SAFE POLICY SELECTION: WHEN TO STOP AND WHAT IT COSTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Testing a trained policy can reveal whether it is safe, but the test itself can cause a safety violation. We study how to search a fixed policy library with predetermined preference scores under a per-episode risk limit. We combine candidate tests with known-safe fallback episodes, guided by risk bounds that remain valid throughout testing. Joint Anytime Certification and Exploration (JACE) certifies a safe recommendation with high probability. Its score is at least that of the fallback and every candidate whose risk lies below the limit by a prescribed margin. We separate the evidence needed for this decision from the extra episodes required to gather it safely. Upper and worst-case lower bounds from the same baseline-only start match up to logarithmic factors. On PointGoal1, JACE uses 37% fewer episodes on average than matched safe round-robin allocation to reach its stopping condition. Controlled experiments show how prioritization gains depend on whether search can focus on a promising candidate or must collect exclusion evidence for every alternative.

1 INTRODUCTION

A library of trained controllers creates two linked questions: which policy can be recommended safely, and when is there enough evidence to stop testing? Each evaluation episode can itself cause a safety violation. Moreover, finding one safe policy does not resolve whether an untested alternative with a higher score should become the new deployment.

Safe exploration and verifiable recommendation are established goals [23, 30, 31, 27]. In our setting, policies and preference scores are fixed before collection. The uncertainty concerns risk, not an unknown reward ranking. The challenge is to obtain enough evidence for the final comparison while controlling the conditional risk of each collection round. Evidence requirements and safe sampling rates must therefore be analyzed together.

JACE couples these decisions through risk bounds that remain valid throughout adaptive testing [7, 32]. It mixes candidate episodes with a known-safe baseline and increases candidate sampling as the bounds tighten. Frozen utilities identify potentially better policies, while a prescribed risk margin determines which alternatives must be ruled out. The resulting stopping condition can be checked without knowing the true risks.

The margin avoids exact classification at the safety boundary. With a 15% risk limit and a two-percentage-point margin, the output must be safe at 15% and match the baseline and all candidates with risk at most 13%. A better policy in the 13–15% band can be omitted, so the margin does not bound utility loss relative to the best safe policy.

Contributions. Our contributions are threefold:

(1). A jointly safe stopping certificate. We introduce JACE and prove joint collection and deployment safety. Its observable stopping rule certifies a safe recommendation whose frozen utility matches or exceeds that of the baseline and every margin-safe candidate. The same rule justifies returning the baseline when no qualifying improvement exists.

(2). Matching bounds for the cost of safe evaluation. From a baseline-only start, we bound total collection cost by local evidence requirements and certificate-dependent waiting. A lower

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.

bound with identical initial risk information matches the worst-case parameter dependence up to logarithmic factors, showing that neither cost can generally be removed.

(3). Evidence for when prioritization helps. On PointGoal1, JACE reduces mean completion cost by 37% relative to matched safe round-robin allocation. CarPush1 separates early certification from the cost of resolving higher-scoring alternatives. Controlled and boundary experiments identify when prioritization helps and how the risk margin affects utility.

2 Problem Setup

Using independent development data, freeze a library $\Pi = \{\pi^0, \dots, \pi^K\}$ with $K \geq 1$ nonbaseline candidates and preference scores $u_0, \dots, u_K$, called utilities. For estimated utilities, the guarantee concerns frozen scores; expected-return guarantees require utility-error control (Corollary A.4). Executing policy j for a complete episode produces a catastrophe indicator $X \sim \text{Bernoulli}(p_j)$. In an episodic control problem, $X = \mathbf{1}\{\sum_{h=1}^{H} C_h > c_{\max}\}$ for nonnegative costs C_h. The fallback, called the baseline π^0 below, has an upper risk bound $0 \leq \beta_0 < \beta < 1$, obtained structurally or by independent calibration with failure budget δ_0. Write $s = \beta - \beta_0$ for its safety slack and $N = K+1$.

Assumption 2.1 (Observation model)**.** Policies and episode laws are fixed. Each policy has an infinite independent potential stream of Bernoulli observations. Adaptive execution reveals its next unused observation. Policy selection uses fresh randomization independent of future observations. All utilities, certificate settings and filtering thresholds are frozen before online collection.

