THE COST OF EVALUATING POLICY UPDATES UNDER CENSORED FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

A policy can appear more accurate simply because its failures are less likely to be observed. We study how to decide whether a proposed policy actually improves expected reward when some rewards are missing. An audit reveals one missing reward, but the corresponding outcome must first be generated. Evaluation therefore has two costs: producing outcomes and inspecting them. For finitely many actions and rewards in $[0, 1]$, we characterize the smallest worst-case estimation error under both budgets. The optimal allocation gives more audits to influential actions, subject to how often their outcomes are available. For actions with too few observations, it reduces reliance on the sampled rewards and includes the remaining uncertainty in the error bound. We also account for estimating missingness and for evaluating a sequence of proposed updates. Experiments in a controlled environment and on fixed Qwen-generated GSM8K responses show when additional evidence resolves more decisions, and why that need not improve the final policy.

1 INTRODUCTION

Suppose a system answers ten questions. Its current policy gets seven right, and all ten answers are checked. A new policy gets only six right, but insufficient evidence causes the verifier to skip three wrong answers. Accuracy on the checked answers rises from $7/10$ to $6/7$, even though accuracy on all answers falls from $7/10$ to $6/10$. Figure 1 illustrates this failure. Comparing the visible scores would favor the worse policy. We instead ask what additional evidence is needed to determine whether an update improves performance on all outcomes.

Several lines of work supply parts of the answer. Missing-outcome methods bound what can be inferred without observing every reward (Manski, 2007). Active testing and active statistical inference choose which additional labels to acquire (Kossen et al., 2021; Zrnic & Candès, 2024). High-confidence policy improvement uses statistical evidence to control the risk of accepting a worse policy (Thomas et al., 2015). Detector-based training motivates studying reward and evidence visibility together (Lang et al., 2024; Cundy & Gleave, 2025; Baker et al., 2025).

The resource constraint we study is that a missing reward can be inspected only after its outcome has been produced. We call that inspection an *audit*. Buying more audits cannot compensate for having too few relevant outcomes. An action may strongly affect the proposed improvement yet occur rarely under the policy collecting evaluation data. How accurately can an update be evaluated with limited interactions to produce outcomes and limited audits to inspect missing rewards?

Our method first holds the proposed policy fixed during evaluation. From ordinary feedback, it estimates which missing rewards could change the decision. It allocates audits in proportion to their influence, subject to each action's expected number of available outcomes. When an action supplies too little evidence, the estimate is pulled toward the middle of its possible range. This is *shrinkage*: it reduces sampling noise, while the error bound explicitly includes the possible bias. The update is accepted only if the resulting confidence interval lies above zero. If the interval lies below zero it is rejected; if it includes zero the evidence is insufficient, and the current policy is kept.

The theory proceeds from uncertainty in ordinary feedback to optimal budget allocation and error control across updates. Experiments then examine the practical tradeoff between cost, the fraction of updates that receive a definite decision, and final policy quality. Grouping proposal steps improves

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

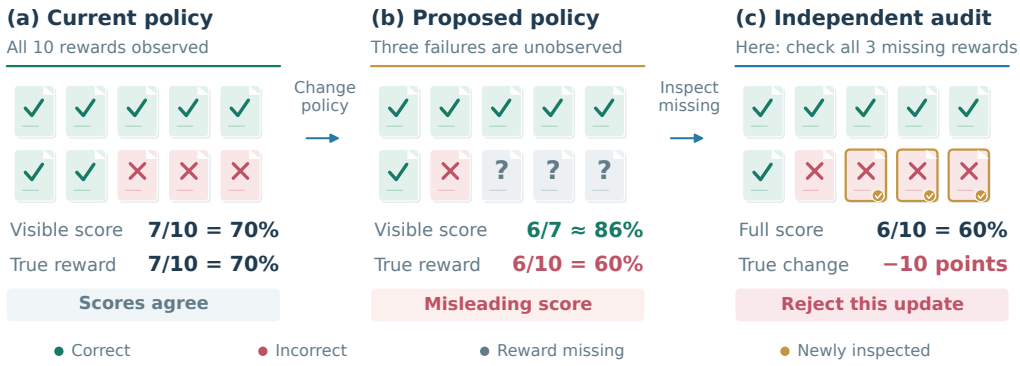

Figure 1: Why visible accuracy can favor a worse policy. In this illustrative ten-outcome example, the current policy has seven correct answers, whereas the candidate has six. Hiding three candidate failures raises its visible accuracy to $6/7$. Checking all three missing outcomes restores its true accuracy, $6/10$, and exposes the decrease. These schematic counts illustrate the problem. With sampled audits, the method uses independent checks and an uncertainty interval to estimate the reward change.

quality at intermediate budgets in the controlled environment. On fixed Qwen-generated GSM8K answers, grouping can resolve more decisions without improving quality.

- **What missing rewards leave unknown.** We characterize the exact range of improvements consistent with ordinary feedback, and give upper and lower bounds on the audits needed to determine the sign of an update (Sections 2–3).
- **How to use both budgets.** We prove that one allocation, capped by outcome availability, is optimal at every confidence scale. Together with shrinkage, it achieves the smallest worst-case estimation error up to universal constants (Section 4).
- **How to evaluate successive updates.** We include uncertainty in the estimated feedback rates and use fresh evaluation data to control incorrect decisions over an adaptive sequence. The experiments quantify the resulting evidence costs and policy quality (Sections 5–6).

2 What can be learned from incomplete feedback?

We compare the expected reward of a current policy with that of a proposed replacement. A policy is a probability distribution over a fixed set of actions. The evaluator can observe ordinary feedback and pay to reveal selected missing rewards. Its output is a positive, negative, or inconclusive assessment of the reward change.

2.1 Actions, observations, and the cost of checking

There are K actions. An action a specifies both how the task is performed and which evidence is presented. After it is selected, the system produces a reward $R \in [0, 1]$ and an indicator G: $G = 1$ means the reward is observed, and $G = 0$ means it is missing. The evaluator sees (a, G, GR), so it can distinguish a missing reward from an observed zero. The joint distribution of (R, G) is fixed for each action, although visibility may depend on the reward itself.

An interaction means selecting an action and producing its outcome. An audit means revealing the reward of an already produced outcome with $G = 0$. For such an outcome, a new coin Q selects it for inspection with probability q_a. The coin is independent of the completed reward: $Q \sim \text{Bernoulli}(q_a)$. Each interaction and each audit costs one unit of its respective resource.

An action can pair a fixed prompting strategy with a reward-dependent reveal rule. Changing their mixture changes aggregate reward and visibility while preserving each action law. Adaptation within an action requires a time-varying model. Our three-prompt experiment uses a fixed model and question population.

2.2 Separating observed rewards from missing rewards

To identify what must still be learned, split the mean reward into observed and missing parts. Let p_a be the probability that the reward is missing, s_a the average observed reward with missing values counted as zero, and h_a the mean among missing rewards. Then the full mean μ_a is

$$p_a = \mathbb{P}(G = 0 \mid a), \qquad s_a = \mathbb{E}[GR \mid a], \qquad h_a = \mathbb{E}[R \mid a, G = 0], \qquad \mu_a = s_a + p_a h_a. \quad (1)$$

Thus s_a is not accuracy conditional on being observed; that conditional mean is $s_a/(1 - p_a)$ when $p_a < 1$. Initially we treat (s, p) as known, to isolate uncertainty about the missing rewards. Section 5 accounts for estimating (s, p) from data. When $p_a = 0$, the choice of h_a has no effect.

Write π_a and π'_a for the current and proposed probabilities of action a, with each policy summing to one. Define $d_a = \pi'_a - \pi_a$. The target is the exact expected reward change $D = \sum_a d_a \mu_a$. A missing reward matters in proportion to both the probability change d_a and missingness p_a. We therefore define its coefficient w_a, total absolute coefficient H, and a reference value m obtained by setting all unknown means to $1/2$:

$$w_a = d_a p_a, \qquad H = \sum_a |w_a|, \qquad m = d^\top (s + p/2), \qquad \mathcal{I} = \{a : w_a \neq 0\}. \quad (2)$$

Here $d^\top x$ means $\sum_a d_a x_a$, and $\mathcal{I}$ contains the actions whose missing rewards affect the update. An unchanged action has $d_a = 0$ and contributes no uncertainty. Since rewards lie in $[0, 1]$, replacing an unknown mean by $1/2$ introduces error at most $1/2$; summing the weighted errors gives the following exact range.

Proposition 1 (Possible improvements before auditing)**.** *For a fixed distribution of ordinary feedback, the possible reward means form exactly $\prod_a [s_a, s_a + p_a]$. The possible improvements form exactly*

$$\mathcal{D}(d) = [m - H/2, m + H/2]. \quad (3)$$

If this entire interval is positive or negative, the update's sign is known without audits. If it contains both signs, reward distributions with identical ordinary feedback require opposite decisions.

Each missing mean can independently range from zero to one, which proves the endpoints (Manski, 2007). For example, swapping the missing success rates of two equally censored actions leaves their observed feedback unchanged while changing which action is better. More ordinary feedback cannot distinguish these cases; checking missing rewards can.

3 How many missing rewards must be checked?

First suppose a fresh missing outcome can be obtained from any chosen action. We count the inspections, postponing the cost of producing those outcomes to Section 4. An efficient rule checks actions according to $|w_a|$, their potential influence on D. For negative w_a, replace reward R by $1 - R$ so that a larger transformed observation always supports a larger improvement. With $H > 0$, the action-selection probabilities ν_a and transformed reward Y satisfy

$$\nu_a = \frac{|w_a|}{H}, \qquad Y = \begin{cases} R, & w_a > 0, \\ 1 - R, & w_a < 0. \end{cases} \qquad D = m + H(\mathbb{E}[Y] - 1/2). \quad (4)$$

Thus the problem reduces to estimating one mean in $[0, 1]$ and comparing it with $1/2 - m/H$. A small true improvement requires a more precise estimate. The following result makes that cost explicit; δ is the allowed error probability, $\gamma = |D|$ is the absolute improvement, and N is the number of audits.

Theorem 2 (Necessary and sufficient audit scale)**.** *Fix p, d, $\delta \in (0, 1/2)$, and $\gamma > 0$. A test using (4), doubling sample sizes, and error probabilities assigned to successive checks returns the correct sign with probability at least $1 - \delta$ using, on the same event,*

$$N = O\left(1 + \frac{H^2}{\gamma^2}\left[\log\frac{1}{\delta} + \log\log\left(e^e + \frac{H}{\gamma}\right)\right]\right) \quad (5)$$

audits. Conversely, for every p, d with $H > 0$ and $0 < \gamma \le H/4$, two reward systems can have identical ordinary feedback but improvements $+\gamma$ and $-\gamma$. Every test correct with probability $1 - \delta$ for all such systems, including tests with adaptive sampling and stopping, satisfies

$$\mathbb{E}_+[N] \ge \frac{H^2}{16\gamma^2} \operatorname{kl}(1-\delta, \delta). \tag{6}$$

Here $\mathbb{E}_+$ is expectation under the positive-improvement system, and $\operatorname{kl}(x, y) = x \log(x/y) + (1 - x) \log((1 - x)/(1 - y))$ measures the difference between Bernoulli distributions.

Both bounds have leading dependence H^2/γ^2: influential missing rewards are expensive to resolve, especially when the improvement is small. The upper bound stops when an interval excludes zero; the lower bound constructs opposite improvements that each audit distinguishes only weakly (Kaufmann et al., 2016). Appendix A.3 proves both statements. These are worst-case bounds: Proposition 1 may already resolve a particular instance. Overall feedback coverage is insufficient to predict cost, because missing rewards on unchanged actions have zero weight.

4 Allocating audits when outcomes are scarce

We now count the interactions needed to produce outcomes. During evaluation, a fixed *behavior policy* selects action a with probability $\rho_a > 0$ for every $a \in \mathcal{I}$. The probability of obtaining a missing reward from this action in one interaction is $r_a = \rho_a p_a$. Let T be the interaction budget, N the expected audit budget, and n_a the expected number of audits assigned to action a. The feasible allocations are

$$\mathcal{A}(T, N) = \{n \ge 0 : n_a \le T r_a,\ \textstyle\sum_a n_a \le N\}. \tag{7}$$

The first constraint says that an outcome can be checked only if it occurs. The second limits total inspection cost. The caps $T r_a$ concern expected counts; a separate stopping rule enforces an absolute spending limit in Section 5.

4.1 Trading sampling noise for bounded bias

Few audits can make an action's estimated missing reward very noisy. Rather than use its full coefficient $|w_a|$, retain a smaller amount $v_a \in [0, |w_a|]$ and replace the rest by the midpoint reward $1/2$. The omitted amount has worst-case error $(|w_a| - v_a)/2$. The sampling uncertainty is proportional to $\sqrt{\sum_a v_a^2/n_a}$. For a confidence scale $z > 0$, choose the retained weights to minimize their sum, and then choose the allocation:

$$\mathcal{M}_z(n) = \min_{0 \le v_a \le |w_a|} \left\{ \frac{1}{2} \sum_a (|w_a| - v_a) + z \sqrt{\sum_a v_a^2/n_a} \right\}, \tag{8}$$

$$\mathcal{R}_z(T, N) = \min_{n \in \mathcal{A}(T,N)} \mathcal{M}_z(n). \tag{9}$$

Here $\mathcal{M}_z$ is the best bias–uncertainty tradeoff at a fixed allocation, and $\mathcal{R}_z$ is its best value under both budgets. Larger z demands stronger protection against sampling error. If $n_a = 0$, take $v_a = 0$ and $0^2/0 = 0$; retaining a nonzero coefficient without observations has infinite sampling cost. This quantity is also called a *modulus*: Appendix A.4 shows that it equals the largest change in the target that is hard to distinguish from the available observations (Donoho, 1994).

The solution has two simple operations. Allocate proportionally to $|w_a|$ until an action reaches its cap, then distribute the remaining budget among the other actions. Next choose how much of each sampled contribution to retain. The scalar λ sets total allocation, while c sets the amount of shrinkage; neither requires observing the missing rewards.

Theorem 3 (One allocation for every confidence scale)**.** *Let $z > 0$ and let $N_0 = \min(N, T \sum_{a \in \mathcal{I}} r_a)$ be the audit budget that can be used in expectation. The capped allocation*

$$n_a^* = \min\{T r_a, \lambda |w_a|\}, \qquad \sum_a n_a^* = N_0 \tag{10}$$

minimizes $\mathcal{M}_z$ for every z. If $N_0 \le 4z^2$, zero auditing attains $\mathcal{R}_z = H/2$. Otherwise the optimal retained weights are

$$v_a^* = \min\{|w_a|, cn_a^*\}, \qquad \sum_a \min\{w_a^2/(c^2 n_a^*), n_a^*\} = 4z^2. \tag{11}$$

In this latter case, every action whose coefficient is reduced has reached its availability cap.

The proportional allocation is classical (Neyman, 1934; Wright, 2020); the result here is that it remains optimal for this bounded-reward error criterion at every confidence scale. Only the retained weights change with z. The equations determine λ and c by one-dimensional searches. Appendix A.4 proves the result and handles zero budgets and saturated caps.

