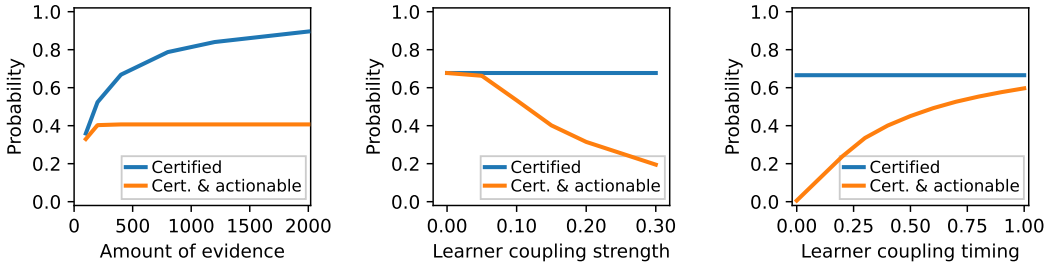
MORE EVIDENCE MAKES ADVICE EASIER TO CERTIFY AND HARDER TO USE

Anonymous authors
Paper under double-blind review

ABSTRACT

The same observations can support a reliability certificate for external advice and update the fallback learner. As evidence accumulates, certification can improve while the advice's learner-relative value shrinks. We call this remaining value actionability, which determines whether the advice is still worth using relative to the fallback. One might expect advice to remain actionable once enough evidence supports certification. We establish that, as confidence tightens, early certification disappears, so nonvanishing success must reach a learner state that preserves enough actionability for certification to finish. The best achievable probability of reaching such a state sets a high-confidence ceiling on certified-and-actionable success. Reachability depends on the advice's remaining advantage over the fallback, how strongly new evidence pushes the fallback toward the advice, and whether a sufficiently favorable random evidence excursion actually occurs. More informative evidence can still lower success by shrinking this advantage.



(a) Certification–actionability divergence

(b) Coupling strength erosion

(c) Coupling timing erosion

Figure 1: Learner feedback can erode true actionability even as certification improves. The orange curve reports the probability that advice is both certified and truly actionable relative to the current fallback. (a) A larger evidence horizon raises certification from 0.359 to 0.896, while certified-and-actionable probability plateaus near 0.407. (b) With the evidence stream fixed, stronger learner updating leaves certification at 0.677 but lowers the joint probability from 0.677 to 0.195. (c) With the same evidence transcript and exactly 120 learner updates, moving the same updates later raises the joint probability from 0.006 to 0.597 while certification remains 0.666.

1 INTRODUCTION

External advice is valuable while a fallback learner still lacks decision-relevant information. This advantage can matter when learner updates are scarce, as illustrated in Appendix A. Evidence collected to evaluate that advice can also train the fallback. The same observations can raise confidence in the advice while shrinking its incremental value. More evidence can make advice easier to certify and harder to use.

We study this conflict as a sequential decision problem. A protocol may deploy advice only after meeting a fixed level-α reliability requirement and while the advice still improves on the current fallback. Meeting the reliability requirement gives *certification*. The advice's remaining gain over the fallback is its *actionability*. Keeping that gain available while certification accumulates is

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.

preservation. Figure 1 shows that stronger certification does not guarantee more usable advice. As the evidence horizon grows, certification continues to rise while certified-and-actionable success levels off. Holding the evidence transcript fixed makes the learner effect explicit. Stronger fallback updating leaves certification at 0.677 while reducing certified-and-actionable success from 0.677 to 0.195. Holding both the transcript and the number of learner updates fixed isolates timing further. Moving the same updates later raises joint success from 0.006 to 0.597, while certification remains 0.666 and final fallback regret changes little. Together, these results show that the same certification evidence can lead to very different chances of usable advice. This raises a central question:

What controls the probability that certification arrives before learning erases actionability?

We show that preservation reachability controls this probability at high confidence. As confidence tightens, early certification disappears, so nonvanishing success must reach a learner state that preserves enough actionability for certification to finish. The best achievable probability of reaching such a state therefore sets the high-confidence ceiling. This reachability depends on the advice's remaining advantage over the fallback, how strongly new evidence pushes the fallback toward the advice, and whether a sufficiently favorable evidence excursion occurs. In a two-query ridge-logistic learner, this ceiling equals the policy optimum, showing that the general preservation bound can be exact. The same model also reveals that more informative evidence can make preservation harder when acquiring that information also changes the downstream learner state. When the refinement preserves the downstream learner state and deployment actionability, the usual information monotonicity returns.

Paper organization. Section 3 defines the certified-and-actionable deployment objective when certification evidence can also update the fallback, including observable control of actionability at deployment. Section 4 develops the preservation-reachability framework through a three-route decomposition, policy-uniform learner-side bounds, high- and finite-confidence success bounds, and a contextual-logistic realization. Section 5 derives an exact ridge-logistic first-passage phase, shows how more informative semantic evidence can lower the optimum by increasing the required preservation depth, and recovers information monotonicity when downstream learner state and actionability are preserved. Section 6 uses fixed-transcript controlled comparisons to isolate the learner-state mechanism: the Gaussian benchmark varies update strength and timing along a common evidence path, while the digits experiment tests the same separation in multiclass classification. Section 7 discusses the implications for certification, evidence reuse, and deployment.

2 Closest Related Work

Sequential certification, controlled sensing, and risk control. Sequential testing determines when evidence suffices under explicit error constraints, while controlled sensing additionally chooses which observations to collect (Chernoff, 1959; Nitinawarat et al., 2013; Deshmukh et al., 2021; Veeravalli et al., 2024). When stopping or continuation is data-dependent, confidence sequences, e-processes, and safe tests retain time-uniform validity (Howard et al., 2021; Ramdas et al., 2023; Grünwald et al., 2024). At the prediction level, risk-control methods extend finite-sample guarantees to prediction sets and calibrated predictive systems (Bates et al., 2021; Angelopoulos et al., 2024; 2025). When the opportunity or value of acting can itself change during information acquisition, stochastic-deadline models incorporate that timing directly (Frazier & Yu, 2007; Alaa & van der Schaar, 2016).

Deferral, mediation, and baseline-relative decision making. Learning to defer routes decisions between models and experts, with work on calibration, scarce expert predictions, dependent errors, and incomplete annotations (Mozannar & Sontag, 2020; Verma & Nalisnick, 2022; Mozannar et al., 2023; Hemmer et al., 2023; Wei et al., 2024; Nguyen et al., 2025). A sequential extension arises in online decision mediation, where expert feedback can resolve current decisions and train later ones (Jarrett et al., 2022). A complementary line asks what can be guaranteed from finite policy data. High-confidence policy evaluation estimates policy values, while safe policy improvement provides guarantees relative to a behavior or baseline policy (Dann et al., 2023; Wienhöft et al., 2023; Sharma et al., 2025). In online learning, conservative contextual bandits impose related baseline-relative constraints as data arrive (Takemura, 2023; Deb et al., 2025).

Feedback and adaptive data reuse. Performative prediction studies learning when deployed predictors alter later data distributions (Perdomo et al., 2020; Jagadeesan et al., 2022; Chen et al., 2024). Adaptive data analysis studies statistical validity when later queries depend on earlier outputs from reused data (Dwork et al., 2015; Marchant & Rubinstein, 2025). Both place the consequences of data use inside the sequential problem.

Controlled reachability and preservation. Stochastic reachability studies controlled probabilities of entering, avoiding, or remaining within prescribed sets (Vinod & Oishi, 2021). When the dynamics are unknown, data-driven formulations estimate these probabilities with finite-sample confidence bounds (Thorpe et al., 2022). For reach-avoid objectives, supermartingale certificates establish guarantees for learned controllers (Zikelić et al., 2023). In our setting, the relevant sets arise from learner states that preserve future advice actionability, and barriers are fixed independently of the risk level and realized policy.

Information ordering. The comparison of experiments introduced by Blackwell (1953) orders information structures by whether one is at least as valuable in every downstream decision problem. Recent work gives geometric characterizations of this ordering (Wu, 2023). When the chosen experiment also induces downstream responses, informative-test models in signaling environments show that information value can change (Weksler & Zik, 2022). Section 5.2 studies an analogous issue when evidence can change the fallback state. Under preservation of downstream learner state and deployment actionability, Section 5.3 recovers monotonicity.

3 Certified-and-actionable deployment

We formalize the decision problem when certification evidence can also update the fallback. Success requires certification and actionability to coexist, so the learner state enters the objective with the certificate path.

A protocol fixes a learner state space $\mathcal{X}$, an evidence-action set $\mathcal{Q}$, observation kernels, and a learner update map. It also fixes one beneficial world e_+, a harmful class $\mathcal{H}_-$, and a level-α certificate family. Let $\mathcal{F}_t$ denote the filtration generated by the realized protocol history through round t, including learner states, selected evidence actions, observations, and any revealed policy randomization. At round t, the learner is $U_{t-1} \in \mathcal{X}$. A predictable policy chooses $Q_t \in \mathcal{Q}$, observes $Y_t \sim P_e(\cdot \mid U_{t-1}, Q_t)$, and updates $U_t = F_t(U_{t-1}, Q_t, Y_t)$. Whether an observation trains the fallback is part of the declared action. A policy can hold out data only when the protocol offers that route.

Let $\Delta(U_t)$ be the true gain of advice over the current fallback on the deployment decision. For a fixed margin $m \geq 0$, define $A_t = \{\Delta(U_t) \geq m\}$. The event A_t records true actionability. It can be evaluated with oracle knowledge even when the policy cannot observe it. A nested level-α certificate is a stopping time τ_C^α satisfying

$$\sup_{e \in \mathcal{H}_-} P_e^\pi(\tau_C^\alpha < \infty) \leq \alpha \qquad \text{for every admissible policy } \pi. \tag{1}$$

Write $C_t^\alpha = \{\tau_C^\alpha \leq t\}$. The overlap $C_t^\alpha \cap A_t$ is the structural target shown by the orange curve in Figure 1. It measures whether certification and true actionability coexist along a trajectory. When A_t is hidden, deployment uses an observable stopped event.

A deployable success event must be observable at stopping. For a bounded stopping time $\tau \leq T$, require

$$S \in \mathcal{F}_\tau, \qquad S \subseteq C_\tau^\alpha \cap A_\tau. \tag{2}$$

The policy-optimal certified-and-actionable probability is

$$P_{\alpha,T}^\star = \sup_{\pi,\tau,S} P_+^\pi(S), \qquad P_\alpha^\star = \sup_T P_{\alpha,T}^\star. \tag{3}$$

The certificate family remains fixed inside this optimization. A protocol can implement actionability through an observable flag $\widehat{A}_t \in \mathcal{F}_t$. If $P_+^\pi(\exists t : \widehat{A}_t \setminus A_t) \leq \beta$ uniformly over admissible policies, any stopped rule supported on $C_\tau^\alpha \cap \widehat{A}_\tau$ has a controlled false-actionability budget. Appendix B.5 gives a generic confidence-set construction; Section 6 gives its Gaussian instance.

4 PRESERVATION REACHABILITY CONTROLS HIGH-CONFIDENCE SUCCESS

We now characterize the learner-side conditions that control equation 3 at high confidence. The analysis tracks states that preserve enough actionability for certification to finish.

4.1 PRESERVATION BARRIERS AND THE THREE-ROUTE DECOMPOSITION

Definition 1 (Preservation barrier)**.** Let L_t collect statistics determined by the learner-consumed transcript. A preservation barrier is a predeclared sequence of measurable sets B_t with $\tau_B^\pi = \inf\{t \geq 0 : L_t \in B_t\}$. The sets may depend on fixed protocol parameters and the advice instance, but not on α, future observations, the realized world, or the selected policy.

For any integer $n \geq 1$, a successful trajectory can follow only three routes. It can certify before time n, hit the preservation barrier before stopping, or remain actionable after time n without hitting the barrier. These routes yield the following decomposition. For a declared barrier B, define its policy-optimal reachability

$$Q(B) = \sup_{T,\pi} P_+^\pi(\tau_B^\pi \leq T). \tag{4}$$

The other two routes are

$$\varepsilon_\alpha(n) = \sup_\pi P_+^\pi(\tau_C^\alpha < n), \tag{5}$$

$$r_B(n) = \sup_{T,\pi,\tau} P_+^\pi(A_\tau, \tau < \tau_B^\pi, \tau \geq n). \tag{6}$$

A barrier is admissible when $r_B(n) \to 0$. Let $\mathfrak{B}$ denote the protocol-level class of admissible barriers.