At round t, history $\mathcal{F}_{t-1}$ determines two distinct choices: the recommended deployment d_t and collection weights w_t. Only the sampled policy $J_t \sim w_t$ is executed, for one complete episode. The recommendation incurs no additional execution. The conditional collection risk is

$$r_t := \mathbb{P}(X_t = 1 \mid \mathcal{F}_{t-1}) = \sum_{j=0}^{K} w_{t,j} p_j. \tag{1}$$

Definition 2.2 (Sampling and deployment safety)**.** The procedure is jointly safe at level $1 - \delta$ if $\mathbb{P}(\forall t \geq 1 :\ p_{d_t} \leq \beta,\ r_t \leq \beta) \geq 1 - \delta$.

Collection safety limits the mixture risk before sampling, not the risk of every candidate that might be drawn. Deployment safety requires the recommended policy itself to be safe. Completion adds a score guarantee relative to the comparison class defined next.

For a prescribed margin $0 < \eta < \beta$, set $\theta = \beta - \eta$. The best utility in the comparison class is $u_\eta = \max\big(\{u_0\} \cup \{u_j : p_j \leq \theta,\ j \geq 1\}\big)$. A correct completion declaration outputs d satisfying $p_d \leq \beta$ and $u_d \geq u_\eta$. The comparator has margin η, whereas the output need only satisfy the original risk limit. Completion differs from first safe identification: it certifies that no margin-safe candidate improves on the output and permits returning the baseline when no such improvement exists. It does not require exact classification of every candidate.

Choosing the margin. Fix the comparison threshold before collection; η is its distance below the deployment risk limit. Smaller margins include more policies but can require more data. Without assumptions linking risk and utility, η cannot bound utility loss against the full safe set. Configure risk bounds as in (7).

3 Joint Anytime Certification and Exploration

JACE combines time-uniform risk bounds, certificate-constrained sampling, and an observable stopping test. We first describe safe allocation, then candidate selection and stopping.

3.1 Risk bounds and safe sampling

For policy j, let $L_j(n)$ and $U_j(n)$ denote the risk bounds after n observations. They change only when that policy is sampled. On a grid $\mathcal{G}$ containing n_g points and $0, \beta, 1$, mix the stakes $\lambda_m =$

2^{-m-1} for integers $m = 0, \ldots, m_{\max}$ using positive weights ω_m that sum to one:

$$K_n^{\mathrm{mix}}(q) = \sum_{m=0}^{m_{\max}} \omega_m \prod_{i=1}^{n} [1 + \lambda_m(q - X_i)]. \tag{2}$$

For $q \le p_j$, each product is a nonnegative supermartingale. The upper certificate hedges this mixture with predictable aGRAPA evidence [32], each with weight $1/2$. A grid point is permanently rejected once its capital crosses $1/\varepsilon_j^U$. The upper bound $U_j(n)$ is the successor of the largest unrejected point, capped at one. The lower certificate applies the geometric mixture to $1 - X_i$ on the reflected grid and sets $L_j(n) = 1 - U_j^{\mathrm{ref}}(n)$.

Fix $0 < \delta < 1$ and $0 \le \delta_0 < \delta$. Before collection, allocate positive certificate budgets with $\delta_0 + \sum_{j=0}^{K}(\varepsilon_j^U + \varepsilon_j^L) \le \delta$. Omit unused baseline lower budgets. Ville's inequality gives joint coverage across local counts and adaptive rounds without a grid union bound. Write $\mathcal{E}_{\mathrm{cov}}$ for this event and cap U_0 at β_0. Appendix A.2 gives the updates and proof.

Safe allocation uses these upper bounds. Let A_t contain candidates not permanently eliminated by $L_j > \beta$, each receiving the following floor to ensure continued sampling:

$$w_{\min} = \rho \min\left\{\frac{1}{K+1}, \frac{s}{K(1-\beta_0)}\right\}, \qquad 0 < \rho < 1. \tag{3}$$

If priority is assigned to h, all other active candidates receive $w_{\min}$ and the baseline receives the remainder. JACE maximizes $w_{t,h}$ subject to the baseline floor and $w_t^\top U(t-1) \le \beta$. With $m = |A_t|$, $b = U_0(t-1)$ and $v = U_h(t-1) > \beta$, this has the closed form

$$w_{t,h} = \min\left\{1 - m w_{\min}, \frac{\beta - b - w_{\min}\sum_{j \in A_t \setminus \{h\}}(U_j - b)}{v - b}\right\}. \tag{4}$$

Eliminated candidates receive no mass. Equation (3) ensures feasibility and yields the following minimum probability of sampling the priority candidate:

$$w_{t,h} \ge \max\left\{w_{\min}, \frac{(1-\rho)s}{U_h(t-1) - \beta_0}\right\}. \tag{5}$$

Evidence therefore supports both recommendation and more efficient collection. Section 4 accounts for the resulting reduction in baseline overhead when bounding completion cost.

