace $\ell^\star$ by the bound from Lemma A1; the same error bound holds. When $p = 0$, no promotion has occurred and $a_t(y) = A_0$, so the desired inequality holds deterministically; set $F_{t,y} = \varnothing$ for these histories. The failure event is also empty when the error allotment is zero.

A union bound gives

$$\mathbb{P}\left(\bigcup_{t,p,y} F_{t,y} \right) \leq \sum_{t \geq 1} \sum_{p=0}^{t} \binom{t}{p} K^p \delta^c_{t,p} \tag{20}$$

$$= \alpha_c \sum_{t \geq 1} w_t \sum_{p=0}^{t} v_p \tag{21}$$

$$\leq \alpha_c. \tag{22}$$

The result follows for the realized decision history. Integrating over (R, D, A_0) gives the unconditional statement.

Under development screening, a history is determined only by its promotion rounds, giving $\binom{t}{p}$ possible histories and removing K^p. The joint statement follows from a union bound over the events of Theorem 1 and this proposition. □

Corollary A1 (Carry-forward between promotions)**.** *Suppose* $p_t \geq 1$ *and let*

$$s(t) = \max\{s \leq t : y_s \neq 0\}$$

be the most recent promotion round. On the event of Proposition A1,

$$V(A_t) - V(A_0) \geq D_{s(t)}.$$

Hence the direct lower bound only needs to be recomputed when a promotion occurs.

Proof. No promotion occurs between rounds $s(t)$ and t, so

$$A_t = A_{s(t)}.$$

Apply Proposition A1 at round $s(t)$. □

Comparison with the running bound. The two bounds summarize cumulative progress differently. The running quantity C_t adds one confidence bound for each promoted step. The direct quantity D_t bounds the single comparison between A_t and A_0. When several promotions accumulate, D_t can be substantially tighter because it pays one confidence margin rather than one margin for every promotion.

Neither bound dominates in every configuration, so both can be reported under their respective error levels.

A.6 Power and the minimum detectable improvement

We next characterize the amount of population gain required to pass the promotion test for fixed comparisons. The pairs below are fixed independently of S. Equivalently, one may fix (R, D, A_0) and a hypothetical decision history as in the proof of Theorem 1. These power statements are not conditional on reaching that history in an adaptive run.

Proposition A2 (Power of the promotion tests)**.** *Consider a fixed candidate–incumbent pair tested at round t after p promotions, with i.i.d. paired differences $Z_1, \dots, Z_n$, population gain Δ, paired second moment*

$$m = \mathbb{E}Z^2,$$

and variance σ^2. Let

$$L = \mathcal{L}_{t,p}, \qquad \beta \in (0, 1).$$

(a) For the Bernstein–KL test, define

$$L_\beta = \ln \frac{2}{\beta},$$

$$m_\beta = m + \sqrt{\frac{2mL_\beta}{n}} + \frac{2L_\beta}{3n},$$

and

$$u = m_\beta + \sqrt{\frac{2m_\beta L}{n}} + \frac{2L}{n}.$$

If

$$\Delta > \gamma + \varepsilon(u, L) + \varepsilon(\sigma^2, L_\beta),$$

then this candidate passes the bounded-score promotion test with probability at least $1 - \beta$.

For pass/fail tasks with fixed $d = \mathbb{P}(Z_i \neq 0) > 0$, the right-hand side of the sufficient condition above is

$$\gamma + \sqrt{\frac{2dL}{n}}(1 + o(1))$$

when $n \to \infty$, $L \to \infty$, $L = o(n)$, and $L_\beta = o(L)$.

(b) For the exact paired sign test with $\gamma = 0$ and $\rho = 1/2$, define

$$c(M, L) = \min\left\{a \in \{0, 1, \dots, M\} : \mathbb{P}\big(\mathrm{Bin}(M, \tfrac{1}{2}) \geq a\big) \leq e^{-L}\right\},$$

with $c(M, L) = \infty$ when the set is empty.

A candidate passes if and only if

$$n_+ \geq c(M, L).$$

For $M \geq 1$ and $n_+ \geq M/2$,

$$M\mathrm{kl}\left(\frac{n_+}{M} \,\middle\|\, \frac{1}{2}\right) \geq L \quad \Longrightarrow \quad \textit{pass} \quad \Longrightarrow \quad M\mathrm{kl}\left(\frac{n_+}{M} \,\middle\|\, \frac{1}{2}\right) \geq L - \frac{1}{2}\ln(2M). \tag{23}$$

In particular, for $M \geq 1$,

$$n_+ - n_- \geq \sqrt{2LM}$$

is sufficient for promotion.

Proof. For part (a), apply Bernstein's inequality (Hoeffding, 1963) to $Y_i = Z_i^2 \in [0,1]$. Here $Y_i - m \le 1$, so the range parameter is $b = 1$, and

$$\mathrm{Var}(Y_i) \le \mathbb{E}Z_i^4 \le \mathbb{E}Z_i^2 = m.$$

Inverting the resulting tail bound, and using $\sqrt{u+v} \le \sqrt{u} + \sqrt{v}$ for $u, v \ge 0$, gives

$$\overline{Z^2} \le m + \sqrt{\frac{2mL_\beta}{n}} + \frac{2L_\beta}{3n} = m_\beta$$

with probability at least $1 - \beta/2$. Applying the same Bernstein inequality to $-Z_i$, now with $b = 2$ and variance σ^2, gives

$$\bar{Z} \ge \Delta - \varepsilon(\sigma^2, L_\beta)$$

with probability at least $1 - \beta/2$.

The inequality

$$\mathrm{kl}(a \,\|\, q) \ge \frac{(q-a)^2}{2q}, \qquad q \ge a,$$

together with the definition of $\hat{s}^+$ implies

$$\hat{s}^+ \le \overline{Z^2} + \sqrt{\frac{2\overline{Z^2}L}{n}} + \frac{2L}{n} \le u.$$

On the intersection of the two high-probability events,

$$\ell \ge \Delta - \varepsilon(\sigma^2, L_\beta) - \varepsilon(u, L) > \gamma.$$

Hence the candidate passes. The asymptotic expression follows from

$$\varepsilon(u, L) = \sqrt{\frac{2mL}{n}}(1 + o(1)).$$

For part (b), the test level is

$$\rho\delta_{t,p} = \frac{\delta_{t,p}}{2} = e^{-L}.$$

Monotonicity of the binomial upper tail gives the characterization through $c(M, L)$. When $M = 0$, $c(M, L) = \infty$ and the test cannot pass; the following calculations assume $M \ge 1$.

For $n_+ \ge M/2$, the Chernoff–Hoeffding bound (Hoeffding, 1963) gives

$$\mathbb{P}\big(\mathrm{Bin}(M, \tfrac{1}{2}) \ge n_+\big) \le \exp\left\{-M\mathrm{kl}\left(\frac{n_+}{M} \,\middle\|\, \frac{1}{2}\right)\right\}.$$

The boundary $n_+ = M/2$ follows from the trivial probability bound by one, and $n_+ = M$ gives equality. This proves the first implication in (23).

The binomial point-mass bound (29) gives, for every integer $0 \le a \le M$,

$$\mathbb{P}\big(\mathrm{Bin}(M, \tfrac{1}{2}) = a\big) \ge \frac{\exp\{-M\mathrm{kl}(a/M \,\|\, 1/2)\}}{\sqrt{2M}}.$$

If the candidate passes, then

$$\frac{\exp\{-M\mathrm{kl}(n_+/M \,\|\, 1/2)\}}{\sqrt{2M}} \le \mathbb{P}\big(\mathrm{Bin}(M, \tfrac{1}{2}) = n_+\big) \le \pi(n_+, M) \le e^{-L}.$$

Taking logarithms yields $M\mathrm{kl}(n_+/M \,\|\, 1/2) \ge L - \frac{1}{2}\ln(2M)$, which is the second implication in (23).

Finally, write

$$n_+ = \frac{M + x}{2}, \qquad x = n_+ - n_-.$$

For $u = x/(2M)$ with $|u| < 1/2$,

$$2u^2 \le \mathrm{kl}\left(\frac{1}{2} + u \,\middle\|\, \frac{1}{2}\right) \le \frac{2u^2}{1-4u^2}. \tag{24}$$

The lower inequality extends to $u = \pm 1/2$ by continuity and gives

$$M\mathrm{kl}\left(\frac{n_+}{M} \,\middle\|\, \frac{1}{2}\right) \ge \frac{x^2}{2M}.$$

Thus $x \ge \sqrt{2LM}$ is sufficient for promotion. □

Asymptotic conditions for Proposition 1. Let $n \to \infty$, keep K and $d \in (0, 1]$ fixed, and set $L = \mathcal{L}_{t,p}$ with $\ln n = o(L)$ and $L = o(n)$. The round and promotion count may vary with n. All candidate–incumbent pairs are fixed independently of S; in the K-candidate statement, each pair has the same probability d of different outcomes. Candidates need not be independent of one another.

Proof of Proposition 1. Let

$$x = \sum_{i=1}^{n} Z_i = n_+ - n_-, \qquad M = \sum_{i=1}^{n} Z_i^2.$$

For pass/fail outcomes,

$$\mathbb{E}x = n\Delta, \qquad \mathrm{Var}(x) \le nd,$$

and

$$\mathbb{E}M = nd, \qquad \mathrm{Var}(M) \le nd.$$

Hence

$$x = n\Delta + O_P(\sqrt{nd}),$$

and

$$M = nd\left(1 + O_P((nd)^{-1/2})\right).$$

Let

$$\tau = \sqrt{2Lnd}.$$

Then

$$\Delta_{\min} = \sqrt{\frac{2dL}{n}} = \frac{\tau}{n}.$$

Since $L \to \infty$,

$$\frac{\tau}{\sqrt{nd}} = \sqrt{2L} \to \infty.$$

Therefore the stochastic fluctuations in x and M are $o_P(\tau)$.

If

$$\Delta \ge (1+\eta)\Delta_{\min},$$

then

$$x \ge (1+\eta)\tau - o_P(\tau),$$

while

$$\sqrt{2LM} = \tau(1 + o_P(1)).$$

Thus the sufficient condition in Proposition A2(b) holds with probability tending to one.

