ACCURACY IS NOT STABILITY: ANSWER TURNOVER IN TEST-TIME SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

More inference compute can repair errors but also discard answers that were already correct. We study these opposing flows, repair and corruption, under nested best-of-N selection. Two populations can share identical accuracy and oracle-coverage curves at every width, yet lose correct answers at different rates. Exact score-record laws explain this gap: if correctness at extreme verifier scores tends to $\ell \in (0, 1)$, each flow from N to M candidates is asymptotic to $(1-N/M)\,\ell(1-\ell)$ as N grows. Adjacent turnover therefore vanishes, whereas multiplicative budget increases retain positive turnover even when accuracy rises monotonically. We derive unbiased finite-pool estimators of both flows with a Rao–Blackwell variance guarantee, and formulate answer preservation as a whole-policy risk constraint. Across five model–task–verifier settings, expanding from one to 32 candidates improves accuracy, yet loses a previously correct answer in 13.9–18.5% of evaluations. At a mean budget of about 2.3 candidates, a preservation policy attains similar observed accuracy to randomized widths while reducing conditional corruption by 12.94 percentage points; an adaptive comparator attains higher accuracy with more corruption. These results separate accuracy improvement from answer preservation and quantify their trade-off.

1 INTRODUCTION

Test-time scaling is usually evaluated by accuracy against compute. Repeated sampling and verification can improve reasoning performance (Snell et al., 2024; Brown et al., 2024), but an accuracy curve hides which answers changed. A gain from 70% to 72% could combine 14 percentage points of repaired errors with 12 points of lost correct answers. The distinction matters when the larger run contains the smaller run's candidates: an imperfect verifier can then discard a correct answer that is still available (Figure 1).

Negative flips after model updates (Yan et al., 2021), self-revision failures (Zeng et al., 2025; Zhou et al., 2026), and reward overoptimization (Gao et al., 2023; Stroebl et al., 2024; Khalaf et al., 2025) show that additional optimization can hurt. We ask what complete scaling curves reveal about preservation along a shared candidate stream. Even accuracy and oracle coverage at *every* width leave paired transitions unidentified: populations that match on both curves can have vanishing or persistent turnover.

Accuracy averages correctness at each score quantile, whereas paired transitions depend on products of correctness probabilities at two quantiles of one input. Exact score-record laws expose this dependence and a measurement-grid effect: when extreme-score correctness converges to a value in $(0, 1)$, adjacent turnover decays like $1/N$, whereas multiplicative budget increases retain positive repair and corruption, even with monotone accuracy gains and an informative verifier.

We make three contributions. First, matching-curve constructions and record laws characterize what population scaling curves fail to identify (Sections 3 and 4). Second, paired subset averaging gives unbiased flow estimates from cached candidate pools, extending established best-of-N estimators (Nakano et al., 2021) with a variance guarantee (Section 5). Third, we separate proposing an answer from replacing the incumbent, and give a finite-sample certificate for the final conditional corruption of a gated policy, COMPUTEGUARD, with Learn-then-Test (Angelopoulos et al., 2021) (Section 6). Experiments measure hidden turnover and evaluate COMPUTEGUARD (Section 7). Together, the theory and measurements separate an accuracy gain from the preservation of already-correct answers.

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

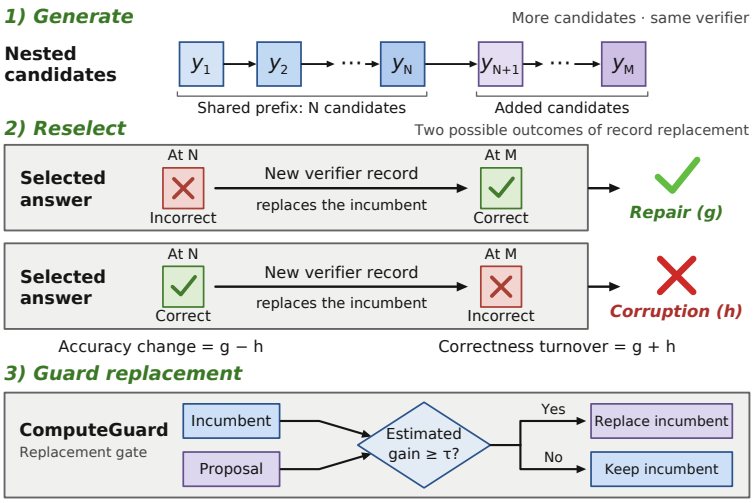


Figure 1: **Accuracy gains can hide answer turnover.** A larger nested pool can repair an error, or replace a correct answer that is still in the pool. Here $g = \bar{g}_{N,M}$ and $h = \bar{h}_{N,M}$ are population repair and corruption. COMPUTEGUARD replaces the incumbent only when the estimated correctness gain clears a margin. Rows show distinct cases. Correctness markers are for exposition and unavailable to the policy; continuation and policy-level calibration are omitted.

2 PAIRED SCALING AND THE CANDIDATE MODEL

Let $x \sim \mathcal{D}$ be an input, $u_x(y) \in \{0, 1\}$ its correctness function, and $C_N = u_x(Y_N)$ the correctness of the answer selected at width N. Throughout the width analysis, nested runs share prefixes of one candidate stream. Write $\bar{A}_N = \mathbb{P}(C_N = 1)$. For $M > N$, define population corruption, repair, and conditional corruption by

$$\bar{h}_{N,M} = \mathbb{P}(C_N = 1, C_M = 0), \quad \bar{g}_{N,M} = \mathbb{P}(C_N = 0, C_M = 1), \quad R^+_{N,M} = \frac{\bar{h}_{N,M}}{\bar{A}_N}. \tag{1}$$

The ratio requires $\bar{A}_N > 0$. Counting transitions gives

$$\bar{A}_M - \bar{A}_N = \bar{g}_{N,M} - \bar{h}_{N,M}, \qquad \mathbb{P}(C_N \neq C_M) = \bar{g}_{N,M} + \bar{h}_{N,M}. \tag{2}$$

We call $\mathbb{P}(C_N \neq C_M)$ *correctness turnover*; changes between two correct answers, or between two incorrect ones, do not count. The coupling is part of the estimand: independently sampled runs answer a different preservation question. All comparisons keep the prompt, decoding distribution, scoring rule, and candidate tie keys fixed while extending the stream.

Conditional candidate model. For each fixed x, candidates (S_i, B_i) are i.i.d., where S_i is a fixed pointwise verifier score and B_i its correctness label. Assume a continuous score CDF F_x and set $U_i = F_x(S_i)$, which is uniform. Define

$$q_x(u) = \mathbb{P}(B_i = 1 \mid U_i = u, x), \qquad p_x = \int_0^1 q_x(u)\, du. \tag{3}$$

Best-of-N selects the largest score. Below, unbarred quantities are conditional on x, which we suppress; bars denote population averages. Let $O_N = \max_{i \le N} B_i$ denote oracle coverage. Since the maximum quantile has density Nu^{N-1},

$$A_N = N \int_0^1 q(u) u^{N-1}\, du, \qquad \mathbb{P}(O_N = 1 \mid x) = 1 - (1 - p_x)^N, \tag{4}$$

and population coverage is $1 - \mathbb{E}_x[(1 - p_x)^N]$. For discrete scores, each candidate receives an independent continuous tie key, reused at every width; the resulting total order defines the quantile model.