Lemma 1 (Three routes)**.** *For every successful stopped event and every integer $n \geq 1$,*

$$S \subseteq \{\tau_C^\alpha < n\} \cup \{\tau_B^\pi \leq \tau\} \cup \{A_\tau, \tau < \tau_B^\pi, \tau \geq n\}. \tag{7}$$

Lemma 1 separates certification from preservation. Early success is controlled by the certificate. Barrier reachability and late actionability are controlled by learner evolution.

4.2 LEARNER-SIDE CONTROL OF PRESERVATION REACHABILITY

We next control the two learner-side routes uniformly over policies. First consider late actionability without a barrier hit. A scalar learner potential is enough when, before B, actionability implies

$$A_t \cap \{\tau_B^\pi > t\} \subseteq \left\{-C_0 + \sum_{s=1}^t Z_s - \gamma_0 t \leq 0\right\}, \tag{8}$$

where $C_0 \geq 0$. Assume, uniformly over admissible actions before the barrier, $\mathbb{E}_+^\pi[Z_t \mid \mathcal{F}_{t-1}] \geq \mu > \gamma_0$, and the centered increment is conditionally σ^2-sub-Gaussian. Standard maximal concentration gives constants $K_B < \infty$ and $\kappa_B > 0$ with

$$r_B(n) \leq K_B e^{-\kappa_B n}. \tag{9}$$

One valid rate is $\kappa_B = (\mu - \gamma_0)^2/(8\sigma^2)$.

The second learner term is barrier reachability itself. Let $G_t = \sum_{s=1}^t V_s$ be a score whose negative excursion is required to hit B. Suppose every barrier hit requires $G_{\tau_B^\pi} \leq -b_B$ and, before the hit,

$$\mathbb{E}_+^\pi[e^{-\eta_B V_t} \mid \mathcal{F}_{t-1}] \leq 1 \tag{10}$$

for some $\eta_B, b_B > 0$, uniformly over admissible actions.

Theorem 1 (Preservation reachability)**.** *Under equation 10 and the depth condition above,*

$$Q(B) \leq e^{-\eta_B b_B}. \tag{11}$$

The exponential supermartingale makes the bound policy-uniform because every admissible action obeys the same reachability condition. These bounds control both learner-side routes.

4.3 HIGH-CONFIDENCE SUCCESS IS BOUNDED BY PRESERVATION REACHABILITY

Theorem 2 (Variational preservation bound)**.** *For every α,*

$$P_\alpha^\star \le \inf_{B\in\mathfrak{B}} \inf_{n\ge 1} \{Q(B) + \varepsilon_\alpha(n) + r_B(n)\}. \tag{12}$$

If $\varepsilon_\alpha(n) \to 0$ for every fixed n as $\alpha \downarrow 0$, then

$$\limsup_{\alpha\downarrow 0} P_\alpha^\star \le \inf_{B\in\mathfrak{B}} Q(B). \tag{13}$$

Theorem 2 is the main structural result. Under the stated high-confidence conditions, early certification and late actionability without a preservation hit vanish. Preservation reachability then sets the ceiling.

Applying the theorem requires a predeclared barrier, a policy-uniform late-actionability bound, and a policy-uniform reachability bound.

The same decomposition gives a finite-confidence rate. For a threshold certificate $C_t^\alpha = \{\sup_{s\le t} E_s \ge 1/\alpha\}$, suppose equation 9 holds and $E_t \le e^{Lt}$ pathwise. Certification before $n_\alpha = \lceil \log(1/\alpha)/L \rceil$ is impossible. For all sufficiently small α,

$$P_\alpha^\star \le Q(B) + K_B \alpha^{\kappa_B/L}. \tag{14}$$

Appendix B gives the stopping-time proofs and a pairwise risk-controlled variant without a fixed crossing statistic. Section 4.4 instantiates all three terms in a scalar learner.

4.4 CONTEXTUAL-LOGISTIC REALIZATION OF THE PRESERVATION FRAMEWORK

A scalar contextual-logistic learner makes all three terms in Theorem 2 explicit. Its certificate and fallback use the same score.

The score G_t is the fallback's signed evidence for the advice label. Under the beneficial world, it has positive drift and moves the fallback toward the advice action. Negative excursions create temporary learner headroom.

Proposition 1 (Scalar contextual-logistic realization)**.** *Let $Y_t \mid X_t, e_\pm \sim \mathrm{Ber}(\mathrm{sigmoid}(\pm\delta X_t))$ with $\delta > 0$. A predictable policy chooses contexts with $0 < x_{\min} \le |X_t| \le x_{\max} < \infty$. Let $\lambda, B > 0$. The fallback is scalar ridge-logistic with prior center $-B$, and advice predicts $+1$. Use zero actionability margin and define $Z_t = X_t(Y_t - 1/2)$, $G_t = \sum_{s=1}^t Z_s$. Then $A_t = \{G_t \le \lambda B\}$. For the likelihood-ratio certificate $\tau_C^\alpha = \inf\{t : \Lambda_t \ge \alpha^{-1}\}$, $\log \Lambda_t = 2\delta G_t$, let $B_h = \{G_t \le -h\}$ for $h > 0$. There are finite constants $K, \kappa > 0$, independent of h, such that*

$$Q(B_h) \le e^{-2\delta h}, \qquad r_{B_h}(n) \le K e^{-\kappa n}. \tag{15}$$

Set $n_\alpha = \left\lceil \frac{\log(1/\alpha)}{\delta x_{\max}} \right\rceil$. Then $\varepsilon_\alpha(n_\alpha) = 0$. For all sufficiently small α, equation 14 gives

$$P_\alpha^\star \le e^{-2\delta h} + K\alpha^{\kappa/(\delta x_{\max})} \qquad \text{for every } h > 0. \tag{16}$$

In particular, $P_\alpha^\star \to 0$ as $\alpha \downarrow 0$. Every successful stopped event also obeys the sharper model-specific bound

$$P_+(S) \le e^{2\delta\lambda B}\alpha. \tag{17}$$

A fixed context gives an exact check. Fix $x \in [x_{\min}, x_{\max}]$ and consider the singleton-action protocol $X_t \equiv x$. Set $p = \mathrm{sigmoid}(\delta x)$ and $\rho = (1-p)/p = e^{-\delta x}$. If $h = dx/2$ with integer $d \ge 1$, then

$$Q_x(B_h) = \rho^d = e^{-2\delta h}. \tag{18}$$

For the same certificate, let $a_\alpha = \left\lceil \frac{\log(1/\alpha)}{\delta x} \right\rceil$, $b = \left\lfloor \frac{2\lambda B}{x} \right\rfloor$. When $a_\alpha > b$, the fixed-context protocol has horizon-supremum success probability

$$P_\alpha^{(x)} = \rho^{a_\alpha - b}. \tag{19}$$

It satisfies $e^{2\delta\lambda B - 2\delta x}\alpha \le P_\alpha^{(x)} \le e^{2\delta\lambda B}\alpha$. This constant-context policy is admissible in the original protocol. Combining its lower bound with equation 17 gives $P_\alpha^\star = \Theta(\alpha)$ as $\alpha \downarrow 0$.

The model instantiates all three terms in Section 4. Early certification is impossible before n_α. Late actionability decays exponentially. Reaching B_h costs at most $e^{-2\delta h}$. The exact likelihood ratio improves the generic finite-confidence rate. Appendix B.7 gives the proofs. Appendix C verifies the same learner-side bounds in a three-item Bradley–Terry learner.

5 EXACT PRESERVATION AND INFORMATION ORDERING IN A RIDGE-LOGISTIC LEARNER

Section 4 gives a general high-confidence ceiling. We now construct a ridge-logistic learner where preservation reachability equals the policy optimum and semantic information can alter learner headroom.

5.1 EXACT FIRST-PASSAGE PHASE

To see when the preservation bound can be exact, we separate two roles of evidence. Query E moves the learner's target-coordinate score and can create the rare excursion that preserves later actionability. Query S supplies late certification evidence, but repeated S queries also train the fallback toward the advice action. A policy sequences these two roles: $x_E = (1, 0), \qquad x_S = (c, \sqrt{1-c^2})$, with $0 < c < 1$ and beneficial-world label probabilities $p_E, p_S > 1/2$. The harmful world flips both logits. The fallback is ridge-logistic MAP with ridge strength $\lambda > 0$ and prior center $(-b, 0)$ with $b > 0$. Advice chooses item 1; the fallback chooses item 2 when its fitted target coordinate is nonpositive. Under unit binary terminal loss, advice gains one in this case and zero otherwise, so every fixed $m \in (0, 1]$ gives the same actionability event. Appendix F records the matched-transcript risk identity.

Let $D_E = 2K_E - N_E$. At the learner boundary $w_1 = 0$, profile the nuisance coordinate and call its S-logit z. Strict convexity gives

$$D_E \le 2\lambda b - \frac{2\lambda c}{1-c^2} z \quad \Longleftrightarrow \quad \text{advice is actionable.} \tag{20}$$

Repeated beneficial-world S queries drive z to $\delta_S = \operatorname{logit}(p_S)$. The remaining canonical headroom is

$$C(c) = \lambda b - \frac{\lambda c \delta_S}{1-c^2}. \tag{21}$$

When $C(c) < 0$, sustained S evidence eliminates actionability. Define the smallest preserving excursion $d(c) = \lceil -2C(c) \rceil, \qquad \rho = \frac{1-p_E}{p_E}, \qquad \delta_E = \operatorname{logit}(p_E)$. Fix the level-$\alpha$ nested certificate $\tau_C^\alpha = \inf\{t : \Lambda_t \ge \alpha^{-1}\}$, where $\log \Lambda_t = \delta_E D_E(t) + \delta_S D_S(t)$.

Theorem 3 (Exact ridge-logistic phase)**.** *For this certificate and any fixed margin $m \in (0, 1]$ under unit binary terminal loss,*

$$\boxed{\lim_{\alpha \downarrow 0} \sup_T P^\star_{\alpha,T}(c) = \begin{cases} 1, & C(c) \ge 0, \\ \rho^{d(c)}, & C(c) < 0. \end{cases}} \tag{22}$$

The optimization allows predictable adaptive E/S querying and bounded stopping.

Nonnegative headroom lets certification accumulate without erasing actionability. With negative headroom, success first needs the depth-d E excursion, which occurs with probability ρ^d.

If $C(c) > 0$, S-only sampling preserves positive headroom while certification grows. At $C(c) = 0$, the centered S score revisits the actionable side after certification; both branches have value one.

If $C(c) < 0$, every success path with nonvanishing high-confidence mass must pay for preservation. Under e_+, the signed E score is a biased random walk with positive drift. Its probability of ever reaching $-d$ is exactly ρ^d. Before that hit, late actionability is exponentially unlikely under every predictable E/S schedule; a stopped likelihood-ratio split removes the remaining no-hit mass. This gives the converse. A matching policy queries E until the depth-d hit, freezes E, and then uses S for certification. Conditional on the hit, later certified actionability occurs with probability tending to one. Appendix D gives the proof.

The order of limits in Theorem 3 lets stronger confidence use a longer horizon. The erosive proof also gives

$$P^\star_{\alpha,T}(c) \le \rho^{d(c)} + K\alpha^{\kappa/L} \tag{23}$$

for all sufficiently small α and every horizon. Appendix D.5 reports the finite-horizon grid.

5.2 More informative semantic evidence can lower the optimum

A richer semantic channel is weakly beneficial when the downstream problem is fixed. Here it can also change the fallback state and consume preservation headroom.

Table 1: Representative points across three consecutive preservation depths on the information-increasing branch. Relative informativeness increases from low to high; finite-horizon values use the staged policy with $T = 240$ and $\alpha = 0.05$.