If

$$\Delta \le (1-\eta)\Delta_{\min},$$

then

$$x \le (1-\eta)\tau + o_P(\tau).$$

On $\{x \le 0\}$ the test cannot pass. On $\{x > 0\}$, the preceding upper bound on x, together with $M/(nd) \to_P 1$ and $L = o(n)$, gives $x/M = o_P(1)$ on this event. Using (24),

$$M\mathrm{kl}\left(\frac{n_+}{M} \,\middle\|\, \frac{1}{2}\right) = \frac{x^2}{2M}(1 + o_P(1)).$$

A pass would require

$$x^2 \geq 2M \left(L - \frac{1}{2} \ln(2M) \right) (1 + o_P(1)) = \tau^2 (1 + o_P(1)),$$

where

$$\frac{\ln(2M)}{L} \to 0$$

by $\ln n = o(L)$. This contradicts

$$x^2 \leq (1 - \eta + o_P(1))^2 \tau^2$$

with probability tending to one.

For fixed K, a union bound over the fixed candidate–incumbent pairs completes the second claim. □

A.7 Lower bound for threshold-based promotion rules

We now show that the dependence on promotion history identified in the main text also appears as a lower bound for a natural class of promotion rules. The construction uses a bounded-score model satisfying Assumptions 1 and 2 exactly.

A.7.1 Model and threshold gates

Model M. A task is an infinite sign sequence

$$W = (g_1, g_2, \ldots) \in \{-1, +1\}^{\mathbb{N}},$$

whose coordinates are independent Rademacher variables.

A system is a pair

$$A = (\vartheta, \beta),$$

where

$$\vartheta \in [-1/4, 1/4], \qquad \beta \in \mathbb{R}^{\mathbb{N}}, \qquad \|\beta\|_1 \leq 1/4,$$

and β has finite support. Define

$$s(A, W) = \frac{1}{2} + \vartheta + \sum_j \beta_j g_j. \tag{25}$$

The score lies in $[0, 1]$ and

$$V(A) = \frac{1}{2} + \vartheta.$$

Hence

$$\Delta(B, A) = \vartheta_B - \vartheta_A.$$

Changing only β leaves population value unchanged. All constructions below start from $A_0 = (0, 0)$. We write e_j for the j-th coordinate vector.

Let

$$S = (W_1, \ldots, W_n)$$

be the evaluation set. Write g_{ij} for coordinate j of task i, and define

$$X_j = n^{-1/2} \sum_{i=1}^{n} g_{ij}.$$

The variables X_j are independent standardized Rademacher sums.

For any systems A and B,

$$Z_i(B, A) = (\vartheta_B - \vartheta_A) + \sum_j (\beta_B - \beta_A)_j g_{ij}. \tag{26}$$

Coordinates shared by A and B cancel. There is no additional evaluation randomness, so repeated evaluation is deterministic.

Definition A1 (Threshold gate)**.** Fix $b \in [0, 1]$. For $z \in \mathbb{R}^n$, define

$$\bar{z} = \frac{1}{n}\sum_i z_i, \qquad \overline{z^2} = \frac{1}{n}\sum_i z_i^2,$$

and

$$T_b(z) = \frac{\sqrt{n}\,\bar{z}}{\sqrt{\overline{z^2} - b\bar{z}^2}}. \tag{27}$$

We set $T_b(0) = 0$. When the denominator vanishes and $\bar{z} \neq 0$, set $T_b(z) = +\infty$ for $\bar{z} > 0$ and $T_b(z) = -\infty$ for $\bar{z} < 0$.

A threshold gate is specified by b and deterministic thresholds

$$\lambda_{t,p} \geq 0.$$

At round t after p previous promotions, it computes

$$T_k = T_b\left(Z_{1:n}(B_{t,k}, A_{t-1})\right).$$

It rejects all candidates when

$$\max_k T_k \leq \sqrt{2\lambda_{t,p}},$$

and otherwise promotes a candidate attaining the maximum.

The case $b = 0$ standardizes by the empirical second moment. The case $b = 1$ uses the empirical variance with divisor n.

All proposers below satisfy Assumptions 1 and 2. They are deterministic and use only previous promotion decisions. We set $\gamma = 0$. Let

$$\mathfrak{F}_t = \{\text{a false promotion occurs by round } t\}.$$

Theorem A1 (First-order lower bound)**.** *For the threshold-gate class in Definition A1, under the conditions of Corollaries A2 and A3 and the corresponding asymptotic regimes in Remark A3, a proposer that observes only previous promotion decisions can cause a false promotion with probability tending to one if the threshold parameters are uniformly below the first-order scale* $\ln(tK)$ *before the first promotion, or* $P\ln(Kt/P)$ *over the relevant rounds after* P *promotions, by a fixed multiplicative factor. In the latter regime, the conditions include* $t = \lfloor \varsigma P \rfloor + 1$, $\varsigma \geq 5$, $P \to \infty$, $K\varsigma \to \infty$, $\ln(K\varsigma) = o(\sqrt{n})$, *and* $P\ln(eK\varsigma) = o(n)$, *as well as* $\lambda_{s,p} < n/34$ *for every* $1 \leq s \leq t-1$ *and* $0 \leq p \leq \min\{P-1, s-1\}$, *as required by Corollary A3.*

The two regimes are established separately in Corollaries A2 and A3; the remaining subsections give the corresponding constructions and proofs.

A.7.2 Auxiliary lemmas

Binomial coefficient and point-mass bounds. For integers $n \geq 2$ and $1 \leq m \leq n-1$,

$$\binom{n}{m} \geq \frac{\exp\{n\ln 2 - n\mathrm{kl}(m/n \,\|\, 1/2)\}}{\sqrt{8m(n-m)/n}}. \tag{28}$$

Consequently, for integers $M \geq 1$ and $0 \leq a \leq M$,

$$\mathbb{P}\big(\mathrm{Bin}(M, \tfrac{1}{2}) = a\big) \geq \frac{\exp\{-M\mathrm{kl}(a/M \,\|\, 1/2)\}}{\sqrt{2M}}. \tag{29}$$

Proof. For integers $j \geq 1$, define

$$A_j = \frac{j!e^j}{j^{j+1/2}}, \qquad r_j = \frac{A_{j+1}}{A_j} = e\left(\frac{j}{j+1}\right)^{j+1/2}.$$

The function $g(x) = (x + 1/2)\ln(1 + 1/x)$ satisfies, for $x > 0$,

$$g'(x) = \ln(1 + 1/x) - \frac{1}{2}\left(\frac{1}{x} + \frac{1}{x+1}\right) < 0.$$

Indeed, strict convexity of $t \mapsto 1/t$ gives $\int_x^{x+1} t^{-1}\,dt < \frac{1}{2}(x^{-1} + (x+1)^{-1})$. Thus $\ln r_j = 1 - g(j)$ is increasing in j. For positive integers m, l, set $F(m,l) = A_{m+l}/(A_m A_l)$. Then

$$\frac{F(m+1,l)}{F(m,l)} = \frac{r_{m+l}}{r_m} \geq 1,$$

and the analogous inequality holds in the second coordinate. Hence

$$F(m,l) \geq F(1,1) = \frac{A_2}{A_1^2} = \frac{1}{\sqrt{8}}.$$

Writing $n = m + l$ and expanding the factorials therefore gives

$$\binom{n}{m} = F(m,l)\sqrt{\frac{n}{ml}}\frac{n^n}{m^m l^l} \geq \frac{\exp\{nH(m/n)\}}{\sqrt{8ml/n}},$$

where $H(x) = -x\ln x - (1-x)\ln(1-x)$. Since $H(x) = \ln 2 - \mathrm{kl}(x \,\|\, 1/2)$, this proves (28).

For $1 \leq a \leq M-1$, divide (28) by 2^M and use $a(M-a) \leq M^2/4$ to obtain (29). For $a = 0$ or $a = M$, both the point mass and $\exp\{-M\mathrm{kl}(a/M \,\|\, 1/2)\}$ equal 2^{-M}, and $\sqrt{2M} \geq 1$, so the bound still holds. □

Lemma A4 (Rademacher tail)**.** *Let S_n be a sum of n independent Rademacher variables. For $n \geq 64$ and*

$$1 \leq x \leq \frac{\sqrt{n}}{8},$$

$$\mathbb{P}(S_n > x\sqrt{n}) \geq \frac{c_R}{x}\exp\left(-\frac{x^2}{2} - \frac{(x^2+5)^2}{9n}\right),$$

where

$$c_R = \frac{e^{-5/2}}{2\sqrt{2}}.$$

Proof. Let

$$B = \frac{S_n + n}{2} \sim \mathrm{Bin}(n, \tfrac{1}{2}).$$

Set

$$m_0 = \left\lfloor \frac{n + x\sqrt{n}}{2} \right\rfloor + 1, \qquad w = \left\lfloor \frac{\sqrt{n}}{2x} \right\rfloor, \qquad m_1 = m_0 + w.$$

Then

$$\{S_n > x\sqrt{n}\} = \{B \geq m_0\}.$$

For $m \geq n/2$, the binomial probability mass is nonincreasing in m, so

$$\mathbb{P}(B \geq m_0) \geq (w+1)\mathbb{P}(B = m_1) \geq \frac{\sqrt{n}}{2x}\mathbb{P}(B = m_1).$$

Applying (28) with $m = m_1$, which satisfies $1 \leq m_1 \leq n-1$, and using $m_1(n-m_1) \leq n^2/4$, we obtain

$$\mathbb{P}(B = m_1) \geq \frac{e^{-n\mathrm{kl}(m_1/n \,\|\, 1/2)}}{\sqrt{2n}}.$$

Writing $m_1/n = (1+v)/2$ and $u = \sqrt{n}\,v$, the assumptions imply

$$x < u \leq x + \frac{2}{\sqrt{n}} + \frac{1}{x}$$

and

$$u^2 \leq x^2 + 5.$$

The expansion of the Bernoulli relative entropy gives

$$n\mathrm{kl}\left(\frac{1+v}{2}\,\middle\|\,\frac{1}{2}\right) \leq \frac{x^2+5}{2} + \frac{(x^2+5)^2}{9n}.$$