4.2 What accuracy can these budgets achieve?

To compare the method with any possible evaluator, let $r_\delta^*(T, N)$ be the smallest fixed error bound that can hold with probability $1 - \delta$ for every hidden reward distribution consistent with ordinary feedback. Interaction and audit budgets are bounded in expectation for every such distribution. The following theorem relates this statistical target to the optimization above. The scales z_U and z_L translate the desired error probability into upper- and lower-bound calculations.

Theorem 4 (Smallest worst-case estimation error)**.** *Fix known ordinary feedback and $0 < \delta \le 1/4$. Put $L = \log(4/\delta)$, $z_U = \sqrt{L/2}$, and $z_L = \sqrt{3\,\mathrm{kl}(1-\delta, \delta)/8}$. Then*

$$\frac{1}{4}\mathcal{R}_{z_L}(T, N) \;\le\; r_\delta^*(T, N) \;\le\; \frac{5}{3}\mathcal{R}_{z_U}(T, N). \tag{12}$$

The upper construction uses exactly T interactions and independent audit coins. Its paid count exceeds $N + \sqrt{2NL} + 2L/3$ with probability at most e^{-L}. On the confidence event, its interval determines the sign whenever $|D| > 10\mathcal{R}_{z_U}(T, N)/3$.

For the upper construction, use $z = z_U$ and inspect each missing outcome from action a with probability $q_a = n_a^*/(Tr_a)$. Define $S = \sum_a (v_a^*)^2/n_a^*$ to measure total sampling variation and $k = \max_a v_a^*/n_a^*$ to bound the largest contribution of one audit. Omitting zero-weight actions, the interval radius is

$$B_{\text{audit}} = \frac{1}{2}\sum_a (|w_a| - v_a^*) + \sqrt{LS/2} + \frac{2}{3}Lk \;\le\; \frac{5}{3}\mathcal{R}_{z_U}(T, N). \tag{13}$$

Its terms cover the bias from shrinkage, accumulated sampling variation, and an unusually large individual contribution. Section 5 specifies the interval's center. If $N_0 \le 4z_U^2$, use the original interval $m \pm H/2$ instead. Without binding availability caps, $\mathcal{R}_z = \min\{H/2, zH/\sqrt{N}\}$. When a cap binds, more interactions can supply information that a larger audit budget alone cannot buy.

5 Using the interval for successive policy updates

For each proposed update, hold the candidate and behavior policy fixed, collect new evaluation data, and estimate its reward change. A positive lower endpoint accepts the candidate; otherwise the current policy is kept. Two-sided coverage supports acceptance and rejection; intervals containing zero are unresolved. Appendix C.2 describes an acceptance-only alternative.

Estimating the reward change. Start with known (s, p) and the design in Section 4. The retained fraction $t_a = v_a^*/|w_a|$ lies in $[0, 1]$. In interaction t, A_t is the selected action and R_t, G_t, Q_t are its reward, visibility and audit indicators. Each inspected missing reward is weighted by the inverse probability of selecting its action and inspecting it. Centering rewards at $1/2$ gives

$$\widehat{D} = m + \frac{1}{T}\sum_{t:G_t=0,Q_t=1} \frac{t_{A_t} d_{A_t}}{\rho_{A_t} q_{A_t}}(R_t - 1/2). \tag{14}$$

The factor t_{A_t} controls shrinkage; the summation index t labels interactions. Only audited missing rewards enter this sum. With radius (13), $\widehat{D} \pm B_{\text{audit}}$ covers the true change at the stated confidence level. If an absolute audit limit prevents completing the planned evaluation, report unresolved rather than use an incomplete estimate.

Estimating missingness before evaluation. In practice, (s, p) must also be learned. An independent preliminary sample gives estimates $\widehat{s}_a, \widehat{p}_a$ whose simultaneous absolute errors are bounded by e_a. For allocation, use the conservative missingness bound $\overline{p}_a = \min(1, \widehat{p}_a + e_a)$: replace w_a by $\overline{w}_a = d_a\overline{p}_a$ and the caps by $T\rho_a\overline{p}_a$. The resulting allocation $\overline{n}_a$ and retained coefficient $\overline{v}_a$ determine $q_a = \overline{n}_a/(T\rho_a\overline{p}_a)$ and $t_a = \overline{v}_a/|\overline{w}_a|$.

For the estimate's center, however, use the empirical midpoint $\widehat{m} = d^\top(\widehat{s} + \widehat{p}/2)$. Insert these quantities in (14), use the radius of the conservative design, and add $\frac{3}{2}\sum_a |d_a|e_a$ for estimation error in the midpoint. This bound also holds for unequal e_a. If no sampling is chosen, use the corresponding calibrated midpoint interval. Appendix A.4 derives these corrections.

To protect a sequence of decisions, let δ_C be the error allowance for the preliminary sample and δ_j that for evaluation j. Choose $\delta_C + \sum_j \delta_j \le \delta$. The feedback distribution for each action stays fixed, so the preliminary estimate can be reused; each candidate receives fresh evaluation data. A union bound then ensures that every resolved decision is correct with probability at least $1 - \delta$, even when later candidates depend on earlier results.

How candidates are generated. Evaluation can be applied to any candidate fixed before its evaluation sample. Our experiments generate candidates by policy-gradient updates. Let θ be policy parameters, $J(\theta) = \sum_a \pi_\theta(a)\mu_a$ their expected reward, and b_t an action-independent baseline chosen from previous observations. Observed rewards receive weight one; an inspected missing reward receives weight $1/q_t$ to account for the missing rewards not inspected. The resulting update estimate is

$$Z_t = \left[G_t + (1 - G_t)\frac{Q_t}{q_t}\right](R_t - b_t)\nabla_\theta \log \pi_{\theta_t}(A_t). \tag{15}$$

Here $q_t > 0$ is the inspection probability for that outcome; the gradient factor is the derivative of its log selection probability. Conditional on the previous history $\mathcal{F}_{t-1}$, $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = \nabla J(\theta_t)$. This inverse-probability correction restores the mean update despite reward-dependent missingness (Hansen & Hurwitz, 1946; Horvitz & Thompson, 1952). A finite update can still be harmful, so it is evaluated separately. Appendix B.6 gives the derivation, and Appendix B.5 studies learning a near-optimal action.

6 Experiments

The experiments ask whether affordable evidence is sufficient to judge an update, and whether making more definite decisions improves the final policy. We study an 18-action controlled environment and three prompting strategies on fixed language-model responses. A resolved decision means that the reward-change interval excludes zero. Retained quality is the expected reward of the policy kept after learning, including occasions when a candidate is rejected or left unresolved. We report means and 95% Student intervals across 30 random seeds.

6.1 Comparisons and cost accounting

Each run starts with a uniform distribution over actions and takes 40 proposal steps. One method evaluates every step; the other evaluates the net change after each group of five steps. Grouped evaluation has eight candidates and five times the per-step evaluation allowance for each, so total proposal and evaluation limits are unchanged. Rejected or unresolved groups are discarded. Preliminary observations used to estimate feedback rates are collected once, independently of proposals and evaluation.

We compare capped allocation with shrinkage, the same allocation without shrinkage, and uniform acquisition with shrinkage. All three make decisions using intervals that include possible shrinkage bias. Three further references accept every proposal: limited-audit corrected feedback, complete feedback, and naturally visible feedback. They show the cost and quality of learning without the interval acceptance rule. The complete-feedback reference pays for every missing proposal reward; none of these three references spends on preliminary estimation or validation, and their decision-resolution rate is undefined.

To compare resource costs, let I denote interaction counts and A audit counts, with subscripts identifying preliminary calibration, proposal generation and validation. Let ℓ be the price of one audit

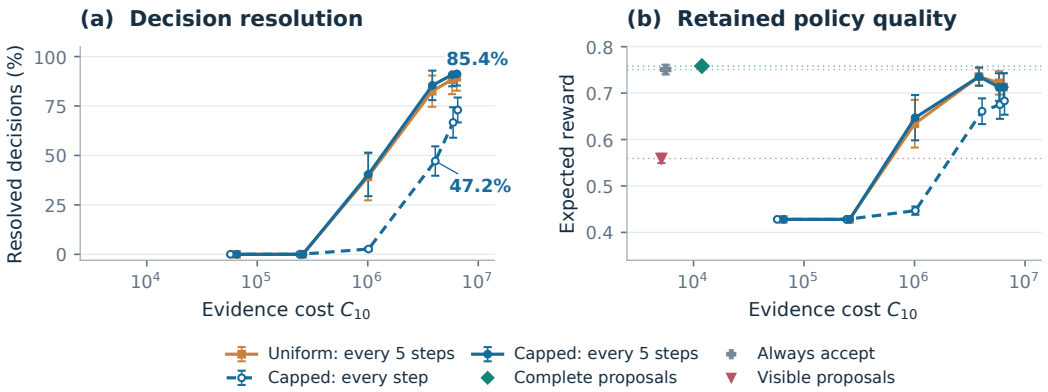

Figure 2: Larger evidence budgets can resolve updates and improve retained reward in controlled learning. Both panels include all six budget settings and 95% intervals across 30 seeds. Isolated markers show the cost of the proposal-only references; dotted lines show their final reward. Removing shrinkage changes no decisions on these learning trajectories.

measured in interaction-cost units. Total cost is

$$C_\ell = I_{\text{calibration}} + I_{\text{proposal}} + I_{\text{validation}} + \ell(A_{\text{proposal}} + A_{\text{validation}}). \tag{16}$$

The main plots use $\ell = 10$; Appendix D also uses 1 and 100. Every validation interaction is charged, including inconclusive evaluations. Increasing the budget provides more evidence within the same 40-step proposal horizon.

6.2 When larger budgets improve the retained policy

The 18 actions combine six reward strategies with three evidence modes, some hiding failures more often than successes. Across six settings, preliminary observations range from 128 to 65,536 per action and validation interactions from 1,000 to 100,000 per proposal step. The largest setting uses 1,179,648 preliminary observations and four million validation interactions per run. Appendix C specifies the setup.

Figure 2 shows that the two smallest settings provide too little evidence to resolve any update. At setting 4, evaluating every five steps resolves 85.4% of candidates and retains reward 0.736; evaluating every step resolves 47.2% and retains 0.661. The paired reward difference is 0.075, with a 95% interval of $[0.045, 0.104]$. Grouping changes both the candidate direction and the error allowance per evaluation, so these results measure their combined effect.

This improvement is costly. The five-step method at setting 4 uses 2,712,576 interactions and about 112,528 audits per run. Complete proposal feedback reaches reward 0.758 with 5,120 interactions and about 679 audits. The distinction is the acceptance rule: none of the interval methods accepts a nonpositive net change in 25,920 evaluated candidates. Accepting every limited-audit proposal instead applies 472 nonpositive single-step changes among 1,200 proposals, or 45 among 240 five-step changes. These counts refer to different units of change.

The learning runs yield identical decisions with and without shrinkage, while some interval widths differ. To examine when it helps, we hold candidates and populations fixed and make one action rare during validation. Figure 3 shows that, at the smallest arrival rate, shrinkage narrows intervals and resolves more decisions than using the full sampled contribution. Appendix D.3 reports all 7,560 evaluations and allocation comparisons. This experiment isolates evaluation accuracy for fixed candidates.

6.3 Visible accuracy can mislead on real model responses

We use Qwen2.5-0.5B-Instruct (Qwen et al., 2025) to answer all 1,319 questions in the GSM8K test split (Cobbe et al., 2021) with three prompts: answer directly, reason step by step, and reason then

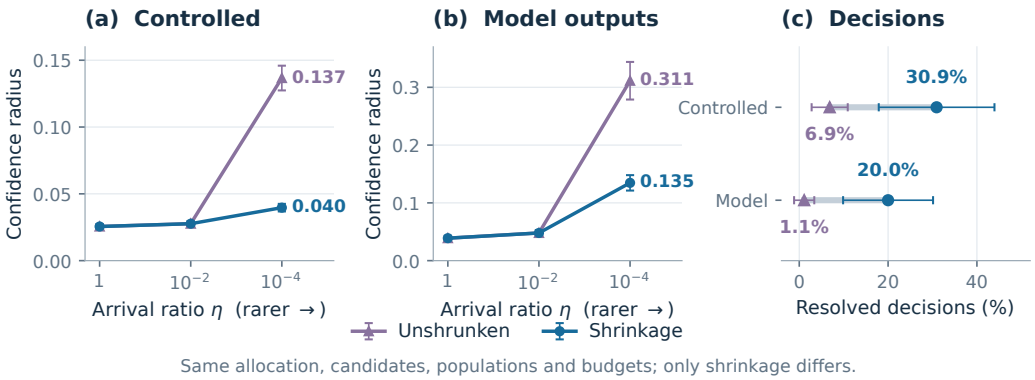

Figure 3: Shrinkage helps when influential outcomes are scarce. One action is selected with probability η/K, where K is the number of actions and η controls scarcity; each action is made rare in turn. Panels (a,b) show interval half-widths; panel (c) shows definite decisions at $\eta = 10^{-4}$. Allocation and budgets are identical for the two methods. Error bars average action rotations within each of 30 seeds before forming 95% Student intervals.

check the arithmetic. The resulting 3,957 responses are generated once and held fixed. Each action selects one prompt strategy. Its reward is one for a numerically correct answer and zero otherwise, under the parser in Appendix C. The three strategies have accuracies 19.3%, 43.0% and 42.5%.

To control missingness, correct responses are revealed with probability 0.9 for all strategies; incorrect responses are revealed with probabilities 0.02, 0.40 and 0.20, respectively. These imposed rates are not given to the learner. It updates the probabilities of selecting prompts, while the model weights and stored responses remain fixed. Preliminary, proposal and evaluation samples are drawn independently with replacement from this population. An audit reveals the correctness of a selected missing response.

The apparent score uses only revealed answers. For policy π, their average reward is $\pi^\top s/\pi^\top(1-p)$ when the denominator, the probability of a revealed answer, is positive. The true population accuracy is $\pi^\top \mu$. This difference is substantial: the visible-feedback reference attains apparent accuracy 0.877 but true accuracy only 0.216 (Figure 4).

At setting 4, five-step evaluation resolves 42.1% of candidates and retains accuracy 0.425, compared with 8.4% and 0.424 for per-step evaluation. Complete proposal feedback reaches 0.426. At setting 3, however, grouping raises decision resolution from 6.0% to 20.8% while lowering accuracy from 0.417 to 0.404. The paired difference is -0.0134, with a 95% interval of $[-0.0237, -0.0032]$. At setting 4 the paired accuracy interval includes zero. More definite decisions therefore need not produce a better final policy.

Generating the responses used 392,961 input tokens, 1,002,781 output tokens and 266.7 seconds of H100 inference. Subsequent draws reuse these responses and incur no additional model inference or distinct annotation. The results characterize evaluation on this fixed population. Appendix D includes all interval errors, unresolved outcomes and accepted nonpositive changes.

6.4 What makes evaluation expensive?

Validation accounts for 94.4% of interactions in the controlled task and 98.9% on model responses at setting 4 (Figure 5 and Table 3). Improving preliminary estimation alone would therefore remove little of the cost. Figure 6 separates grouping's effect on definite decisions from its effect on retained quality.

At setting 4, this implementation purchases about 166 times as many audits as complete proposal feedback. The extra evidence screens updates before retention; the reference accepts every change. Deployment value depends on the downstream cost of adverse updates, a separate quantity from the evidence costs measured here.