$d(c)$	c	Relative informativeness	$p_S(c)$	High-confidence optimum	Finite horizon
1	0.65	Low	0.702	**0.429**	0.428
2	0.67	Medium	0.704	**0.184**	0.184
3	0.75	High	0.712	**0.079**	0.079

Consider

$$\delta_S(c) = c\delta_E + \sqrt{1-c^2}\,\delta_0, \qquad 0 < \delta_0 < \delta_E, \tag{24}$$

with $p_S(c) = \mathrm{sigmoid}(\delta_S(c))$. On $c \in (c_{\mathrm{info}}, c_{\mathrm{peak}})$, $c_{\mathrm{info}} = \frac{\delta_E^2-\delta_0^2}{\delta_E^2+\delta_0^2}$, $c_{\mathrm{peak}} = \frac{\delta_E}{\sqrt{\delta_E^2+\delta_0^2}}$, we have $p_S(c) > p_E$ and $p_S(c)$ increases with c. The symmetric two-world Bernoulli channel is therefore strictly increasing in Blackwell order. At the same time, $C'(c) < 0$, so the same change can force a larger required depth.

Corollary 1 (Blackwell improvement with a lower certified-and-actionable ceiling)**.** *Take $c_{\mathrm{info}} < c_1 < c_2 < c_{\mathrm{peak}}$ with $C(c_1), C(c_2) < 0$. If $d(c_2) > d(c_1)$, then the semantic S channel at c_2 strictly Blackwell-dominates the channel at c_1, while*

$$\limsup_{\alpha\downarrow 0} \sup_T P^{\star}_{\alpha,T}(c_2) < \limsup_{\alpha\downarrow 0} \sup_T P^{\star}_{\alpha,T}(c_1).$$

Table 1 illustrates Corollary 1 across three consecutive preservation depths. Along the information-increasing branch, greater semantic informativeness can raise the required preservation depth and lower the high-confidence optimum. Because $p_S(c)$ and $C(c)$ vary continuously and strictly on the stated interval, this ordering holds over neighborhoods within each depth regime rather than only at isolated parameter points.

5.3 Downstream-preserving information monotonicity

If the richer experiment preserves learner state and deployment actionability, the usual Blackwell argument applies (Blackwell, 1953).

Proposition 2 (Downstream-preserving information monotonicity)**.** *Suppose a richer experiment has a world-independent causal Blackwell projection to a base experiment. Suppose every base action and stopping rule remains available. A copied policy preserves the projected learner-state law, the deployment actionability status, and the certificate or risk coefficients. Then, for every finite horizon,*

$$P^{\star,\,\mathrm{rich}}_{\alpha,T} \geq P^{\star,\,\mathrm{base}}_{\alpha,T}.$$

The proof projects the richer transcript to the base transcript and ignores residual information. Every feasible base trajectory then has an outcome-equivalent implementation in the richer experiment. Appendix E gives the formal copy argument. Corollary 1 changes the ordering only when richer evidence consumes enough learner headroom to cross a preservation depth.

6 Fixed-transcript checks of the learner-state mechanism

The theory identifies learner-state evolution as the mechanism that can separate certification from actionability. One question remains: does this separation persist when the realized certificate path itself is held fixed? Answering it requires a controlled comparison rather than another distributional specialization of the theory. We therefore reuse the same certification transcript and vary only how the fallback consumes that evidence. The Gaussian benchmark varies update strength and timing, while the digits experiment repeats the comparison in multiclass classification.

6.1 Update strength and timing

The Gaussian benchmark isolates the learner-state mechanism by holding the certification evidence fixed while varying how the fallback consumes it. This lets us test whether certificate-time actionability can change even when the certificate path is unchanged.

Figure 1 uses a Gaussian benchmark whose full learner law appears in Appendix G. Panel (a) lets certification and the learner move as the evidence horizon grows. Panel (b) keeps the evidence path fixed and changes learner-update strength. Panel (c) is the tighter timing control: it fixes $T = 400$, the full evidence transcript, and exactly $K = 120$ learner updates. Every arm uses the same step sequence $\eta(u) = \frac{0.65}{u+5}$, where u counts previous learner updates. Only the wall-clock location of the 120-round block changes. Moving the block from rounds 1 through 120 to rounds 281 through 400 raises the true certified-and-actionable probability from 0.006 to 0.597, while certification stays 0.666. Across all eleven timing arms, mean final fallback regret remains between 0.097 and 0.099. The contrast isolates learner state at certification from the evidence path, update count, and eventual fallback quality.

For deployment, the same Gaussian confidence sequence gives a conservative observable actionability flag. If $\bar{r}_t$ is the running mean and $\|\theta - \bar{r}_t\| \le \rho_t(\beta)$, then

$$\underline{\Delta}_t = \|x_t\|^2 - \|a\|^2 + 2(a - x_t)^\top \bar{r}_t - 2\|a - x_t\|\rho_t(\beta) \tag{25}$$

is a lower bound on true gain. At $T = 400$, certified-and-observably-actionable probability is 0.073, versus 0.407 with true actionability; at $T = 2000$ it is 0.079. Appendix G gives the construction and its anytime-valid guarantee.

6.2 Digits with a fixed certificate path

The Digits experiment tests the same mechanism in multiclass classification. Within each partition, all arms share the same certification stream and certificate path; only the frequency of fallback updates changes.

We use the 10-class optical digit data (Alpaydin & Kaynak, 1998). A fixed multinomial logistic classifier supplies advice; an averaged online logistic SGD model is the fallback. Each of 100 stratified partitions uses one shared certification stream. A Bernoulli likelihood-ratio e-process tests whether advice accuracy on the certification pool exceeds 0.90 at level $\alpha = 0.01$. The three arms differ only in fallback updating: held out, every fourth example, or every example.

Table 2: Digits protocol check over 100 paired partitions. Each row uses the same certification transcript within a partition.

Update rule	Certified	True actionable	Monitor flag	Final fallback acc.
Held out	0.59	0.59	0.59	0.718
Every fourth	0.59	0.59	0.49	0.853
Immediate	0.59	0.29	0.20	0.908

Certification stays at 0.59 in every arm, while immediate updating lowers certificate-time true actionability to 0.29 (Table 2). Among the 59 certifying partitions, mean advice-minus-fallback test gain is 0.218, 0.158, and 0.091 under held-out, every-fourth, and immediate updating. The analytic models supply the policy-uniform learner bounds; this classification check isolates the same fixed-transcript learner-state effect.

Appendix I repeats the comparison in Model Picker (Karimi et al., 2021); Appendix A studies the finite-budget value of preservation. Together, the analytic realization and fixed-transcript checks show the same separation. Certification follows the evidence path, while preservation also depends on the learner state created as that evidence is consumed.

7 DISCUSSION

This work studies whether external advice can remain actionable long enough to be certified when the same evidence also updates the fallback. We show that preservation reachability controls certified-and-actionable success at high confidence. As confidence tightens, early certification disappears, and success depends on reaching a learner state that preserves enough actionability for certification to finish. In the ridge-logistic learner, this ceiling equals the policy optimum. The same model shows that more informative evidence can lower certified-and-actionable success when it changes the downstream learner state. Information monotonicity returns when the downstream learner state and deployment actionability are preserved.

The results show why certification and preservation must be considered together under evidence reuse. Certification follows the evidence path, while preservation also depends on the learner state created as that evidence is consumed. The fixed-transcript checks make this distinction explicit. Update strength and timing can change certificate-time actionability without changing the certificate path. Evidence routing therefore affects whether certified advice remains actionable at deployment. This suggests choosing update strength and timing together with the certificate when certification evidence also trains the fallback.

The current framework requires advice to be certified before deployment. It does not cover settings that use advice immediately and constrain its use as evidence arrives. Extending the analysis to such settings would connect certification with protocols that control advice during deployment. Our results also characterize the high-confidence ceiling without resolving how fast certification accumulates relative to the loss of actionability. A rate-level analysis could identify when certification can finish before learning erases the advice's remaining gain, especially at finite confidence.

AI USE STATEMENT

Generative AI tools were used to aid and polish writing, retrieve and discover related work, and support research ideation and execution. They were also used to draft portions of the manuscript and assist with proving and checking mathematical claims. The authors independently verified all AI-assisted outputs, including mathematical and experimental claims, and take full responsibility for the paper.

REPRODUCIBILITY STATEMENT

We provide complete proofs of the main theoretical results in the appendices, together with additional derivations. The appendices document experimental protocols, data splits, evaluation procedures, model configurations, and additional results used in the paper.

REFERENCES

Ahmed M. Alaa and Mihaela van der Schaar. Balancing suspense and surprise: Timely decision making with endogenous information acquisition. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2910–2918, 2016.

Ethem Alpaydin and Cenk Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998. doi: 10.24432/C50P49.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

Yatong Chen, Wei Tang, Chien-Ju Ho, and Yang Liu. Performative prediction with bandit feedback: Learning through reparameterization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7298–7324. PMLR, 2024.

Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3): 755–770, 1959. doi: 10.1214/aoms/1177706205.

Chris 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.

Rohan Deb, Mohammad Ghavamzadeh, and Arindam Banerjee. Conservative contextual bandits: Beyond linear representations. In *The Thirteenth International Conference on Learning Representations*, 2025.

Aditya Deshmukh, Venugopal V. Veeravalli, and Srikrishna Bhashyam. Sequential controlled sensing for composite multihypothesis testing. *Sequential Analysis*, 40(2):259–289, 2021. doi: 10.1080/07474946.2021.1912525.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi: 10.1126/science.aaa9375.

Peter I. Frazier and Angela J. Yu. Sequential hypothesis testing under stochastic deadlines. In *Advances in Neural Information Processing Systems*, volume 20, pp. 465–472, 2007.

Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024. doi: 10.1093/jrsssb/qkae011.

Patrick Hemmer, Lukas Thede, Michael Vössing, Johannes Jakubik, and Niklas Kühl. Learning to defer with limited expert predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6002–6011, 2023. doi: 10.1609/aaai.v37i5.25742.

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.

Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9760–9785. PMLR, 2022.

Daniel Jarrett, Alihan Hüyük, and Mihaela van der Schaar. Online decision mediation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1790–1805, 2022. doi: 10.52202/068431-0131.

Mohammad Reza Karimi, Nezihe Merve Gürel, Bojan Karlaš, Johannes Rausch, Ce Zhang, and Andreas Krause. Online active model selection for pre-trained classifiers. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 307–315. PMLR, 2021.

Neil Marchant and Benjamin Rubinstein. Adaptive data analysis for growing data. In *Advances in Neural Information Processing Systems*, volume 38, pp. 106662–106696, 2025. doi: 10.52202/085713-3217.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020.

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Who should predict? Exact algorithms for learning to defer to humans. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10520–10545. PMLR, 2023.

Cuong Nguyen, Thanh-Toan Do, and Gustavo Carneiro. Probabilistic learning to defer: Handling missing expert annotations and controlling workload distribution. In *The Thirteenth International Conference on Learning Representations*, 2025.

Sirin Nitinawarat, George K. Atia, and Venugopal V. Veeravalli. Controlled sensing for multihypothesis testing. *IEEE Transactions on Automatic Control*, 58(10):2451–2464, 2013. doi: 10.1109/TAC.2013.2261188.

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.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

Abhishek Sharma, Leo Benac, Sonali Parbhoo, and Finale Doshi-Velez. Decision-point guided safe policy improvement. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2935–2943. PMLR, 2025.

Kei Takemura. Contextual conservative interleaving bandits. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33468–33489. PMLR, 2023.

Adam J. Thorpe, Kendric R. Ortiz, and Meeko M. K. Oishi. State-based confidence bounds for data-driven stochastic reachability using Hilbert space embeddings. *Automatica*, 138:110146, 2022. doi: 10.1016/j.automatica.2021.110146.

Venugopal V. Veeravalli, Georgios Fellouris, and George V. Moustakides. Quickest change detection with controlled sensing. *IEEE Journal on Selected Areas in Information Theory*, 5:1–11, 2024. doi: 10.1109/JSAIT.2024.3362324.

Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22184–22202. PMLR, 2022.

Abraham P. Vinod and Meeko M. K. Oishi. Stochastic reachability of a target tube: Theory and computation. *Automatica*, 125:109458, 2021. doi: 10.1016/j.automatica.2020.109458.