Combining the inequalities gives the result. □

Lemma A5 (Uniform second moment of selected probes)**.** *Let n, M, P be positive integers with $P \le M$, and let $(g_{ij})_{i\le n, j\le M}$ be independent Rademacher variables. For $J \subseteq [M]$ with $|J| = P$, define*

$$Q_J = \frac{1}{nP}\sum_{i=1}^{n}\left(\sum_{j\in J} g_{ij}\right)^2.$$

For every $\varepsilon > 0$,

$$\mathbb{P}\left(\max_{|J|=P} Q_J \ge 1+\varepsilon\right) \le \binom{M}{P}\exp\left\{-\frac{n}{2}\left[\varepsilon - \ln(1+\varepsilon)\right]\right\},$$

and therefore

$$\mathbb{P}\left(\max_{|J|=P} Q_J \ge 1+\varepsilon\right) \le \left(\frac{eM}{P}\right)^P \exp\left\{-\frac{n\varepsilon^2}{4(1+\varepsilon)}\right\}.$$

Proof. Fix J and set

$$Y_i = P^{-1/2}\sum_{j\in J} g_{ij}.$$

Then

$$\mathbb{E}e^{\mu Y_i} = \cosh(\mu/\sqrt{P})^P \le e^{\mu^2/2}.$$

If $\xi \sim N(0,1)$ is independent, then for $0 \le s < 1/2$,

$$\mathbb{E}e^{sY_i^2} = \mathbb{E}e^{\sqrt{2s}\xi Y_i} \le \mathbb{E}e^{s\xi^2} = (1-2s)^{-1/2}.$$

Markov's inequality with

$$s = \frac{\varepsilon}{2(1+\varepsilon)}$$

gives

$$\mathbb{P}\left(\sum_i Y_i^2 \ge n(1+\varepsilon)\right) \le \exp\left\{-\frac{n}{2}\left[\varepsilon - \ln(1+\varepsilon)\right]\right\}.$$

Apply a union bound over the $\binom{M}{P}$ possible sets J and use

$$\binom{M}{P} \le (eM/P)^P$$

and

$$\varepsilon - \ln(1+\varepsilon) \ge \frac{\varepsilon^2}{2(1+\varepsilon)}.$$

□

Lemma A6 (Calibration of a probe round)**.** *Let $n \ge 3$, $b \in [0,1]$, and $\tau > 0$. For $r \in [0, 1/4]$ and $x \in [-1,1]$, define*

$$h_r(x) = \frac{\sqrt{n}(r+x)}{\sqrt{D_r(x)}}, \qquad D_r(x) = 1 + r^2 + 2rx - b(r+x)^2.$$

Then the following properties hold.

1. *If $g \in \{-1,1\}^n$ has empirical mean x, then*
$$T_b\big(\tau(r\mathbf{1} + g)\big) = h_r(x).$$
2. *For fixed r, $h_r(x)$ is strictly increasing in x. For $x \in (-1,1)$, it is continuous and strictly increasing in r.*
3. *For $x \in [0,1]$,*
$$h_{1/4}(x) \ge \frac{\sqrt{n}}{\sqrt{17}}.$$

4. Let

$$\Lambda_n = \{-1, -1 + 2/n, \dots, 1\}.$$

For $u \in \Lambda_n \cap (0, 1)$ and $u^- = u - 2/n$, every threshold

$$h_0(u) \le \theta < h_{1/4}(u)$$

admits an $r \in (0, 1/4]$ such that

$$h_r(x) > \theta \quad \Longleftrightarrow \quad x \ge u$$

for every $x \in \Lambda_n$.

Proof. For $z = \tau(r\mathbf{1} + g)$,

$$\bar{z} = \tau(r + x), \qquad \overline{z^2} = \tau^2(1 + r^2 + 2rx),$$

which proves the first claim.

Direct differentiation gives

$$\partial_x h_r(x) = \sqrt{n}(1 + rx)D_r(x)^{-3/2} > 0$$

and

$$\partial_r h_r(x) = \sqrt{n}(1 - x^2)D_r(x)^{-3/2} > 0$$

for $|x| < 1$. This proves the monotonicity statements.

The third claim follows from

$$h_{1/4}(x) \ge h_{1/4}(0) = \frac{\sqrt{n}}{4\sqrt{1 + (1 - b)/16}} \ge \frac{\sqrt{n}}{\sqrt{17}}.$$

For the final claim, let

$$f(r) = h_r(u), \qquad \varphi(r) = h_r(u^-).$$

Both functions are continuous and strictly increasing and satisfy

$$\varphi(r) < f(r).$$

The intermediate value argument therefore produces an r for which

$$\varphi(r) < \theta < f(r).$$

Monotonicity in x gives the desired cutoff. □

A.7.3 THE COST OF REPEATED CANDIDATE TESTING

Theorem A2 (Neutral candidates defeat thresholds below $\ln(tK)$)**.** *In model* M*, let (b, λ) be a threshold gate. Let $n \ge 64$, $K, t \ge 1$, and define*

$$N = tK, \qquad \Lambda_t = \max_{s \le t} \lambda_{s,0}, \qquad x = \max\{1, \sqrt{2\Lambda_t}\}.$$

Assume $x \le \sqrt{n}/8$.

Consider the non-adaptive proposer that submits

$$B_{s,k} = \left(0, \frac{1}{4} e_{(s-1)K+k}\right).$$

Every candidate has zero population gain. If

$$q = \mathbb{P}(S_n > x\sqrt{n}),$$

then

$$\mathbb{P}(\mathfrak{F}_t^c) \le (1 - q)^N \le \exp\left\{-\frac{c_R N}{x} e^{-x^2/2 - (x^2+5)^2/(9n)}\right\}.$$

Proof. Until a promotion occurs, the incumbent remains A_0. Candidate (s, k) depends on a fresh probe coordinate $j = (s-1)K + k$. By Lemma A6,

$$T_k \geq X_j$$

whenever $X_j \geq 0$.

Since

$$\sqrt{2\lambda_{s,0}} \leq \sqrt{2\Lambda_t} \leq x,$$

any probe with $X_j > x$ forces a promotion. All such promotions are false because every candidate is population neutral.

The $N = tK$ probe coordinates are independent. Hence

$$\mathbb{P}(\mathfrak{F}_t^c) \leq (1-q)^N \leq e^{-Nq}.$$

Apply Lemma A4. □

Corollary A2. *In the setting of Theorem A2, suppose*

$$\Lambda_t \leq (1-\eta)\ln N$$

for a fixed $\eta \in (0, 1]$, with $N \geq 2$ and $\ln N \leq n/128$. Then

$$\mathbb{P}(\mathfrak{F}_t) \geq 1 - \exp\left\{-c_R e^{-1/2} \frac{N^\eta}{\sqrt{1+2\ln N}} e^{-(2\ln N+6)^2/(9n)}\right\}.$$

In particular,

$$\mathbb{P}(\mathfrak{F}_t) \to 1$$

whenever

$$tK \to \infty, \qquad \frac{\ln(tK)}{n} \to 0.$$

Proof. Apply Theorem A2 with

$$x \leq \max\{1, \sqrt{2\ln N}\}.$$

The stated bound follows by substituting the tail estimate from Lemma A4. Under the asymptotic conditions, the exponent diverges. □

Remark A2. The result depends on the thresholds used throughout the tested rounds. More generally, if a set of rounds $\mathcal{R} \subseteq [t]$ satisfies

$$\lambda_{s,0} \leq (1-\eta)\ln(|\mathcal{R}|K) \qquad \text{for } s \in \mathcal{R},$$

the proposer can submit neutral probes only in those rounds. The same argument then applies with

$$N = |\mathcal{R}|K.$$

A.7.4 The cost of promotion history

Fix

$$a \in (0, \sqrt{n}/8].$$

Let

$$\ell = \min\{v \in \sqrt{n}\,\Lambda_n : v \geq a\}, \qquad u = \ell/\sqrt{n},$$

and define

$$q = \mathbb{P}(S_n \geq a\sqrt{n}), \qquad \pi = 1 - (1-q)^K.$$

For a target number P of promotions and $\eta' > 0$, set

$$t_0 = \left\lceil \frac{(1+\eta')P}{\pi} \right\rceil, \qquad M = Kt_0.$$

We construct a proposer that uses previous promotion decisions to identify probe directions favored by the evaluation set. In a shifted-probe round, a successful probe is attached to a genuine positive shift. A neutral-probe promotion is already false. If no false promotion has occurred, after P

genuine promotions the proposer combines the selected neutral probes into one final candidate whose population gain is zero.

Let

$$\tau = \frac{1}{8P}.$$

The proposer keeps the selected probe indices in a set J, initially empty, and uses fresh indices $j_{s,k} = (s-1)K + k$. At round s, let $p = |J|$ and

$$\theta_{s,p} = \sqrt{2\lambda_{s,p}}.$$

Adaptive proposer $\mathcal{P}_B$. The proposer acts as follows.

1. If a false promotion has already occurred or the final candidate has already been submitted, submit copies of the incumbent.
2. If $p = P$, submit K copies of
$$B^{\mathrm{fin}} = A_{s-1} + \left(0, \tau \sum_{j \in J} e_j\right).$$
3. If
$$\theta_{s,p} < h_0(u),$$
submit neutral probes
$$B_{s,k} = A_{s-1} + (0, \tau e_{j_{s,k}}).$$
4. If
$$h_0(u) \le \theta_{s,p} < h_{1/4}(u),$$
choose r_s using Lemma A6 and submit
$$B_{s,k} = A_{s-1} + (r_s\tau, \tau e_{j_{s,k}}).$$
5. Otherwise submit copies of the incumbent.