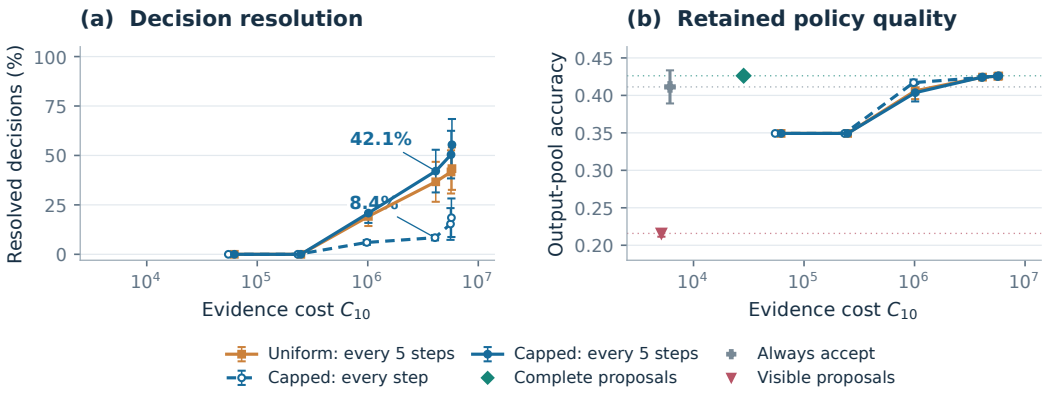

Figure 4: Evaluation on 3,957 fixed Qwen responses to 1,319 GSM8K questions. Policies change the mixture of three prompts. Cost counts draws from this fixed population and requests for missing labels; quality is population accuracy. Error bars use 30 seeds of feedback and proposal randomness. Model generation occurs once and is counted separately.

7 Related work

What incomplete feedback permits. Missing-outcome analysis bounds conclusions supported by incomplete data (Manski, 2007; Altschuler et al., 2019). Selective-label and missing-reward models study learning under specified observation mechanisms (Wei, 2021; Mahrooghi et al., 2025). Label-efficient prediction and costly-observation bandits balance cumulative reward against observation cost (Cesa-Bianchi et al., 2005; Tucker et al., 2023). We instead evaluate a particular update, counting both interactions and inspections.

Which observations to inspect. Active testing corrects evaluation bias (Kossen et al., 2021); active statistical inference uses uncertainty in predictive errors to choose labels (Zrnic & Candès, 2024; Li et al., 2025). These methods already cap sampling probabilities at one. Our allocation combines constrained variance minimization (Neyman, 1934; Wright, 2020) with bounded-reward error and shrinkage under two budgets (Donoho, 1994). The empirical comparisons isolate these operations. Predictive methods offer complementary comparisons under shared accounting for fitting, tuning and acquisition.

When to accept a policy change. Policy learning uses logged rewards (Swaminathan & Joachims, 2015), and high-confidence improvement controls acceptance risk (Thomas et al., 2015). Multiple-policy evaluation shares observed trajectories across given policies (Dann et al., 2023). Our fresh-data construction allocates error probability across adaptive candidates (Kaufmann et al., 2016; Garivier & Kaufmann, 2016; Howard et al., 2021), providing a reference point for future comparisons with data reuse. Policy-gradient and baseline theory supplies the proposals (Williams, 1992; Greensmith et al., 2004).

8 Conclusion

Missing outcomes must be produced before they can be checked. Accounting for their influence and availability, and reducing reliance on scarce observations, attains the smallest worst-case estimation error up to constants. Preliminary estimation and fresh evaluation extend the guarantee to successive updates. Experiments show that reliable decisions can be costly, and more definite decisions need not improve final quality.

Stable action-specific distributions, valid preliminary bounds and independent audits define the guarantee. The three-prompt experiment isolates missingness on fixed responses. Changing feedback and data reuse provide directions for extension.

AI USE STATEMENT

In preparing the final version of this paper, large language models were used for language polishing and the analysis and summarization of experimental results. All other aspects of the work, including problem formulation, theoretical development, experimental design and implementation, scientific interpretation, and final manuscript decisions, are the responsibility of the human authors.

REPRODUCIBILITY STATEMENT

Appendices A–B give the proofs and supporting theory; Appendices C–D specify the experiments, accounting conventions, and additional results. The supplementary package provides configurations, aggregate and seed-level measurements, and a numerical verification script. It supports checking reported summaries, rather than rerunning the complete experiments. Full implementation and generated responses may be requested from the authors through the anonymous OpenReview discussion during review, or from the corresponding author after identities are released, subject to access approval and applicable data terms.

REFERENCES

Lina Alkarmi, Ziyuan Huang, and Mingyan Liu. When in doubt, abstain: The impact of abstention on strategic classification. In *Game Theory and AI for Security*, volume 16224 of *Lecture Notes in Computer Science*, pp. 124–144. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-08067-7_7. URL https://link.springer.com/chapter/10.1007/978-3-032-08067-7_7.

Jason Altschuler, Victor-Emmanuel Brunel, and Alan Malek. Best arm identification for contaminated bandits. *Journal of Machine Learning Research*, 20(91):1–39, 2019. URL https://jmlr.org/papers/v20/18-395.html.

Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025. doi: 10.48550/arXiv.2503.11926. URL https://arxiv.org/abs/2503.11926.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, 2005. doi: 10.1109/TIT.2005.847729. URL https://cesa-bianchi.di.unimi.it/Pubblicazioni/J22.pdf.

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. URL https://arxiv.org/abs/2110.14168.

Chris Cundy and Adam Gleave. Preference learning with lie detectors can induce honesty or evasion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 129495–129553. Curran Associates, Inc., 2025. doi: 10.52202/085713-4313. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/bc14bd40e1186727722e23616ee2aeb9-Abstract-Conference.html.

Christoph Dann, Mohammad Ghavamzadeh, and Teodor V. Marinov. Multiple-policy high-confidence policy evaluation. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9470–9487. PMLR, 2023. URL https://proceedings.mlr.press/v206/dann23a.html.

David L. Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):238–270, 1994. doi: 10.1214/aos/1176325367. URL https://doi.org/10.1214/aos/1176325367.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning*

Research, pp. 998–1027. PMLR, 2016. URL https://proceedings.mlr.press/v49/garivier16a.html.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004. URL https://jmlr.org/papers/v5/greensmith04a.html.

Morris H. Hansen and William N. Hurwitz. The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41(236):517–529, 1946. doi: 10.1080/01621459.1946.10501894. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1946.10501894.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pp. 111–122. ACM, 2016. doi: 10.1145/2840728.2840730. URL https://doi.org/10.1145/2840728.2840730.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483446.

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. doi: 10.1214/20-AOS1991. URL https://doi.org/10.1214/20-AOS1991.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL https://jmlr.org/papers/v17/kaufman16a.html.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763. PMLR, 2021. URL https://proceedings.mlr.press/v139/kossen21a.html.

Leon Lang, Davis Foote, Stuart Russell, Anca Dragan, Erik Jenner, and Scott Emmons. When your AIs deceive you: Challenges of partial observability in reinforcement learning from human feedback. In *Advances in Neural Information Processing Systems*, volume 37, pp. 93240–93299. Curran Associates, Inc., 2024. doi: 10.52202/079017-2959. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a995960dd0193654d6b18eca4ac5b936-Abstract-Conference.html.

Puheng Li, Tijana Zrnic, and Emmanuel J. Candès. Robust sampling for active statistical inference. In *Advances in Neural Information Processing Systems*, volume 38, pp. 68727–68756. Curran Associates, Inc., 2025. doi: 10.52202/085713-2312. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/638947056421498360 4d1ac81631c2c5-Abstract-Conference.html.

Ilia Mahrooghi, Mahshad Moradi, Sina Akbari, and Negar Kiyavash. Multi-armed bandits with missing outcomes. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 2844–2875. PMLR, 2025. URL https://proceedings.mlr.press/v286/mahrooghi25a.html.

Charles F. Manski. Minimax-regret treatment choice with missing outcome data. *Journal of Econometrics*, 139(1):105–115, 2007. doi: 10.1016/j.jeconom.2006.06.006. URL https://www.sciencedirect.com/science/article/pii/S0304407606001047.

Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–606, 1934. doi: 10.1111/j.2397-2335.1934.tb04184.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.2397-2335.1934.tb04184.x.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609. PMLR, 2020. URL https://proceedings.mlr.press/v119/perdomo20a.html.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 814–823. PMLR, 2015. URL https://proceedings.mlr.press/v37/swaminathan15.html.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388. PMLR, 2015. URL https://proceedings.mlr.press/v37/thomas15.html.

Aaron D. Tucker, Caleb Biddulph, Claire Wang, and Thorsten Joachims. Bandits with costly reward observations. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2147–2156. PMLR, 2023. URL https://proceedings.mlr.press/v216/tucker23a.html.

Dennis Wei. Decision-making under selective labels: Optimal finite-domain policies and beyond. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11035–11046. PMLR, 2021. URL https://proceedings.mlr.press/v139/wei21a.html.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696. URL https://link.springer.com/article/10.1007/BF00992696.

Tommy Wright. A general exact optimal sample allocation algorithm: With bounded cost and bounded sample sizes. *Statistics & Probability Letters*, 165:108829, 2020. doi: 10.1016/j.spl.2020.108829. URL https://www.sciencedirect.com/science/article/pii/S0167715220301322.

Tianrun Zhao, Xiaojie Mao, and Yong Liang. Online strategic classification with noise and partial feedback. In *Advances in Neural Information Processing Systems*, volume 38, pp. 156297–156351. Curran Associates, Inc., 2025. doi: 10.52202/085713-5222. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e50e253e21cbcdcd200394f61d73acc8-Abstract-Conference.html.

Tijana Zrnic and Emmanuel J. Candès. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62993–63010. PMLR, 2024. URL https://proceedings.mlr.press/v235/zrnic24a.html.

A FULL PROOFS OF THE MAIN RESULTS

We prove the identified improvement interval, the audit bounds, and the optimal allocation and shrinkage results in the order used by the main text. Appendix B gives the supporting optimization and learning results. Experimental details and additional evidence appear in Appendices C and D.

A.1 COMPLETE ASSUMPTIONS AND NOTATION

The set Δ_K contains nonnegative K-dimensional probability vectors that sum to one. For each action, the joint distribution of reward and visibility is fixed over time, except where a result explicitly concerns a conditional score identity. Each interaction uses fresh environment randomness after its action is selected. The policy may depend on all previous natural records and audits. An independent coin drawn after the outcome determines whether to reveal a missing reward. In the results with both interaction and audit budgets, the behavior distribution is fixed within each validation episode. It may change between episodes, provided each new candidate is frozen before its validation data are collected.

Symbol	Meaning
$a \in [K]$	Composite action, including quality and evidence presentation
R, G	Bounded reward and natural reveal indicator
Q, q	Independent audit coin and its logged probability
μ_a	True reward mean $\mathbb{E}[R \mid a]$
p_a, s_a, h_a	Missing probability, visible reward mass, and hidden conditional mean
π, π', d	Current and proposed policies and their difference
D	Exact finite improvement $d^\top \mu$
w_a, H, m	Hidden contrast coefficient $d_a p_a$, total absolute weight, and midpoint
ρ_a, r_a	Behavior probability and closed-outcome arrival probability $\rho_a p_a$
N, T	Paid audit and interaction budgets
$n_a, \widehat{h}_a$	Audit quota or expected allocation, and a hidden-mean estimate
γ, ϵ, δ	Improvement margin, learning error, and failure probability
$\mathrm{kl}(x, y)$	Bernoulli relative entropy $x \log(x/y) + (1-x)\log((1-x)/(1-y))$

Table 1: Notation. The meaning of n_a as an expected allocation or an integer quota is specified in each theorem.

An observable law for action a consists of p_a and the full distribution of the naturally revealed reward. Necessarily $0 \le s_a \le 1 - p_a$. Fixing only these moments is enough to compute the bounds, but the indistinguishability constructions below preserve the entire observable law. For $p_a = 0$ we set $h_a = 1/2$ by convention. An action with missing rewards affects the comparison only if $p_a > 0$ and $d_a \neq 0$.

Uniform δ-correctness for the separated sign problem means that, on every instance with $|D| \ge \gamma$, the procedure returns the correct sign with probability at least $1 - \delta$. An unresolved output counts as a failure for this task. A confidence interval at a fixed budget instead controls the probability of an incorrect resolved decision, with its resolution rate reported separately. The sign-testing lower bounds concern the first task, and the experiments report interval validity and resolution. Theorem 4 concerns uniform estimation with a fixed error radius.

Finite contexts. The contrast results extend to finitely many observed contexts with fixed distribution P_C. Set $d_{ca} = P_C(c)(\pi'(a \mid c) - \pi(a \mid c))$ and $\rho_{ca} = P_C(c)\rho(a \mid c)$. These quantities sum to zero and one, respectively. Treat each (c, a) pair as a stratum. The finite-action PAC learner in Theorem 8 assumes that actions can be selected deliberately for acquisition. The experiments in this manuscript use one context per population.

A.2 Identification of the improvement

A.2.1 Proof of Proposition 1

To find the range of possible rewards, separate the observed and missing contribution for each action:

$$\mu_a = \mathbb{E}[GR \mid a] + \mathbb{E}[(1-G)R \mid a] \tag{17}$$
$$= s_a + p_a h_a, \qquad 0 \le h_a \le 1. \tag{18}$$

Hence $\mu_a \in [s_a, s_a + p_a]$. Conversely, retain the gate probability and the complete open-reward distribution. On $G = 0$, assign any distribution on $[0,1]$ with mean $h_a = (\mu_a - s_a)/p_a$ when $p_a > 0$. A Bernoulli distribution suffices. This change is invisible to the natural channel. The choices can be made independently across actions, proving that the full product box is attained.

To measure the uncertainty around the midpoint, write $h_a = 1/2 + u_a$, where $u_a \in [-1/2, 1/2]$ is the unknown deviation. Substitution gives

$$D = \sum_a d_a(s_a + p_a/2) + \sum_a d_a p_a u_a = m + \sum_a w_a u_a. \tag{19}$$

The maximum is attained by $u_a = \operatorname{sign}(w_a)/2$ and equals $m + H/2$; the minimum uses the opposite signs. Intermediate values follow by convex mixing of these hidden distributions. A strictly positive interval guarantees a positive change in every compatible system. If the interval contains a positive and a negative point, the corresponding hidden laws have the same natural observations and opposite correct decisions.

A.3 Audit bounds for deciding an update's sign

A.3.1 The mixture reduction

Suppose $H > 0$ and sample an action with probability $|w_a|/H$. Its closed reward has mean h_a. For the transformed reward in (4),

$$H\mathbb{E}[Y] = \sum_{w_a > 0} w_a h_a + \sum_{w_a < 0} |w_a|(1 - h_a) \tag{20}$$
$$= \sum_a w_a h_a + \sum_{w_a < 0} |w_a|. \tag{21}$$

Also $\sum_{w_a<0} |w_a| = (H - \sum_a w_a)/2$. Therefore

$$m + H(\mathbb{E}[Y] - 1/2) = d^\top(s + p/2) + \sum_a w_a h_a - \frac{1}{2}\sum_a w_a \tag{22}$$
$$= d^\top s + \sum_a w_a h_a = D. \tag{23}$$

The reduction uses fresh hidden outcomes. Drawing a new action and a new closed-stratum outcome each time makes the transformed observations independent and identically distributed in $[0,1]$.