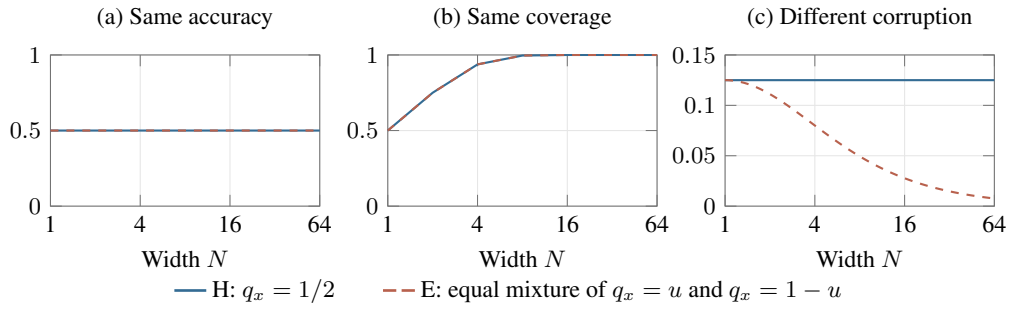


Figure 2: **Two complete curves still miss stability.** The populations of corollary 1 have identical accuracy and oracle coverage at every width, yet their corruption under $N \to 2N$ differs: $1/8$ for H versus $N/\{2(N+1)^2\} \to 0$ for E. Repair equals corruption in both. All curves are exact.

3 What Scaling Curves Cannot Identify

Proposition 1 (Identical accuracy curves, different stability)**.** *Fix $a \in (0,1)$. For every $\kappa \in [0, a(1-a)]$, there is a population of conditionally i.i.d. candidates satisfying*

$$\bar{A}_N = a \quad \textit{for every } N, \qquad \bar{h}_{N,M} = \bar{g}_{N,M} = \frac{M-N}{M}\kappa \quad \textit{for every } N < M. \tag{5}$$

Consequently, the complete population accuracy curve does not identify paired corruption or repair.

Proof. Take $q_x(u) \equiv L_x$ with $\mathbb{E}L_x = a$. Every selected candidate has conditional correctness L_x, and a new block of $M - N$ candidates wins with probability $(M-N)/M$, independently of correctness. Thus each population flow is $(M-N)\mathbb{E}[L_x(1-L_x)]/M$. The constant $L_x = a$ gives $\kappa = a(1-a)$, a Bernoulli $L_x \in \{0,1\}$ with mean a gives $\kappa = 0$, and mixtures attain every intermediate value. □

Oracle coverage can differ in this construction; the next one also matches it.

Corollary 1 (Matching both accuracy and oracle coverage)**.** *Consider population H with $q_x(u) = 1/2$ for every input, and population E with $q_x(u) = u$ on half of the inputs and $q_x(u) = 1 - u$ on the other half. In both populations, for every N,*

$$\bar{A}_N = 1/2, \qquad \mathbb{P}(O_N = 1) = 1 - 2^{-N}. \tag{6}$$

Their doubling flows nevertheless satisfy

$$\bar{h}^{\mathrm{H}}_{N,2N} = \bar{g}^{\mathrm{H}}_{N,2N} = \frac{1}{8}, \qquad \bar{h}^{\mathrm{E}}_{N,2N} = \bar{g}^{\mathrm{E}}_{N,2N} = \frac{N}{2(N+1)^2} \longrightarrow 0. \tag{7}$$

Proof. All inputs have $p_x = 1/2$; in population E, the conditional accuracies $N/(N+1)$ and $1/(N+1)$ average to $1/2$. Population H is the case $L_x \equiv 1/2$ of proposition 1. In E, eq. (8) with $M = 2N$ gives corruptions for $q = u$ and $q = 1 - u$ that sum to $N/(N+1)^2$; repair equals corruption because $\bar{A}_N$ is constant. □

The distinction is heterogeneity across inputs. With $\bar{q}(u) = \mathbb{E}_x q_x(u)$ and $Q_2(u,v) = \mathbb{E}_x[q_x(u)q_x(v)]$, the population-averaged corruption integrand is $\bar{q}(u) - Q_2(u,v)$, and the repair integrand is $\bar{q}(v) - Q_2(u,v)$. The moments $\bar{A}_N/N$ identify $\bar{q}$ almost everywhere, but not Q_2. By contrast, the complete *conditional* curve of an input identifies q_x almost everywhere, and hence its flows; our obstruction concerns population curves only.

4 Record Flows and the Scale of Turnover

Write $K = M - N$. The old winner's quantile U and the added block's winner quantile V are independent, with densities Nu^{N-1} and Kv^{K-1}. Correctness can change only if $V > U$.

Proposition 2 (Exact paired record flows). *For every $M > N$ under the conditional model,*

$$h_{N,M} = NK \int_0^1 u^{N-1} q(u) \int_u^1 v^{K-1}[1 - q(v)]\, dv\, du,$$
$$g_{N,M} = NK \int_0^1 u^{N-1}[1 - q(u)] \int_u^1 v^{K-1} q(v)\, dv\, du. \tag{8}$$

They satisfy $g_{N,M} - h_{N,M} = A_M - A_N$ and $g_{N,M} + h_{N,M} \le K/M$.

The integrands multiply the old winner's correctness, the new winner's incorrectness, and their joint density; exchanging the correctness factors gives repair. Symmetry among the M candidates gives $\mathbb{P}(V > U) = K/M$. Setting $K = 1$ yields the adjacent flows $h_N = h_{N,N+1}$ and $g_N = g_{N,N+1}$.

A finite-budget bound. Suppose $q(u) \to \ell$ as $u \uparrow 1$, and define $\omega_q(t) = \sup_{u \in (t,1)} |q(u) - \ell|$. Concentration of the maximum near one gives

$$|A_N - \ell| \le \omega_q(t) + t^N. \tag{9}$$

The limit ℓ describes correctness at extreme verifier scores; it need not bound accuracy at finite widths. Conditional on the new block winning, (U, V) has density $MNu^{N-1}v^{K-1}$ on $\{u < v\}$, and $U \le t$ has probability

$$B_{N,M}(t) = \frac{Mt^N - Nt^M}{M - N} \le t^N[1 + N(1 - t)]. \tag{10}$$

Proposition 3 (Finite-budget turnover bound). *For either $f_{N,M} = h_{N,M}$ or $f_{N,M} = g_{N,M}$ and every $t \in (0, 1)$,*

$$\left| \frac{M}{M - N} f_{N,M} - \ell(1 - \ell) \right| \le 2\omega_q(t) + B_{N,M}(t). \tag{11}$$

On $\{U > t\}$, both correctness factors are within $\omega_q(t)$ of their limits; the remaining probability is $B_{N,M}(t)$, whose bound in eq. (10) is free of M, so the result is uniform in the expansion size.