If a candidate is promoted in a probe round, its probe index is added to J. The proposer can identify a false promotion from its own construction: neutral probes and the final candidate have zero population gain. Thus the construction depends only on previous promotion decisions and satisfies Assumption 2. Every submitted system remains in M: before the final round, $\|\beta\|_1 \le P\tau = 1/8$ and $0 \le \vartheta \le P\tau/4 = 1/32$, while the final candidate has $\|\beta\|_1 \le 2P\tau = 1/4$.

Theorem A3 (Adaptive proposer defeats slowly growing thresholds)**.** *In model M, let (b, λ) be a threshold gate. With the parameters defined above, fix $\varepsilon > 0$ and assume*

(B1)
$$\lambda_{s,p} < \frac{n}{34}$$
for all $s \le t_0$ and $p \le \min\{P-1, s-1\}$;

(B2)
$$\lambda_{s,P} \le \frac{a^2 P}{2(1+\varepsilon)}$$
for all $P + 1 \le s \le t_0 + 1$.

Then the proposer $\mathcal{P}_B$ satisfies

$$\mathbb{P}(\mathfrak{F}^c_{t_0+1}) \le \exp\left\{-\frac{\eta'^2 P}{2(1+\eta')}\right\} + \binom{M}{P} \exp\left\{-\frac{n}{2}\big(\varepsilon - \ln(1+\varepsilon)\big)\right\}.$$

Proof. First suppose $n \in \{1, 2\}$. Since $0 < a \le \sqrt{n}/8$, the smallest positive value of a standardized Rademacher sum is $\sqrt{n}$, so $\ell = \sqrt{n}$ and $u = 1$. Moreover,

$$h_0(1) = h_{1/4}(1) = \frac{\sqrt{n}}{\sqrt{1-b}},$$

where the value is interpreted as $+\infty$ when $b = 1$, consistently with the convention for T_b. For every $s \le t_0$, condition (B1) gives

$$\theta_{s,0} = \sqrt{2\lambda_{s,0}} < \sqrt{n/17} < h_0(1).$$

Thus, until the first false promotion, the proposer submits only neutral probes and the promotion count remains zero. A fresh probe satisfies $X_j \ge a$ if and only if all its n Rademacher coordinates equal $+1$. On this event its statistic is $\sqrt{n}/\sqrt{1-b}$ (or $+\infty$ when $b = 1$), which exceeds the threshold. Hence, if at least one of the K probes in a round has this property, a false promotion occurs. Here $q = 2^{-n}$ and $\pi = 1 - (1-q)^K$. Independence of the fresh probe blocks and the definition of t_0 therefore yield

$$\mathbb{P}(\mathfrak{F}^c_{t_0+1}) \le (1-\pi)^{t_0} \le \exp(-\pi t_0) \le \exp\big(-(1+\eta')P\big) \le \exp\left\{-\frac{\eta'^2 P}{2(1+\eta')}\right\}.$$

The last inequality follows from $\eta'^2 \le 2(1+\eta')^2$. Since the second term in the stated bound is nonnegative, this proves the theorem for $n \in \{1, 2\}$.

Henceforth assume $n \ge 3$. Let $\mathcal{G}_s$ contain all probe coordinates introduced through round s and all gate randomness up to that round. Fresh probe coordinates in round $s+1$ are independent of $\mathcal{G}_s$.

For a probe or neutral round, let

$$X^{(s)} = \max_k X_{j_{s,k}}.$$

Then

$$\mathbb{P}(X^{(s)} \ge a) = \pi.$$

If $\theta_{s,p} < h_0(u)$, a probe with $X^{(s)} \ge a$ crosses the threshold even though its population gain is zero. Hence the round produces a false promotion with conditional probability at least π.

If

$$h_0(u) \le \theta_{s,p} < h_{1/4}(u),$$

Lemma A6 chooses r_s so that promotion occurs exactly when

$$X^{(s)} \ge a.$$

The promoted candidate then has population gain

$$r_s \tau > 0,$$

so this promotion is genuine.

Condition (B1) guarantees that, until a false promotion or the P-th selected probe, every round is one of these two types. Define $I_s = \mathbf{1}\{X^{(s)} \ge a\}$ using the fresh coordinates for all $s \le t_0$; these indicators are independent Bernoulli variables with success probability π. On the event that no false promotion occurs and fewer than P probes are selected by t_0, each success $I_s = 1$ must have produced a genuine selected probe. Consequently, Chernoff's inequality gives

$$\mathbb{P}\big(\mathfrak{F}^c_{t_0} \cap \{\text{fewer than } P \text{ selected probes by } t_0\}\big) \le \mathbb{P}\left(\sum_{s=1}^{t_0} I_s < P\right) \le \exp\left\{-\frac{\eta'^2 P}{2(1+\eta')}\right\}.$$

Suppose instead that P genuine probe promotions have occurred. The final candidate has paired differences

$$z = \tau \sum_{j \in J} g_{\cdot j}$$

and population gain zero. On the event

$$\max_{|J|=P} Q_J < 1 + \varepsilon,$$

we have

$$\bar{z} \geq \tau n^{-1/2} Pa$$

and

$$\overline{z^2} = \tau^2 P Q_J.$$

Therefore

$$T_b(z) > \frac{Pa}{\sqrt{P(1+\varepsilon)}} = a\sqrt{\frac{P}{1+\varepsilon}}.$$

Condition (B2) makes this larger than the promotion threshold, so the final neutral candidate is promoted.

The only two ways to avoid a false promotion are therefore failure to collect P probes by t_0 or failure of the uniform second-moment event. Apply Lemma A5. □

Corollary A3 (First-order promotion-history rate)**.** *Let $\eta \in (0,1]$, $K, P \geq 1$, and $\varsigma \geq 5$ with $K\varsigma \geq e^8$. Define*

$$t = \lfloor \varsigma P \rfloor + 1$$

and

$$\Lambda_K(\varsigma) = \ln(K\varsigma) - \frac{1}{2}\ln\big(2\ln(K\varsigma)\big) - 6.$$

Assume the sample size satisfies

$$\sqrt{2\Lambda_K(\varsigma)} \leq \frac{\sqrt{n}}{8}, \qquad \frac{(2\Lambda_K(\varsigma)+5)^2}{9n} \leq \frac{1}{2},$$

and that $\lambda_{s,p} < n/34$ for all $s \leq t-1$ and $p \leq \min\{P-1, s-1\}$. If

$$\lambda_{s,P} \leq (1-\eta)\Lambda_K(\varsigma) P$$

for all $P+1 \leq s \leq t$, then a decision-only proposer satisfies

$$\mathbb{P}(\mathfrak{F}_t) \geq 1 - e^{-P/4} - \exp\left\{ P\ln(eK\varsigma) - \frac{n\eta^2}{8} \right\}.$$

Proof. Apply Theorem A3 with

$$a = \sqrt{2\Lambda_K(\varsigma)}, \qquad \eta' = 1, \qquad \varepsilon = \eta.$$

The two sample-size conditions allow us to apply Lemma A4 and bound its correction term by $1/2$. Since $\Lambda_K(\varsigma) \leq \ln(K\varsigma)$, this gives

$$q \geq \frac{c_R}{a}\exp\{-\Lambda_K(\varsigma) - 1/2\} \geq \frac{e^3}{2\sqrt{2}\,K\varsigma} \geq \frac{6}{K\varsigma}.$$

Hence

$$\pi = 1 - (1-q)^K \geq 1 - e^{-6/\varsigma} \geq \frac{3.49}{\varsigma},$$

which implies

$$t_0 = \left\lceil \frac{2P}{\pi} \right\rceil \leq \frac{2\varsigma P}{3.49} + 1 \leq \varsigma P.$$

The conditions of Theorem A3 follow. Finally use

$$\binom{M}{P} \leq (eK\varsigma)^P$$

and

$$\frac{n\eta^2}{4(1+\eta)} \geq \frac{n\eta^2}{8}.$$

□

Corollary A4 (Linear growth in the promotion count is insufficient)**.** *Fix $c > 0$ and an integer $p_0 \geq 0$. Suppose*

$$\lambda_{t,p} \leq cp$$

for all $p \geq p_0$, and assume

$$\sup_{t,p<p_0} \lambda_{t,p} = o(n).$$

The latter condition is omitted when $p_0 = 0$. Along any sequence satisfying

$$n \to \infty, \qquad P \to \infty, \qquad P \ln(eK) = o(n), \qquad \ln K = o(\sqrt{n}),$$

there exists a decision-only proposer that causes a false promotion within $O(P)$ rounds with probability tending to one.

Proof. Choose ς large enough that

$$\varsigma \geq e^8, \qquad \Lambda_1(\varsigma) \geq 2c$$

and apply Corollary A3 with $\eta = 1/2$. For sufficiently large P,

$$cP \leq \frac{1}{2}\Lambda_K(\varsigma)P.$$

The pre-promotion thresholds are $o(n)$ because $P = o(n)$, and the two sample-size conditions in Corollary A3 hold because $\ln K = o(\sqrt{n})$. Also $P \ln(eK\varsigma) = o(n)$, so both error terms in that corollary converge to zero. □

Remark A3 (Comparison with REUSE)**.** For

$$P + 1 \leq t \leq \varsigma P + 1,$$

the allocation used by REUSE satisfies

$$\mathcal{L}_{t,P} \leq P \ln(eK\varsigma) + O\left(\ln(K\varsigma P) + \ln\frac{1}{\alpha}\right).$$

Thus the history-dependent contribution per promotion is at most $\ln(eK\varsigma)$. The lower bound in Corollary A3 has scale

$$\Lambda_K(\varsigma) = \ln(K\varsigma) - O(\ln\ln(K\varsigma) + 1)$$

per promotion. These history-dependent rates agree to first order as $K\varsigma \to \infty$. This comparison concerns the history term, not the additional time- and error-level terms in $\mathcal{L}_{t,P}$.

For example, along sequences with $P \to \infty$, $K\varsigma \to \infty$, $\ln(K\varsigma) = o(\sqrt{n})$, and $P \ln(eK\varsigma) = o(n)$, the two remainder terms in Corollary A3 tend to zero for every fixed $\eta \in (0, 1]$. Under its threshold conditions, the false-promotion probability therefore tends to one. Together with Corollary A2, this gives the two regimes summarized in Theorem A1. The thresholds in these lower bounds are controlled over the stated ranges of rounds; the results do not assert that every individual threshold must exceed the stated rate.