Zixi Wei, Yuzhou Cao, and Lei Feng. Exploiting human-AI dependence for learning to defer. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52484–52499. PMLR, 2024.

Ran Weksler and Boaz Zik. Informative tests in signaling environments. *Theoretical Economics*, 17 (3):977–1006, 2022. doi: 10.3982/TE4461.

Patrick Wienhöft, Marnix Suilen, Thiago D. Simão, Clemens Dubslaff, Christel Baier, and Nils Jansen. More for less: Safe policy improvement with stronger performance guarantees. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 4406–4415, 2023. doi: 10.24963/ijcai.2023/490.

Wenhao Wu. A geometric Blackwell's order. *Economics Letters*, 226:111082, 2023. doi: 10.1016/j.econlet.2023.111082.

Djordje Zikelić, Mathias Lechner, Thomas A. Henzinger, and Krishnendu Chatterjee. Learning control policies for stochastic systems with reach-avoid guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11926–11935, 2023. doi: 10.1609/aaai.v37i10.26407.

A FINITE-BUDGET OPTION VALUE OF PRESERVED ACTIONABILITY

This appendix studies a finite-budget question: can already-certifiable advice substitute for fallback learning on one part of the task, so that a fixed learner-update budget can be redirected elsewhere? The scarce resource is the number of fallback updates, not the number of observed labels. We compare policies at matched update budgets and characterize how the advantage changes as the budget grows.

A.1 TWO-REGION PROTOCOL

Each replicate contains two independent binary logistic regions. Region A is advice-covered: $X_A \sim N(0, I_5)$, labels follow a fixed logistic model, and a fixed adviser predicts from a perturbed version of the true parameter. Region B is unresolved: $X_B \sim N(0, I_{50})$, labels follow an independent logistic model, and no advice is available. Every replicate has $3{,}000$ labeled rounds, exactly $1{,}500$ from each region in uniformly random order. The two region models, adviser, labeled stream, and independent $5{,}000$-example evaluation sets are fixed within a replicate and shared across policies.

Only Region-A examples update the advice certificate. For adviser-correctness indicators Z_s, we use the Bernoulli likelihood-ratio e-process for

$$H_0 : p_{\mathrm{adv}} \le 0.83 \qquad \text{versus} \qquad p_1 = 0.88.$$

The primary threshold is $1/0.05$; $\alpha = 0.01$ provides a secondary sensitivity analysis on the same path. Certificate timing is independent of how the fallback spends its update budget.

Both regional fallbacks are online logistic learners initialized at zero. One learner update costs one unit, and the update-count schedule is

$$\eta(u) = \frac{0.18}{\sqrt{1 + u/100}},$$

where u is the number of updates already consumed by that policy. The budget grid is

$$K \in \{20, 40, 80, 160, 240, 320, 400, 480, 640, 800, 1000, 1200, 1400\}.$$

Every policy uses exactly K updates.

Immediate reuse. The first K labeled rounds are used for fallback learning, so Region-A labels both contribute to certification and train fallback A, while Region-B labels train fallback B.

Preserve and reallocate. Region-A labels still contribute to the same certificate but never update fallback A. All K learner updates are spent on the first K Region-B labels. This deliberately preserves the adviser–fallback gap on the covered region and reallocates the same update budget to the unresolved region.

No-advice control. This arm uses the same learner states as Immediate reuse but never deploys advice. It isolates whether the adviser provides substitution value rather than merely changing the accounting of performance.

At the final horizon, Immediate and Preserve deploy the adviser on A if its certificate has crossed; otherwise they use fallback A. Region B always uses fallback B. Deployed accuracy equally weights the two regions. At the first certificate time, advice is actionable when adviser accuracy on the fixed A evaluation set is strictly larger than the policy's fallback-A accuracy at that same time.

A.2 FINITE-BUDGET RESULTS

Results aggregate 400 paired realizations under prespecified randomization. The primary budget is $K = 160$. Immediate reuse spends on average 79.78 of its 160 updates on Region A and 80.22 on Region B; Preserve spends all 160 updates on B. Region-B fallback accuracy is 0.658 for Immediate and 0.692 for Preserve, a paired difference of $+0.034$ with bootstrap 95% interval $[0.031, 0.036]$. Because the certified adviser can cover Region A, this reallocation also raises deployed accuracy from 0.771 to 0.788, a paired difference of $+0.017$ with interval $[0.016, 0.018]$. The no-advice control is 0.764 at the same budget.

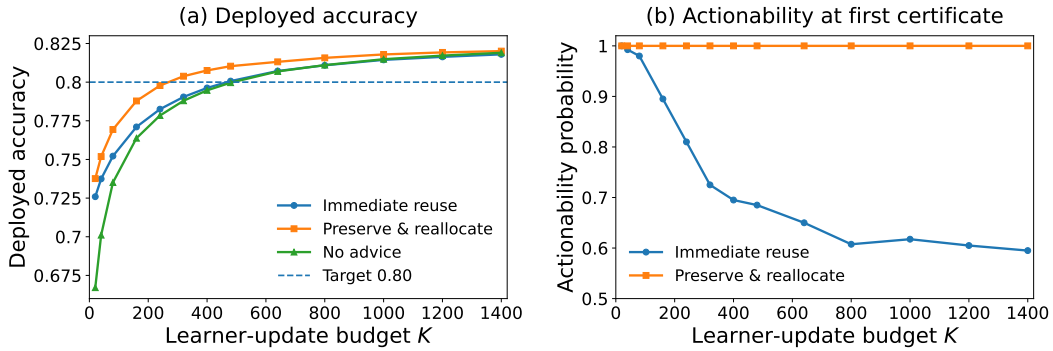


Figure 2: Finite-budget value of preserved actionability. Left: deployed accuracy versus the number of fallback learner updates. On the budget grid, Preserve reaches deployed accuracy 0.80 at $K = 320$, compared with $K = 480$ for Immediate reuse and $K = 640$ for the no-advice control. Right: under Immediate reuse, the probability that advice is still actionable at its first $\alpha = 0.05$ certificate decreases as more fallback updates can be spent before certificate arrival; Preserve keeps fallback A fixed and remains actionable at that arrival time.

The target deployed accuracy 0.80 and the complete budget grid were fixed before evaluation. The smallest grid budget reaching that target is $K = 320$ for Preserve, $K = 480$ for Immediate, and $K = 640$ for No advice. On this grid, Preserve reaches the same target with 160 fewer learner updates than Immediate, a 33.3% reduction relative to $K = 480$. The comparison uses the first grid point that reaches 0.80.

The mechanism is visible at certificate arrival. For Immediate reuse, the probability that advice remains actionable at the first $\alpha = 0.05$ certificate is 1.000 at $K = 20$, 0.895 at $K = 160$, 0.725 at $K = 320$, 0.650 at $K = 640$, and 0.595 at $K = 1400$. Preserve remains actionable in all 400 realizations because fallback A is deliberately not trained. Tightening the certificate to $\alpha = 0.01$ delays median arrival from 421.5 to 693 rounds and produces still lower Immediate actionability at moderate and large budgets.

The performance advantage shrinks as the update budget grows. Preserve minus Immediate deployed accuracy is $+0.017$ at $K = 160$, $+0.013$ at $K = 320$, $+0.006$ at $K = 640$, $+0.004$ at $K = 1000$, and $+0.002$ at $K = 1400$. The last gap is about one eighth of the primary-budget gap. This attenuation matches a finite-budget opportunity-cost interpretation: preserving an advice-covered region frees scarce learner updates for unresolved decisions, while the large-budget difference becomes small.

A.3 Design validation

The design, primary budget $K = 160$, budget grid, confidence levels, target accuracy 0.80, bootstrap rule, and randomization scheme were fixed before evaluation. The construction enforces exact update counts, certificate-path identity across policies and budgets, paired-stream identity, target-crossing definitions, and the prescribed large-budget comparison. Under $20{,}000$ null paths, empirical e-process boundary crossings agree with the nominal levels within sampling variation.

B Proofs for the certification–preservation race

B.1 A minimal obstruction from evidence reuse

The simplest model shows why evidence reuse can be costly. Labels have the same certification value under training and holdout, but reuse changes the fallback benchmark and can eliminate actionable use. Advice chooses item 1. In e_+, labels satisfy $Y_t \sim \mathrm{Ber}(p)$ with $p > 1/2$; in e_-, they satisfy $Y_t \sim \mathrm{Ber}(1-p)$. Let $K_n = \sum_{t=1}^{n} Y_t$. The fallback chooses item 1 only after a strict majority favors

it. Advice remains actionable in e_+ exactly when $K_n \le n/2$. On every such history,

$$\frac{P_+(K_n = k)}{P_-(K_n = k)} = \left(\frac{p}{1-p}\right)^{2k-n} \le 1. \tag{26}$$

Proposition 3 (Same evidence, vanishing actionable-use power)**.** *Let τ be any bounded stopping time and let $S \in \mathcal{F}_\tau$ be any randomized advice-use event supported on positive actionability. If $P_-(S) \le \alpha$, then $P_+(S) \le \alpha$. If the same labels are held out from fallback training, the beneficial-world certification probability for the corresponding likelihood-ratio certificate approaches one as the horizon grows.*

Proof of Proposition 3. A stopped history contains the labels and world-independent policy randomization revealed before stopping. If its length is n and k labels equal one, positive actionability requires $k \le n/2$. equation 26 then gives $dP_+/dP_- \le 1$ on every stopped history in S. Predictable action, randomization, and stopping kernels cancel in the stopped likelihood ratio. Summing over S yields $P_+(S) \le P_-(S) \le \alpha$.

For held-out evidence, the fallback never changes. Under e_+ the log-likelihood ratio has positive drift, so its running maximum crosses every fixed boundary almost surely. The supremum over finite horizons therefore approaches one. □

Proposition 3 isolates the obstruction. Histories that preserve actionability cannot accumulate a favorable likelihood ratio under e_+. Holding labels out preserves their certification content without moving the fallback benchmark.

B.2 Three-route decomposition and variational bound

Proof of Lemma 1. Take $\omega \in S$. Certification gives $\tau_C^\alpha(\omega) \le \tau(\omega)$. If $\tau_C^\alpha < n$, the first set in equation 7 holds. Otherwise $\tau \ge n$. If $\tau_B^\pi \le \tau$, the second set holds. If not, success gives A_τ while $\tau < \tau_B^\pi$, which is the third set. □

Proof of Theorem 2. Fix an admissible barrier B and $n \ge 1$. Lemma 1 and the definitions of $Q(B)$, $\varepsilon_\alpha(n)$, and $r_B(n)$ give

$$P_\alpha^\star \le Q(B) + \varepsilon_\alpha(n) + r_B(n).$$

Take the two infima. For the high-confidence statement, fix B and n, send $\alpha \downarrow 0$, then send $n \to \infty$. Finally take the infimum over B. □

B.3 Learner-progress and reachability bounds

For equation 8, write $R_t = \sum_{s=1}^t (Z_s - \mathbb{E}[Z_s \mid \mathcal{F}_{s-1}])$. On the event in the left side,

$$R_t \le C_0 - (\mu - \gamma_0)t.$$

For $t \ge 2C_0/(\mu - \gamma_0)$, this implies $R_t \le -(\mu - \gamma_0)t/2$. Conditional sub-Gaussian concentration and a union bound over $t \ge n$ give

$$r_B(n) \le K_B e^{-\kappa_B n}, \qquad \kappa_B = \frac{(\mu - \gamma_0)^2}{8\sigma^2}$$

for a finite K_B.

Proof of Theorem 1. By equation 10, $M_t := e^{-\eta_B G_{t \wedge \tau_B^\pi}}$ is a nonnegative supermartingale. Every barrier hit has $M_{\tau_B^\pi} \ge e^{\eta_B b_B}$. Ville's inequality gives

$$P_+^\pi(\tau_B^\pi \le T) \le e^{-\eta_B b_B}$$

for every policy and finite T. Taking the supremum proves equation 11. □

B.4 Choosing a preservation barrier

The barrier is an analysis object, but its construction need not start from guesswork. A useful route is to work backward from the learner's actionability boundary.