Theorem 1 (Turnover across scaling intervals). *If $q(u) \to \ell \in [0, 1]$ as $u \uparrow 1$, then for every integer sequence $M_N > N$,*

$$\frac{M_N}{M_N - N} h_{N,M_N} \to \ell(1 - \ell), \qquad \frac{M_N}{M_N - N} g_{N,M_N} \to \ell(1 - \ell). \tag{12}$$

Thus for $\ell \in (0, 1)$, $h_N \sim g_N \sim \ell(1 - \ell)/N$. If $M_N/N \to \lambda > 1$, then

$$h_{N,M_N}, g_{N,M_N} \to (1 - \lambda^{-1})\ell(1 - \ell). \tag{13}$$

If, in addition, q is absolutely continuous, $q' \in L^1(0, 1)$, and

$$q'(u) = -c\beta(1 - u)^{\beta-1} + o((1 - u)^{\beta-1}) \quad \text{as } u \uparrow 1 \quad (c \ne 0,\ \beta > 0), \tag{14}$$

then

$$A_N - \ell \sim c\Gamma(\beta + 1)N^{-\beta}, \tag{15}$$
$$\Delta_N := A_{N+1} - A_N \sim -c\beta\Gamma(\beta + 1)N^{-\beta-1}. \tag{16}$$

For $\ell \in (0, 1)$, gross adjacent turnover exceeds $|\Delta_N|$ by a factor of order N^β.

The first claim follows from proposition 3 by letting $N \to \infty$ at fixed $t < 1$ and then $t \uparrow 1$; the rates follow from beta integrals. The theorem separates a nearly flat curve from stable decisions: adjacent turnover vanishes, while multiplicative expansions keep replacing successes with errors and errors with successes.

Corollary 2 (Population turnover). *If $q_x(u) \to \ell_x$ as $u \uparrow 1$ for almost every x, the normalized population flows in eq.* (12) *converge to $\mathbb{E}_x[\ell_x(1 - \ell_x)]$. In particular,*

$$(N + 1)\bar{h}_N, (N + 1)\bar{g}_N \to \mathbb{E}_x[\ell_x(1 - \ell_x)]. \tag{17}$$

With $a = \mathbb{E}\ell_x$ and $\sigma_\ell^2 = \mathrm{Var}(\ell_x)$, the geometric-limit flows are $(1 - \lambda^{-1})\{a(1 - a) - \sigma_\ell^2\}$.

Dominated convergence applies because the normalized conditional flows lie in $[0, 1]$. At fixed $a \in (0, 1)$, the limiting conditional corruption therefore ranges over $[0, (1 - \lambda^{-1})(1 - a)]$; both endpoints are attained by the constructions in proposition 1.

Persistent turnover with an informative verifier. Take $q(u) = 0.2+0.6u$, so correctness increases strictly with score. Then $A_N = 0.8 - 0.6/(N+1)$ increases at every width, while theorem 1 gives $h_{N,2N}, g_{N,2N} \to 0.08$. A well-ordered but imperfect verifier can thus retain 16 percentage points of doubling turnover even as the accuracy gain tends to zero. Without monotonicity of q, the score ordering constrains no finite accuracy pattern: for any $d \in \mathbb{R}^{n_0}$, a $[0,1]$-valued polynomial q with $p_x = 1/2$ realizes $A_{n+1} - A_n = \gamma d_n$ for $n = 1, \dots, n_0$ and some $\gamma > 0$. The construction perturbs $q \equiv 1/2$ within the span of $\phi_n(u) = (n+1)u^n - nu^{n-1}$, which are linearly independent, integrate to zero, and satisfy $A_{n+1} - A_n = \int_0^1 q\phi_n$. This concerns attainable patterns, not their amplitude.

5 Estimating Paired Flows from Finite Pools

For one input, draw m candidates, sort them by increasing score, and write c_r for the correctness at rank r. The standard subset-averaged best-of-N estimator (Nakano et al., 2021; Gao et al., 2023) is

$$\widehat{A}_N = \sum_{r=N}^{m} \frac{\binom{r-1}{N-1}}{\binom{m}{N}} c_r. \tag{18}$$

The same principle underlies pass@k estimation (Chen et al., 2021). For paired flows, draw a uniform old subset of size N and then a uniform expansion of size $K = M - N$ from its complement.

Proposition 4 (Paired finite-pool estimators)**.** *For $1 \le N < M \le m$, let*

$$W_{r,s} = \binom{r-1}{N-1}\binom{s-1-N}{K-1}, \qquad Z = \binom{m}{N}\binom{m-N}{K}. \tag{19}$$

Under the conditional model, unbiased estimators are

$$\widehat{h}_{N,M} = Z^{-1} \sum_{r=N}^{m-1} \sum_{s=\max\{M,r+1\}}^{m} W_{r,s} c_r (1 - c_s), \tag{20}$$

$$\widehat{g}_{N,M} = Z^{-1} \sum_{r=N}^{m-1} \sum_{s=\max\{M,r+1\}}^{m} W_{r,s} (1 - c_r) c_s. \tag{21}$$

For every realized pool, $\widehat{g}_{N,M} - \widehat{h}_{N,M} = \widehat{A}_M - \widehat{A}_N$.

For old maximum rank r and new maximum rank $s > r$, the binomial factors count the other $N-1$ old candidates below rank r and the other $K-1$ expansion candidates below rank s outside the old subset. Averaging over all such choices proves the formulas; an i.i.d. pool makes each choice distributed as a fresh nested experiment. After sorting, weighted suffix sums evaluate each budget pair in $O(m)$ time.

Let $H = \mathbf{1}\{C_N = 1, C_M = 0\}$ be the chronological corruption indicator and let $\mathcal{G}$ contain the input and the unordered scored and labeled pool, including tie keys. At fixed x, exchangeability yields $\widehat{h}_{N,M} = \mathbb{E}[H \mid \mathcal{G}]$, and therefore

$$\mathrm{Var}(\widehat{h}_{N,M}) \le \mathrm{Var}(H) = h_{N,M}(1 - h_{N,M}), \tag{22}$$

with an analogous bound for repair; the comparison also holds over heterogeneous inputs. Averaging the estimators across inputs, we estimate conditional corruption by $\widehat{\bar{h}}/\widehat{\bar{A}}_N$, which is consistent as the number of independent input–stream records grows when $\bar{A}_N > 0$, but need not be unbiased. Drawing more subsets from one fixed pool adds no independent observations; input-level uncertainty resamples prompts, preserving their repeated streams and candidate subsets.

6 Preserving an Incumbent under a Compute Budget

We use the paired objective to separate candidate generation from incumbent replacement. Let z denote the incumbent's observable state, v a proposal's state, $\widehat{q}$ a correctness predictor fitted on

development data, and $\widehat{P}_x$ a fitted proposal distribution. The observable state includes the evaluation setting and score bin. A margin τ gives the acceptance rule

$$D_\tau(z, v) = \mathbf{1}\{\widehat{q}(v) - \widehat{q}(z) \geq \tau\}. \tag{23}$$

Define one-step plug-in repair and corruption scores and their difference, the estimated net gain, by

$$\begin{aligned}
\widehat{g}_\tau(z) &= (1 - \widehat{q}(z)) \int \widehat{q}(v) D_\tau(z, v)\, d\widehat{P}_x(v), \\
\widehat{h}_\tau(z) &= \widehat{q}(z) \int (1 - \widehat{q}(v)) D_\tau(z, v)\, d\widehat{P}_x(v), \\
\widehat{V}_\tau(z) &= \int [\widehat{q}(v) - \widehat{q}(z)] D_\tau(z, v)\, d\widehat{P}_x(v).
\end{aligned} \tag{24}$$

These scores motivate a myopic continuation rule. Under the conditional model, conditioning on the full prompt and current quantile would make the analogous probabilities exact, but a fitted score-only model need not remain calibrated after adaptive selection. The resulting policy, COMPUTEGUARD (Algorithm 1), therefore uses the scores only for decisions and relies on separate whole-policy calibration for its guarantee.