3.2 Candidate Selection and Stopping

At each decision, the eligible set is $I_t = \{0\} \cup \{j \in A_t : U_j \le \beta\}$. Deploy its largest-utility policy, breaking ties by the smallest index. The unresolved higher-scoring candidates form $V_t = \{j \in A_t : U_j > \beta,\ u_j > u_{d_t},\ L_j \le \theta\}$.

The priority set V_t contains candidates that remain uncertified, have higher utility than the current recommendation, and have not been screened out of the margin comparison. Exclusion from V_t is not permanent elimination from A_t: active candidates still receive their sampling floors. Initially $d_1 = 0$, $U_j = 1$ and $L_j = 0$, so only utility distinguishes candidates and $V_1 = \{j : u_j > u_0\}$.

For nonempty sorted $V_t = \{v_1 < \cdots < v_{|V_t|}\}$, take the rotating reference $\kappa_t = v_{1+((t-1) \bmod |V_t|)}$. Among candidates in V_t with $U_j \le U_{\kappa_t}$, prioritize the largest threshold log-capital. When V_t is empty, *baseline fallback* assigns active candidates their floors and the remaining mass to the baseline. Appendix A.5 describes which candidates these filters retain.

The resulting recommendation satisfies the observable completion condition precisely when every higher-utility candidate has been screened beyond the comparison threshold:

$$\mathsf{Q}_t: \quad d_t \in I_t \quad \text{and} \quad L_j(t-1) > \theta \ \text{ for every } u_j > u_{d_t}. \tag{6}$$

The equivalence $\mathsf{Q}_t \iff V_t = \emptyset$ holds pathwise. Algorithm 1 records completion but continues collection for fixed-window evaluation. Applications may stop immediately instead. Strict-identification experiments follow a separate continuation rule described in Appendix B.6.

Algorithm 1 JACE: Joint Anytime Certification and Exploration

Input: frozen library Π, utilities $u_0, \dots, u_K$, risk limits $\beta_0 < \beta$, threshold $\theta = \beta - \eta$, floor $w_{\min}$, precommitted certificates
Initialize $L_j \leftarrow 0$, $U_j \leftarrow 1$ for every $j = 0, \dots, K$, set $U_0 \leftarrow \beta_0$ and $A_1 \leftarrow \{1, \dots, K\}$
for $t = 1, 2, \dots$ **do**
  // Recommend a certified policy
  $I_t \leftarrow \{0\} \cup \{j \in A_t : U_j \le \beta\}$
  $d_t \leftarrow \min \arg\max_{j \in I_t} u_j$
  Deploy d_t as the recommendation without executing an additional episode
  // Allocate safe collection or declare completion
  $V_t \leftarrow \{j \in A_t : U_j > \beta,\ u_j > u_{d_t},\ L_j \le \theta\}$
  Set $w_{t,j} \leftarrow 0$ for eliminated candidates $j \notin A_t \cup \{0\}$
  if $V_t \neq \emptyset$ **then**
    $\kappa_t \leftarrow$ rotating index in sorted V_t
    $h \leftarrow \min \arg\max_{j \in V_t : U_j \le U_{\kappa_t}} \log K^{j,\mathrm{hed}}(\beta)$
    Set w_t by (4), retaining floors on all other active candidates
  else
    Record completion with recommendation d_t
    $w_{t,j} \leftarrow w_{\min}$ for $j \in A_t$
    $w_{t,0} \leftarrow 1 - |A_t| w_{\min}$
  end if
  // Execute one collection episode and update its certificates
  Draw $J_t \sim w_t$ and execute one complete episode of π^{J_t}
  Update only the certificates of J_t
  $A_{t+1} \leftarrow \{j \in A_t : L_j \le \beta\}$
end for

Theorem 3.1 (Safety and valid stopping)**.** *Under Assumption 2.1 and the stated budget allocation, Algorithm 1 is jointly safe. On one event of probability at least* $1 - \delta$*, every declaration satisfying* Q_t *has* $p_{d_t} \le \beta$ *and* $u_{d_t} \ge u_\eta$*. From the first such declaration onward, baseline-fallback collection satisfies* $r_t \le \beta_0 + \rho s < \beta$*.*

Appendix A.4 proves the result by showing that valid lower certificates cannot exclude a higher-utility margin-safe policy. On joint coverage, certified policies remain active and deployment utility is nondecreasing, so an empty priority set stays empty. Time-uniform coverage justifies repeated checks without spending additional confidence budget.