A.3.2 A sequential sign test

First return a sign if the natural interval in (3) excludes zero. If $H = 0$, $D = m$ is known; a zero value can be returned as equality. Otherwise inspect the estimate after sample sizes $n_j = 16 \cdot 2^j$, where $j = 0, 1, \ldots$ indexes successive checks. To keep the total error probability below δ despite repeated checks, assign error allowance δ_j to check j and choose its interval half-width r_j as

$$\delta_j = \frac{6\delta}{\pi^2(j+1)^2}, \qquad r_j = H\sqrt{\frac{\log(2/\delta_j)}{2n_j}}. \tag{24}$$

Here π in the denominator is the numerical circle constant, not a policy. After n_j audits, let $\overline{Y}_{n_j}$ be their transformed sample mean and form $\widehat{D}_j = m + H(\overline{Y}_{n_j} - 1/2)$. Return positive if $\widehat{D}_j - r_j > 0$,

negative if $\widehat{D}_j + r_j < 0$, and otherwise continue. A fixed audit budget stops the procedure at its last completed look and returns unresolved if necessary.

Hoeffding's inequality and $\sum_j \delta_j = \delta$ imply

$$\mathbb{P}\left(\forall j,\ |\widehat{D}_j - D| \le r_j\right) \ge 1 - \delta. \tag{25}$$

On this event every returned sign is correct. Let $\gamma = |D| > 0$. If $r_j < \gamma/2$, then $|\widehat{D}_j| > r_j$, so the test has stopped by look j. It is enough that

$$n_j > \frac{2H^2}{\gamma^2} \log \frac{\pi^2 (j+1)^2}{3\delta}. \tag{26}$$

The geometric schedule increases by a factor of two. Solving this inequality gives a constant multiple of $1 + (H^2/\gamma^2)[\log(1/\delta) + \log\log(e^e + H/\gamma)]$. One way to see this is to put $x = \max\{1, H^2/\gamma^2\}$, note that the stopping look is $O(\log(x \log(1/\delta)) + 1)$, and substitute this logarithmic look count into $\log(j+1)$. This proves the stated high-probability sample bound. The budget experiments instead use the fixed-horizon construction in Section 5.

A.3.3 The adaptive audit lower bound

To show why some updates require many audits, construct two systems with identical observed rewards. Fix p, d and set every observed reward to $1/2$. This makes $s_a = (1 - p_a)/2$ and sets the comparison midpoint to zero:

$$m = \sum_a d_a \left((1 - p_a)/2 + p_a/2\right) = \tfrac{1}{2} \sum_a d_a = 0. \tag{27}$$

Make the systems differ only in the missing rewards. Choose their Bernoulli means to shift the improvement by opposite amounts:

$$h_a^{\pm} = \tfrac{1}{2} \pm \operatorname{sign}(w_a)\gamma/H. \tag{28}$$

Substituting these missing means verifies that the two true reward changes have opposite signs:

$$D_{\pm} = \sum_a w_a (h_a^{\pm} - 1/2) = \pm\gamma. \tag{29}$$

The natural feedback law is the same in both worlds. For $x = \gamma/H \le 1/4$,

$$\operatorname{kl}(1/2 + x, 1/2 - x) = 2x \log \frac{1 + 2x}{1 - 2x} \tag{30}$$

$$\le \frac{8x^2}{1 - 4x^2} \le 16x^2. \tag{31}$$

The first inequality follows by integrating the derivative of $\log((1+z)/(1-z))$, or bounding it by $2z/(1-z^2)$ at $z = 2x$. Complementing a Bernoulli outcome leaves this divergence unchanged, so each relevant audited action has the same bound.

Let $\mathcal{H}_\tau$ be the complete stopped observable transcript, including actions, gates, natural rewards, audit coins, and revealed audit rewards. At each history, the algorithm uses the same conditional action, auditing, and stopping rules under the two worlds. Those kernels contribute zero conditional relative entropy. Natural records also contribute zero. The chain rule therefore gives

$$\operatorname{KL}(\mathbb{P}_+^{\mathcal{H}_\tau}, \mathbb{P}_-^{\mathcal{H}_\tau}) = \sum_a \mathbb{E}_+[N_a] \operatorname{kl}(h_a^+, h_a^-) \le \frac{16\gamma^2}{H^2} \mathbb{E}_+[N]. \tag{32}$$

For integrable stopping times this follows by truncation and the stopped-history change-of-measure lemma of Kaufmann et al. (2016). If the expected audit count is infinite, the lower bound already holds. Apply data processing to the event E that the output is positive. Uniform correctness gives $\mathbb{P}_+(E) \ge 1 - \delta$ and $\mathbb{P}_-(E) \le \delta$. Bernoulli relative entropy is minimized over these probabilities at $(1 - \delta, \delta)$ for $\delta < 1/2$. Thus

$$\frac{16\gamma^2}{H^2} \mathbb{E}_+[N] \ge \operatorname{kl}(1 - \delta, \delta), \tag{33}$$

which proves Theorem 2. The lower bound charges only independent new hidden rewards. Re-reading an already revealed value cannot increase transcript information.

A.4 Optimal error under both resource budgets

We first solve the deterministic allocation and shrinkage problem in Theorem 3. We then construct a randomized estimator and prove the information lower bound for uniform estimation in Theorem 4. Both statistical bounds concern a fixed error radius under the same interaction and audit budgets.

A.4.1 Zero coordinates and the modulus dual

Write $a_i = |w_i|$ within this proof, and $C_i = Tr_i$ for the available audit counts. We use the perspective convention $v_i^2/0 = +\infty$ when $v_i \neq 0$, and $0^2/0 = 0$. A coordinate with $C_i = 0$ contributes the unremovable term $a_i/2$ to the modulus; remove such coordinates from the norm calculations and add their constant at the end. Coordinates with $a_i = 0$ need no samples. If $N = 0$ or all available caps vanish, the result is $H/2$. For $z = 0$, the definition gives $\mathcal{M}_0(n) = \frac{1}{2}\sum_{i:n_i=0} a_i$; the statistical constructions below use $z > 0$.

To connect the error criterion to an information lower bound, let $u_a \in [0, 1/2]$ represent a possible change in a missing mean. The weighted sum measures its effect on the target, while $\sum_a n_a u_a^2$ limits how easily the available audits distinguish that change. Convex duality gives the equivalent form

$$\mathcal{M}_z(n) = \max_{\substack{0\le u_a\le 1/2 \\ \sum_a n_a u_a^2 \le z^2}} \sum_a |w_a| u_a. \tag{34}$$

This is the maximum representation corresponding to the minimum in (8). We prove their equality next.

For fixed n, let $B = [-1/2, 1/2]^K$ and $E = \{u : \sum_i n_i u_i^2 \le z^2\}$. Symmetry permits the maximizer of $a^\top u$ to be nonnegative. The support function of their intersection equals the infimal convolution of their support functions:

$$\mathcal{M}_z(n) = \inf_v \left\{ \tfrac{1}{2}\|a - v\|_1 + z\sqrt{\sum_i v_i^2/n_i} \right\} \tag{35}$$

$$= \min_{0\le v\le a} \left\{ \tfrac{1}{2}\sum_i (a_i - v_i) + z\sqrt{\sum_i v_i^2/n_i} \right\}. \tag{36}$$

For positive n and $z > 0$, the box and ellipsoid have intersecting interiors, so convex support-function duality has no gap. Zero coordinates follow from the stated conventions, or by a limit from positive n. Projecting each v_i onto $[0, a_i]$ reduces both terms, proving the second equality. This is the bounded linear-functional optimal-recovery formulation (Donoho, 1994) specialized to the present audit information.

There is also an explicit verification. If $\sum_i n_i/4 \le z^2$, the whole box fits within the information ellipsoid, so $\mathcal{M}_z = H/2$, attained in the dual by $v = 0$. Otherwise choose $c > 0$ such that

$$\sum_i \min\{a_i^2/(c^2 n_i), n_i\} = 4z^2. \tag{37}$$

The primal optimizer is $u_i = \min\{1/2, a_i/(2cn_i)\}$, with $u_i = 1/2$ when $n_i = 0$. Its information constraint is tight. The dual optimizer is $v_i = \min\{a_i, cn_i\}$. Setting $S = \sum_i v_i^2/n_i$ gives $S = 4c^2 z^2$, and

$$\tfrac{1}{2}\sum_i (a_i - v_i) + z\sqrt{S} = \tfrac{1}{2}\sum_{a_i > cn_i} a_i + \sum_{a_i \le cn_i} \frac{a_i^2}{2cn_i} = \sum_i a_i u_i. \tag{38}$$

Thus the primal and dual values agree directly.

A.4.2 A common optimal allocation for all confidence scales

To optimize the allocation and retained weights separately, first fix v. On actions with positive availability caps, let $\phi(v)$ be the smallest variance term achievable by any feasible allocation:

$$\phi(v) = \min_{n\in\mathcal{A}(T,N)} \sum_i v_i^2/n_i. \tag{39}$$

Each quadratic perspective is jointly convex in (v_i, n_i), so minimizing over the convex allocation set preserves convexity in v. Also $\phi(tv) = t^2\phi(v)$, $\phi(-v) = \phi(v)$, and $\phi(v) > 0$ for nonzero v. Its unit sublevel set is convex and symmetric; its gauge is $\sqrt{\phi(v)}$, which is therefore a norm. Equivalently, homogeneity and convexity show that the square root satisfies the triangle inequality.

Since minimization over n and v can be interchanged, the global design problem is

$$\min_{0\le v\le a}\left\{\tfrac{1}{2}\sum_i (a_i - v_i) + z\sqrt{\phi(v)}\right\}. \tag{40}$$

This is a convex problem in v. The original two-variable bias-plus-square-root expression need not be jointly convex; the norm reduction is the step that supplies global optimality.

We next identify when sampling is no better than retaining the original uncertainty interval. The usable expected budget is $N_0 = \min(N, \sum_i C_i)$. Cauchy–Schwarz bounds the sampling term from below:

$$\sqrt{\sum_i v_i^2/n_i} \ge \frac{\sum_i v_i}{\sqrt{\sum_i n_i}} \ge \frac{\sum_i v_i}{\sqrt{N_0}}. \tag{41}$$

If $N_0 \le 4z^2$, the objective in (40) is at least $H/2 + (z/\sqrt{N_0} - 1/2)\sum_i v_i \ge H/2$. The choice $v = 0$, requiring no audits, attains this value.

Now assume $N_0 > 4z^2$. Let $n_i^* = \min(C_i, \lambda a_i)$ use budget N_0, as given by constrained Neyman allocation (Neyman, 1934; Wright, 2020). First suppose $N < \sum_i C_i$, so λ is finite and the budget constraint is active. At $c = 1/\lambda$, every coordinate satisfies $a_i/n_i^* \ge c$, and the left side of the root equation in (11) equals $N_0 > 4z^2$. It decreases continuously to zero as $c \to \infty$, with a unique root above $1/\lambda$. Set $v_i^* = \min(a_i, cn_i^*)$.

If $v_i^* < a_i$, then $a_i/n_i^* > c > 1/\lambda$. Such a coordinate must have $n_i^* = C_i$: an unsaturated coordinate would have ratio exactly $1/\lambda$. Moreover $\lambda v_i^* = \lambda c C_i > C_i$. On an unshrunken coordinate $v_i^* = a_i$. Thus

$$n_i^* = \min(C_i, \lambda v_i^*) \tag{42}$$

for every coordinate, so n^* is also a minimizer of $\phi(v^*)$. With positive v^*, that allocation is unique. The envelope derivative of $\sqrt{\phi}$ is

$$\partial_i\sqrt{\phi(v^*)} = \frac{v_i^*}{n_i^*\sqrt{S}}, \qquad S = \sum_i (v_i^*)^2/n_i^*. \tag{43}$$

The root gives $\sqrt{S} = 2zc$. At every interior coordinate $v_i^* < a_i$, the derivative of (40) is $-1/2 + zcn_i^*/(n_i^*\sqrt{S}) = 0$. At an upper-bound coordinate $v_i^* = a_i \le cn_i^*$, that derivative is nonpositive. These are the complete convex optimality conditions on the box. Therefore v^* is globally optimal and n^* minimizes $\mathcal{M}_z$.

If $N \ge \sum_i C_i$, all coordinates receive their maximum possible expected samples. This allocation minimizes $\phi(v)$ for every v, so the fixed-allocation dual solution above proves the same result. We have shown that the allocation n^* is common to every z; only its optimal shrinkage varies. In the nontrivial branch $N_0 > 4z^2$, every partially shrunken coordinate is at its arrival cap. The low-budget branch permits shrinkage of every coordinate to zero.

Without active caps, $n_i^* = Na_i/H$. The primal maximizer has equal coordinates until it reaches the box boundary, which gives

$$\mathcal{R}_z(T, N) = \min\{H/2, zH/\sqrt{N}\}. \tag{44}$$

With a zero arrival cap, the corresponding half-width is retained as a fixed ambiguity contribution. These cases explain the smooth sampling regime and the bounded-uncertainty regime within a single expression.

A.4.3 Bernoulli implementation and the 5/3 radius

For the high-budget design, set $q_i = n_i^*/(Tr_i)$ and draw the audit coin independently on each closed outcome. Define the signed shrunken estimator

$$\widehat{D}_v = m + \sum_{t=1}^{T}\sum_i \mathbf{1}\{A_t = i, G_t = 0, Q_t = 1\}\frac{\operatorname{sign}(w_i)v_i^*}{n_i^*}(R_t - 1/2). \tag{45}$$

Coordinates with $n_i^* = 0$ have $v_i^* = 0$ and are omitted. The expectation is $m + \sum_i \text{sign}(w_i) v_i^* (h_i - 1/2)$. Its distance from D is at most $b = \frac{1}{2}\sum_i (a_i - v_i^*)$.

To control random error, let X_t be the contribution of interaction t to (45). These contributions are independent because the behavior policy is fixed. Define S to bound their total variance and k to bound the magnitude of one contribution:

$$S = \sum_i (v_i^*)^2 / n_i^*, \qquad k = \max_i v_i^* / n_i^*. \tag{46}$$

Each $X_t \in [-k/2, k/2]$, so $|X_t - \mathbb{E}X_t| \le k$. Since the probability of an audited action-i outcome is $r_i q_i = n_i^*/T$,

$$\sum_{t=1}^{T} \text{Var}(X_t) \le \sum_{t=1}^{T} \mathbb{E}[X_t^2] \le T \sum_i \frac{n_i^*}{T} \frac{(v_i^*)^2}{4(n_i^*)^2} = S/4. \tag{47}$$

Bernstein's inequality gives, except with probability $2e^{-L}$,

$$|\widehat{D}_v - D| \le b + \sqrt{LS/2} + \frac{2}{3}Lk =: R_B. \tag{48}$$

Set $z = \sqrt{L/2}$. The root identity gives $\sqrt{S} = 2zc$ and $k \le c$. Hence

$$R_B \le b + \frac{5}{3} z\sqrt{S} \le \frac{5}{3}\mathcal{R}_z(T, N). \tag{49}$$

In the low-budget branch, the midpoint interval of radius $H/2$ is deterministically valid and requires no audit. With $L = \log(4/\delta)$, the error probability above is at most $\delta/2$. Accepting a positive update when $\widehat{D}_v - R_B > 0$ is always valid on the confidence event; resolving a true sign on that event is guaranteed when $|D| > 2R_B$. The sufficient condition $|D| > 10\mathcal{R}_z/3$ in the theorem follows.