Algorithm 1. COMPUTEGUARD with frozen (τ, μ)

Input: x, fitted $\widehat{q}$ and $\widehat{P}_x$, candidate cap M.
1. Generate and store the initial candidate and its state z; set $n = 1$.
2. While $n < M$ and $\widehat{V}_\tau(z) > \mu$:
   Generate a proposal with state v; increment n.
   If $D_\tau(z, v) = 1$, replace the incumbent and set $z = v$.
3. Return the incumbent.

The cost unit is one generated candidate, including rejected proposals, and μ is its utility penalty. The controller uses 20 score-quantile bins fitted on development candidates. For bin b, $\widehat{q}_b = (k_b + 1)/(n_b + 2)$ and $\widehat{p}_b = n_b / \sum_c n_c$, where k_b of n_b candidates are correct; no monotonicity is imposed. These setting-level bin masses serve as $\widehat{P}_x$ and ignore prompt-specific variation. If the incumbent lies in bin $b(z)$, continuation uses

$$\widehat{V}_\tau(z) = \sum_{b=1}^{20} \widehat{p}_b \, (\widehat{q}_b - \widehat{q}_{b(z)}) \, \mathbf{1}\{\widehat{q}_b - \widehat{q}_{b(z)} \geq \tau\}. \tag{25}$$

The policy family contains $J = 20$ pairs, with $\tau \in \{0.06, 0.10, 0.13, 0.16\}$ and $\mu \in \{0, 0.010, 0.020, 0.030, 0.040\}$, each capped at $M = 32$ candidates.

Whole-policy calibration. Before observing calibration outcomes, freeze the predictors, policies, selection rule, and multiple-testing procedure. On n independent calibration records, run each policy once, sharing the initial answer and candidate stream within each record. Let $C_{0,i}$ be the correctness of record i's initial answer (C_1 in width notation). Write

$$I_+ = \{i : C_{0,i} = 1\}, \quad n_+ = |I_+|, \quad K_j = \sum_{i \in I_+} \mathbf{1}\{C_{\pi_j, i} = 0\}, \quad R_j^+ = \mathbb{P}(C_{\pi_j} = 0 \mid C_0 = 1). \tag{26}$$

For $K_j < n_+$, the one-sided Clopper–Pearson bound is $U_j^+(\gamma) = \mathrm{Beta}^{-1}(1 - \gamma; K_j + 1, n_+ - K_j)$; set it to 1 when $K_j = n_+$ (Clopper and Pearson, 1934).

Proposition 5 (Learn-then-Test preservation certificate)**.** *Conditional on development data, assume that calibration records (inputs with their generation randomness) are i.i.d., that $\mathbb{P}(C_0 = 1) > 0$, and that deployment follows the same law. Either rule below controls the selected policy's risk:*

1. Bonferroni: *retain every policy with $U_j^+(\delta/J) \leq \alpha$.*
2. Fixed sequence: *in a development-fixed order, retain policies up to, but excluding, the first with $\mathbb{P}\{\mathrm{Bin}(n_+, \alpha) \leq K_j\} > \delta$.*

Table 1: **Expanding from 1 to 32 candidates in five settings.** Rates are percentages; ΔA and oracle-coverage gain ΔCov. are percentage points, with $\Delta A = g - h$ computed from the displayed rates. Conditional corruption is $R^+ = h/A_1$.

Setting	ΔA	Repair g	Corruption h	R^+	$g + h$	ΔCov.
S1/T-A/V1	12.70	27.30	14.60	26.29	41.90	44.45
S1/T-B/V1	10.43	25.45	15.02	35.80	40.47	57.68
S2/T-C/V1	11.10	25.00	13.90	42.03	38.90	64.95
S2/T-D/V2	6.63	24.20	17.57	34.92	41.77	49.55
S3/T-A/V2	3.10	21.60	18.50	32.46	40.10	42.98

Selecting any retained policy, with the initial-answer policy as fallback if $n_+ = 0$ or none is retained, gives

$$\mathbb{P}_{\mathrm{cal}}(R^+_{\widehat{\pi}} \leq \alpha) \geq 1 - \delta. \tag{27}$$

Conditional on n_+, each K_j is binomial. Bonferroni gives simultaneous coverage; fixed sequence bounds the probability of retaining the first unsafe policy. Neither argument requires independence across policies or monotonicity of risk in τ. We thus apply the existing Learn-then-Test framework (Angelopoulos et al., 2021) to final conditional corruption; the effective sample size is the number of initially correct calibration records. The guarantee controls final loss, allowing intermediate errors that are later repaired; it is neither a per-input nor an anytime guarantee. Choosing between the two procedures after inspecting their outputs requires an additional multiplicity correction.

For fixed-sequence testing, policies are ordered on development data: those with development risk at most 0.8α come first, ranked by development utility, accuracy minus 0.002 times candidate count, followed by the rest in the same order. Among retained policies, the highest development utility wins, with ties broken by ascending lexicographic order of (τ, μ). All fitting, ordering, and baseline cost matching precede calibration.

7 Empirical Evaluation

We ask how much turnover accompanies accuracy gains, what preservation costs at a similar candidate budget, and how the calibration bound relates to the policy's final loss. Tables report empirical estimates without overbars; g and h are unconditional, whereas R^+ conditions on initial correctness.

Settings and statistical units. In the supplied setting map, generators S1–S3 are Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct (Qwen et al., 2024), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024); tasks T-A–T-D are GSM8K (Cobbe et al., 2021), MATH-500 (Hugging Face H4, n.d.), BoolQ (Clark et al., 2019), and ARC-Challenge (Clark et al., 2018); verifiers V1 and V2 are generator likelihood and a pointwise likelihood judge. We evaluate five settings: S1/T-A/V1, S1/T-B/V1, S2/T-C/V1, S2/T-D/V2, and S3/T-A/V2. Section D details measurement units and policy parameters. The turnover study reports $4{,}000$ records per setting, each with two 32-candidate streams, at widths $\{1, 2, 4, 8, 16, 32\}$. The policy study reports $50{,}000$ equally weighted, setting-specific records. Records count evaluations, not distinct questions.

Pairing and uncertainty. Nested widths reuse a candidate stream, and all policies share the initial answer and candidate stream within a record, so every comparison is paired. Candidate counts include the initial answer and rejected proposals. The tables report point estimates; the comparisons below are therefore descriptive and assert neither statistical equivalence nor significance. Section D.1 specifies a task-stratified prompt-cluster bootstrap with $1{,}000$ draws that preserves repeated evaluations and cross-setting prompt linkage.

7.1 Turnover despite higher accuracy

Accuracy improves in every setting, yet 13.90–18.50% of evaluations lose a previously correct answer (Table 1); conditional on initial correctness, the loss is 26.29–42.03%. Repair exceeds corruption, so

Table 2: **Policy comparison on** $50{,}000$ **reported evaluation records.** Accuracy, conditional corruption R^+, and unconditional repair are percentages. Candidate counts measure generation only. Cal. U^+ is the Bonferroni-adjusted upper bound on R^+ from calibration counts, valid under proposition 5. Entries are point estimates.

Policy	Accuracy	R^+	Repair	Candidates	Cal. U^+
Initial answer	46.45	0.00	0.00	1.00	$0^\dagger$
Best-of-32	56.91	33.13	25.84	32.00	–
Fitted-q best-of-32	56.87	31.87	25.22	32.00	–
Random width	50.57	20.35	13.57	2.33	–
Adaptive raw-score rule	52.28	19.00	14.66	2.23	–
Gate without stopping	51.95	10.44	10.35	32.00	–
COMPUTEGUARD	50.54	7.41	7.53	2.32	0.0842

†Zero by construction.

the net gain is positive, but gross turnover remains 38.90–41.90%. The gain in oracle coverage thus reflects available solutions, not preservation of the selected one.