4 The Cost of Safe Search

How many episodes are needed to reach a valid stopping decision? Define $\tau_Q = \inf\{t \ge 1 : \mathsf{Q}_t\}$ and $T_Q = \tau_Q - 1$, including baseline episodes. For comparison, τ_0 marks first nonbaseline certification, and τ_{dec} denotes a general algorithm's declaration decision. These endpoints differ: if no candidate has higher utility, the baseline already satisfies the objective without further testing.

Handling the safety boundary. Let h be the maximum grid spacing and $\lambda_{\min}$ the smallest geometric stake. Use a stored grid point θ, its strict successor $q_\theta = \min\{q \in \mathcal{G} : q > \theta\}$, and

$$\eta_{\mathrm{eff}} = \beta - q_\theta > 0, \qquad h \le \eta_{\mathrm{eff}}, \qquad \lambda_{\min} \le \eta_{\mathrm{eff}}/10. \tag{7}$$

For each candidate j, write $g_j = \max\{\beta - p_j, p_j - q_\theta\} \ge \eta_{\mathrm{eff}}/2$, $H_j = p_j(1 - p_j) + g_j$, and $R_j = H_j/g_j^2$. The gap g_j measures separation from at least one of two exits: certification at β or screening beyond θ. Their overlap keeps this gap positive even when $p_j = \beta$, as Figure 1 illustrates. Thus completion does not require separation from the deployment threshold alone. The strict successor implements $L_j > \theta$. A sufficient grid includes both thresholds and satisfies $\eta \ge 2h$.

Expected total completion cost. Let $I_+ = \{j \ge 1 : u_j > u_0\}$ and $K_+ = |I_+| > 0$. Write $\varepsilon_{\min}$ and $\omega_{\min}$ for the smallest positive certificate budget and geometric weight. Set $R_{\max} = \max_{j \in I_+} R_j$ and $\Lambda = \log \frac{16 n_g}{\varepsilon_{\min} \omega_{\min}} + 2 \log(e R_{\max})$.

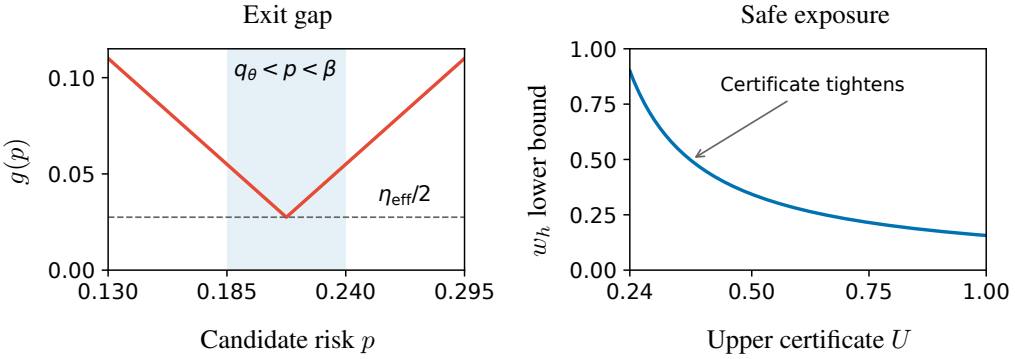


Figure 1: **Overlapping thresholds keep the exit gap positive, while tighter certificates permit more candidate sampling.** The left panel shows the exit gap, including at $p = \beta$. The right shows the priority-probability lower bound in (5). Here $\beta = 0.24$, $q_\theta = 0.185$, $\beta_0 = 0.08$, $\rho = 0.1$.

Theorem 4.1 (Expected completion cost)**.** *Under Assumption 2.1 and (7), Algorithm 1 satisfies the following bound for every admissible frozen configuration:*

$$\mathbb{E}T_Q \le \frac{C\Lambda}{1-\rho}\left[\underbrace{\sum_{j\in I_+}\frac{H_j}{g_j^2}}_{\text{evidence acquisition}} + \underbrace{\frac{1}{s}\sum_{j\in I_+}\left(\frac{H_j}{g_j}+\log(eR_j)\right)}_{\text{safety-induced waiting}}\right] \tag{8}$$

Here C is universal, and the expectation includes certificate-failure paths. If $K_+ = 0$, the initial decision completes the search with $\tau_Q = 1$ and $T_Q = 0$.