The total paid count A_T is binomial with mean $\sum_i n_i^* \le N$. Its upper tail satisfies

$$\mathbb{P}\left(A_T > N + \sqrt{2NL} + \frac{2}{3}L\right) \le e^{-L}. \tag{50}$$

With a hard audit limit, use this estimator if validation completes within the limit and return unresolved otherwise. A union bound with the estimator event gives combined failure probability at most $3e^{-L} \le \delta$. For a prescribed integer limit B, use the design expectation $N' = (B - \sqrt{2BL} - 2L/3)_+$ and stop acquiring at B. If $N' = 0$, use the midpoint-only branch. This construction can sample a stratum with expected count below one because it uses Bernoulli audits rather than integer quotas for each coordinate.

A.4.4 Information lower bound for uniform estimation

Define $r_\delta^*(T, N)$ as in Theorem 4, with the budget bounds holding over the compatible hidden laws. Fix a reference law with hidden Bernoulli rewards of mean $1/2$, and let n_i be the algorithm's expected audits under that law. The natural channel is held fixed. The availability argument in Appendix B.3 gives $n \in \mathcal{A}(T, N)$ even for adaptive audit choices and stopping.

Suppose an estimator is uniformly accurate to radius η with probability $1 - \delta$. Put $\kappa = \text{kl}(1 - \delta, \delta)$ and $z_L = \sqrt{3\kappa/8}$. If $\mathcal{M}_{z_L}(n) > 4\eta$, choose a feasible maximizing vector $u \in [0, 1/2]^K$ and alternative hidden Bernoulli rewards with means

$$h_i^1 = 1/2 + \text{sign}(w_i) u_i / 2. \tag{51}$$

Its contrast differs from the reference by $\sum_i a_i u_i / 2 > 2\eta$. All alternative means lie in $[1/4, 3/4]$. Therefore

$$\sum_i n_i \,\text{kl}(1/2, h_i^1) \le \frac{16}{3} \sum_i n_i (u_i/2)^2 \tag{52}$$

$$\le \frac{4}{3} z_L^2 = \kappa/2. \tag{53}$$

On the other hand, the event that the estimate lies within η of the reference target has reference probability at least $1-\delta$ and alternative probability at most δ, because the two accuracy intervals are disjoint. Data processing and stopped-history change of measure require relative entropy at least κ, a contradiction. Thus $\mathcal{M}_{z_L}(n) \le 4\eta$ and $\mathcal{R}_{z_L}(T,N) \le 4\eta$. Taking the infimum over valid estimators proves the left side of (12).

This lower bound and the upper construction concern the same uniform fixed-radius estimation task. For $\delta \le 1/4$, $\kappa \ge \log(4/\delta)/6$, so $z_U/z_L \le \sqrt{8}$. Also $\mathcal{M}_{cz}(n) \le c\mathcal{M}_z(n)$ for $c \ge 1$, directly from its bias-plus-noise dual. Minimizing over n preserves this inequality. The two sides of (12) therefore differ by a universal factor in the attainable error radius at fixed budgets. Near a bounded-uncertainty threshold, this radius comparison is not a pointwise assertion about the ratio of budgets required to cross that threshold.

For the centered natural law $m = 0$, the same argument gives a separated-sign lower bound. Let $0 < \gamma \le H/8$ and take reference means $h_i^0 = 1/2 + \mathrm{sign}(w_i)\gamma/H$, so that $D_0 = \gamma$. Define $n_i = \mathbb{E}_0[N_i]$ under this reference. If $\mathcal{M}_{z_L}(n) > 4\gamma$, a maximizing vector u gives alternative means $h_i^1 = h_i^0 - \mathrm{sign}(w_i)u_i/2$ with $D_1 < -\gamma$. Both sets of means lie in $[1/4, 3/4]$, and their transcript KL is at most $\kappa/2$ by the bound above. Uniform correctness would require KL at least κ, so $\mathcal{R}_{z_L} \le 4\gamma$ is necessary. An unresolved output counts as failure for this separated-sign task. Fixed-budget interval evaluations report validity and resolution separately.

Computing and interpreting the design. The allocation and shrinkage rules require scalar searches in λ and c, using the monotonicity in (10) and (11). Each search step scans the active coordinates once. Changing the confidence scale requires recomputing shrinkage but leaves the allocation unchanged. The design uses missingness and the proposed contrast, before hidden rewards are acquired.

For fixed natural feedback, behavior policy, and budgets, replacing d by αd with $\alpha > 0$ scales m, H, and $\mathcal{R}_z$ by α, leaving the allocation and retained fractions t_a unchanged. The estimate and radius scale together: reducing step size along the same direction does not make its sign harder to determine. Grouped proposals in Section 6 can instead change the direction and place weight on outcomes with different arrival rates.

A.4.5 Finite calibration of the shrinkage estimator

We must also account for error in the preliminary estimates. On the event that all estimates are accurate, assume $|\widehat{s}_i - s_i| \le e_i$ and $|\widehat{p}_i - p_i| \le e_i$, where e_i is the error bound for action i. Set the conservative missingness bound $\overline{p}_i = \min(1, \widehat{p}_i + e_i)$ and empirical midpoint $\widehat{m} = d^\top(\widehat{s} + \widehat{p}/2)$. The midpoint error is bounded by

$$\begin{aligned} |\widehat{m} - m| &= \left| \sum_i d_i \big[(\widehat{s}_i - s_i) + (\widehat{p}_i - p_i)/2 \big] \right| \\ &\le \sum_i |d_i| \big(|\widehat{s}_i - s_i| + |\widehat{p}_i - p_i|/2 \big) \le \frac{3}{2} \sum_i |d_i| e_i. \end{aligned} \tag{54}$$

This coordinatewise bound permits unequal e_i. The empirical $\widehat{p}_i$ defines the midpoint; the upper bound $\overline{p}_i$ enters the design below. Design with $\overline{a}_i = |d_i|\overline{p}_i$, caps $T\rho_i\overline{p}_i$, expected allocations $\overline{n}_i = T\rho_i\overline{p}_i q_i$, and shrinkage $t_i = \overline{v}_i/\overline{a}_i$. The executable estimator is

$$\widehat{m} + \sum_{t=1}^{T} \sum_i \mathbf{1}\{A_t = i, G_t = 0, Q_t = 1\} \frac{t_i d_i}{T\rho_i q_i} (R_t - 1/2). \tag{55}$$

The weights use the recorded audit probabilities and the design computed from calibration. On the calibration event, the shrinkage bias is at most $\frac{1}{2}\sum_i \overline{a}_i(1-t_i)$, the variance is at most $\frac{1}{4}\sum_i \overline{v}_i^2/\overline{n}_i$, and the centered range is at most $\max_i \overline{v}_i/\overline{n}_i$. Centering an interval at (55) with the Bernstein radius of the barred design plus $\frac{3}{2}\sum_i |d_i|e_i$ therefore gives coverage at the allocated confidence level.

On the same event, the expected paid count is $T\sum_i \rho_i p_i q_i \le \sum_i \overline{n}_i$. If calibration fails with probability δ_C, the unconditional expectation without truncation is bounded by $N + T\delta_C$. A hard audit limit bounds realized spending on every path, and the union of the calibration, estimator, and

budget-tail events controls the decision error. Because the action-conditional law is stable, one simultaneous calibration event can be reused across a candidate sequence at failure cost δ_C. Each fresh validation episode receives its own failure allocation.

B Additional theory

The following results connect update evaluation to optimization without audits, learning from censored feedback, fixed audit quotas, and finite-action learning. The proofs state the required observations and their costs.

Strategic classification, abstention, and partial feedback model responses to evaluation rules (Hardt et al., 2016; Alkarmi et al., 2026; Zhao et al., 2025), and performative prediction studies distributions affected by deployment (Perdomo et al., 2020). The stable composite-action model isolates how a policy's evidence choices affect learning and the observations needed to evaluate a change in true reward.

B.1 A no-audit optimization criterion

Without audits, one can still choose a policy that minimizes its worst possible shortfall from the best action. This shortfall is its regret. Let $L_a = s_a$ and $U_a = s_a + p_a$ be the smallest and largest possible mean rewards. For a fixed policy π and comparator action j, the worst shortfall is

$$\sup_{\mu \in \prod_a [L_a, U_a]} (\mu_j - \pi^\top \mu) = (1 - \pi_j) U_j - \sum_{i \neq j} \pi_i L_i. \tag{56}$$

The coefficient of μ_j is nonnegative and all other coefficients are nonpositive, so the stated endpoints attain the supremum. Taking the maximum over j gives the worst compatible regret. Introduce a scalar t that bounds the shortfall against every action j. Minimizing this bound gives the linear program

$$\min_{\pi \in \Delta_K, t} \quad t \tag{57}$$

$$\text{subject to} \quad t \geq (1 - \pi_j) U_j - \sum_{i \neq j} \pi_i L_i, \qquad j \in [K] \tag{58}$$

computes the smallest worst-case regret. A policy that is ϵ-optimal for every compatible reward vector exists exactly when this optimum is at most ϵ. This is a direct finite-action form of minimax-regret treatment choice with missing outcomes (Manski, 2007).

A two-action lower bound follows by taking $p_1 = p_2 = p$ and open reward $1/2$. Let hidden means be $(1/2 + u, 1/2 - u)$ in one world and their exchange in the other, with $0 < u < 1/2$. True means are $(1/2 + pu, 1/2 - pu)$ or the exchanged vector. All natural records have the same distribution for every adaptive strategy. A no-audit algorithm therefore has the same output law in both worlds. If its mean probability of choosing action 1 is x, its expected regrets in the two worlds are $2pu(1 - x)$ and $2pux$, so the larger is at least pu. Thus the learning error remains positive without audits, even for adaptive action selection.

B.2 Why missing feedback changes the baseline effect

With complete feedback, subtracting an action-independent baseline changes policy-gradient variance but not its mean. We show why that fails when missing rewards are simply discarded. Let $G_b(\theta)$ denote the mean update using only observed rewards and baseline b; this vector is distinct from the visibility indicator G. Changing the baseline by a constant c changes that mean in proportion to the derivative of the observation probability:

$$\begin{aligned} G_b(\theta) &= \mathbb{E}_{a \sim \pi_\theta}[G(R - b) \nabla_\theta \log \pi_\theta(a)], \\ G_{b+c}(\theta) - G_b(\theta) &= -c \nabla_\theta \mathbb{P}_{\pi_\theta}(G = 1). \end{aligned} \tag{59}$$

B.2.1 Derivation of the baseline identity

Let $s_a = \mathbb{E}[GR \mid a]$ and $o_a = \mathbb{P}(G = 1 \mid a) = 1 - p_a$. At a fixed parameter value with an action-independent baseline b,

$$G_b(\theta) = \sum_a \pi_\theta(a)(s_a - bo_a)\nabla_\theta \log \pi_\theta(a) \tag{60}$$

$$= \sum_a (s_a - bo_a)\nabla_\theta \pi_\theta(a). \tag{61}$$

Here the baseline is held fixed when computing the score update. Subtracting the expressions at $b + c$ and b gives

$$G_{b+c} - G_b = -c \sum_a o_a \nabla_\theta \pi_\theta(a) \tag{62}$$

$$= -c\nabla_\theta \sum_a \pi_\theta(a) o_a = -c\nabla_\theta \mathbb{P}_{\pi_\theta}(G = 1). \tag{63}$$

Under complete feedback, $o_a = 1$ for every action and this derivative vanishes because probabilities sum to one. The gated update retains the derivative of its observation probability.

For a tabular softmax policy, θ_i is the logit of action i, and $\nabla_{\theta_i} \log \pi(a) = \mathbf{1}\{a = i\} - \pi_i$. To see the effect on one action, consider deterministic visibility $g_a \in \{0, 1\}$ and deterministic reward $R_a^0 \in [0, 1]$. Its observed centered reward is $f_a = g_a(R_a^0 - b)$. Component i of the mean update is then

$$G_{b,i} = \pi_i \left(f_i - \sum_a \pi_a f_a \right). \tag{64}$$

Consequently, the logit of a hidden action with $f_i = 0$ increases or decreases according to the average visible contribution. Its update depends on the baseline through that average, while the complete-feedback mean update is baseline-invariant.

B.3 Information constraints and quota designs

To decide between opposite improvements, the inspected rewards must contain enough information to distinguish them. We first bound that information under both budgets, then derive a simpler quadratic condition and a procedure that checks a fixed number of missing rewards from each action.

Theorem 5 (Information and availability)**.** *Let $H > 0$ and $0 < \gamma \le H/4$. Use the centered natural law and reference hidden means $h_a^0 = 1/2 + \operatorname{sign}(w_a)\gamma/H$. For every test that returns the correct sign with probability at least $1 - \delta$ on each instance with $|D| \ge \gamma$, let N_a be the number of action-a audits and $n_a = \mathbb{E}_0[N_a]$ its expectation under the reference system. The allocation obeys $\sum n_a \le N$ and $n_a \le Tr_a$. Every alternative hidden-mean vector h that changes the improvement by at least 2γ must be distinguishable, so*

$$\inf_{\substack{h \in [0,1]^{\mathcal{I}} \\ \sum_a w_a(h_a^0 - h_a) \ge 2\gamma}} \sum_a n_a \operatorname{kl}(h_a^0, h_a) \;\ge\; \operatorname{kl}(1 - \delta, \delta). \tag{65}$$

Corollary 6 (Local quadratic design)**.** *When $0 < \gamma \le \min_{a \in \mathcal{I}} |w_a|/16$, the information requirement has a simpler quadratic consequence. Let $\Phi(T, N)$ be the smallest full-weight sampling-variance term among feasible positive allocations:*

$$\Phi(T, N) = \min_{\substack{n_a > 0,\ n_a \le Tr_a \\ \sum_a n_a \le N}} \sum_{a \in \mathcal{I}} w_a^2/n_a. \tag{66}$$

The expected budgets of every uniformly δ-correct sign test satisfy

$$\Phi(T, N) \le 64\gamma^2/[3 \operatorname{kl}(1 - \delta, \delta)]. \tag{67}$$

B.3.1 EXACT INFORMATION CONDITION

Fix the reference law h^0 in Theorem 5 and define the expected allocation $n_a = \mathbb{E}_0[N_a]$ before choosing an alternative. For any alternative h with $\sum_a w_a(h_a^0 - h_a) \geq 2\gamma$, its contrast is at most $-\gamma$. Its natural observations remain identical to the reference observations. The stopped chain rule and the positive-output event give

$$\sum_a n_a \operatorname{kl}(h_a^0, h_a) \geq \operatorname{kl}(1-\delta, \delta). \tag{68}$$

The statement holds for every alternative, so it holds for their infimum.