Across the five settings and five doubling steps, the reported summary identifies 8 of 25 intervals (32%) with an absolute accuracy change below one percentage point. The $16 \to 32$ interval in S2/T-D/V2 shows the cancellation: repair is 9.80% and corruption 9.90%, so 19.70 points of turnover yield a -0.10-point accuracy change; conditional corruption is 17.35%. This finite-width example illustrates eq. (2); it does not estimate the asymptotic limit in theorem 1. This threshold applies to point estimates and is not an equivalence test.

7.2 PRESERVATION AT A SIMILAR CANDIDATE BUDGET

Best-of-32 ranks by raw score and fitted-q best-of-32 by $\widehat{q}$. Gate without stopping uses the same gate and generates 32 candidates. The adaptive raw-score rule accepts raw-score records and stops when $\widehat{V}_0(z) \leq 0.020$ (Section D).

Preservation and repair move together. Random width draws 3 candidates with probability 0.329 and 2 otherwise, a mixture set to match the development-set candidate cost of COMPUTEGUARD. Its test mean is 2.33, versus 2.32 for COMPUTEGUARD. Conditional corruption falls from 20.35% to 7.41%, a 12.94-point reduction, while accuracy changes from 50.57% to 50.54%. At initial accuracy 46.45%, this amounts to about 6.01 fewer unconditional corruptions per 100 evaluations, offset by 6.04 fewer repairs. Similar accuracy thus reflects a shift in both flows, not an absence of answer changes (Figure 3).

Ranking, replacement, and stopping. At 32 candidates, ranking by $\widehat{q}$ reduces conditional corruption only from 33.13% to 31.87%. The adaptive raw-score rule reaches 52.28% accuracy and 19.00% conditional corruption at 2.23 candidates. Relative to it, COMPUTEGUARD lowers corruption by 11.59 points and accuracy by 1.74 points; since both its acceptance and continuation rules differ, this comparison assesses the operating points jointly. Within the fixed gate, adding stopping reduces the mean count from 32 to 2.32, corruption by 3.03 points, and accuracy by 1.41 points. These comparisons trace an accuracy–preservation trade-off; tokens, scorer overhead, and latency are not matched.

Calibration. The reported calibration sample has $n = 2{,}000$ records. For COMPUTEGUARD at $(\tau, \mu) = (0.13, 0.020)$, it contains 56 final corruptions among 919 initially correct records (6.09%). With $\alpha = \delta = 0.10$ and $J = 20$, the Bonferroni-adjusted bound is $U^+(\delta/J) \approx 0.0842 < \alpha$; the smaller unadjusted value $U^+(\delta) \approx 0.0723$ ignores policy selection. By proposition 5, the certificate is valid if the policy family and development fits were frozen and calibration records are i.i.d. from the deployment law; the counts alone cannot verify these conditions.

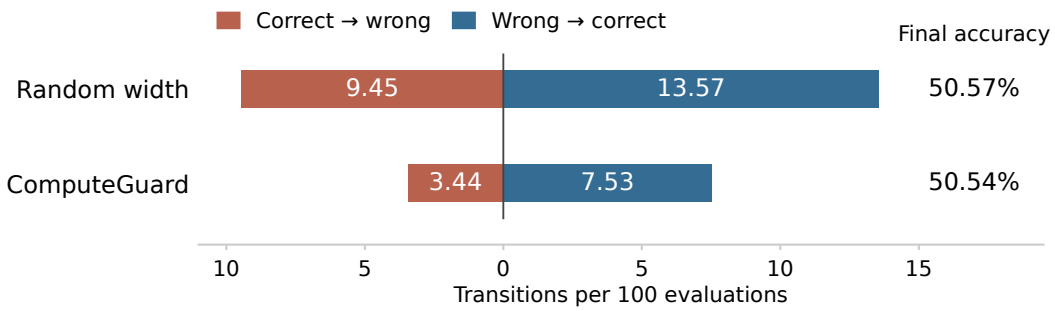


Figure 3: **Similar accuracy, different transition rates.** At mean candidate counts of 2.33 (random width) and 2.32 (COMPUTEGUARD), both policies improve on the 46.45% initial accuracy, but COMPUTEGUARD reduces repair and corruption together. Bars are unconditional point estimates from Table 2, with corruption computed as A_0R^+.

7.3 ESTIMATOR VALIDATION

Exact enumeration verifies the paired finite-pool formulas on all 2,688 label-vector/budget cases of a seven-candidate pool (2^7 label vectors times 21 budget pairs); another 1,020 polynomial checks verify the record identities (Section C). These checks complement the unbiasedness and variance results in proposition 4 and eq. (22). The reported comparison of finite-pool estimates from stream A with nested measurements on stream B covers five settings and $\binom{6}{2} = 15$ width pairs: 75 comparisons per metric. Mean absolute differences are 0.53, 0.48, and 0.37 percentage points for accuracy change, repair, and corruption. Stream B has its own sampling error, so aggregate agreement is a diagnostic, not a coverage test.

8 RELATED WORK AND LIMITATIONS

Sampling and verification enable test-time scaling (Snell et al., 2024; Brown et al., 2024); reward overoptimization and verifier limitations explain diminishing returns (Gao et al., 2023; Stroebl et al., 2024; Khalaf et al., 2025). Negative flips in model updates (Yan et al., 2021) and sequential overthinking (Zeng et al., 2025; Zhou et al., 2026; Ghosal et al., 2025) concern related events under different couplings. Our paired-budget laws show that turnover can persist while accuracy increases and is not identified even by complete accuracy and coverage curves.

Our estimators extend best-of-N subset averaging (Nakano et al., 2021), and Learn-then-Test supplies the calibration guarantee (Angelopoulos et al., 2021). Selective answering (Jurayj et al., 2025; Angelopoulos et al., 2024) controls response risk, and thought calibration (Wu et al., 2025) controls reasoning effort; our constraint instead concerns losing an initially correct answer. Our baselines do not establish superiority over these distinct algorithms.

Limitations. The record laws require conditionally i.i.d. candidates and fixed pointwise scoring. Adaptive generation, listwise judges, and self-consistency (Wang et al., 2023) require other selection models, although the bookkeeping identity still holds. The score-only controller can miss prompt heterogeneity and is not claimed to be optimal. Its certificate covers final corruption on the specified mixture of settings, not intermediate states, individual inputs, or distribution shifts. Candidate counts omit verifier cost and answer length, so the comparisons do not establish end-to-end compute savings.

9 CONCLUSION

Scaling curves do not determine preservation. Record laws and paired finite-pool estimators separate repair from corruption across budget scales, and whole-policy calibration bounds final corruption. The reported policy comparisons quantify an accuracy–preservation trade-off. Whenever computation can replace an existing answer, repair and corruption belong alongside accuracy.