The proof is in Appendix A.8. The bound covers the full run from a baseline-only start, including instances without a safe improvement. It holds for each frozen configuration. Appendix A.1 specifies the additional integrability condition for averaging over random development or calibration settings.

From observations to evaluation cost. Let $M_j = \inf\{n : U_j(n) \le \beta \text{ or } L_j(n) > \theta\}$ be the local count needed to resolve candidate j, and let A_∞ count nonempty-priority rounds. Every priority candidate lies in I_+ with count $n < M_j$. Conditional on all potential observation streams, each priority attempt samples it with probability at least $(1-\rho)s/[\max\{\beta, U_j(n)\} - \beta_0]$. Sampling increases the count permanently, and floor samples only shorten the wait, giving the total bound

$$\mathbb{E}T_Q \le \mathbb{E}A_\infty \le \frac{1}{(1-\rho)s}\sum_{j\in I_+}\mathbb{E}\sum_{n<M_j}[\max\{\beta, U_j(n)\} - \beta_0]. \tag{9}$$

Equation (9) converts local evidence requirements into collection time by charging observations at their current certificate states. Combining local exit tails with variance-sensitive widths yields (8). A fixed initial rate would instead apply the initial safety penalty to every required observation. Since $H_j \le \beta + 2g_j$, the explicit configuration in Corollary A.7 gives

$$\mathbb{E}T_Q = \widetilde{O}\left(\frac{K_+}{1-\rho}\left[\frac{\beta}{\eta_{\text{eff}}^2} + \frac{\beta}{s\eta_{\text{eff}}}\right]\right). \tag{10}$$

The first term captures risk discrimination, while the second captures the additional cost of limited safe exposure. With fixed grid, stakes and instance, and budgets proportional to δ/K, dependence on $\log(1/\delta)$ is linear. The bound covers any predictable priority within V_t retaining the same certificates, filters, floors and allocation rule. Evidence ordering is evaluated empirically against matched priorities.

A computable cost bound. Allocation and stopping use observable certificates, not the unknown risks in the analysis. Replacing each R_j by $\overline{R} := \eta_{\text{eff}}^{-2} + 2\eta_{\text{eff}}^{-1}$ gives the following computable bound with an explicit, though loose, constant.

Corollary 4.2 (Bound using known parameters)**.** *Under the conditions of Theorem 4.1 with $K_+ > 0$, set $L_0 = \log[4/(\varepsilon_{\min}\omega_{\min})]$. The expected total completion cost then satisfies*

$$\mathbb{E}T_Q \le 1 + \frac{512\overline{R}}{w_{\min}}\left[L_0 + \log(2K_+) + 1\right]. \tag{11}$$

Appendix A.11 obtains this explicit constant from sampling floors. Like (8), this bounds total cost from initialization, not remaining time conditional on an online history. Experiments test the scaling trends rather than the unspecified constant in (8).

4.1 A MATCHING LOWER BOUND

Are both cost components unavoidable? Every method starts with one structural baseline of risk β_0 and utility zero, and M unobserved utility-one candidates. On P^0 all candidates have risk $\beta+\eta$, whereas on P^j only candidate j has risk $\beta-\eta$. Initial information is identical, and one-step Bernoulli feedback supplies all new risk information. Set $d_\eta = \mathrm{kl}(\beta+\eta, \beta-\eta)$ for this comparison.

Theorem 4.3 (Safe completion lower bound)**.** *Let $0 < \beta < 1/2$, $0 < \eta \le \beta/2$, $0 \le \beta_0 < \beta$ and $0 < \delta < 1/4$. Any method that, on every instance above, jointly ensures safe collection before declaration and a finite declaration with $p_d \le \beta$, $u_d \ge u_\eta$ with probability at least $1-\delta$, satisfies*

$$\mathbb{E}^0(\tau_{\mathrm{dec}} - 1) \ge \frac{M(s+\eta)\mathrm{kl}(1-\delta,\delta)}{s\, d_\eta} \ge \frac{M}{128}\left[\frac{\beta}{\eta^2} + \frac{\beta}{s\eta}\right]\log\frac{1}{\delta}. \tag{12}$$

The proof is in Appendix A.14. Under P^0, safe collection limits total candidate exposure to $s/(s+\eta)$. A correct no-improvement declaration must rule out every P^j, requiring a confidence logarithm for each candidate. The stopped-likelihood argument couples these information requirements to the shared exposure cap, yielding both cost terms in (12) [13, 26].