1. Isolate a learner score. Choose an observable scalar or low-dimensional statistic $G_t = g(L_t)$ that controls the fallback decision. Rewrite actionability as a restriction on this score. In the contextual-logistic model, for example, actionability is $G_t \le \lambda B$; in the canonical model, profiling the nuisance coordinate reduces the learner boundary to equation 20.

2. Identify the erosive drift. Establish the beneficial-world drift under every admissible evidence action. If ordinary learner evolution pushes G_t away from the actionable region, late actionability without an opposing excursion is a concentration event. The required uniformity is over actions, not only over one fixed policy. A learner-neutral action can invalidate this step and remove a strict preservation ceiling.

3. Put the barrier at the compensating excursion. Choose a level set B_h reached only after an excursion large enough to create future actionability slack. The smallest useful level is model dependent. In the ridge-logistic model it is the depth-d E excursion. In contextual logistic learning it is a negative level of G_t. In a multidimensional learner, the same step can use a level set of a Lyapunov function rather than a coordinate.

4. Bound the hit and the no-hit tail separately. For reachability, search for a nonnegative process $M_{t\wedge\tau_B}$ that is a supermartingale under every admissible action and satisfies $M_{\tau_B} \ge m_B$ on a hit. Ville's inequality gives $Q(B) \le M_0/m_B$. Exponential potentials yield Theorem 1; other Lyapunov supermartingales give the same interface. Separately, concentration of the learner score on $\{\tau_B > t\}$ supplies $r_B(n) \to 0$.

These steps are a construction principle, not a universal barrier-search algorithm. Their role is to reduce a new learner to two verifiable statements: late actionability is rare without the declared hit, and the hit itself has a policy-uniform reachability bound. The contextual-logistic proof below and the canonical proof in Appendix D instantiate the same recipe with different scores.

B.5 Anytime-valid observable actionability

True actionability is useful for structural analysis even when the deployment policy cannot observe it. A confidence sequence gives a generic observable replacement whenever the unknown deployment parameter admits time-uniform inference.

Proposition 4 (Confidence-set actionability)**.** *Suppose the deployment gain is $\Delta(U_t;\vartheta)$ for an unknown parameter $\vartheta_\star$. Let $\Theta_t \in \mathcal{F}_t$ be a confidence sequence satisfying*

$$\sup_\pi P_+^\pi(\exists t : \vartheta_\star \notin \Theta_t) \le \beta.$$

Define

$$\underline{\Delta}_t = \inf_{\vartheta\in\Theta_t} \Delta(U_t;\vartheta), \qquad \widehat{A}_t = \{\underline{\Delta}_t \ge m\}.$$

Then $\widehat{A}_t \in \mathcal{F}_t$ and

$$\sup_\pi P_+^\pi\Big(\exists t : \widehat{A}_t \setminus A_t\Big) \le \beta. \tag{27}$$

Every stopped deployment rule based on $C_\tau^\alpha \cap \widehat{A}_\tau$ is fully observable and keeps the original level-α certification guarantee together with a level-β false-actionability guarantee.

Proof. On the simultaneous coverage event, $\vartheta_\star \in \Theta_t$ for every t. Then $\underline{\Delta}_t \le \Delta(U_t;\vartheta_\star)$ and $\widehat{A}_t \subseteq A_t$ simultaneously. Any false observable declaration requires confidence-sequence failure, whose probability is at most β. □

The Gaussian rule in Section 6 is this construction with

$$\Theta_t = \{\vartheta : \|\vartheta - \bar{r}_t\| \le \rho_t(\beta)\}.$$

Minimizing squared-loss gain over this ball gives exactly equation 25. The structural theorems continue to use A_t because oracle access enlarges the feasible class for an upper bound. Proposition 4 supplies a deployable observable subset when a confidence sequence is available.

B.6 FINITE-CONFIDENCE AND RISK-CONTROLLED VARIANTS

If the certificate statistic obeys $E_t \le e^{Lt}$ pathwise, the crossing $E_t \ge 1/\alpha$ cannot occur before $n_\alpha = \lceil \log(1/\alpha)/L \rceil$. Insert this split in Theorem 2 and use equation 9 to obtain equation 14.

The same learner-side argument does not require a fixed crossing rule. For a beneficial/harmful pair, let

$$\Lambda_t^\pi = \frac{dP_+^\pi}{dP_-^\pi}\Big|_{\mathcal{F}_t}.$$

If $S \subseteq A_\tau$ and $P_-^\pi(S) \le \alpha$, split success by preservation and by whether $\Lambda_\tau \le R$. The low-likelihood no-hit term is at most $R\alpha$ by stopped change of measure. If $\Lambda_t \le e^{Lt}$, the high-likelihood term requires $\tau \ge \log R/L$ and is bounded by $KR^{-\kappa/L}$. Optimizing gives a remainder of order $\alpha^{\kappa/(\kappa+L)}$.

For a deterministic deadline D, restrict $\tau \le D$ and replace $Q(B)$ by

$$Q_D(B) = \sup_\pi P_+^\pi(\tau_B^\pi \le D).$$

The same decomposition gives

$$P_{\alpha,D}^\star \le \inf_{B\in\mathfrak{B}} \min_{1\le n\le D} \{Q_D(B) + \varepsilon_\alpha(n) + r_B(n)\}.$$

B.7 SCALAR CONTEXTUAL-LOGISTIC REALIZATION

For Proposition 1, the fallback minimizes

$$F_t(\theta) = \frac{\lambda}{2}(\theta + B)^2 + \sum_{s=1}^{t} \{\log(1 + e^{X_s\theta}) - Y_s X_s \theta\}.$$

Let $G_t = \sum_{s\le t} X_s(Y_s - 1/2)$. Strict convexity and

$$F_t'(0) = \lambda B - G_t$$

show that the fitted parameter is nonpositive exactly when $G_t \le \lambda B$. With zero actionability margin, this is the actionability event for advice label $+1$.

For one contextual observation,

$$\log \frac{P_+(Y_t \mid X_t)}{P_-(Y_t \mid X_t)} = 2\delta X_t(Y_t - 1/2).$$

Predictable context selection cancels from the likelihood ratio. At every time t,

$$\log \Lambda_t = 2\delta G_t. \tag{28}$$

The process $\Lambda_t = dP_+^\pi/dP_-^\pi|_{\mathcal{F}_t}$ is a nonnegative P_-^π martingale. Ville's inequality makes

$$\tau_C^\alpha = \inf\{t : \Lambda_t \ge \alpha^{-1}\}$$

a level-α certificate for every predictable context policy.

Each log-likelihood increment is at most $\delta x_{\max}$. For

$$n_\alpha = \left\lceil \frac{\log(1/\alpha)}{\delta x_{\max}} \right\rceil,$$

we have $\tau_C^\alpha \ge n_\alpha$ pathwise. This gives $\varepsilon_\alpha(n_\alpha) = 0$.

On every actionable stopped history, $G_\tau \le \lambda B$. Any successful stopped event S satisfies $P_-(S) \le \alpha$. Stopped change of measure gives

$$P_+(S) = E_-[\Lambda_\tau \mathbf{1}_S] \le e^{2\delta\lambda B} P_-(S) \le e^{2\delta\lambda B}\alpha.$$

This proves equation 17.

We next verify the two learner-side terms. Under e_+,

$$E[Z_t \mid \mathcal{F}_{t-1}] = |X_t|\{\mathrm{sigmoid}(\delta|X_t|) - 1/2\} \geq \mu > 0,$$

where one may take

$$\mu = x_{\min}\{\mathrm{sigmoid}(\delta x_{\min}) - 1/2\}.$$

Conditional on the past, Z_t lies in an interval of length at most $x_{\max}$. Its centered increment is sub-Gaussian with variance proxy $\sigma^2 = x_{\max}^2/4$.

Let $M_t = \sum_{s=1}^{t}\{Z_s - E[Z_s \mid \mathcal{F}_{s-1}]\}$. If $G_t \leq \lambda B$ and $t \geq 2\lambda B/\mu$, then

$$M_t \leq \lambda B - \mu t \leq -\frac{\mu t}{2}.$$

A conditional sub-Gaussian tail bound and a union bound over $t \geq n$ give constants $K < \infty$ and $\kappa > 0$ with

$$P_+^\pi(\exists t \geq n : G_t \leq \lambda B) \leq Ke^{-\kappa n}.$$

The constants depend on the fixed protocol parameters and are independent of h. Since the event defining $r_{B_h}(n)$ is a subset of the event above,

$$r_{B_h}(n) \leq Ke^{-\kappa n}.$$

For $h > 0$, define $\tau_h = \inf\{t : G_t \leq -h\}$. equation 28 gives

$$\exp(-2\delta G_t) = \frac{dP_-^\pi}{dP_+^\pi}\Big|_{\mathcal{F}_t}.$$

This is a nonnegative P_+^π martingale. Ville's inequality yields

$$P_+^\pi(\tau_h < \infty) \leq e^{-2\delta h}$$

uniformly over predictable context policies. This proves equation 15.

The likelihood-ratio certificate also satisfies the growth condition in equation 14 with $L = \delta x_{\max}$. Substitution gives

$$P_\alpha^\star \leq e^{-2\delta h} + K\alpha^{\kappa/(\delta x_{\max})}$$

for every $h > 0$ and all sufficiently small α. Sending $h \to \infty$ gives $P_\alpha^\star \leq K\alpha^{\kappa/(\delta x_{\max})}$. This bound tends to zero as $\alpha \downarrow 0$. This proves equation 16 and the high-confidence claim in Proposition 1.

Fixed-context specialization. Fix $x \in [x_{\min}, x_{\max}]$, set $X_t \equiv x$, and write

$$p = \mathrm{sigmoid}(\delta x), \qquad \rho = \frac{1-p}{p} = e^{-\delta x}, \qquad D_t = \frac{2G_t}{x} = \sum_{s=1}^{t}(2Y_s - 1).$$

Under e_+, D_t is a nearest-neighbor random walk with upward probability $p > 1/2$. For $h = dx/2$ with integer $d \geq 1$, the barrier event is the first hit of $-d$. The standard biased-walk hitting equation gives

$$P_+(\tau_{-d} < \infty) = \rho^d = e^{-2\delta h}.$$

This proves equation 18. It also shows that the reachability bound in Theorem 1 can be attained in a singleton-action protocol.

The certificate condition is $D_t \geq a_\alpha$, where

$$a_\alpha = \left\lceil \frac{\log(1/\alpha)}{\delta x} \right\rceil.$$

Actionability is $D_t \leq b$, where

$$b = \left\lfloor \frac{2\lambda B}{x} \right\rfloor.$$

Assume $a_\alpha > b$. The positive-drift walk hits a_α almost surely. After that hit, the probability of ever reaching b is $\rho^{a_\alpha - b}$. Truncating this return time at a deterministic horizon and sending the horizon to infinity gives

$$P_\alpha^{(x)} = \rho^{a_\alpha - b}.$$

This proves equation 19.

The ceiling and floor operations satisfy

$$a_\alpha - b \le \frac{\log(1/\alpha)}{\delta x} - \frac{2\lambda B}{x} + 2.$$

It follows that

$$P_\alpha^{(x)} \ge e^{2\delta\lambda B - 2\delta x}\alpha.$$

The upper bound in equation 17 has the same linear rate. Since $x \in [x_{\min}, x_{\max}]$, the constant policy $X_t \equiv x$ is admissible in the original protocol. Therefore $P_\alpha^\star = \Theta(\alpha)$ as $\alpha \downarrow 0$.

C Additional learner realizations

Proposition 1 gives the non-ranking realization used in the main text. We verify the same preservation conditions in a three-item Bradley–Terry learner (Bradley & Terry, 1952).