AI USE STATEMENT

In this work, we used generative AI tools (large language models) solely for language polishing, including improving grammar, clarity, and readability. These tools were not used to generate substantive content or to perform any research tasks, including developing methods or theoretical frameworks, formulating claims or hypotheses, constructing proofs, designing experiments, generating or processing data, analyzing or interpreting results, writing code, searching the literature, translating text, or preparing figures. All AI-assisted revisions were reviewed by the authors to ensure that they preserved the intended meaning and technical accuracy. We take full responsibility for the final content of this work, including all claims, results, and references.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021. https://arxiv.org/abs/2110.01052.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. https://openreview.net/forum?id=33XGfHLtZg.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024. https://arxiv.org/abs/2407.21787.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. https://arxiv.org/abs/2107.03374.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. https://arxiv.org/abs/1803.05457.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*, 2019. https://arxiv.org/abs/1905.10044.

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. https://doi.org/10.1093/biomet/26.4.404.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. https://arxiv.org/abs/2110.14168.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866, 2023. https://proceedings.mlr.press/v202/gao23h.html.

Soumya Suvra Ghosal, Souradip Chakraborty, Avinash Reddy, Yifu Lu, Mengdi Wang, Dinesh Manocha, Furong Huang, Mohammad Ghavamzadeh, and Amrit Singh Bedi. Does thinking more always help? Mirage of test-time scaling in reasoning models. In *Advances in Neural Information Processing Systems*, 2025. https://arxiv.org/abs/2506.04210.

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

Hugging Face H4. MATH-500 dataset release. https://huggingface.co/datasets/HuggingFaceH4/MATH-500. Accessed September 2026.

William Jurayj, Jeffrey Cheng, and Benjamin Van Durme. Is that your final answer? Test-time scaling improves selective question answering. In *Annual Meeting of the Association for Computational Linguistics*, 2025. https://arxiv.org/abs/2502.13962.

Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio du Pin Calmon. Inference-time reward hacking in large language models. In *Advances in Neural Information Processing Systems*, 2025. https://arxiv.org/abs/2506.19248.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. https://arxiv.org/abs/2112.09332.

Qwen, An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. https://arxiv.org/abs/2412.15115.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. https://arxiv.org/abs/2408.03314.

Benedikt Stroebl, Sayash Kapoor, and Arvind Narayanan. Inference scaling fLaws: The limits of LLM resampling with imperfect verifiers. *arXiv preprint arXiv:2411.17501v1*, 2024. https://arxiv.org/abs/2411.17501v1.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. https://arxiv.org/abs/2203.11171.

Menghua Wu, Cai Zhou, Stephen Bates, and Tommi Jaakkola. Thought calibration: Efficient and confident test-time scaling. *arXiv preprint arXiv:2505.18404*, 2025. https://arxiv.org/abs/2505.18404.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. https://arxiv.org/abs/2011.09161.

Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities? *arXiv preprint arXiv:2502.12215*, 2025. https://arxiv.org/abs/2502.12215.

Shu Zhou, Rui Ling, Junan Chen, Xin Wang, Tao Fan, and Hao Wang. When more thinking hurts: Overthinking in LLM test-time compute scaling. *arXiv preprint arXiv:2604.10739*, 2026. https://arxiv.org/abs/2604.10739.

A Additional Proof Details

Record flows and the quantile model. Fix x. Conditional on the candidate quantiles, correctness labels are independent Bernoulli variables with means $q(U_i)$. The maximum of N quantiles has density Nu^{N-1}, giving the accuracy formula in eq. (4); marginal candidate correctness is i.i.d. Bernoulli with mean p_x, giving coverage. For $K = M - N$, the two block maxima are independent. Corruption requires $u < v$ and labels $(1, 0)$, giving the first integral in eq. (8); labels $(0, 1)$ give repair. The identity follows by averaging $C_M - C_N$, and the bound follows because any change requires the overall maximum to lie among the K new candidates, an event of probability K/M. For score atoms, an independent uniform tie key T gives $U = F_x(S^-) + T[F_x(S) - F_x(S^-)]$. This quantile is uniform and preserves the score order; reusing the key preserves nesting.

Identification and matching coverage. The constructions in proposition 1 are proved in the main text. For the conditional identification statement, equal accuracy curves imply $\int_0^1 (q_1 - q_2)u^k \, du = 0$ for every integer $k \geq 0$. Polynomials are uniformly dense in $C[0, 1]$, and $q_1 - q_2$ is integrable, so its integral against every continuous function is zero. Thus $q_1 = q_2$ almost everywhere. The same argument identifies $\bar{q}$ from population accuracies, but not the cross-products $\mathbb{E}_x[q_x(u)q_x(v)]$ appearing in paired flows.

For the remaining algebra in corollary 1, integrating first over the old maximum for $q(u) = u$ gives

$$h^{\uparrow}_{N,M} = \frac{NK}{N+1} \int_0^1 v^M (1 - v) \, dv = \frac{NK}{(N+1)(M+1)(M+2)}. \tag{28}$$

Bookkeeping gives $g^{\uparrow}_{N,M} = h^{\uparrow}_{N,M} + K/[(N+1)(M+1)]$. Replacing q by $1 - q$ exchanges the flows, so their equal mixture has $\bar{h}_{N,M} = \bar{g}_{N,M} = K(2N + M + 2)/[2(N+1)(M+1)(M+2)]$. Setting $M = 2N$ proves the stated formula. Mixing this population with the constant-q population preserves both complete curves and attains every limiting doubling flow in $[0, 1/8]$.

Finite-budget bound and turnover limits. For eq. (9), split $\mathbb{E}[q(U)] - \ell$ at $U = t$: the lower event has probability t^N, and the upper event contributes at most $\omega_q(t)$. For proposition 3, divide the joint block-maximum density on $u < v$ by K/M. This gives the probability density $w_{N,M}(u, v) = MNu^{N-1}v^{K-1}$. Direct integration gives its lower-tail probability

$$\mathbb{P}_w(u \leq t) = t^N \left[1 + \frac{N}{K}(1 - t^K)\right] \leq t^N[1 + N(1 - t)], \tag{29}$$

using $1 - t^K \leq K(1 - t)$. On $u > t$, also $v > t$, and the difference between $q(u)[1 - q(v)]$ and $\ell(1 - \ell)$ is at most $2\omega_q(t)$. On the complement it is at most one. Since $(M/K)h_{N,M}$ is the expectation of this product under w, the bound follows; exchanging the labels proves it for repair.

For every fixed $t < 1$, the bound in eq. (29) vanishes as $N \to \infty$, uniformly over $K \geq 1$. First taking that limit and then $t \uparrow 1$ proves eq. (12) for every sequence $M_N > N$, including expansions whose replacement probability tends to zero. Multiplying by $(M_N - N)/M_N$ yields the adjacent and geometric regimes. Positive asymptotic equivalence requires $\ell \in (0, 1)$; at the endpoints the normalized flows converge to zero.

Accuracy rates and population averaging. Integrating eq. (14) from u to one gives $q(u) - \ell = c(1 - u)^{\beta} + o((1 - u)^{\beta})$. For any $\epsilon > 0$, choose $t < 1$ so that the remainder on $(t, 1)$ is bounded by $\epsilon(1 - u)^{\beta}$. Its contribution to accuracy is bounded by $\epsilon N\mathrm{B}(N, \beta + 1)$, plus an $O(t^N)$ contribution below t. The leading term and the corresponding derivative integral are

$$cN\mathrm{B}(N, \beta + 1) \sim c\Gamma(\beta + 1)N^{-\beta}, \qquad -c\beta\mathrm{B}(N + 1, \beta + 1) \sim -c\beta\Gamma(\beta + 1)N^{-\beta-1}. \tag{30}$$

For the second expression, use eq. (31) and bound the derivative remainder by $\epsilon(1 - u)^{\beta-1}$ above t; the contribution below t is at most $t^N\|q'\|_1$. Letting $\epsilon \downarrow 0$ proves both rates. Combining them with the adjacent flow limit gives $(h_N + g_N)/|\Delta_N| \sim 2\ell(1 - \ell)N^{\beta}/(|c|\beta\Gamma(\beta + 1))$.