The lower bound applies to the threshold-gate class in Definition A1. REUSE's exact test is not literally a member of this class, but its standardized critical value has the same first-order scale in the regime of Proposition 1. The attack uses non-binary bounded scores; the comparison does not by itself establish a lower bound for every pass/fail promotion rule. It identifies the first-order history-dependent requirement within the stated threshold-gate class.

B EXPERIMENTAL DETAILS

This appendix describes the protocol of the live experiments, the proposer and its prompts, the compared methods, and additional results.

Table 2: Protocol shared by all methods in the live experiments.

Quantity	Setting
Data	
Training set D_{tr}	50,000 examples
Development set D_{dev}	20,000 examples
Evaluation pool $\mathcal{E}$	60,000 examples
Held-out set H	50,000 examples
Evaluation-set size n	2,000 or 10,000
Seeds	30 at each evaluation-set size, seeds 3 to 32
Search	
Rounds T	200
Candidates per round K	8
Changed fields per candidate	at most 2
Model families	histogram gradient boosting, logistic regression
Preprocessing	no, standard, or robust scaling, optional selection of 10, 20, or 30 features
Feature subsets	all, topographic, hillshade, soil, topographic and soil
Starting system A_0	gradient boosting, all features, depth 3, 100 iterations, learning rate 0.1
Promotion target	classification accuracy
Proposer	
Model	Qwen2.5-7B-Instruct served by vLLM
Sampling	temperature 0.8, top-p 0.95, top-k 20, repetition penalty 1.05
Requests per round	at most 3, each asking for 8 configurations
REUSE	
Error levels	$\alpha = 0.05$, $\alpha_c = 0.05$, $\gamma = 0$, $\rho = 1/2$
Allocation weights	$w_t = 1/[t(t+1)]$ and $v_p = 1/[(p+1)(p+2)]$
Draft threshold	$z_{\mathrm{dev}} \geq 1.645$

B.1 PROTOCOL

Covertype contains 581,012 examples with 54 features, and we label cover type 2 as positive. A single random permutation partitions the data into the disjoint blocks D_{tr}, D_{dev}, $\mathcal{E}$, and H of Section 3.1, and the remaining examples are unused. For each seed, S is a random sample of n examples from $\mathcal{E}$, and the 2,000-example set is contained in the 10,000-example set of the same seed. All methods use the same evaluation set within a seed. Table 2 summarizes the protocol. Every configuration is fitted once on D_{tr} with fitting seed 0, and its predictions are cached so that a repeated configuration reuses the same fitted model. The experiments use NVIDIA A100 GPUs with 40 GB of memory, Python 3.11.16, scikit-learn 1.9.1, and vLLM 0.29.0.

B.2 PROPOSER AND PROMPT TEMPLATES

At every round, Qwen2.5-7B-Instruct (Yang et al., 2024), served through vLLM (Kwon et al., 2023), receives one system message and one user message. The system message describes the task, the configuration schema, and the allowed values, and it ends with a feedback paragraph that differs between REUSE and the compared methods. The user message contains the parent configuration, its accuracy on D_{dev}, its error rates on 14 slices of D_{dev}, and the previous ten rounds of search history. Each request asks for eight complete JSON configurations. Invalid proposals, proposals equal to the parent, and proposals that change more than two fields read by the parent's or the proposal's model family are discarded, and up to three requests are made per round. REUSE also discards configurations already evaluated in the run. Unfilled slots are replaced by copies of the parent, which cannot be promoted. Table 3 lists the information available to the proposer, and the templates follow.

System message. The following part is shared by all methods.

```
You are an expert machine-learning engineer improving a scikit-learn pipeline. The task is
   binary classification on forest cover-type data: 54 columns (10 topographic and hillshade
   measurements, 4 wilderness-area indicators, 40 soil-type indicators) and 50,000 training
   rows. Every configuration you propose is fitted from scratch on those same rows and
   judged by its accuracy.
```

Table 3: Information in the prompt of the proposer. Only REUSE withholds evaluation accuracies on S, so its proposer learns about S solely through promotion decisions. SICA-style additionally sees up to ten archive members with the highest evaluation accuracy.

Information in the prompt	Compared methods	REUSE
Parent configuration	✓	✓
Accuracy and slice error rates on D_{dev}	✓	✓
Accuracies of past proposals on D_{dev}	✓	✓
Evaluation accuracies on S	✓	✗
Promotion decisions	✓	✓
Development-draft updates and submissions	✗	✓
Accuracies on H	✗	✗

```
A configuration is a JSON object with exactly these eleven keys:
- "model_family": "hist_gbm" (histogram gradient boosting) or "logreg" (logistic regression)
- "learning_rate": number in [0.02, 0.3], hist_gbm only
- "max_depth": integer in [2, 6], hist_gbm only
- "max_iter": integer in [50, 300], boosting iterations, hist_gbm only
- "min_samples_leaf": integer in [5, 100], hist_gbm only
- "l2_regularization": number in [0, 1], hist_gbm only
- "C": number in [0.001, 100], inverse regularization strength, logreg only
- "scaling": "none", "standard" or "robust"
- "feature_selection": "none", "kbest_10", "kbest_20" or "kbest_30" (ANOVA F-test over the
    columns kept)
- "feature_subset": "all", "topo", "soil", "hillshade" or "topo_soil" -- fixed column
    presets: topo is elevation, aspect, slope and the three distance columns; hillshade is
    the three hillshade columns; soil is the 40 soil-type indicators; topo_soil is both; all
    is every column, including the four wilderness-area indicators
- "seed": 0

Rules:
- Every object must carry all eleven keys. A key you leave out is not "unchanged": it falls
    back to the default and silently undoes earlier work. Copy the current configuration and
    edit only the fields you mean to change.
- A value outside its range, an unknown key or malformed JSON discards that candidate.
- Keys that do not apply to the model family you chose are kept and ignored by the fit.
- A proposal that changes more than two fields is discarded unseen, so keep every candidate
    within two changes of the current configuration.
- Each proposal's accuracy on the development set is reported back to you in the next round,
    with its difference from the current configuration's. Build on the changes that raised it
    and abandon the ones that lowered it.

Answer with one short line of reasoning and then a JSON array of configurations in a ```json
    code block. If the current configuration were the starting one, a reply proposing a
    slower learning rate and a scaled, screened variant would look like this:

```json
[
  {"model_family": "hist_gbm", "learning_rate": 0.05, "max_depth": 3, "max_iter": 100,
   "min_samples_leaf": 20,
   "l2_regularization": 0.0, "C": 1.0, "scaling": "none", "feature_selection": "none",
   "feature_subset": "all",
   "seed": 0},
  {"model_family": "hist_gbm", "learning_rate": 0.1, "max_depth": 3, "max_iter": 100,
   "min_samples_leaf": 20,
   "l2_regularization": 0.0, "C": 1.0, "scaling": "standard", "feature_selection":
   "kbest_20",
   "feature_subset": "all", "seed": 0}
]
```
```

B.3 COMPARED METHODS

All compared methods share the proposer, the candidate space, and D_{tr} with REUSE, and they differ in parent selection and promotion rule. Every run keeps an archive of distinct fitted configurations. The incumbent is the deployed configuration, and the parent is the configuration the proposer edits in the next round. Table 4 summarizes the methods, and the reported system of every method is the incumbent after round T.

Empirical best-of-K edits the incumbent and promotes the candidate with the largest strictly positive gain on S. SICA-style follows the archive-based search of SICA (Robeyns et al., 2025). It also edits

Table 4: Compared methods in the live experiments. Select., Repeat., and Adapt. indicate whether a rule controls errors from within-round selection, repeated testing, and adaptive dependence, as in Table 1.

Method	Feedback from S	Promotion rule	Select.	Repeat.	Adapt.
Empirical best-of-K	scores	largest positive gain on S	✗	✗	✗
SICA-style	scores	largest positive gain on S	✗	✗	✗
DGM-style	scores	largest positive gain on S	✗	✗	✗
Elite	scores	largest positive gain on S	✗	✗	✗
Niche-elite	scores	largest positive gain on S	✗	✗	✗
PACE-style	scores	McNemar test at level 0.05	✗	✗	✗
Bonferroni-style	scores	McNemar test at level $0.05/(TK)$	✓	✓	✗
REUSE (ours)	decisions only	exact paired test with (4)	✓	✓	✓

Table 5: Additional summaries of the live experiments over 30 seeds at each evaluation-set size, in percentage points. Diff. is the seed-matched difference in final population improvement between a method and REUSE, reported as mean ± standard error, so a positive value means the method ends higher. Worst is the smallest population improvement among all promotions of a method, pooled over runs.

	Diff. vs REUSE		Worst ↑	
Method	$n = 2{,}000$	$n = 10{,}000$	$n = 2{,}000$	$n = 10{,}000$
Empirical best-of-K	$+0.02 \pm 0.25$	$+0.00 \pm 0.15$	-0.89	-0.61
SICA-style	-0.66 ± 0.29	-0.63 ± 0.19	-0.63	-0.42
DGM-style	-3.13 ± 0.31	-3.16 ± 0.23	-1.44	-0.21
Elite	-0.30 ± 0.28	-0.34 ± 0.18	-1.12	-0.44
Niche-elite	-0.56 ± 0.34	-0.82 ± 0.19	-1.13	-0.41
PACE-style	-0.57 ± 0.34	$+0.04 \pm 0.14$	-0.34	-0.18
Bonferroni-style	-4.67 ± 0.50	-0.29 ± 0.17	0.86	0.41
REUSE (ours)	–	–	1.20	0.34

the incumbent, and its prompt additionally lists up to ten archive members with the highest evaluation accuracy. DGM-style follows the parent selection of DGM (Zhang et al., 2026). Archive member i is chosen as parent with probability proportional to

$$\frac{\sigma(40(a_i - a_0 - 0.08))}{1 + e_i},$$

where a_i is its evaluation accuracy, a_0 is that of A_0, e_i is the number of rounds in which it has been parent, and σ is the logistic function. Elite draws its parent uniformly from the four archive members with the highest evaluation accuracy, and Niche-elite groups configurations by model family and feature subset, draws one group uniformly, and uses its best member. These two methods follow the population-based search of AlphaEvolve (Novikov et al., 2025). The five methods above share the promotion rule of Empirical best-of-K.