For the availability restriction, let $C_a(\tau)$ be the number of closed action-a arrivals by the stopping time. Pathwise, $N_a \leq C_a(\tau)$. The event that round t is reached is measurable before drawing that round's action and outcome. By conditional expectation and Tonelli's theorem,

$$\mathbb{E}_0[C_a(\tau)] = \sum_{t\geq 1} \mathbb{E}_0[\mathbf{1}\{\tau \geq t\}\mathbf{1}\{A_t = a, G_t = 0\}] \tag{69}$$

$$= \sum_{t\geq 1} r_a \mathbb{P}_0(\tau \geq t) = r_a \mathbb{E}_0[\tau]. \tag{70}$$

Hence $n_a \leq T r_a$ and $\sum_a n_a \leq N$ for the stated expected budgets. If $r_a = 0$ on a consequential stratum, the corresponding alternatives receive no information through this behavior policy. The theorem assumes positive arrival rates on $\mathcal{I}$.

B.3.2 LOCAL QUADRATIC NECESSARY CONDITION

Assume $0 < \gamma \leq \min_{a\in\mathcal{I}} |w_a|/16$. First every n_a on $\mathcal{I}$ must be positive. If $n_a = 0$, alter only that hidden mean by $2\gamma/w_a$. This changes the contrast by 2γ and keeps the hidden mean in $[0, 1]$, but gives zero transcript divergence, contradicting (65).

To construct a difficult alternative with improvement reduced by exactly 2γ, spread the change across actions according to their weights and audit counts. Let S normalize that change, u_a be the reduction of the missing mean, and h_a^1 the alternative mean:

$$S = \sum_a \frac{w_a^2}{n_a}, \qquad u_a = \frac{2\gamma w_a}{n_a S}, \qquad h_a^1 = h_a^0 - u_a. \tag{71}$$

Then $\sum_a w_a u_a = 2\gamma$. Since $S \geq w_a^2/n_a$,

$$|u_a| \leq \frac{2\gamma}{|w_a|} \leq \frac{1}{8}, \qquad |h_a^0 - 1/2| = \gamma/H \leq \frac{1}{16}. \tag{72}$$

Thus every $h_a^1 \in [1/4, 3/4]$, and it is a valid alternative. The elementary bound $\operatorname{kl}(x, y) \leq (x - y)^2/[y(1 - y)]$ follows from $\log z \leq z - 1$. It yields

$$\sum_a n_a \operatorname{kl}(h_a^0, h_a^1) \leq \frac{16}{3} \sum_a n_a u_a^2 \tag{73}$$

$$= \frac{16}{3} \frac{4\gamma^2}{S^2} \sum_a \frac{w_a^2}{n_a} = \frac{64\gamma^2}{3S}. \tag{74}$$

Combining this upper bound with the information lower bound gives $S \leq 64\gamma^2/[3\operatorname{kl}(1 - \delta, \delta)]$. The allocation lies in the feasible set of (66), so $\Phi(T, N) \leq S$ and Corollary 6 follows. The simpler local condition can be replaced by direct feasibility of the constructed displacements u_a.

B.3.3 CHECKING FIXED COUNTS AND COMPLETING THEM IN TIME

The next construction checks a prescribed integer number n_a of missing rewards from each relevant action. It must both obtain those rewards within T interactions and estimate them accurately. Let $\widehat{h}_a$ be their sample mean, S the sum of squared weighted range lengths, and γ the minimum improvement to distinguish. The two conditions below address availability and estimation error, respectively.

Theorem 7 (Finite-horizon quota guarantee). *For known (s, p), fix integers $n_a \geq 1$, $S = \sum_{a\in\mathcal{I}} w_a^2/n_a$, and audit each action's first n_a closed outcomes. If*

$$Tr_a \geq \max\{2n_a, 8\log(2|\mathcal{I}|/\delta)\} \quad (a \in \mathcal{I}), \qquad S \leq 2\gamma^2/\log(4/\delta), \tag{75}$$

the sign of $\widehat{D} = d^\top s + \sum_a w_a \widehat{h}_a$ is correct with probability at least $1-\delta$ for $|D| \geq \gamma$, within T interactions and $\sum_a n_a$ audits. A quota failure may return unresolved.

Fix integer quotas and imagine an infinite independent stream from behavior ρ. For each action, select the first n_a closed outcomes in this stream. The selected rewards have their action's conditional hidden law. Conditional on all action and gate records, they are independent within and across actions. Dependence between G and R is absorbed into that conditional law.

Each term $w_a R_{ai}/n_a$ has range length $|w_a|/n_a$. The sum of squared lengths is

$$\sum_a n_a \left(\frac{|w_a|}{n_a}\right)^2 = S. \tag{76}$$

Hoeffding's inequality therefore gives

$$\mathbb{P}(|\widehat{D} - D| \geq \gamma) \leq 2\exp(-2\gamma^2/S). \tag{77}$$

The second condition of (75) makes this probability at most $\delta/2$.

At the fixed horizon T, the number $C_a(T)$ of closed action-a arrivals has distribution $\mathrm{Binomial}(T, r_a)$. These counts need not be independent across actions. If $Tr_a \geq 2n_a$, the lower-tail Chernoff bound implies

$$\mathbb{P}(C_a(T) < n_a) \leq \mathbb{P}(C_a(T) < Tr_a/2) \leq \exp(-Tr_a/8). \tag{78}$$

The first condition of (75) makes this at most $\delta/(2|\mathcal{I}|)$. A union bound controls any missing quota by $\delta/2$. On the complementary event the finite-horizon estimator agrees with the infinite-stream estimator. Its sign is correct unless the reward deviation event occurs. This proves the joint $1-\delta$ guarantee. Returning unresolved on quota failure gives the same probability of producing a correct decision within the resource limits.

The theorem concerns sign decisions in the separated class $|D| \geq \gamma$. For a decision that may remain unresolved, choose an audit error allowance $\delta_A \in (0, 1)$ and use the interval half-width

$$r_A = \sqrt{\tfrac{1}{2} S \log(2/\delta_A)}. \tag{79}$$

It accepts positive improvement only when $\widehat{D} - r_A > 0$, and symmetrically for a negative decision. The experiments use this sign-exclusion criterion with the calibrated Bernstein radius and report both resolution and error rates.

B.3.4 Solution of the constrained design

Set $c_a = Tr_a > 0$. The objective $\sum w_a^2/n_a$ is strictly convex and decreasing on positive coordinates. If $N \geq \sum c_a$, the optimum is $n_a = c_a$. Otherwise all available audit budget is used. Introduce nonnegative multipliers η for the total budget and ζ_a for each action cap. The constrained minimum is characterized through the Lagrangian

$$L(n, \eta, \zeta) = \sum_a \frac{w_a^2}{n_a} + \eta\left(\sum_a n_a - N\right) + \sum_a \zeta_a(n_a - c_a), \tag{80}$$

where $\eta, \zeta_a \geq 0$. For an unsaturated coordinate, complementary slackness gives $\zeta_a = 0$, and stationarity gives $n_a = |w_a|/\sqrt{\eta}$. Saturated coordinates have $n_a = c_a$. Writing $\lambda = 1/\sqrt{\eta}$ proves

$$n_a^* = \min(c_a, \lambda|w_a|). \tag{81}$$

The sum is continuous and increasing in λ until all coordinates saturate, so the budget equation determines the allocation. It can be solved by scalar bisection. Dividing by $Tr_a = T\rho_a p_a$ yields the audit probabilities $q_a^* = n_a^*/(Tr_a)$.

Without saturation, $n_a = N|w_a|/H$. Substitution gives

$$\sum_a \frac{w_a^2}{n_a} = \frac{H}{N}\sum_a |w_a| = \frac{H^2}{N}. \tag{82}$$

The allocation is feasible exactly when $N|d_a|p_a/H \le T\rho_a p_a$ for all relevant a. Cancelling $p_a > 0$ gives

$$N/T \le H/\max_{a\in\mathcal{I}}(|d_a|/\rho_a). \tag{83}$$

Integer quotas can be rounded upward with at most $|\mathcal{I}|$ extra audits; the acquisition horizon uses the rounded counts. Caps $Tr_a/2$ leave the factor-of-two arrival margin in Theorem 7. The global Bernoulli implementation avoids these per-coordinate integer quotas.

B.4 Leaving low-impact missing rewards unchecked

Some actions matter too little to justify waiting for their missing outcomes. Fix a set J of actions to audit before observing their audit rewards, and use the midpoint $1/2$ for every other missing mean. Let $\widehat{h}_a$ be the audited sample mean and H_{J^c} the total omitted weight. This gives the estimate

$$\widehat{D}_J = d^\top(s + p/2) + \sum_{a\in J} w_a(\widehat{h}_a - 1/2), \qquad H_{J^c} = \sum_{a\notin J}|w_a|. \tag{84}$$

The omitted part of the true contrast has absolute value at most $H_{J^c}/2$. Weighted Hoeffding for the retained part gives, with probability at least $1-\delta_A$,

$$|\widehat{D}_J - D| \le \frac{H_{J^c}}{2} + \sqrt{\frac{1}{2}\log\frac{2}{\delta_A}\sum_{a\in J}\frac{w_a^2}{n_a}}. \tag{85}$$

The sum of a deterministic identification bound and a stochastic radius therefore supports the same three-way decision. This result holds for every bounded reward law and every fixed subset J with obtainable quotas.

Omitting a coordinate with small weight can keep acquisition costs bounded as its arrivals become rare. To make this distinction explicit, use three actions with all rewards missing, $p = (1,1,1)$. Let b set the main probability transfer, η the small weight moved to the third action, and κ its behavior probability. Choose

$$d = (-b, b-\eta, \eta), \qquad \rho = ((1-\kappa)/2, (1-\kappa)/2, \kappa), \qquad 0 < b < 1/2. \tag{86}$$

For $0 < \eta < \min\{b, \gamma/4\}$ and sufficiently small $\kappa > 0$, $\rho + d$ is a valid proposed policy. Auditing the first two actions and assigning midpoint $1/2$ to the third incurs deterministic error at most $\eta/2$. The interaction cost for the retained coordinates stays bounded as $\kappa \downarrow 0$. A positive quota on the third action requires waiting for its arrivals, and the unshrunken quadratic bound contains the diverging term $\eta^2/(T\kappa)$. The bounded-mean formulation accounts for this difference. The local quadratic corollary applies when each relevant weight is large enough relative to the decision margin.

B.5 Estimating feedback rates and learning a near-optimal action

We now ask how many observations are needed to choose an action whose mean reward is within ϵ of the best, with failure probability at most δ. This is the probably approximately correct (PAC) learning goal. Before giving the learning procedure, we bound the error introduced by estimating the ordinary feedback rates. For an evaluation with fixed audit quotas, estimate natural feedback from a separate sample. Suppose $|\widehat{s}_a - s_a| \le e_a$ and $|\widehat{p}_a - p_a| \le e_a$ hold simultaneously. Let $\widehat{h}_a$ be the mean of n_a independent audited missing rewards, and let δ_A be their error allowance. For $\widehat{D} = \sum_a d_a(\widehat{s}_a + \widehat{p}_a\widehat{h}_a)$, separately bound preliminary-estimation error and audit-sampling error. Adding the two bounds gives the valid radius

$$\underbrace{2\sum_a |d_a| e_a}_{\text{natural-feedback calibration}} + \underbrace{\sqrt{\frac{1}{2}\log\frac{2}{\delta_A}\sum_a \frac{d_a^2\widehat{p}_a^2}{n_a}}}_{\text{independent audit uncertainty}}. \tag{87}$$

Theorem 8 (Learning with calibrated feedback)**.** *For $\epsilon \in (0,1)$ and $\delta \in (0,1/2)$, there is a finite-action learner that, with probability at least $1-\delta$, returns an ϵ-optimal action and uses*

$$O\left(\frac{\log(K/\delta)}{\epsilon^2}\sum_{a:p_a>3\epsilon/8} p_a^2\right) \text{ paid audits.} \tag{88}$$

The initial calibration uses $O(K\epsilon^{-2}\log(K/\delta))$ natural interactions. On the same success event, acquisition uses $O(\epsilon^{-2}\log(K/\delta)\sum_{a:p_a>3\epsilon/8} p_a)$ interactions. For common p, $K \geq 2$, $\epsilon \leq p/16$, and $\delta \leq 1/4$, the corresponding lower bound is $\Omega(Kp^2\epsilon^{-2}\log(1/\delta))$ expected audits.

The construction estimates (s,p) without audits, covers sufficiently small missing masses by their identification intervals, and estimates the remaining hidden means at precision proportional to $\epsilon/\overline{p}_a$. It then selects the largest reconstructed reward mean. The construction below specifies every sample count and finite acquisition horizon. The lower bound uses competing hidden Bernoulli means and indistinguishable natural laws. This theorem measures the price of restoring learnability when natural interactions and paid labels have different costs.

B.5.1 Proof of the calibrated contrast radius

Condition on a calibration data set satisfying the simultaneous bounds in Section 5. Conditional audit quotas may be selected from this data. Let $\bar{h}_a$ be the audit mean (the same estimator denoted $\widehat{h}_a$ above), with $0 \leq \bar{h}_a \leq 1$. Separate error in the observed contribution, the missing probability, and the missing mean:

$$\widehat{D} - D = \sum_a d_a(\widehat{s}_a - s_a) + \sum_a d_a(\widehat{p}_a - p_a)h_a + \sum_a d_a\widehat{p}_a(\bar{h}_a - h_a). \tag{89}$$

The first two terms have combined absolute value at most $2\sum_a |d_a|e_a$. Conditional on the calibration sample, the last term is a weighted sum of independent bounded audit outcomes. Its sum of squared range lengths is $\sum_a d_a^2\widehat{p}_a^2/n_a$. Hoeffding gives the second term of (87) with conditional failure probability at most δ_A. Adding the calibration failure probability gives the total error guarantee. The policy difference can be chosen from earlier training data or from a simultaneous natural-calibration event, but it is frozen before these validation audits.

B.5.2 An explicit finite-action acquisition algorithm

The following learner proves Theorem 8 with explicit audit counts and interaction horizons.

1. First estimate every action's observed contribution and missing probability accurately enough for the final ϵ target. Set the per-estimate tolerance $e = \epsilon/16$ and the preliminary count per action to
$$n_0 = \left\lceil 128\epsilon^{-2}\log(12K/\delta)\right\rceil. \tag{90}$$
For each action, collect n_0 unaudited interactions. Let $\widehat{s}_a$ be the mean of GR and $\widehat{p}_a$ the mean of $1-G$. Set $\overline{p}_a = \min(1, \widehat{p}_a + e)$ and $\underline{p}_a = \max(0, \widehat{p}_a - e)$.
2. If $\overline{p}_a \leq \epsilon/2$, assign $\widehat{h}_a = 1/2$ and purchase no label for that action.
3. For each remaining action, n_a is the number of missing rewards to inspect and T_a is the maximum number of interactions spent obtaining them. Use the upper missingness bound to set estimation precision and the lower bound to allow enough time for acquisition:
$$n_a = \left\lceil 8\overline{p}_a^2\epsilon^{-2}\log(6K/\delta)\right\rceil, \quad T_a = \left\lceil \frac{\max\{2n_a, 8\log(3K/\delta)\}}{\underline{p}_a}\right\rceil. \tag{91}$$
Select that action for at most T_a interactions and audit its first n_a closed outcomes. If its quota is obtained, let $\widehat{h}_a$ be the hidden sample mean. If any quota is unavailable, record acquisition failure and return an arbitrary action.
4. When all requested samples have been obtained, choose the action with the largest reconstructed full reward mean:
$$\widehat{a} \in \arg\max_a \left(\widehat{s}_a + \widehat{p}_a\widehat{h}_a\right). \tag{92}$$

All horizons are finite and computable. Indeed, an audited action has $\widehat{p}_a + e > \epsilon/2$, so $\underline{p}_a = \widehat{p}_a - e > 3\epsilon/8$. This argument does not require the calibration event.