For corollary 2, each normalized conditional flow is an expectation of a product in $[0, 1]$. Dominated convergence therefore gives the population limit $\mathbb{E}_x[\ell_x(1 - \ell_x)] = a(1 - a) - \sigma_\ell^2$. Also $\bar{A}_N \to a$ by bounded convergence. When $a > 0$, dividing the geometric flow limit by a gives limiting conditional corruption. The inequality $0 \leq \sigma_\ell^2 \leq a(1 - a)$ and the constructions in proposition 1 establish its sharp range. No uniform tail convergence across inputs is needed for these bounded limits.

Accuracy shape requires a separate assumption. The score maximum increases with width, but correctness along that ordering need not. For absolutely continuous q,

$$\Delta_N = \int_0^1 u^N(1-u)q'(u)\,du. \tag{31}$$

If q' is nonnegative below a threshold t and nonpositive above it, then $\Delta_m \le t^{m-n}\Delta_n$ for $m > n$; once accuracy stops increasing, it cannot resume. Without that shape restriction, even the finite sign pattern is unrestricted.

Theorem 2 (Arbitrary finite scaling patterns)**.** *Fix $K \ge 1$, a vector $d \in \mathbb{R}^K$, and $\varepsilon_0 \in (0, 1/2)$. There exist a polynomial $q : [0,1] \to [\varepsilon_0, 1-\varepsilon_0]$ and a scalar $\gamma > 0$ such that*

$$A_{n+1} - A_n = \gamma d_n, \qquad n = 1, \dots, K, \quad \textit{and} \quad p_x = 1/2. \tag{32}$$

Hence any finite pattern of improvements, declines, and plateaus can occur while oracle coverage equals $1 - 2^{-N}$.

To construct q, let $k_n(u) = (n+1)u^n - nu^{n-1}$. These polynomials are linearly independent and have zero integral. Their positive-definite Gram matrix gives a polynomial f with $\int_0^1 fk_n = d_n$ and $\int_0^1 f = 0$. A sufficiently small $\gamma > 0$ makes $q = 1/2 + \gamma f$ lie in the required interval. This proves the result without changing candidate coverage. The common rescaling can make increments small; the conclusion concerns shape, not a fixed improvement size. Neither record ordering nor a first observed decline supplies a universal stopping guarantee.

Finite patterns and the single-peak condition. The Gram-matrix construction in the proof of theorem 2 is completed by choosing $0 < \gamma \le (1/2 - \varepsilon_0)/\|f\|_\infty$ when $f \ne 0$; any positive γ works when $f = 0$. For the sufficient shape condition, integration by parts has zero boundary terms because $u^N(1-u)$ vanishes at both endpoints, proving eq. (31). For $m > n$, multiply its integrand by u^{m-n}. Below t, the integrand is nonnegative and this multiplier is at most t^{m-n}; above t, the signs and inequality reverse. Consequently $\Delta_m \le t^{m-n}\Delta_n$, proving that a nonpositive increment cannot be followed by a positive one.

Finite-pool estimation and variance. For proposition 4, condition on the sorted pool. There are $Z = \binom{m}{N}\binom{m-N}{K}$ equally likely old/expansion subset pairs. If the old maximum is rank r and the expansion maximum is $s > r$, choose the remaining old ranks in $\binom{r-1}{N-1}$ ways. All N old ranks lie below s, leaving $s - 1 - N$ available lower expansion ranks and $\binom{s-1-N}{K-1}$ choices. Feasibility gives exactly the limits in eqs. (20) and (21). Multiplying by the appropriate labels counts corruption or repair. Unconditionally, any such value-independent selection from an i.i.d. pool has the law of a fresh nested experiment, proving unbiasedness. The old subset and its union with the expansion are uniform subsets of sizes N and M; averaging their correctness difference proves the sample-level identity.

Let $\mathcal{G}$ contain the input and unordered scored-and-labeled pool, including tie keys. Conditional on $\mathcal{G}$, all chronological orders are equally likely. Hence the subset averages equal $\mathbb{E}[H \mid \mathcal{G}]$ and $\mathbb{E}[G \mid \mathcal{G}]$. The law of total variance proves eq. (22), both conditional on an input and over its population. The terms within a pool remain dependent; averaging over independent inputs supplies the sampling units.

B Whole-Policy Calibration

Proof of proposition 5. Condition on development data, fixing the predictors, finite policy family, order, and $\alpha, \delta \in (0,1)$. Calibration input-and-randomness tuples are i.i.d.; deployment uses the same law. Given $n_+ > 0$, the qualifying tuples are independent draws from this law conditional on initial correctness. Thus, for each policy, $K_j \mid n_+ \sim \mathrm{Bin}(n_+, R_j^+)$. Shared streams correlate policies, but preserve this marginal binomial law.

Clopper–Pearson coverage gives $\mathbb{P}\{R_j^+ > U_j^+(\delta/J) \mid n_+\} \le \delta/J$. A union bound makes all J bounds valid simultaneously with probability at least $1 - \delta$. On that event, every retained policy has risk at most α, irrespective of subsequent selection among them.

For fixed sequence, let $p_j = \mathbb{P}\{\mathrm{Bin}(n_+, \alpha) \le K_j\}$. The binomial lower-tail test is valid at $R_j^+ = \alpha$ and conservative for $R_j^+ > \alpha$, by stochastic monotonicity. Therefore every unsafe policy satisfies $\mathbb{P}(p_j \le \delta \mid n_+) \le \delta$. If any policy is unsafe, let $j_\star$ be the first one in the fixed order. Selecting any unsafe policy requires passing its test, because testing stops at the first failure. The probability of an unsafe selection is thus at most δ, without assumptions about dependence between policies or risk monotonicity along the order. If none is unsafe, the claim is immediate.

The initial-answer fallback has identically zero corruption and is used when no policy passes or $n_+ = 0$. Averaging the conditional guarantee over n_+ proves eq. (27). □

Fixed sequence certifies the predeclared target α; its selected pointwise bound is not itself a selection-valid confidence bound. Bonferroni bounds are simultaneous. For $n_+ = 919$ and $K = 56$, the calibration corruption proportion is $K/n_+ \simeq 0.06093580 = 6.09358\%$. The bounds are $U^+(0.10) \simeq 0.07232299$ and $U^+(0.10/20) \simeq 0.08417955$. These are probability upper bounds; the corresponding percentages are approximately 7.2323% and 8.4180%.

C Exact Checks

For $q(u) = \sum_k a_k u^k$, accuracy is $A_N = \sum_k a_k N/(N+k)$, and polynomial integration evaluates the paired flows exactly. Rational-arithmetic checks verify $1{,}020$ distinct adjacent and doubling cases for $q = u$, $1-u$, $0.05 + 1.35u - 0.90u^2$, and 0.60, with old widths $N = 1, \ldots, 128$. The overlap at $(N, M) = (1, 2)$ is counted once for each law. Exhaustive enumeration of 2^7 label vectors and $\binom{7}{2}$ budget pairs verifies all $2{,}688$ finite-pool cases and their sample-level identities. For the quadratic law, $p = 0.425$, $\ell = 0.5$, and $\Delta_N = 0.45(5-N)/[(N+1)(N+2)(N+3)]$. Accuracy is maximal at both $N = 5$ and 6, with value $149/280$; adjacent flows decay as $0.25/N$, whereas doubling flows converge to 0.125.