C.1 Three-item Bradley–Terry geometry

Let the centered item-utility basis be

$$a = (1, -1, 0)/\sqrt{2}, \qquad r = (1, 1, -2)/\sqrt{6}.$$

Use edge $(1, 2)$ for the target decision and edge $(1, 3)$ for diagnostic evidence. Their feature vectors are

$$x_{12} = \sqrt{2}\, a, \qquad x_{13} = \frac{1}{\sqrt{2}}a + \sqrt{\frac{3}{2}}r,$$

so their cosine is $1/2$. With isotropic ridge coefficient λ_3 and centered prior $u_0 = -b_3 a$, the MAP estimate stays in $\mathrm{span}\{a, r\}$. Write $u = u_a a + u_r r$ and $w = \sqrt{2}(u_a, u_r)$. Then

$$x_{12}^\top u = w_1, \qquad x_{13}^\top u = \frac{1}{2}w_1 + \frac{\sqrt{3}}{2}w_2.$$

This is the two-query ridge-logistic geometry of Section 5 at $c = 1/2$, with

$$b = \sqrt{2}b_3, \qquad \lambda = \lambda_3/2.$$

Let N_{ij} and K_{ij} denote query and favorable-outcome counts on edge (i, j), and set $D_{12} = 2K_{12} - N_{12}$. Under e_+, $Y_t \sim \mathrm{Ber}(p_E)$ on edge $(1, 2)$ and $Y_t \sim \mathrm{Ber}(p_S)$ on edge $(1, 3)$, with $p_E, p_S > 1/2$. Under e_-, each probability is replaced by its complement. Write $\delta_E = \log(p_E/(1 - p_E))$ and $\delta_S = \log(p_S/(1 - p_S))$.

Advice ranks item 1 above item 2. We use zero actionability margin. Advice is actionable when the fitted target contrast satisfies $w_1 \le 0$. At $w_1 = 0$, let z be the fitted $(1, 3)$ logit. Profiling the nuisance coordinate gives

$$\frac{\lambda_3}{2}z + \frac{3}{4}\{N_{13}\,\mathrm{sigmoid}(z) - K_{13}\} = 0. \tag{29}$$

Strict convexity gives the exact actionability condition

$$D_{12} \le \sqrt{2}\lambda_3 b_3 - \frac{2\lambda_3}{3}z. \tag{30}$$

Define

$$C_3 = \frac{\lambda_3 b_3}{\sqrt{2}} - \frac{\lambda_3}{3}\delta_S, \qquad d_3 = \lceil -2C_3 \rceil,$$

and consider the erosive case $C_3 < 0$.

C.2 POLICY-UNIFORM PRESERVATION BOUNDS

Let

$$B_{d_3,t} = \{D_{12}(t) \le -d_3\}, \qquad B_{d_3} = (B_{d_3,t})_{t\ge 0},$$

with hitting time $\tau_{d_3} = \inf\{t : D_{12}(t) \le -d_3\}$.

Proposition 5 (Three-item Bradley–Terry realization)**.** *Let $\rho = (1-p_E)/p_E$. For every predictable policy choosing between edges $(1,2)$ and $(1,3)$,*

$$Q(B_{d_3}) \le \rho^{d_3}.$$

There are constants $K < \infty$ and $\kappa > 0$, independent of the policy, such that

$$r_{B_{d_3}}(n) \le Ke^{-\kappa n}.$$

One may take $\kappa = \mu_3^2/2$, where

$$\mu_3 = \min\left\{p_E - \frac{1}{2}, \frac{1}{2}(p_S - q_{-,3})\right\} > 0$$

and

$$z_{-,3} = \frac{3b_3}{\sqrt{2}} + \frac{3(d_3-1)}{2\lambda_3}, \qquad q_{-,3} = \mathrm{sigmoid}(z_{-,3}).$$

Proof. For a target-edge query, set $V_t = +1$ after a favorable outcome and $V_t = -1$ otherwise. Set $V_t = 0$ on the diagnostic edge. With $\eta = \log(1/\rho)$,

$$\mathbb{E}_+[e^{-\eta V_t} \mid \mathcal{F}_{t-1}] = 1$$

for either predictable edge choice. The process $\exp(-\eta D_{12}(t))$ is a nonnegative martingale. Ville's inequality gives

$$P_+^\pi(\tau_{d_3} \le T) \le e^{-\eta d_3} = \rho^{d_3}$$

for every finite T and every policy. Taking the suprema gives $Q(B_{d_3}) \le \rho^{d_3}$.

It remains to control actionability before the barrier. Since $d_3 - 1 < -2C_3$, we have $z_{-,3} < \delta_S$ and $q_{-,3} < p_S$. Define

$$W_t = K_{12} - \frac{N_{12}}{2} - \frac{\lambda_3 b_3}{\sqrt{2}} + \frac{1}{2}\{K_{13} - q_{-,3}N_{13}\}.$$

Before the barrier, $D_{12} \ge -d_3 + 1$. Combining this with equation 30 gives $z \le z_{-,3}$. Using equation 29,

$$W_t \le \frac{N_{13}}{2}\{\mathrm{sigmoid}(z) - q_{-,3}\} \le 0.$$

Every actionable pre-barrier history lies in $\{W_t \le 0\}$.

Write $W_t = -C_0 + \sum_{s=1}^t Z_s$, where $C_0 = \lambda_3 b_3/\sqrt{2}$ and

$$Z_t = Y_t - \frac{1}{2} \quad \text{on } (1,2), \qquad Z_t = \frac{1}{2}(Y_t - q_{-,3}) \quad \text{on } (1,3).$$

Every admissible action has conditional mean at least $\mu_3 > 0$. Hoeffding's lemma gives the common conditional sub-Gaussian proxy $\sigma^2 = 1/4$. For all large enough t, pre-barrier actionability forces the centered sum below $-\mu_3 t/2$. The tail bound is

$$P_+^\pi(A_t, \tau_{d_3} > t) \le e^{-\mu_3^2 t/2}.$$

A union bound over $t \ge n$ gives $r_{B_{d_3}}(n) \le Ke^{-\kappa n}$ with $\kappa = \mu_3^2/2$. □

The one-step log-likelihood increment has magnitude at most $L = \max\{\delta_E, \delta_S\}$. For the likelihood-ratio certificate, certification cannot occur before $\lceil \log(1/\alpha)/L \rceil$. Theorem 2 gives

$$P_\alpha^\star \le \rho^{d_3} + K\alpha^{\kappa/L}$$

for all sufficiently small α. This verifies the two learner-side conditions in a two-dimensional pairwise-ranking model. Section 5 establishes the exact policy-optimal equality for the ridge-logistic model.

D EXACT RIDGE-LOGISTIC ANALYSIS

For reference, $x_E = (1, 0)$, $x_S = (c, \sqrt{1-c^2})$, and $\delta_q = \mathrm{logit}(p_q)$; e_- flips the beneficial-world logits. The fallback is ridge-logistic MAP with prior center $(-b, 0)$. The fixed certificate is $\tau_C^\alpha = \inf\{t : \Lambda_t \geq \alpha^{-1}\}$ with $\log \Lambda_t = \delta_E D_E(t) + \delta_S D_S(t)$.

D.1 LEARNER BOUNDARY

Let $s = \sqrt{1-c^2}$. With N_E, N_S query counts and K_E, K_S positive-label counts, define $D_E = 2K_E - N_E$ and $D_S = 2K_S - N_S$. At $w_1 = 0$, the profiled S-logit z solves

$$\lambda z + s^2\{N_S \,\mathrm{sigmoid}(z) - K_S\} = 0. \tag{31}$$

The profiled objective is strictly convex. Its derivative at zero is

$$\lambda b - \frac{D_E}{2} - \frac{\lambda c}{s^2} z.$$

The fallback stays on item 2 exactly when $D_E \leq 2\lambda b - 2\lambda c z/(1-c^2)$. Under repeated S sampling in e_+, $z \to \delta_S$. The limiting headroom is $C(c) = \lambda b - \lambda c \delta_S/(1-c^2)$, with $d = \lceil -2C(c) \rceil$ when $C(c) < 0$.

D.2 CANONICAL PRESERVATION PROBABILITY

For each E query, set $V = +1$ when $Y = 1$ and $V = -1$ otherwise. Let $D_E(n) = \sum_{s=1}^n V_s$ and $\tau_{-d}^E = \inf\{n : D_E(n) = -d\}$. With $\rho = (1-p_E)/p_E$ and $\eta = \log(1/\rho)$,

$$\mathbb{E}_+[e^{-\eta V}] = 1.$$

The biased-walk hitting probability solves

$$h(k) = p_E h(k+1) + (1-p_E)h(k-1), \quad h(-d) = 1, \quad h(k) \to 0,$$

so $h(k) = \rho^{k+d}$. Starting at zero,

$$P_+(\tau_{-d}^E < \infty) = \rho^d. \tag{32}$$

D.3 POLICY-UNIFORM LATE ACTIONABILITY

Lemma 2 (Uniform no-hit tail)**.** *Assume $C(c) < 0$. There are constants $K < \infty$, $\kappa > 0$, and n_0 such that every predictable E/S policy satisfies*

$$P_+\big(\exists t \geq n : A_t,\ \tau_{-d}^E > N_E(t)\big) \leq K e^{-\kappa n} \tag{33}$$

for every $n \geq n_0$.

Proof. Before the depth-d hit, $D_E \geq -d + 1$. On an actionable history, the boundary above gives

$$z \leq z_- := \frac{(1-c^2)(2\lambda b + d - 1)}{2\lambda c}.$$

Because $d - 1 < -2C(c)$, the definition of $C(c)$ gives $z_- < \delta_S$.

Expose two independent potential streams before the policy starts: an iid S stream with success probability p_S and an iid E stream with success probability p_E. A predictable policy reveals only prefixes. A random schedule changes the revealed prefix lengths $N_S(t)$ and $N_E(t)$, but not either potential-stream law.

Choose $\gamma \in (0, 1)$ so that

$$\frac{2c\gamma}{1-\gamma} < 2p_E - 1.$$

First suppose $N_S(t) \geq \gamma t$. From equation 31,

$$\frac{K_S}{N_S} = \mathrm{sigmoid}(z) + \frac{\lambda z}{(1-c^2)N_S}.$$

Choose q_S with $\mathrm{sigmoid}(z_-) < q_S < p_S$. For all sufficiently large t, actionability and $N_S(t) \geq \gamma t$ imply $K_S/N_S \leq q_S$. If $K_S(m)$ denotes successes in the first m potential S outcomes, a Bernoulli Chernoff bound yields

$$P_+\big(\exists t \geq n : A_t, \tau^E_{-d} > N_E(t), N_S(t) \geq \gamma t\big) \leq \sum_{m \geq \lceil \gamma n \rceil} e^{-m D_{\mathrm{Ber}}(q_S \| p_S)}.$$

The sum is geometric; D_{Ber} denotes Bernoulli relative entropy.

Now suppose $N_S(t) < \gamma t$. From equation 31, $z \geq -(1-c^2)N_S/\lambda$. Actionability then gives

$$D_E \leq 2\lambda b + 2cN_S.$$

Write $m = N_E(t) > (1-\gamma)t$. Pick $\epsilon > 0$ so that

$$q_E := \frac{2c\gamma}{1-\gamma} + \epsilon < 2p_E - 1, \qquad \bar{p}_E := \frac{1+q_E}{2} < p_E.$$

For all sufficiently large t, actionability implies $D_E(m)/m \leq q_E$. A second Bernoulli Chernoff bound gives

$$P_+\big(\exists t \geq n : A_t, \tau^E_{-d} > N_E(t), N_S(t) < \gamma t\big) \leq \sum_{m \geq \lceil (1-\gamma) n \rceil} e^{-m D_{\mathrm{Ber}}(\bar{p}_E \| p_E)}.$$

Both geometric tails are independent of the predictable schedule. Combining them proves equation 33. □

D.4 Converse and achievability

Let S be any successful deployment in the erosive branch, and let H be the event that the E score reaches $-d$ before deployment. By equation 32, $P_+(H) \leq \rho^d$. On H^c, split by the stopped beneficial-to-harmful likelihood ratio Λ_τ. For any $R > 1$,

$$P_+(S) \leq \rho^d + P_+(S \cap H^c, \Lambda_\tau \leq R) + P_+(S \cap H^c, \Lambda_\tau > R).$$

Stopped change of measure and the level-α risk constraint give

$$P_+(S \cap H^c, \Lambda_\tau \leq R) = E_-[\Lambda_\tau \mathbf{1}_{S \cap H^c, \Lambda_\tau \leq R}] \leq R\alpha.$$

For the fixed certificate stated above, (Λ_t) is a mean-one nonnegative martingale under e_- for every predictable E/S policy. Ville's inequality gives $P^\pi_-(\tau^\alpha_C < \infty) \leq \alpha$. Its one-step increments are bounded by $L := \max\{\delta_E, \delta_S\}$. A value $\Lambda_\tau > R$ therefore requires $\tau \geq \log R/L$. Lemma 2 gives, for large enough R,

$$P_+(S \cap H^c, \Lambda_\tau > R) \leq KR^{-\kappa/L}.$$

Therefore

$$P_+(S) \leq \rho^d + R\alpha + KR^{-\kappa/L}. \tag{34}$$

Taking $R = \alpha^{-1/2}$ makes both remainder terms vanish. The supremum over predictable policies and bounded stopping times gives

$$\limsup_{\alpha \downarrow 0} \sup_T P^\star_{\alpha,T}(c) \leq \rho^d.$$

For this fixed nested likelihood-ratio family, certification cannot occur before $\lceil \log(1/\alpha)/L \rceil$. Applying Lemma 2 at that time gives $P^\star_{\alpha,T}(c) \leq \rho^d + K\alpha^{\kappa/L}$, uniformly in T.