PACE-style and Bonferroni-style edit the incumbent and test all K candidates against it with the one-sided McNemar test (Dietterich, 1998),

$$z_{\mathrm{MC}} = \frac{n_+ - n_-}{\sqrt{n_+ + n_-}}, \qquad p_{\mathrm{MC}} = 1 - \Phi(z_{\mathrm{MC}}),$$

with n_+ and n_- as in Section 2.2, Φ the standard normal distribution function, and $p_{\mathrm{MC}} = 1$ when $n_+ + n_- = 0$. PACE-style tests every candidate at level 0.05, and Bonferroni-style at level $0.05/(TK) = 0.05/1600$. If several candidates pass, the one with the largest gain on S is promoted.

B.4 ADDITIONAL RESULTS

Table 5 complements Table 1 with seed-matched differences in final population improvement and the smallest population improvement among the promotions of each method.

Table 6: Sensitivity of REUSE to its allocation (4) at $n = 2{,}000$ over 10 new seeds. We change one of ρ or w_t or v_p at a time from the default. The variable t indexes the rounds where S is queried and the harmonic weight uses $c \approx 0.590747$. The columns for S tests, Prom., Pop. Impr. and C_T are means $\pm$ standard error over seeds. The False / all column sums false and all promotions over the 10 runs. Diff. is the seed-matched difference from the default run. Changed counts the runs in which the default allocation would have made at least one querying or promotion decision differently. C_T re-prices the promotions of the same-seed default run at the certificate level $(1 - \rho)\delta_{t,p}$. All improvements are in percentage points.

Setting	S tests per run	Prom. per run	False / all	Pop. Impr. (pp)	Diff. (pp)	Changed runs	C_T at default's promotions (pp)
Default	4.30 ± 0.37	2.90 ± 0.28	**0**/29	6.38 ± 0.23	N/A	N/A	0.74 ± 0.11
Split ρ							
$\rho = 0.1$	4.00 ± 0.30	3.00 ± 0.15	**0**/30	7.23 ± 0.36	$+0.85 \pm 0.40$	9/10	0.89 ± 0.11
$\rho = 0.25$	3.90 ± 0.18	2.60 ± 0.16	**0**/26	6.52 ± 0.27	$+0.14 \pm 0.37$	4/10	0.84 ± 0.11
$\rho = 0.75$	4.80 ± 0.33	3.00 ± 0.26	**0**/30	6.52 ± 0.35	$+0.14 \pm 0.29$	1/10	0.59 ± 0.10
$\rho = 0.9$	4.70 ± 0.33	3.00 ± 0.21	**0**/30	6.52 ± 0.26	$+0.14 \pm 0.32$	5/10	0.44 ± 0.09
Submission weight w_t							
harmonic	4.60 ± 0.34	3.10 ± 0.23	**0**/31	6.94 ± 0.21	$+0.56 \pm 0.21$	2/10	0.63 ± 0.10
flat	3.00 ± 0.26	2.30 ± 0.15	**0**/23	6.51 ± 0.33	$+0.12 \pm 0.43$	10/10	0.31 ± 0.06
geometric	4.30 ± 0.21	3.10 ± 0.23	**0**/31	6.64 ± 0.21	$+0.26 \pm 0.35$	1/10	0.80 ± 0.12
Promotion weight v_p							
front-loaded	4.90 ± 0.50	2.90 ± 0.18	**0**/29	7.11 ± 0.37	$+0.73 \pm 0.47$	8/10	0.65 ± 0.10
flat	4.00 ± 0.26	2.80 ± 0.20	**0**/28	6.85 ± 0.33	$+0.47 \pm 0.39$	7/10	0.49 ± 0.09
geometric	4.00 ± 0.30	3.00 ± 0.26	**0**/30	6.83 ± 0.42	$+0.45 \pm 0.42$	0/10	0.79 ± 0.11

B.4.1 ROBUSTNESS AND ABLATIONS

The theoretical guarantee of Theorem 1 holds for any allocation weights w_t and v_p and split ρ fixed before querying S. We tested the sensitivity of REUSE to these choices by running 11 different allocations on 10 fresh seeds at $n = 2{,}000$ (Table 6). REUSE successfully controlled the error across all configurations and made exactly zero false promotions out of 317 total promotions. The final population improvement remained stable and within the natural variation of the proposer. The running certificate C_T also responded naturally to the split ρ.

The core guarantee of REUSE is highly robust to changes in the search heuristics. To demonstrate this reliability we evaluated an alternative variant that queries S at every step. In this setup the proposer edits the incumbent directly and the candidate with the highest development accuracy is tested on S at level $\rho\delta_{t,p}$ with $K = 1$. The proposer receives only the promotion decision. Even with this more conservative search structure the variant completely avoided false promotions across 30 seeds at both evaluation set sizes (Table 7). It made 30 strictly positive promotions at $n = 2{,}000$ and 100 at $n = 10{,}000$. These results confirm that the error control mechanism fundamentally prevents false promotions and maintains rigorous validity regardless of the specific search mechanics or allocation parameters used.

We further isolated the role of the exact test on S by ablating it. We ran REUSE at $n = 10{,}000$ on 10 new seeds using a smaller development set of 500 examples and a lowered pre-screening threshold of $z_{\text{dev}} > 0$ to deliberately allow poor candidates to reach the evaluation set. We compared this setup with a baseline that promotes every submitted candidate without testing it on S. The proposer of the untested version is told that each submitted candidate was tested and accepted so the two versions differ only in the application of the test. Figure 4 illustrates every evaluated candidate under these conditions. REUSE rejected all 9 candidates with population improvement at most zero and made no false promotion among its 47 accepted updates. Without the test the process committed 6 false promotions across 5 of the 10 runs with population improvements dropping down to -0.56 pp. The two tested variants achieved very similar final population improvements, with no statistically significant difference between them. This confirms the test on S is strictly necessary and effective at filtering out the false improvements that pass the initial development screen.

Table 7: False promotions of the alternative REUSE variant and of the seven compared methods over the 30 seeds at each evaluation set size. Prom. and False are identical to the metrics in Table 1. Runs is the number of runs out of 30 with at least one false promotion.

	$n = 2{,}000$			$n = 10{,}000$		
Method	Prom.	False ↓	Runs ↓	Prom.	False ↓	Runs ↓
Empirical best-of-K	13.0	75	30/30	14.2	71	28/30
SICA-style	12.3	72	30/30	12.5	62	26/30
DGM-style	6.4	34	22/30	7.2	18	15/30
Elite	14.3	89	27/30	13.7	59	25/30
Niche-elite	12.3	71	28/30	12.4	44	24/30
PACE-style	6.4	4	4/30	8.1	3	3/30
Bonferroni-style	1.3	0	0/30	6.5	0	0/30
Alternative variant	1.0	**0**	**0/30**	3.3	**0**	**0/30**

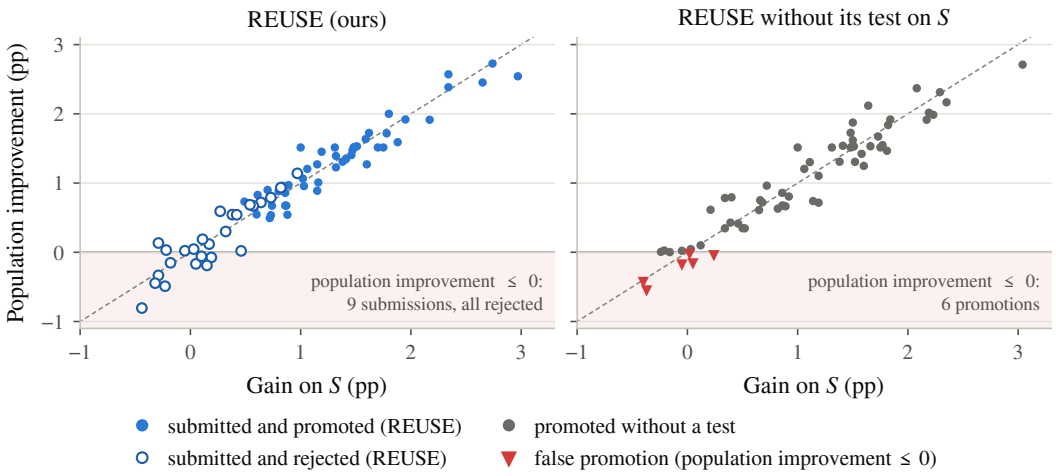


Figure 4: Performance of REUSE with and without its test on S at $n = 10{,}000$ over 10 seeds using a 500 example development set and pre-screening threshold $z_{\text{dev}} > 0$. Each point is a submitted candidate placed at its gain on S and its population improvement over the incumbent estimated on H. The shaded band marks population improvement at most zero and the dashed line is equality. The left panel shows REUSE whose test on S promotes the filled points and rejects the open circles. The right panel shows the same rule without the test which promotes every submitted candidate. Red triangles are false promotions.

B.4.2 PER-SEED RESULTS

Tables 8 and 9 list every run: for each method and seed, the final population improvement and the numbers of false and of all promotions.

Table 8: Per-seed results at $n = 2{,}000$ for the 30 seeds of Table 1. For each of the eight methods, a cell gives the final population improvement of the deployed system over A_0 in percentage points, estimated on the held-out set H, followed in parentheses by the run's false promotions and all promotions; nonzero false counts are in red. Mean averages the final population improvement over the 30 seeds and Total sums the promotion counts. Mean and the false counts in Total equal the corresponding columns of Table 1, and the total promotions divided by 30 give its Prom. column.