D Experimental Protocol and Aggregate Sources

The empirical source consists of aggregate tables, calibration counts, and a supplied model–task–scorer map. Result identifiers follow the original table rows. The turnover study reports $4{,}000$ evaluations per setting, the policy study $50{,}000$ evaluations overall, and calibration $2{,}000$ records overall. These are evaluation counts, not counts of distinct questions. Exact run revisions and question IDs are required to link the supplied configuration to the measured records.

Table 3: **Supplied setting map retained for reference.** Names follow the supplied setting map; setting IDs and row order follow the reported result tables. Checkpoint and dataset revision hashes are not included in the source.

Setting ID	Generator	Task	Scorer
S1/T-A/V1	Qwen2.5-7B-Instruct	GSM8K	V1
S1/T-B/V1	Qwen2.5-7B-Instruct	MATH-500	V1
S2/T-C/V1	Qwen2.5-14B-Instruct	BoolQ	V1
S2/T-D/V2	Qwen2.5-14B-Instruct	ARC-Challenge	V2
S3/T-A/V2	Llama-3.1-8B-Instruct	GSM8K	V2

Evaluation units and pairing. The turnover study reports two 32-candidate streams for each evaluation record and widths $\{1, 2, 4, 8, 16, 32\}$. Nested widths share prefixes within a stream; the initial answer counts toward the candidate budget. Five settings and five adjacent doubling intervals yield 25 setting–interval comparisons. Policy comparisons share the question, initial answer, and candidate sequence within each record. Candidate cost includes every generated proposal, including rejected ones. Repeated evaluations of a question remain in the same prompt cluster; shared task/question IDs receive common resampling multiplicities across settings. The aggregate source does not determine the number of distinct questions or their development, calibration, and test allocation.

Generation and scoring specification. The supplied generation configuration specifies temperature 0.7, top-$p = 0.95$, a $2{,}048$-token output cap, and master seed 42. It is retained as configuration information; execution logs are needed to establish its correspondence to the reported measurements. V1 is mean untempered generator log-probability over generated non-special tokens. V2 uses Qwen2.5-14B-Instruct to compare Yes/No continuation likelihoods for a question and candidate, without a reference answer. The controller bins raw scores and estimates correctness probabilities through eq. (25); the gate acts on differences of fitted correctness, not on normalized raw scores. Exact prompts, answer parsers, checkpoint revisions, and seed derivation belong to the run record.

Policy family and operating point. The original family consists of the 20 pairs

$$\tau \in \{0.06, 0.10, 0.13, 0.16\}, \qquad \mu \in \{0, 0.010, 0.020, 0.030, 0.040\}, \qquad M = 32.$$

The reported COMPUTEGUARD operating point is $(\tau, \mu) = (0.13, 0.020)$. It accepts a proposal when its fitted correctness exceeds the incumbent's by at least 0.13, and continues sampling while $\widehat{V}_{0.13}(z) > 0.020$, up to 32 candidates. The gate-only comparator uses the same $\tau = 0.13$ and always generates 32 candidates. The adaptive raw-score comparator accepts raw-score records and stops at $\widehat{V}_0(z) \le 0.020$. The family, score fit, ordering, and selection rule must be fixed before calibration, as specified in Section 6.

Calibration counts. For the reported operating point, the calibration summary gives $K = 56$ final corruptions among $n_+ = 919$ initially correct records. The empirical proportion is 6.09358%. The single-policy one-sided 90% upper bound is 0.07232299, and the 20-policy Bonferroni bound is 0.08417955. This calculation uses the original family size and the original selected policy. The interpretation in proposition 5 additionally requires a development-frozen family and i.i.d. calibration records from the deployment input-and-randomness law. The supplied counts do not establish the sampling law, split allocation, or freezing order.

A/B estimator comparison. The source comparison covers five settings and all $\binom{6}{2} = 15$ width pairs, giving 75 comparisons for each metric. The reported MAEs are 0.53, 0.48, and 0.37 percentage points for accuracy change, repair, and corruption. The 25 adjacent doubling intervals used for the near-flat summary are a different comparison set. Reconstructing the A/B diagnostic requires the independent pool/stream records and the per-comparison estimates; no missing outcomes are imputed from the three MAEs.

D.1 PAIRED UNCERTAINTY PROTOCOL

The analysis protocol uses $1{,}000$ bootstrap replicates and analysis seed 2026, separately from the supplied generation seed 42. It resamples question IDs within each task, retaining all repetitions and paired outcomes for a sampled question. Settings sharing a task/question ID share the bootstrap multiplicity. Streams are averaged within questions, questions within settings, and settings receive equal weight. Conditional corruption is pooled mean corruption divided by pooled mean initial correctness. Paired differences are computed within common draws; percentile endpoints give descriptive 95% intervals, and bootstrap standard deviations give SEs. Undefined zero-denominator draws are counted explicitly.

This analysis has not been applied to the retained measurements because the original records are absent. Earlier SEs and intervals remain in the historical archive and are not displayed as revised results. The 12.94-point difference and $8/25 = 32\%$ near-flat summary remain reported point summaries.

D.2 TRACEABILITY OF EMPIRICAL CLAIMS

Table 4: **Sources of retained numerical claims.** Paths refer to fields in `results/manuscript_aggregates.json`. No missing per-record values are imputed.

Claim	Available aggregate source and interpretation
$1 \to 32$ flows	`table1_and_table4`: five rows; $\Delta A = g - h$ and turnover $g + h$ follow the displayed rates. In S2/T-D/V2, $24.20 - 17.57 = 6.63$.
Policy operating points	`table2`: seven policies; checks verify $A_\pi = A_0 + g - A_0 R^+$ within rounding.
12.94-point reduction	$20.35 - 7.41$, from random width and COMPUTEGUARD in `table2`; a descriptive difference.
Near-flat intervals	`plateau_summary`: $8/25 = 32\%$; the example has $9.80 - 9.90 = -0.10$ and turnover 19.70.
Calibration bound	`calibration`: $K = 56$, $n_+ = 919$, $J = 20$; $56/919 \simeq 6.09358\%$ and the adjusted beta quantile gives 0.08417955.
A/B agreement	`q2_summary`: aggregate MAEs only; the 75 per-setting/budget comparisons are absent.

The $8/25 = 32\%$ near-flat frequency is retained as a reported aggregate summary; the complete interval table is not included in the available records. The 12.94-point contrast follows directly from $20.35 - 7.41$, whereas its paired uncertainty requires shared prompt-level outcomes. Transcribed paired SEs are not used to assert significance. The displayed 6.63-point gain in S2/T-D/V2 is calculated from the displayed repair and corruption rates, rather than from recovered unrounded outcomes.

The turnover rows imply a mean initial accuracy of approximately 47.57% through h/R^+; the policy table reports 46.45%. Different streams can give different estimates, so the figures do not form an algebraic contradiction. Verifying a common evaluation mixture requires the two studies' prompt IDs, initial counts, and weights. The manifest preserves this distinction rather than forcing the studies to share an inferred initial count.