For achievability, query E until its score first reaches $-d$. Use a cutoff that diverges with the horizon and leaves a diverging number of later rounds. On no hit, do not deploy. After the hit, freeze E and query S. The hit probability converges to ρ^d.

If $d > -2C(c)$, the post-hit boundary has strict slack above δ_S. Conditional on a hit, the profiled estimator converges to δ_S, advice is eventually actionable, and the fixed certificate crosses almost surely. If $d = -2C(c)$, the post-hit condition is a centered bounded nondegenerate iid random-walk inequality. Its partial sums have $\limsup = +\infty$ and $\liminf = -\infty$ almost surely, so both signs are visited infinitely often while the fixed certificate remains open. Conditional success again tends to one. The harmful-world crossing risk is at most α. The staged construction attains ρ^d.

When $C(c) > 0$, S-only sampling leaves strict asymptotic actionability slack and certifies almost surely. At $C(c) = 0$, the same centered-walk recurrence implies that the S score revisits the actionable side almost surely on an unbounded horizon. These cases complete Theorem 3.

D.5 EXACT FINITE-HORIZON COMPUTATION

For a finite horizon and binary observations, the nested-certificate control problem can be solved exactly on count states (N_E, K_E, N_S, K_S) by backward dynamic programming. Available actions are query E, query S, deploy on a certified-and-actionable state, or stop. The state annotation uses the exact profiled boundary in equation 31.

The staged construction already tracks both witness ceilings at ordinary confidence once the horizon is long enough:

c	α	T	staged	ceiling
0.65	0.05	240	0.428	0.429
0.67	0.05	240	0.184	0.184

The policy optimum can be far above the limiting phase at shorter horizons because early certification is still available. For $c = 0.67$, the exact count-state program gives:

α	T	DP optimum	staged	ceiling
0.05	20	0.613	0.123	0.184
0.01	40	0.586	0.169	0.184
0.001	40	0.484	0.156	0.184
10^{-6}	40	0.159	0.061	0.184

For $\alpha = 0.01$, the staged value is 0.184 at $T = 240$. These calculations separate the high-confidence preservation ceiling from the finite-confidence early-certification route.

D.6 WHAT PERSISTS UNDER PERTURBATIONS

The exact value ρ^d uses two special features of the canonical model: a skip-free Bernoulli score and a learner boundary whose useful depth is the integer $d = \lceil -2C(c) \rceil$. The first-passage bottleneck is less fragile than this closed form.

Parameter perturbations. Let the canonical parameters vary continuously and write $u = -2C(c)$. If $u \notin \mathbb{Z}$, then $d = \lceil u \rceil$ is constant on a neighborhood of the parameter point. The hitting value ρ^d then changes continuously with p_E while the depth is fixed. A depth crossing occurs only when u reaches an integer. The witness pair in Appendix E lies on opposite sides of such a crossing, and each witness lies inside a nonzero-width constant-depth interval.

Approximate optimization. Let F_t denote the ridge objective and $w_t^\star$ its exact minimizer. Since F_t is λ-strongly convex, any approximate solution $\widetilde{w}_t$ satisfying $\|\nabla F_t(\widetilde{w}_t)\| \le \varepsilon_t$ obeys

$$\|\widetilde{w}_t - w_t^\star\| \le \varepsilon_t/\lambda.$$

If the deployment score is L_A-Lipschitz in w, exact and approximate actionability agree whenever the exact score has slack exceeding $L_A\varepsilon_t/\lambda$. Small optimization error preserves the phase away from actionability and depth boundaries. Near a boundary, even a small error can change the required depth.

Continuous score increments. If the E score has continuous or non-lattice increments with positive beneficial-world drift, preservation is still a downcrossing problem. The exact probability becomes the corresponding ruin probability $\psi(d) = P_+(\tau_{-d} < \infty)$ rather than ρ^d. When an adjustment coefficient $\eta > 0$ satisfies $E_+[e^{-\eta V_t} \mid \mathcal{F}_{t-1}] \le 1$, Ville's inequality gives $\psi(d) \le e^{-\eta d}$. Overshoot generally removes the integer staircase.

More actions and other learners. The variational theorem already allows arbitrary declared evidence actions. Its conclusion continues to apply when every admissible action satisfies the required no-hit progress and reachability bounds. Extra actions can change the exact optimum: a learner-neutral certifying route can remove the strict ceiling, while another action can create a

cheaper path to the barrier. The contextual-logistic and Bradley–Terry constructions satisfy the same learner-side requirements outside the canonical geometry. Their exact law generally differs from ρ^d.

These cases separate the stable statement from the model-specific one. The need to reach a rare learner state can survive perturbations; the Bernoulli staircase and the closed-form value ρ^d belong to the canonical realization.

E Information-order details

E.1 Semantic ordering and depth crossing

The family used in Section 5.2 has $\delta_S(c) = c\delta_E + \sqrt{1-c^2}\,\delta_0$ with $0 < \delta_0 < \delta_E$. Its derivative is $\delta_S'(c) = \delta_E - c\delta_0/\sqrt{1-c^2}$. Here $c_{\text{info}} = (\delta_E^2 - \delta_0^2)/(\delta_E^2 + \delta_0^2)$ and $c_{\text{peak}} = \delta_E/\sqrt{\delta_E^2 + \delta_0^2}$. The derivative is positive below c_{peak}, while $\delta_S(c_{\text{info}}) = \delta_E$. On $(c_{\text{info}}, c_{\text{peak}})$, $p_S(c) > p_E$ and increases with c. For symmetric Bernoulli experiments, a world-independent bit flip maps any channel with parameter $p_2 > p_1 > 1/2$ to the weaker p_1 channel, establishing strict Blackwell dominance.

Let $\Gamma(c) = c\delta_S(c)/(1-c^2)$. Direct differentiation gives

$$\Gamma'(c) = \frac{2\delta_E c}{(1-c^2)^2} + \frac{\delta_0}{(1-c^2)^{3/2}} > 0,$$

so $C'(c) < 0$. If the integer depth increases between c_1 and c_2, Theorem 3 gives the strict reversal in Corollary 1.

For $p_E = 0.7$, $\delta_0 = \text{logit}(0.6)$, $b = 0.5$, $\lambda = 1$,

$$p_S(0.65) = 0.702, \quad C(0.65) = -0.467, \quad d(0.65) = 1,$$
$$p_S(0.67) = 0.704, \quad C(0.67) = -0.556, \quad d(0.67) = 2.$$

The resulting values are 0.429 and 0.184. The stronger channel can be garbled to the weaker one by flipping its binary output with probability 0.005. The first depth crossing on this information-increasing branch solves $C(c) = -1/2$ at $c = 0.658$. $c = 0.65$ lies inside a nonzero depth-one interval and $c = 0.67$ lies inside a nonzero depth-two interval. The reported reversal persists under small perturbations of either witness.

E.2 Downstream-preserving simulation

Proof of Proposition 2. Take any feasible base policy. In the richer experiment, apply the world-independent garbling to obtain the projected base history and ignore the residual signal. Action and stopping inclusion reproduce the base control. The copied projected history reproduces the downstream learner law. Copying the deployment actionability status and certificate/risk coefficients reproduces the same success and feasibility events. The richer experiment therefore contains an outcome-equivalent copy of every base feasible policy. Taking the supremum proves the claim. □

The learner-state and deployment-status conditions are both needed. A richer signal can change a history-conditioned deployment status even without learner updating; a learner update can also change actionability while leaving the semantic channel ordered. The proposition covers only refinements that keep the downstream decision problem fixed.

F Matched-transcript comparator

In the terminal binary specialization, the matched-transcript fallback receives the same queries, labels, learner updates, internal randomization, and stopping time as the advice policy, and changes only the terminal action. If S_π is the event that the advice policy selects advice while the current fallback selects item 2, the beneficial-world pathwise gain is 1 on S_π and 0 otherwise. Unit binary loss gives

$$R_+(\pi) - R_+^{\text{F}}(\pi) = -P_+(S_\pi), \qquad R_-(\pi) - R_-^{\text{F}}(\pi) = P_-(S_\pi).$$

The harmful-world excess-risk budget is exactly the harmful advice-selection budget. This identity links the certificate formulation to the matched-transcript risk contract used in the canonical phase.

G CONTINUOUS BENCHMARK DETAILS

The benchmark uses dimension $D = 5$, $\|\theta\|^2 = D$, Gaussian observations $r_t = \theta + \epsilon_t$, and advice $a = \theta + \sqrt{qD}\, v$ with $q \sim \mathrm{Unif}(0.05, 0.80)$ and a random unit vector v. The baseline learner uses exposure parameter $\lambda_L = 0.15$ and $t_0 = 5$. The reliability certificate is a time-uniform Gaussian confidence-ball test of a fixed-reference advice gain of at least 0.5 at level $\alpha = 0.01$. True actionability requires advice to beat the current learner by margin 0.2 under squared loss.

Let $g_t = \lambda_L/(t + t_0)$. The learner update is $x_t = (1 - g_t)x_{t-1} + g_t r_t$ with $x_0 = 0$. Its error satisfies

$$x_t - \theta = P_t(x_0 - \theta) + \sum_{s=1}^{t} w_{s,t}\epsilon_s, \qquad P_t = \prod_{j=1}^{t}(1 - g_j), \qquad w_{s,t} = g_s \prod_{j=s+1}^{t}(1 - g_j).$$

The covariance is $V_t I_D$, where $V_0 = 0$ and $V_t = (1 - g_t)^2 V_{t-1} + g_t^2$. Conditional on q, $\|x_t - \theta\|^2/V_t$ is noncentral-χ^2_D, giving an analytic actionability probability. At horizons $t = 50, 100, 200, 400, 800, 1200, 2000$, the largest absolute discrepancy between the analytic law and 10,000 Monte Carlo paths is 0.007.

The confidence ball uses the spending sequence $\gamma_t(\gamma) = 6\gamma/(\pi^2 t^2)$ and radius

$$\rho_t(\gamma) = \frac{1}{\sqrt{t}}\sqrt{\chi^2_{D,1-\gamma_t(\gamma)}},$$

with unit observation noise. Let $\bar{r}_t = t^{-1}\sum_{s \le t} r_s$. Certification occurs when

$$2a^\top \bar{r}_t - \|a\|^2 - 2\|a\|\rho_t(\alpha) \ge 0.5.$$

Observable actionability uses the same ball with $\gamma = \beta$. Its lower bound is $\underline{\Delta}_t = \|x_t\|^2 - \|a\|^2 + 2(a - x_t)^\top \bar{r}_t - 2\|a - x_t\|\rho_t(\beta)$. On simultaneous coverage, $\underline{\Delta}_t \le \Delta_t$ at every inspected time. With $\beta = 0.05$, observable certified-and-actionable probability is 0.073 at $T = 400$ and 0.079 at $T = 2000$.

The learner-exposure sweep uses common paths and changes only λ_L. Across exposure levels 0 through 0.30, certification stays 0.677 while true certified-and-actionable probability falls from 0.677 to 0.195.