Seed	Empirical best-of-K	SICA-style	DGM-style	Elite	Niche-elite	PACE-style	Bonferroni-style	REUSE (ours)
3	7.14 (1/9)	5.76 (3/11)	4.65 (1/8)	7.14 (3/13)	7.17 (3/20)	7.22 (0/9)	0.00 (0/0)	7.24 (0/3)
4	7.25 (2/12)	7.24 (3/14)	5.26 (0/5)	7.25 (0/9)	7.24 (2/9)	5.75 (0/5)	3.21 (0/2)	5.73 (0/3)
5	7.06 (3/14)	7.41 (3/16)	2.77 (2/7)	4.65 (4/15)	6.81 (1/14)	3.79 (0/5)	0.00 (0/0)	7.26 (0/4)
6	6.70 (3/16)	5.55 (2/11)	2.33 (2/4)	6.70 (5/19)	6.70 (6/19)	6.72 (1/8)	4.11 (0/2)	6.70 (0/4)
7	7.14 (3/15)	5.92 (2/14)	3.21 (1/9)	5.92 (5/18)	6.50 (3/13)	4.65 (0/4)	0.86 (0/1)	10.00 (0/5)
8	7.24 (4/14)	7.24 (3/13)	2.82 (1/5)	7.24 (8/20)	7.24 (2/10)	7.24 (0/6)	3.93 (0/2)	7.24 (0/3)
9	7.02 (2/11)	7.24 (1/9)	4.03 (2/10)	7.02 (2/15)	4.65 (2/11)	7.24 (0/9)	0.00 (0/0)	7.24 (0/4)
10	7.10 (6/19)	5.40 (9/19)	4.43 (0/6)	5.75 (4/13)	7.27 (4/14)	7.17 (0/8)	5.76 (0/3)	5.06 (0/2)
11	5.45 (2/13)	5.75 (2/10)	1.86 (0/4)	7.37 (3/12)	7.34 (0/10)	6.52 (0/6)	1.51 (0/1)	5.76 (0/3)
12	7.37 (3/16)	6.64 (2/14)	4.30 (2/6)	7.37 (3/15)	5.24 (5/14)	7.37 (0/7)	0.00 (0/0)	10.80 (0/3)
13	7.37 (1/9)	5.32 (3/10)	4.65 (0/6)	6.84 (1/11)	6.15 (0/10)	7.24 (0/7)	4.65 (0/3)	5.92 (0/3)
14	7.34 (1/8)	7.34 (1/13)	4.92 (0/4)	6.64 (2/14)	5.77 (2/11)	7.34 (0/8)	4.11 (0/3)	7.41 (0/4)
15	7.22 (3/16)	6.18 (3/11)	3.89 (1/10)	6.70 (1/10)	5.67 (2/8)	6.55 (1/7)	4.65 (0/2)	4.80 (0/2)
16	7.34 (2/14)	7.34 (2/12)	2.73 (2/8)	7.02 (2/16)	7.34 (1/14)	7.14 (0/5)	2.73 (0/2)	6.40 (0/3)
17	6.80 (2/10)	7.37 (1/14)	2.84 (1/7)	7.37 (1/13)	7.37 (1/11)	7.37 (0/7)	0.00 (0/0)	6.31 (0/2)
18	7.14 (2/14)	4.65 (3/13)	4.11 (2/6)	5.52 (4/15)	5.71 (2/12)	7.24 (0/8)	1.20 (0/1)	6.84 (0/4)
19	6.77 (7/16)	6.84 (2/13)	4.06 (1/6)	5.29 (6/16)	7.24 (2/12)	5.14 (1/6)	0.00 (0/0)	5.72 (0/2)
20	5.87 (2/14)	6.84 (1/10)	4.23 (0/4)	5.29 (2/10)	7.25 (2/15)	7.14 (0/7)	2.19 (0/1)	6.45 (0/3)
21	7.37 (2/12)	5.75 (3/13)	5.23 (1/8)	7.27 (0/12)	5.80 (5/15)	7.14 (0/7)	4.11 (0/2)	5.80 (0/3)
22	7.24 (1/10)	7.27 (1/12)	2.77 (0/3)	6.18 (3/14)	5.26 (1/9)	6.49 (1/7)	4.11 (0/2)	5.76 (0/3)
23	7.37 (1/9)	7.37 (1/11)	4.56 (1/5)	7.37 (0/12)	5.68 (3/11)	5.72 (0/5)	4.65 (0/2)	9.67 (0/4)
24	7.24 (5/19)	5.82 (2/10)	4.43 (0/6)	7.10 (3/16)	5.76 (2/8)	4.65 (0/3)	5.65 (0/3)	7.10 (0/3)
25	7.34 (1/15)	5.65 (1/11)	3.35 (2/5)	7.27 (3/16)	7.34 (4/14)	6.64 (0/6)	2.73 (0/1)	6.84 (0/3)
26	6.80 (5/15)	5.52 (3/9)	4.65 (2/9)	7.12 (2/12)	7.05 (4/16)	6.08 (0/5)	3.21 (0/2)	6.45 (0/3)
27	6.70 (3/12)	7.14 (2/12)	4.30 (1/6)	6.70 (5/17)	7.14 (3/14)	7.24 (0/6)	2.95 (0/1)	7.02 (0/3)
28	6.95 (1/11)	3.49 (4/11)	3.51 (1/8)	7.14 (2/11)	6.95 (2/14)	7.12 (0/8)	0.00 (0/0)	7.18 (0/4)
29	7.14 (4/12)	6.70 (3/15)	2.92 (1/4)	6.81 (3/13)	6.45 (2/8)	7.24 (0/6)	4.07 (0/2)	10.00 (0/5)
30	7.34 (1/10)	7.05 (1/14)	4.65 (4/8)	7.06 (5/21)	5.00 (1/10)	6.50 (0/7)	0.00 (0/0)	7.34 (0/3)
31	7.24 (1/13)	7.05 (3/12)	4.11 (2/8)	7.24 (3/15)	7.05 (3/16)	5.68 (0/7)	0.00 (0/0)	7.10 (0/3)
32	7.25 (1/12)	5.94 (2/13)	5.14 (1/8)	7.14 (4/16)	5.71 (1/8)	4.11 (0/4)	0.00 (0/0)	7.41 (0/3)
Mean	7.04	6.36	3.89	6.72	6.46	6.45	2.35	7.02
Total	(75/390)	(72/370)	(34/193)	(89/429)	(71/370)	(4/193)	(0/38)	(0/97)

Table 9: Per-seed results at $n = 10{,}000$ for the 30 seeds of Table 1. For each of the eight methods, a cell gives the final population improvement of the deployed system over A_0 in percentage points, estimated on the held-out set H, followed in parentheses by the run's false promotions and all promotions; nonzero false counts are in red. Mean averages the final population improvement over the 30 seeds and Total sums the promotion counts. Mean and the false counts in Total equal the corresponding columns of Table 1, and the total promotions divided by 30 give its Prom. column.

Seed	Empirical best-of-K	SICA-style	DGM-style	Elite	Niche-elite	PACE-style	Bonferroni-style	REUSE (ours)
3	7.17 (1/17)	4.65 (0/5)	2.92 (1/7)	7.41 (4/20)	7.24 (2/10)	7.24 (0/9)	7.10 (0/7)	6.70 (0/5)
4	7.37 (1/13)	7.44 (2/14)	5.27 (1/8)	7.27 (3/16)	5.61 (0/7)	7.37 (0/7)	7.24 (0/7)	7.25 (0/8)
5	6.64 (2/14)	7.37 (2/13)	2.19 (0/5)	6.45 (2/8)	5.80 (0/9)	7.24 (0/9)	7.10 (0/7)	7.37 (0/6)
6	7.41 (2/15)	7.34 (2/13)	5.46 (0/7)	7.24 (1/9)	6.57 (2/20)	7.27 (0/10)	4.65 (0/4)	7.34 (0/7)
7	7.41 (3/17)	5.92 (1/13)	4.65 (1/8)	6.62 (4/21)	5.18 (0/13)	7.24 (0/9)	7.05 (0/7)	7.24 (0/6)
8	7.34 (3/13)	7.34 (2/13)	2.55 (0/4)	7.34 (1/15)	5.46 (1/12)	7.22 (0/8)	6.45 (0/6)	6.62 (0/6)
9	5.75 (2/11)	7.24 (5/15)	4.65 (1/8)	7.27 (1/13)	7.37 (2/15)	7.17 (0/9)	7.24 (0/7)	7.24 (0/6)
10	7.26 (0/10)	7.14 (3/15)	4.71 (0/11)	7.24 (2/21)	6.02 (0/12)	7.24 (0/8)	6.45 (0/7)	7.34 (0/6)
11	7.41 (2/16)	4.46 (3/9)	2.73 (1/9)	5.76 (0/9)	7.34 (3/18)	7.05 (1/9)	6.52 (0/6)	6.45 (0/6)
12	7.34 (3/14)	7.37 (1/17)	3.27 (0/5)	6.52 (2/12)	6.24 (2/11)	7.10 (0/8)	7.24 (0/7)	7.10 (0/5)
13	7.34 (2/13)	6.41 (1/10)	4.65 (0/6)	7.34 (2/15)	4.60 (1/8)	7.34 (0/8)	7.24 (0/7)	7.34 (0/6)
14	7.12 (2/12)	7.14 (4/15)	4.19 (2/9)	5.52 (3/11)	7.37 (4/14)	7.24 (1/8)	7.24 (0/7)	7.24 (0/5)
15	7.37 (1/13)	6.60 (3/15)	4.52 (1/9)	7.37 (1/14)	5.65 (2/11)	7.41 (0/11)	5.65 (0/6)	6.70 (0/5)
16	7.24 (1/8)	7.34 (1/16)	5.24 (0/7)	7.10 (3/16)	5.72 (0/7)	7.24 (0/6)	7.24 (0/5)	7.05 (0/6)
17	7.12 (2/11)	6.28 (0/10)	2.54 (1/6)	7.41 (1/17)	7.37 (1/12)	7.05 (0/7)	5.65 (0/5)	7.26 (0/7)
18	7.24 (1/10)	5.00 (3/11)	4.79 (0/6)	7.24 (0/14)	5.94 (1/9)	7.10 (0/7)	7.37 (0/7)	7.10 (0/6)
19	7.24 (4/18)	7.24 (4/18)	3.25 (1/8)	7.27 (0/9)	7.24 (2/16)	7.24 (0/10)	7.37 (0/7)	7.05 (0/6)
20	7.37 (4/18)	7.34 (2/16)	5.27 (1/10)	7.37 (2/16)	5.92 (2/13)	7.24 (0/7)	7.24 (0/7)	6.48 (0/6)
21	7.24 (3/16)	7.24 (2/15)	2.92 (0/5)	6.01 (3/13)	7.24 (1/15)	7.14 (0/8)	7.05 (0/7)	10.00 (0/6)
22	7.24 (3/13)	5.92 (2/12)	4.65 (0/7)	5.72 (1/12)	5.76 (1/10)	7.34 (0/8)	7.34 (0/7)	6.40 (0/6)
23	7.12 (6/19)	6.02 (0/6)	4.24 (3/10)	5.87 (2/13)	6.84 (3/18)	7.34 (0/8)	7.24 (0/6)	6.84 (0/6)
24	7.41 (5/18)	5.75 (3/13)	2.73 (0/5)	7.24 (3/14)	5.92 (1/14)	7.24 (0/7)	7.02 (0/7)	7.10 (0/6)
25	7.24 (1/17)	5.72 (1/10)	4.52 (0/10)	7.34 (0/8)	5.94 (1/13)	7.02 (0/8)	7.24 (0/7)	9.08 (0/6)
26	7.34 (3/16)	7.24 (4/16)	4.70 (1/8)	7.24 (3/17)	7.24 (3/13)	7.41 (1/9)	6.60 (0/6)	6.84 (0/4)
27	7.10 (2/10)	6.40 (1/7)	4.43 (0/7)	7.14 (3/12)	7.24 (2/10)	7.24 (0/6)	7.37 (0/6)	7.26 (0/6)
28	7.25 (2/12)	6.70 (2/12)	4.65 (1/9)	5.58 (2/11)	7.25 (0/10)	7.24 (0/7)	7.24 (0/7)	6.45 (0/6)
29	7.14 (6/21)	7.24 (2/9)	3.21 (0/3)	7.24 (2/11)	6.24 (2/14)	7.34 (0/8)	7.37 (0/6)	7.37 (0/6)
30	7.28 (3/15)	5.76 (3/11)	4.52 (0/5)	7.37 (0/13)	7.28 (1/12)	7.10 (0/8)	5.82 (0/5)	7.22 (0/5)
31	7.12 (1/15)	5.94 (0/11)	2.77 (1/5)	7.37 (5/19)	5.18 (3/13)	7.26 (0/8)	7.37 (0/7)	7.17 (0/7)
32	7.25 (0/11)	7.37 (3/16)	4.65 (1/10)	5.71 (3/11)	6.40 (1/14)	7.14 (0/8)	7.24 (0/7)	7.14 (0/5)
Mean	7.19	6.56	4.03	6.85	6.37	7.23	6.90	7.19
Total	(71/426)	(62/376)	(18/217)	(59/410)	(44/373)	(3/243)	(0/194)	(0/177)