B.5.3 Correctness and explicit resource bounds

First control the two preliminary estimates for each of K actions. Their error tolerance is $e = \epsilon/16$; the chosen n_0 makes Hoeffding's bound small enough to sum over all $2K$ estimates:

$$\mathbb{P}(|\widehat{s}_a - s_a| > e) \le 2e^{-2n_0 e^2} \le \frac{\delta}{6K}, \qquad \mathbb{P}(|\widehat{p}_a - p_a| > e) \le \frac{\delta}{6K}. \tag{93}$$

A union bound gives a calibration event E_C of probability at least $1 - \delta/3$. On this event, $\underline{p}_a \le p_a \le \overline{p}_a$.

Next require the missing-mean error to be at most $\epsilon/(4\overline{p}_a)$, so that its contribution to full-reward error is at most $\epsilon/4$. Each prescribed audit count gives

$$\mathbb{P}\left(|\widehat{h}_a - h_a| > \frac{\epsilon}{4\overline{p}_a} \,\middle|\, \text{calibration}\right) \le 2\exp\left(-\frac{n_a \epsilon^2}{8\overline{p}_a^2}\right) \le \frac{\delta}{3K}. \tag{94}$$

The simultaneous hidden-mean event has failure probability at most $\delta/3$. On E_C, the expected number of closed outcomes during the allowed horizon is $T_a p_a \ge T_a \underline{p}_a \ge \max\{2n_a, 8\log(3K/\delta)\}$. The Chernoff argument from Theorem 7 gives total quota failure probability at most $\delta/3$.

On all three good events, the reward estimate uses $\widehat{p}_a$, while $\overline{p}_a$ sets the audit precision. Thus $|\widehat{p}_a - p_a| \le e$ and $p_a \le \overline{p}_a$. For an audited action,

$$|\widehat{s}_a + \widehat{p}_a\widehat{h}_a - \mu_a| \le |\widehat{s}_a - s_a| + |\widehat{p}_a - p_a|\,|\widehat{h}_a| + p_a|\widehat{h}_a - h_a| \tag{95}$$

$$\le 2e + \epsilon/4 = 3\epsilon/8. \tag{96}$$

For a skipped action, $p_a \le \overline{p}_a \le \epsilon/2$ and $|h_a - 1/2| \le 1/2$, giving error at most $e + e/2 + p_a/2 \le 11\epsilon/32 < 3\epsilon/8$. To translate accurate reward estimates into a good action choice, define $\widehat{\mu}_a = \widehat{s}_a + \widehat{p}_a\widehat{h}_a$. Let a^* maximize the true means and $\widehat{a}$ maximize these estimates. Comparing their two estimation errors gives

$$\mu_{a^*} - \mu_{\widehat{a}} \le (\mu_{a^*} - \widehat{\mu}_{a^*}) + (\widehat{\mu}_{\widehat{a}} - \mu_{\widehat{a}}) \le 3\epsilon/4 < \epsilon. \tag{97}$$

The combined success probability is at least $1 - \delta$.

On E_C, an audited action satisfies $p_a > 3\epsilon/8$, because $\overline{p}_a \le p_a + 2e = p_a + \epsilon/8$. For such an action,

$$\overline{p}_a \le p_a + \epsilon/8 \le 4p_a/3, \qquad \underline{p}_a \ge p_a - \epsilon/8 \ge 2p_a/3. \tag{98}$$

Consequently its audit count is $O(p_a^2\epsilon^{-2}\log(K/\delta))$, and its horizon is $O(p_a\epsilon^{-2}\log(K/\delta))$. Rounding contributes a constant per audited action and is absorbed because $p_a/\epsilon > 3/8$. Summing over the audited subset proves the two costs in Theorem 8. The Kn_0 initial natural interactions are charged separately. When all missing probabilities are small enough, the procedure covers their effect with the identification bound and uses no audits on the calibration event.

B.5.4 The common-mass learning lower bound

Fix $K \ge 2$, a common missing probability p, and $\epsilon \le p/16$. Every action has deterministic open reward $1/2$. In a reference instance, hidden means are

$$h_1 = 1/2 + 2\epsilon/p, \qquad h_a = 1/2 \quad (a \ge 2). \tag{99}$$

Action 1 has true mean $1/2 + 2\epsilon$ and every other action has mean $1/2$. It is the unique ϵ-optimal action. For every $a \ge 2$, construct an alternative that changes only h_a to $1/2 + 4\epsilon/p$. Then action a is uniquely ϵ-optimal, with mean $1/2 + 4\epsilon$. The full natural law is unchanged in every instance.

For a uniformly (ϵ, δ)-correct learner, the event of selecting action 1 has reference probability at least $1 - \delta$ and alternative probability at most δ. Only audits of action a distinguish that alternative. Hence

$$\mathbb{E}_0[N_a]\,\mathrm{kl}(1/2, 1/2 + 4\epsilon/p) \ge \mathrm{kl}(1 - \delta, \delta). \tag{100}$$

The alternative mean is at most $3/4$. Applying $\mathrm{kl}(x, y) \le (x-y)^2/[y(1-y)]$ gives

$$\mathrm{kl}(1/2, 1/2 + 4\epsilon/p) \le \frac{256\epsilon^2}{3p^2}. \tag{101}$$

Sum over $a = 2, \dots, K$ to obtain

$$\mathbb{E}_0[N] \ge \frac{3(K-1)p^2}{256\epsilon^2}\,\mathrm{kl}(1-\delta, \delta). \tag{102}$$

This is $\Omega(Kp^2\epsilon^{-2}\log(1/\delta))$ in the usual small-confidence regime, for example $\delta \le 1/4$. For all $\delta < 1/2$, the displayed binary-KL expression is the precise bound. The construction explicitly gives one alternative for each competing action, so its factor K follows from distinct information requirements.

B.6 Linear correction, variance, and adaptive histories

Let $\mathcal{F}_{t-1}$ contain the observed history. Before a new action is drawn, the policy and baseline are fixed functions of this history. After the action and gate exist, the audit probability may depend on permitted pre-audit information, but its random coin is independent of the completed reward conditional on that information. Positivity is needed wherever a closed reward has a nonzero coefficient in the update.

To correct the unequal chance of observing each reward, give observed rewards weight one and inspected missing rewards weight $1/q$. The combined weight is $W = G + (1-G)Q/q$. For a completed action, visibility indicator and reward (A, G, R), independence of the audit coin makes its conditional mean exactly one:

$$\mathbb{E}[W \mid \mathcal{F}_{t-1}, A, G, R] = G + (1-G) = 1. \tag{103}$$

The score $S(A) = \nabla \log \pi(A)$ is the derivative of the action's log selection probability. With scalar baseline b chosen from past observations, iterated expectation shows that this weighting recovers the complete-feedback mean gradient:

$$\begin{aligned}
\mathbb{E}[W(R-b)S(A) \mid \mathcal{F}_{t-1}] &= \mathbb{E}[(R-b)S(A) \mid \mathcal{F}_{t-1}] && (104)\\
&= \sum_a \nabla\pi(a)\mu_a - b\nabla\sum_a \pi(a) && (105)\\
&= \nabla J(\theta). && (106)
\end{aligned}$$

This argument allows arbitrary reward dependence in the natural gate. With a predictable action-dependent baseline b_a, add the exactly computable correction $\sum_a \pi(a) b_a S(a)$. Baseline values are control variates held fixed in this calculation.

To quantify the extra variance caused by missing rewards, compare two estimators of D. Draw A from behavior distribution ρ, and fix action-specific baselines b_a before that draw. Let F use the full reward and X use the independently audited reward:

$$F = d^\top b + \frac{d_A}{\rho_A}(R - b_A), \qquad X = d^\top b + \frac{d_A}{\rho_A} W(R - b_A). \tag{107}$$

Both have expectation D. Subtracting them isolates the randomness introduced by checking only some missing rewards:

$$X - F = \frac{d_A}{\rho_A}(1-G)(Q/q_A - 1)(R - b_A). \tag{108}$$

This difference has conditional mean zero given (A, G, R), and therefore zero covariance with F. As $\mathrm{Var}(Q/q_A) = 1/q_A - 1$,

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{Var}(F) + \mathbb{E}[(X-F)^2] && (109)\\
&= \mathrm{Var}(F) + \mathbb{E}\left[(1-G)\frac{d_A^2}{\rho_A^2}(R-b_A)^2\left(\frac{1}{q_A} - 1\right)\right]. && (110)
\end{aligned}$$

This is the active-inference residual variance identity (Zrnic & Candès, 2024) with naturally revealed outcomes retained. At $b_a = 1/2$, its additional variance is bounded by

$$\frac{1}{4}\sum_a \frac{p_a d_a^2}{\rho_a}(1/q_a - 1), \tag{111}$$

with equality for hidden rewards on $\{0, 1\}$. Minimizing at audit rate $B = \sum_a \rho_a p_a q_a$ gives the quota design's stationarity condition. In an unsaturated regime, $q_a = c|d_a|/\rho_a$ gives

$$B = cH, \qquad \frac{1}{4}\sum_a \frac{p_a d_a^2}{\rho_a}\left(\frac{1}{q_a} - 1\right) = \frac{H^2}{4B} - \frac{1}{4}\sum_a \frac{p_a d_a^2}{\rho_a}. \tag{112}$$

The complete-feedback variance $\mathrm{Var}(F)$ remains in (110). For general vector scores, minimizing a reward-uniform upper bound on the trace of additional covariance gives audit probabilities proportional to the score norm, with the same saturation operation. The budget experiments fix the proposal audit probability at 0.04 and optimize the separate validation design.

For validation episode j, condition on the observed history and newly frozen policy difference. With known natural feedback, the fresh validation data give conditional failure probability at most δ_j. Taking expectations and a union bound gives total failure probability at most $\sum_j \delta_j$. With finite calibration, this argument applies on the simultaneous calibration event, whose failure probability δ_C is added once. Choosing $\delta_C + \sum_j \delta_j \le \delta$ therefore makes every resolved decision correct with probability at least $1 - \delta$, including adaptively proposed candidates. Evaluating several candidates on a shared validation sample requires a simultaneous confidence construction (Dann et al., 2023; Howard et al., 2021).

C Experimental setup

C.1 Populations and observable feedback

To separate reward quality from evidence presentation, the controlled environment crosses six reward strategies with three visibility rules. The simulator `EvidenceEnvironment` generates one fixed instance using seed 711. We use context 0 throughout the budget study, so there are 18 actions and one stable reward distribution per action. The instance is generated once, independently of every learning seed. The supplementary environment manifest lists all 18 reward means and reveal rates; the full simulator implementation is available on request.

For the model-output study, the generation program loads the unmodified public Qwen2.5-0.5B-Instruct checkpoint and all 1,319 questions in the `main/test` split of GSM8K. It uses the checkpoint's chat template, a common helpful-assistant system message, greedy decoding, a 512-token generation limit, bfloat16 arithmetic, and batches of 128. The direct prompt requests only the final numerical answer. The reasoning prompt requests a step-by-step solution. The check prompt additionally requests an arithmetic check. Every prompt asks for a final answer following `####`. Supplementary metadata gives repository revisions, library versions, and token counts. The system message is "You are a helpful assistant." The user message contains the question, two newline characters, and one of these instructions:

- Direct: "Give only the final numerical answer. Write it as #### followed by the number."
- Reason: "Solve the problem step by step. End with #### followed by the final numerical answer."
- Check: "Solve the problem step by step, then check the arithmetic. End with #### followed by the final numerical answer."

The generated responses, extracted answers and full implementation may be requested from the authors.

The parser takes the last numeric expression after the final answer marker, or the last numeric expression in the response when no marker is present. It normalizes commas, decimal notation, scientific notation, and fractions using decimal arithmetic. Reward is one when the normalized answer equals

the normalized reference answer and zero otherwise, including unparseable outputs. Generation-limit outcomes are retained. This is an automatic numerical-answer metric rather than a judgment of the quality of the written reasoning.

The three model-output actions have success reveal probability 0.9 and failure reveal probabilities $(0.02, 0.40, 0.20)$ for direct, reasoning, and check prompts. These rates define the feedback intervention; the evaluator is not given them. An outcome is selected before feedback and auditing. For proposals, the code samples actual stored output indices. For calibration and validation, it draws the exact multinomial and binomial sufficient statistics of independent samples with replacement from the same empirical population. The environment's complete labels are used to draw feedback and report population quality; hidden unaudited labels do not enter the update or decision rule. The study measures randomness in feedback and policy learning conditional on the fixed questions and outputs.

C.2 Proposals and validation

For seed $s \in \{200, \dots, 229\}$, proposal, validation, and calibration streams use seeds $723001 + s$, $924001 + s$, and $163001 + s$. All methods start with zero logits. Each of 40 proposal steps samples a batch of 128 actions from its current softmax policy and uses the mean score gradient with baseline 0.75 and logit step size 8. Limited-audit proposals use independent audit probability 0.04 and inverse-probability correction. The complete reference audits every missing proposal outcome; the visible reference audits none and uses the visible score signal. Both accept all proposals. All costs incurred by these references are counted.

For decisions every five steps, five provisional updates form one candidate. If the resulting finite change is rejected or unresolved, all five provisional steps are discarded. The intermediate policies are used only for proposal generation. The retained policies are the endpoints of accepted groups; no claim is made that every provisional step within a group improves reward.

Each interval-based run first collects n_0 natural observations per action. With $\delta_C = 0.01$, it uses $e = \sqrt{\log(4K/\delta_C)/(2n_0)}$ and the upper missing probabilities in Section 5. The remaining failure probability 0.04 is divided equally across 40 or eight evaluations. At cadence k, a validation round has $T = kT_0$ interactions and hard audit cap $B = kB_0$. To reduce the chance that random audits exceed the hard limit B, reserve a margin before allocating them. With $L = \log(4/\delta_j)$, use expected design budget

$$N' = \max\{0, B - \sqrt{2BL} - 2L/3\}. \tag{113}$$

The capped method applies `global_design` with this budget and the calibrated arrival caps. The unshrunken ablation uses the same allocation and sets retained weights to their full values. The uniform comparator assigns a common audit probability to relevant missing outcomes and optimizes shrinkage for that fixed allocation. Each method uses its own bias-aware Bernstein radius plus $1.5\|d\|_1 e$.

If the hard cap prevents completion, the decision is unresolved and no over-budget estimate is returned. Positive lower endpoints accept a candidate; negative upper endpoints reject it. Every other case is unresolved. Rejected and unresolved candidates leave the retained policy unchanged. The complete fixed validation horizon is included in the cost even for a zero-audit decision. Table 2 gives all prespecified settings; they form a joint resource sweep rather than an isolated causal comparison of calibration and validation.