The timing intervention in Figure 1(c) uses a separate fixed-budget update rule. It fixes $T = 400$, $N = 6000$, the full evidence transcript, and exactly $K = 120$ fallback updates per arm. The u-th consumed learner observation uses

$$\eta(u) = \frac{0.65}{u + 5}, \qquad x \leftarrow x + \eta(u)(r_t - x),$$

so each arm receives the same update-step sequence. For delay index $j = 0, \ldots, 10$, the contiguous update block starts at zero-based round $28j$ and has length 120. Timing depends only on j and the round number. Certification paths are identical across all arms.

Delay	Update rounds	$P(\text{cert})$	$P(\text{cert} \cap A)$	Final regret
0.0	1–120	0.666	0.006	0.097
0.1	29–148	0.666	0.122	0.097
0.2	57–176	0.666	0.238	0.098
0.3	85–204	0.666	0.335	0.099
0.4	113–232	0.666	0.401	0.098
0.5	141–260	0.666	0.451	0.097
0.6	169–288	0.666	0.492	0.097
0.7	197–316	0.666	0.526	0.098
0.8	225–344	0.666	0.553	0.097
0.9	253–372	0.666	0.577	0.098
1.0	281–400	0.666	0.597	0.098

The final-regret range is narrow despite the large change in certificate-time actionability. The sweep therefore changes update timing rather than the evidence transcript, update count, or the learner's update-step sequence.

H DIGITS PROTOCOL

We use the 1,797 examples in the optical Digits data set and divide pixel values by 16. Each of 100 deterministic stratified partitions allocates 400 examples to advice training, 80 to fallback initialization, 400 to the certification pool, 500 to an actionability-monitor pool, and 417 to a final test set. The stratified partitions are fixed deterministically before comparing update arms. The advice model is multinomial logistic regression with $C = 1$. The fallback is an averaged online logistic-regression learner initialized on the 80 fallback examples.

A length-600 certification stream is sampled with replacement from the fixed certification pool using prespecified randomness. The same sampled indices and labels are used in every update arm. A Bernoulli likelihood-ratio e-process compares $p_1 = 0.97$ with the composite null $p \leq p_0 = 0.90$ and crosses at $1/\alpha$ with $\alpha = 0.01$. The update arms use no certification examples, every fourth example, or every example. Within a round, the selected learner update is applied and then the common certificate statistic is updated.

True actionability at first certification requires the advice test accuracy to exceed current fallback test accuracy by at least 0.08. The monitor proxy uses paired correctness differences on the independent 500-example monitor split and declares actionability when $\bar{z} - 1.96\,\mathrm{SE}(z) \geq 0.08$. Across the 100 partitions, mean advice test accuracy is 0.943 and mean initial fallback test accuracy is 0.718. Certificate crossing probability is 0.59 in all arms, with median crossing time 182 among crossings. Across the 59 certifying partitions, the paired held-out minus immediate advice-gain difference is 0.127 with standard error 0.007 and interval $[0.113, 0.140]$. The partitions share the same underlying dataset; these summaries describe partition-level variation.

I MODEL PICKER WITH AN ANYTIME-VALID DESIGN-BASED CERTIFICATE

We evaluate Model Picker with an anytime-valid level-α reliability certificate under the adaptive query rule (Karimi et al., 2021), rather than using hindsight best-model identification as an evidence-resolution criterion. This provides an external validation in which Shared and Frozen realizations have identical evidence and certificate paths; only fallback reuse differs.

I.1 PROTOCOL AND RELIABILITY CLAIM

We use the official CIFAR-10 prediction matrix with 80 fixed classifiers. A fixed split allocates 2,000 examples to adviser selection, 2,000 to an actionability monitor, and 6,000 to a fixed active certification population. The selection split chooses model 79. Its accuracies are 0.916 on selection, 0.919 on the active population, and 0.921 on the monitor. The closest active-population competitor is model 78, which trails the adviser by 0.004 accuracy. The Frozen fallback excludes the adviser and has monitor accuracy 0.897.

For competitor k, define the fixed active-population loss gap

$$\mu_k = E_{\mathrm{pool}}[\ell_k - \ell_a].$$

The reliability claim is $\mu_k > -0.02$ for every $k \neq a$: the selected adviser is within two percentage points of the best fixed classifier on the active population. The primary confidence level is $\alpha = 0.05$; $\alpha = 0.01$ uses the same trajectories as a stricter sensitivity analysis.

At wall-clock round t, an active-population index I_t is drawn uniformly with replacement. Let $\mathcal{G}_t = \sigma(\mathcal{F}_{t-1}, I_t)$. Model Picker observes the 80 predictions at I_t and chooses q_t with its published scale-1 rule, so q_t is $\mathcal{G}_t$-measurable. Conditional on $\mathcal{G}_t$, let $Q_t \sim \mathrm{Ber}(q_t)$ be the query indicator and define

$$D_t^k = \ell_k(I_t) - \ell_a(I_t) \in \{-1, 0, 1\}, \qquad X_t^k = \frac{Q_t D_t^k}{q_t}.$$

When $q_t = 0$, all classifiers agree, so $D_t^k = 0$ for every pair and we set $X_t^k = 0$. Under this design, $E[D_t^k \mid \mathcal{F}_{t-1}] = \mu_k$ and $E[X_t^k - D_t^k \mid \mathcal{G}_t] = 0$, hence $E[X_t^k \mid \mathcal{F}_{t-1}] = \mu_k$.

The pairwise e-process uses a fixed $\lambda = 0.025$. Write

$$\psi(\lambda; c, v) = \frac{\lambda^2 v}{2(1 - \lambda c/3)}.$$

For pair k, let r_k be the fraction of active-population predictions on which k disagrees with the adviser. The population-draw term $D_t^k - \mu_k$ has one-sided bound 2 and conditional variance at most r_k given $\mathcal{F}_{t-1}$. On pair-disagreement rounds, the query-noise term $X_t^k - D_t^k$ has one-sided bound $c_t = \max\{1, 1/q_t - 1\}$ and conditional variance at most $(1 - q_t)/q_t$ given $\mathcal{G}_t$; both are zero on agreement rounds. Model Picker gives $q_t \geq \sqrt{\log(80)/(2t)}$ on disagreement rounds. Through 25,000 rounds this keeps $\lambda c_t/3 < 1$. Let $\Psi_{k,t}$ be the sum of the two Bernstein–Freedman penalties ψ over rounds. For

$$H_{0,k} : \mu_k \leq -0.02, \qquad E_{k,t} = \exp\left\{ \lambda \sum_{s \leq t} (X_s^k + 0.02) - \Psi_{k,t} \right\}$$

is a nonnegative e-process. A pair is permanently passed after its first crossing of $1/\alpha$. The adviser is certified when every competitor pair has passed. If the global claim is false, at least one pair null is true; global certification forces that true-null e-process to cross $1/\alpha$. Ville's inequality therefore gives global anytime type-I error at most α without a multiplicity correction.

The fixed horizon is 25,000 wall-clock rounds, with 200 realizations under prespecified randomization. Results are reported at 100, 200, 300, 400, 500, 600 queried labels. Larger label counts are omitted because they are not attained uniformly across realizations within the fixed wall-clock horizon.

I.2 Shared versus Frozen fallback reuse

The Shared fallback is a Hedge ensemble over the 79 non-adviser classifiers with learning rate 0.25 and updates on every queried label. Frozen receives no queried-label updates. Within each realization, the two arms share sampled indices, query coins, queried labels, Model Picker weights, and every pairwise e-process. The only difference is whether the queried label updates the fallback. Actionability is evaluated on the disjoint monitor by the zero-margin event that adviser accuracy exceeds current fallback accuracy.

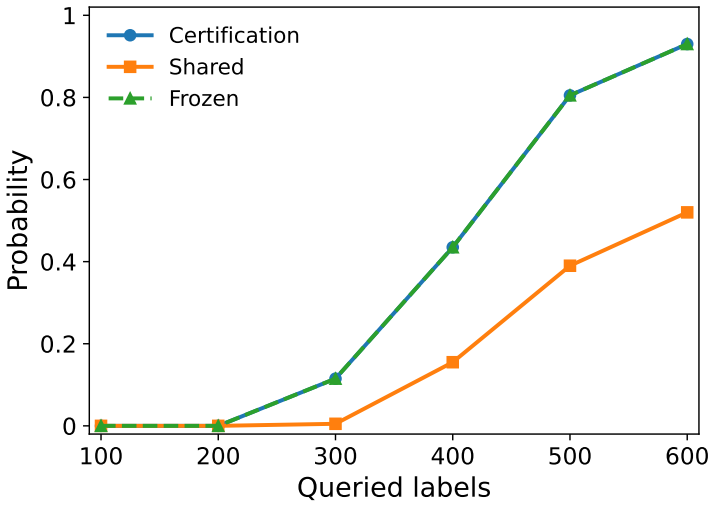


Figure 3: Anytime-valid Model Picker certificate on the official CIFAR-10 prediction matrix at $\alpha = 0.05$. Certification is identical in Shared and Frozen. At 500 queried labels, 0.805 of realizations are certified; the certified-and-actionable probability is 0.390 in Shared and 0.805 in Frozen. At 600 labels the corresponding values are 0.930, 0.520, and 0.930.

Across the 200 realizations, certificate probability rises from 0.115 at 300 queried labels to 0.805 at 500 and 0.930 at 600. Shared actionability at those queried-label counts is 0.510, 0.540, and 0.565. The joint probability is only 0.005, 0.390, and 0.520, while Frozen joint success equals certification because its fallback remains below the adviser.

The first-certificate view removes dependence on a chosen evaluation point. By the fixed horizon, the $\alpha = 0.05$ certificate appears in $199/200 = 0.995$ realizations. Its median arrival is 420 queried labels, with interquartile range 356.5–483. At the first valid certificate, Frozen certified-and-actionable success is 0.995 and Shared success is 0.365. The paired loss fraction is 0.630 with Wilson 95% interval $[0.561, 0.694]$. Among certifying realizations, the adviser remains actionable in Shared only 0.367 of the time. Mean Shared fallback monitor accuracy at first certification is 0.927, above the fixed adviser's 0.921.

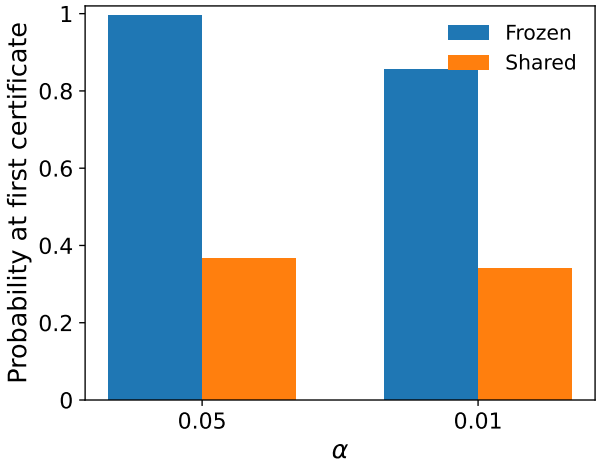


Figure 4: Certified-and-actionable success at the first anytime-valid certificate. The Shared and Frozen certificate times are identical within each realization. Tightening the certificate from $\alpha = 0.05$ to 0.01 delays arrival but leaves a large reuse gap.

With $\alpha = 0.01$, the certificate appears by the fixed horizon in 0.855 of realizations. Median arrival is 557 queried labels. Shared success at first certification is 0.340 versus 0.855 in Frozen, a gap of 0.515 with Wilson 95% interval $[0.446, 0.583]$.

I.3 Validity safeguards

The construction preserves the published Model Picker query rule, exponential-weights update, inverse-probability identity, zero-query agreement convention, Bernstein-domain condition, split separation, and Hedge fallback definition. As a negative-control validation, a false 0.02-optimality claim yielded no global certificates in 500 independent null realizations at either $\alpha = 0.05$ or 0.01. The monitor labels are used only for offline actionability evaluation; they never enter Model Picker, the e-process, or fallback updates.

Separate controlled interventions on the same Model Picker asset are consistent with this certified analysis. Under those interventions, stronger reuse lowers identified-and-actionable overlap while holding the acquisition trace fixed, and delaying fallback reuse recovers the overlap. These comparisons isolate update strength and timing. Figures 3–4 establish the stronger point needed here: the evidence side can be a genuine anytime-valid certificate rather than hindsight identification.