B.4.3 TRAJECTORIES

Figure 5 separates the trajectories of Figure 2 by method and adds standard errors across seeds.

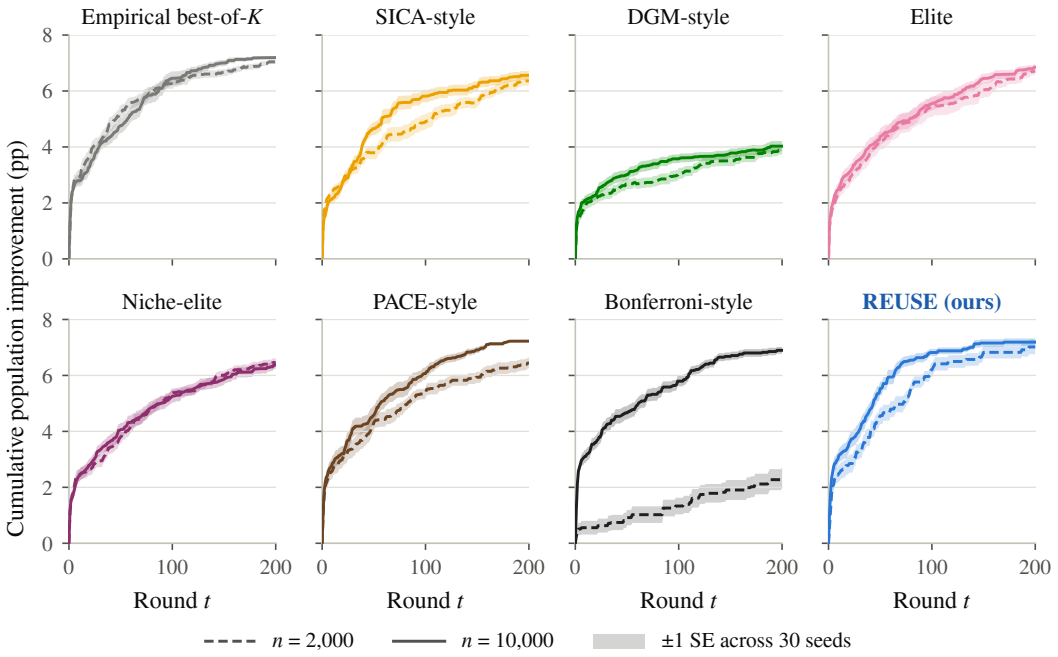


Figure 5: Mean population improvement of the deployed system over A_0 by round, estimated on H and averaged over 30 seeds, with one panel per method; dashed lines use $n = 2{,}000$ and solid lines use $n = 10{,}000$. Shaded bands show ± 1 standard error across the 30 seeds.

B.4.4 PROMOTIONS AND FINAL PERFORMANCE

Figure 6 plots the final population improvement of every run against its number of promotions. REUSE reaches a final population improvement comparable to that of Empirical best-of-K (Table 5) with far fewer promotions, a median of 3 against 13 per run at $n = 2{,}000$ and 6 against 14 at $n = 10{,}000$. No run of the seven compared methods improves on A_0 by more than 7.44 pp, whereas four runs of REUSE at $n = 2{,}000$ and two at $n = 10{,}000$ improve on it by 9.08 to 10.80 pp.

B.4.5 CERTIFICATES

REUSE records the running certificate C_T of Section 2.2 and the direct certificate D_T of Section A.5. Both are lower confidence bounds computed from the reused evaluation set alone, so they are reported separately from the population-gain estimates on H. Table 10 compares them with the realized population improvement, and none of the 274 per-promotion lower bounds ℓ that enter C_T exceeds the population improvement of its promotion. The direct certificate is larger than the running certificate in all 60 runs, because it pays one confidence margin for the whole trajectory rather than one per promotion.

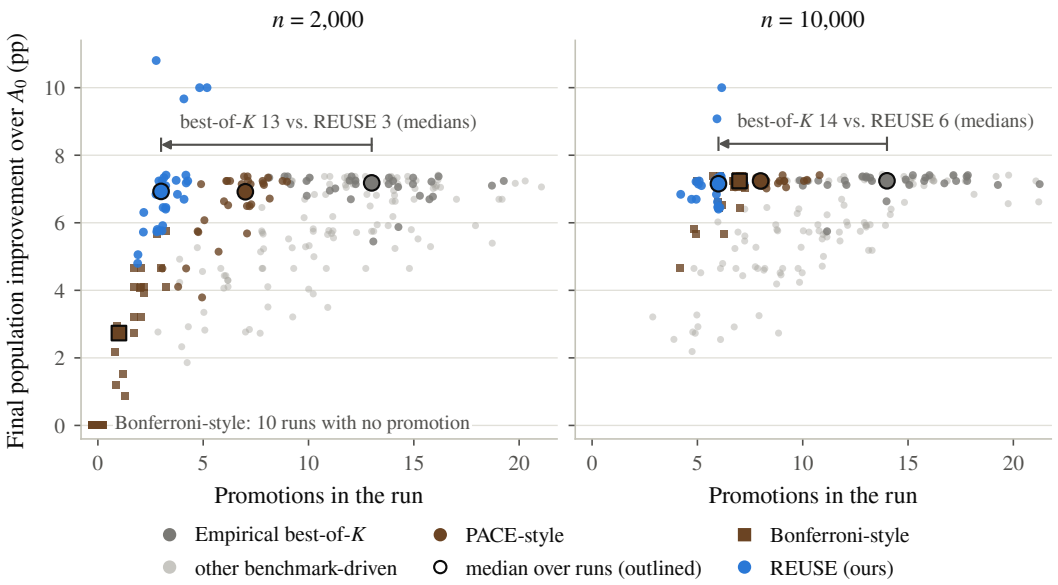


Figure 6: Final population improvement over A_0 against the number of promotions in the run, estimated on H, one point per run with small horizontal jitter, for the 30 runs of each method at each evaluation-set size. Light gray points are the other benchmark-driven frameworks (SICA-style, DGM-style, Elite and Niche-elite), and outlined markers place Empirical best-of-K, PACE-style, Bonferroni-style and REUSE at their median number of promotions and median final population improvement. The arrow joins the medians of Empirical best-of-K and REUSE.

Table 10: Certificates of cumulative improvement for REUSE, over the 30 runs at each evaluation-set size. The final population improvement over A_0 is the mean over runs, estimated on H as in Table 1. C_T and D_T are reported as the mean over runs with the range across runs in brackets. Valid gives, for C_T and for D_T, the number of runs (of 30) in which the certificate is at most the realized final population improvement. The final population improvement, C_T and D_T are in percentage points.

n	Final population improvement	Promotions per run	Running certificate C_T	Direct certificate D_T	Valid C_T / D_T
2,000	7.02	3.2	0.92 [0.13, 1.76]	3.10 [1.16, 5.49]	30 / 30
10,000	7.19	5.9	2.67 [2.06, 3.88]	5.33 [4.34, 6.92]	30 / 30