Acceptance-only evaluation. The two-sided intervals report both positive and negative decisions. When the objective is acceptance alone, the upper tail of the estimation error determines false acceptance. Keeping the same allocation, shrinkage, calibration correction and hard-cap reserve, replace L in the Bernstein radius by $L_+ = \log(2/\delta_j)$. On the calibration event, the one-sided estimator failure probability is at most $e^{-L_+} = \delta_j/2$; the existing audit-count reserve contributes at most $\delta_j/4$. Accept when this lower bound is positive and otherwise retain the current policy. At the reported $\delta_j = 0.001$ and 0.005, the square-root sampling term decreases by 4.3% and 5.3%, respectively; the bias and calibration terms are unchanged. These are analytical radius comparisons at fixed designs. The reported trajectories use two-sided intervals, and end-to-end cost savings depend on the resulting acceptance sequence.

Table 2: Budget settings. n_0 is calibration observations per action; T_0 and B_0 are validation interactions and the audit cap per proposal step. For five-step decisions, both validation limits are multiplied by five.

Setting	n_0	T_0	B_0
1	128	1,000	64
2	512	4,000	256
3	2,048	16,000	1,024
4	8,192	64,000	4,096
5	32,768	100,000	4,096
6	65,536	100,000	4,096

C.3 Accounting and uncertainty

Each population has 1,260 trajectories: three evaluation methods, six settings, two cadences, and 30 seeds, plus three proposal references at both cadences. A trajectory has 40 single-step or eight five-step candidate decisions. Across both populations, there are 51,840 interval evaluations and 8,640 reference proposals. Repeated evaluations within a trajectory are not independent statistical replicates.

Totals count calibration once, all proposal batches and validation horizons, and all paid audits. The repeated calibration field in raw logs must not be summed across rounds. Quality is evaluated on the fixed population. Decision rate includes positive and negative decisions, with incomplete evaluations unresolved; it is undefined for proposal-only references. Student intervals use 30 seed-level values, and cadence comparisons use paired seed differences.

D Additional budget results

D.1 Where evidence is spent and what grouping changes

Figure 5 decomposes interactions and paid audits at the operating points in Table 3. Validation dominates the interaction count for the interval methods; complete proposal feedback spends only on proposals. Audits are shown separately because they carry their own price. The bars do not combine replay requests with one-time model inference.

Figure 6 compares the two evaluation cadences at every budget setting. Each point is a paired difference within seed, with all other run settings matched. Grouping changes both the candidate direction and the number of confidence allocations, so these effects are interpreted jointly. In model-output replay at setting 3, more resolved decisions coexist with lower retained accuracy.

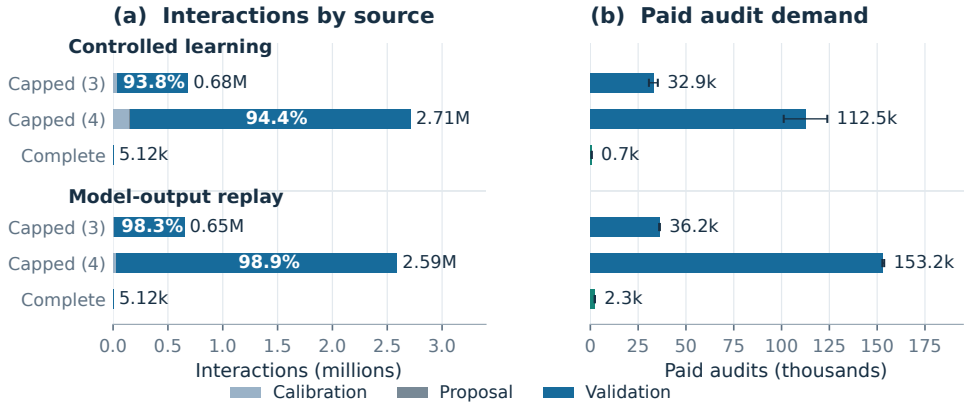

Figure 5: Evidence cost has two components. Left: calibration, proposal, and validation interactions; internal percentages give the validation share. Right: paid audits, with 95% Student intervals across 30 seeds. Capped rows evaluate every five steps at settings 3 and 4. The complete-feedback reference purchases all missing proposal labels and performs no validation.

Table 3: Mean counts per seed, in thousands. Calibration and validation count interactions; audits include proposal and validation labels. Capped rows evaluate every five steps, with the setting in parentheses. Complete is the proposal reference; quality uses the corresponding fixed population.

Method	Calibration	Validation	Audits	Quality
Controlled learning				
Capped (3)	36.9	640.0	32.9	0.647
Capped (4)	147.5	2560.0	112.5	0.736
Complete	0.0	0.0	0.7	0.758
Model-output replay				
Capped (3)	6.1	640.0	36.2	0.404
Capped (4)	24.6	2560.0	153.2	0.425
Complete	0.0	0.0	2.3	0.426

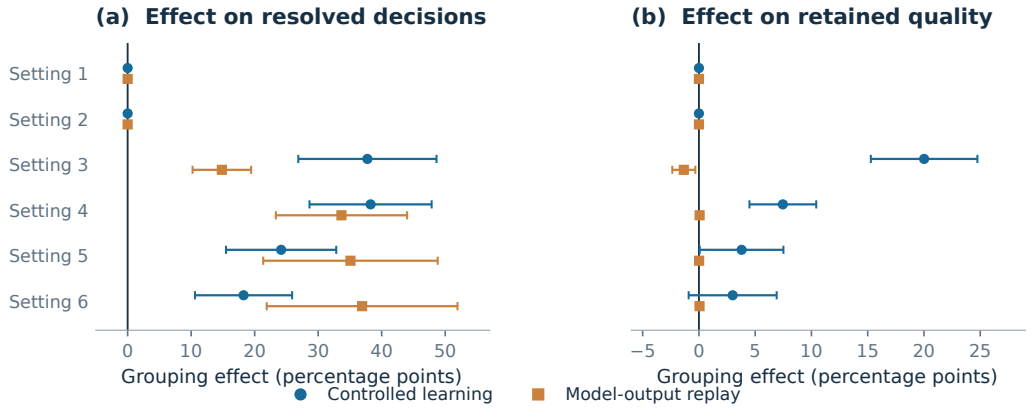

Figure 6: More resolved decisions need not yield better final policies. Paired differences between evaluation every five steps and every step for capped allocation, across all six budget settings. Error bars are 95% Student intervals over 30 seed differences. Quality differences are expressed in percentage points of bounded expected reward or output-pool accuracy. The vertical line denotes no change.

D.2 COMPLETE OPERATING-POINT RECORDS

Table 4 reports selected operating points and the proposal references. All six settings appear in the figures; the summary CSV also includes the unshrunken ablation and every cadence. Five-step controlled-task quality is 0.713 at setting 6 versus 0.736 at setting 4. Each budget produces its own sequence of accepted groups and subsequent proposals, so within-run improvement is compatible with differing final quality across budgets.

Table 4: Measured operating points. Cadence is proposal steps per decision. Counts are means per seed except nonpositive accepted changes, which are totals across 30 seeds. I denotes all interactions; A includes proposal and validation audits. References without intervals have no decision rate.

Method	Setting/cadence	I	A	Resolved	Quality	Nonpos.
Controlled learning						
Capped	6/1	5,184,768	132,910	73.0%	0.683	0
Capped	3/5	681,984	32,950	40.4%	0.647	0
Capped	4/5	2,712,576	112,528	85.4%	0.736	0
Capped	6/5	5,184,768	121,785	91.2%	0.713	0
Uniform	4/5	2,712,576	111,257	82.5%	0.736	0
Accept every	–/1	5,120	47	–	0.751	472
Complete	–/1	5,120	679	–	0.758	157
Visible only	–/1	5,120	0	–	0.559	237
Model-output replay						
Capped	6/1	4,201,728	151,079	18.5%	0.426	0
Capped	3/5	651,264	36,175	20.8%	0.404	0
Capped	4/5	2,589,696	153,202	42.1%	0.425	0
Capped	6/5	4,201,728	157,419	55.4%	0.426	0
Uniform	4/5	2,589,696	153,202	36.7%	0.424	0
Accept every	–/1	5,120	100	–	0.411	880
Complete	–/1	5,120	2,329	–	0.426	394
Visible only	–/1	5,120	0	–	0.216	843

In controlled learning, the interval-based methods return 0 nonpositive accepts, 0 incorrect negative decisions, and 0 intervals missing the true change across 25,920 evaluations. There are 17,331 unresolved decisions, including 0 incomplete validations at the hard audit cap. In model-output replay, the interval-based methods return 0 nonpositive accepts, 0 incorrect negative decisions, and 0 intervals missing the true change across 25,920 evaluations. There are 23,205 unresolved decisions, including 0 incomplete validations at the hard audit cap. The stored model corpus includes 197 generation-limit outputs and 0 unparseable outputs, all retained in evaluation.



Figure 7: Audit-price sensitivity for the controlled learning environment. Columns use interaction-equivalent audit prices 1, 10, and 100. The data, calibration costs, and proposal horizon are unchanged. Error bars are 95% intervals over 30 seeds.



Figure 8: Audit-price sensitivity for the fixed model-output population. The horizontal coordinate counts replay requests and paid label requests. The one-time model-generation cost is common to these analyses and is reported in the main text.

D.3 SENSITIVITY TO SCARCE ARRIVALS

We isolate validation behavior while preserving the two reward populations, prompting strategies, feedback mechanisms, and proposal rule. For each of the same 30 seeds, we freeze the first five-step candidate from the limited-audit proposal reference, starting at the uniform policy. These 60 candidates are taken directly from the recorded budget runs. We evaluate every candidate for each possible scarce action b, using

$$\rho_b = \eta/K, \qquad \rho_a = (1-\eta/K)/(K-1) \quad (a \neq b), \qquad \eta \in \{1, 10^{-2}, 10^{-4}\}. \tag{114}$$

Thus every action is made rare in turn, independently of its reward or proposal weight. The reward law stays fixed, and only the validation behavior changes. Each evaluation uses the fourth setting's five-step resources: $n_0 = 8{,}192$ calibration observations per action, $T = 320{,}000$ validation interactions, and hard audit limit $B = 20{,}480$. Calibration and episode error allocations are 0.01 and 0.005. These are individual evaluation comparisons, not one adaptive sequence across the sensitivity grid.

The capped and unshrunken methods use the same allocation. The latter sets all active retained fractions to one. The uniform method uses a common audit probability and optimizes shrinkage. A fourth allocation ablation first assigns $\widetilde{n}_a = N'|d_a|\overline{p}_a / \sum_i |d_i|\overline{p}_i$ ignoring caps, then clips to $n_a = \min\{\widetilde{n}_a, T\rho_a\overline{p}_a\}$ and optimizes shrinkage. It leaves any clipped budget unallocated. This defines a feasible comparator with the same calibrated Bernstein construction. It is not an implementation of Zrnic & Candès (2024) or Li et al. (2025), whose acquisition probabilities already respect unit bounds.

Table 5 reports all three arrival settings and all four methods. At $\eta = 10^{-4}$ in the controlled population, shrinkage reduces the mean radius from 0.137 to 0.040 and raises the resolved fraction from 6.9% to 30.9%. In the model-output population, the corresponding radii are 0.311 and 0.135, and resolved fractions are 1.1% and 20.0%. At $\eta = 1$, the capped and unshrunken methods coincide. All 7,560 evaluations completed within the hard audit limit; none produced a wrong sign or an interval missing its target. These finite observations complement the coverage argument rather than strengthen its nominal confidence level.

At $\eta = 10^{-4}$, clipping leaves about 1,107 audits of the mean design allowance unallocated in the controlled population and 6,645 in the model-output population. These are design budgets computed with $\overline{p}$, not realized paid counts. Reallocation tightens the radius slightly, but its observed decision advantage is small: the resolved fractions match in the controlled population and differ by 1.1 percentage points in model-output replay. Uniform acquisition with shrinkage also approaches the capped method at the smallest arrival rate. The observed gain is concentrated in shrinkage under scarce arrivals; the allocation refinements have smaller effects on decisions.

The sensitivity program reads the original proposal records and uses calibration seed $163001 + s$. Validation uses NumPy's seed sequence $[493001, s, j, b]$, where j indexes the three arrival settings. All methods share the seed for a given candidate and behavior. Counts are exact multinomial/binomial sufficient statistics, as in the budget study. The 18 or three action rotations share each seed's candidate and calibration draw, so confidence intervals first average over rotations and then use 30 seed-level means. The supplement supplies these seed means; rotation-level records and the experiment program may be requested from the authors. This experiment measures frozen-candidate evaluation, not retained quality over further learning or generalization beyond the fixed output population.

Table 5: Arrival sensitivity of frozen proposals. Radius and resolved fraction (%) are means ± 95% Student half-widths over 30 seed averages, each averaging all scarce-action rotations. All rows use the same calibration and validation allowances.

η	Method	Radius	Resolved (%)
Controlled learning population ($K = 18$)			
1	Capped + shrinkage	0.026 ± 0.002	53.1 ± 18.3
1	Capped, unshrunken	0.026 ± 0.002	53.1 ± 18.3
1	Clipped uncapped	0.026 ± 0.002	53.1 ± 18.3
1	Uniform + shrinkage	0.027 ± 0.002	52.0 ± 18.4
10^{-2}	Capped + shrinkage	0.028 ± 0.002	50.7 ± 18.3
10^{-2}	Capped, unshrunken	0.028 ± 0.002	50.7 ± 18.3
10^{-2}	Clipped uncapped	0.028 ± 0.002	50.6 ± 18.1
10^{-2}	Uniform + shrinkage	0.040 ± 0.003	33.7 ± 13.8
10^{-4}	Capped + shrinkage	0.040 ± 0.003	30.9 ± 13.0
10^{-4}	Capped, unshrunken	0.137 ± 0.009	6.9 ± 4.1
10^{-4}	Clipped uncapped	0.040 ± 0.003	30.9 ± 13.0
10^{-4}	Uniform + shrinkage	0.042 ± 0.003	28.7 ± 12.5
Fixed model-output population ($K = 3$)			
1	Capped + shrinkage	0.039 ± 0.004	95.6 ± 7.1
1	Capped, unshrunken	0.039 ± 0.004	95.6 ± 7.1
1	Clipped uncapped	0.039 ± 0.004	95.6 ± 7.1
1	Uniform + shrinkage	0.039 ± 0.004	95.6 ± 7.1
10^{-2}	Capped + shrinkage	0.048 ± 0.005	85.6 ± 10.2
10^{-2}	Capped, unshrunken	0.048 ± 0.005	85.6 ± 10.2
10^{-2}	Clipped uncapped	0.048 ± 0.005	85.6 ± 10.2
10^{-2}	Uniform + shrinkage	0.088 ± 0.009	20.0 ± 10.6
10^{-4}	Capped + shrinkage	0.135 ± 0.013	20.0 ± 10.1
10^{-4}	Capped, unshrunken	0.311 ± 0.033	1.1 ± 2.3
10^{-4}	Clipped uncapped	0.136 ± 0.014	18.9 ± 10.2
10^{-4}	Uniform + shrinkage	0.135 ± 0.013	20.0 ± 10.1