Corollary A.7 specifies grid, stakes, weights and budgets with logarithmic configuration overhead. For $K_+ = K = M$ in the range of Theorem 4.3, the bounds share a baseline-only start and match worst-case parameter scales up to logarithmic factors. The information conditions are detailed in Appendix A.9. For the distinct goal of finding any safe nonbaseline policy, Appendix C introduces the multiscale procedure JACE-MS and gives separate search-time bounds.

5 EXPERIMENTS

Experiments compare completion costs, scaling and identification; Appendix B.5 examines margin sensitivity.

5.1 EXPERIMENTAL SETUP

Libraries and settings. PointGoal1 and CarPush1 each use 16 OSRL policies [19] and ten paired runs. The synthetic library has six policies and 200 paired seeds. For these libraries, we fix $\delta = 0.05$, $\rho = 0.1$, $\eta = 0.02$, $T = 40{,}000$ and $\beta = 0.15$, except $\beta = 0.20$ on CarPush1. Control development, assessment and online episodes are independent. Reported costs measure additional online collection after library preparation and baseline qualification.

Metrics and methods. The primary metric is capped completion cost $\min\{T_Q, T\}$, including baseline episodes. A valid completion outputs a safe policy with utility at least u_η; we check this against known synthetic risks or independent control-policy assessment intervals. JACE, JACE-UCB, Safe-RR and Safe-Uniform share certificates, filters, floors and stopping. Their priorities use evidence ordering, minimum risk lower bounds, rotation and feasible uniform allocation.

5.2 COMPLETION EFFICIENCY

Completion gains. Table 1 and Figure 2 show the lowest mean costs for JACE on PointGoal1 and synthetic data. On PointGoal1, it uses 5,413 episodes versus 8,616 for Safe-RR, a 37% reduction. Relative to JACE-UCB, JACE saves 1,278 episodes on average on PointGoal1 and 178 on synthetic data. All PointGoal1 and synthetic runs complete, with 10/10 assessment-supported PointGoal1 outputs and 198/200 valid synthetic outputs. Appendix B.3 reports paired intervals.

Screening targets completion. Disabling both priority filters yields earlier certification on PointGoal1 and synthetic data, but later completion and more catastrophes. On CarPush1, screening raises completion from 0/10 to 7/10 and reduces JACE's capped cost by 11.6% and full-window catastrophes by 7.8%. First certification accepts any safe nonbaseline policy, even if its

Table 1: **Certification, completion and catastrophes.** Means ± nominal 95% Student half-widths. Costs are capped at 40,000 episodes; events cover the full window. Each control library uses ten paired runs, synthetic data 200 paired seeds. Red/blue mark the two lowest means.

Method	PointGoal1			CarPush1			Synthetic		
	First cert. ↓	Completion ↓	Events ↓	First cert. ↓	Completion ↓	Events ↓	First cert. ↓	Completion ↓	Events ↓
JACE	3,494 ± 1,829	5,413 ± 1,273	2,565 ± 38	3,995 ± 1,673	35,351 ± 3,233	5,910 ± 377	1,922 ± 182	9,540 ± 517	2,701 ± 42
JACE-UCB	5,724 ± 2,573	6,691 ± 1,743	2,613 ± 60	7,945 ± 2,186	35,341 ± 3,199	5,921 ± 386	2,455 ± 139	9,718 ± 511	2,709 ± 41
Safe-RR	8,611 ± 945	8,616 ± 938	2,674 ± 46	13,780 ± 3,514	35,795 ± 3,097	5,936 ± 362	3,240 ± 186	10,057 ± 515	2,730 ± 42
Safe-Uniform	9,147 ± 739	9,147 ± 739	2,700 ± 43	15,638 ± 3,123	35,778 ± 3,105	5,937 ± 362	3,739 ± 191	10,272 ± 512	2,744 ± 42
JACE-unfiltered	691 ± 370	9,859 ± 1,814	4,192 ± 49	8,442 ± 8,498	40,000†	6,412 ± 97	562 ± 60	11,554 ± 555	5,307 ± 13
UCB-unfiltered	902 ± 168	9,563 ± 458	2,928 ± 86	8,451 ± 2,304	37,134 ± 2,708	6,378 ± 60	1,094 ± 84	10,638 ± 506	5,278 ± 44

Note: All PointGoal1 outputs have assessment support (10/10); 198/200 synthetic outputs are valid. CarPush1 completion is 7/10 for the first four methods and 0/10, 6/10 for the unfiltered variants, with assessment support for all completed outputs. †: all ten runs are censored.

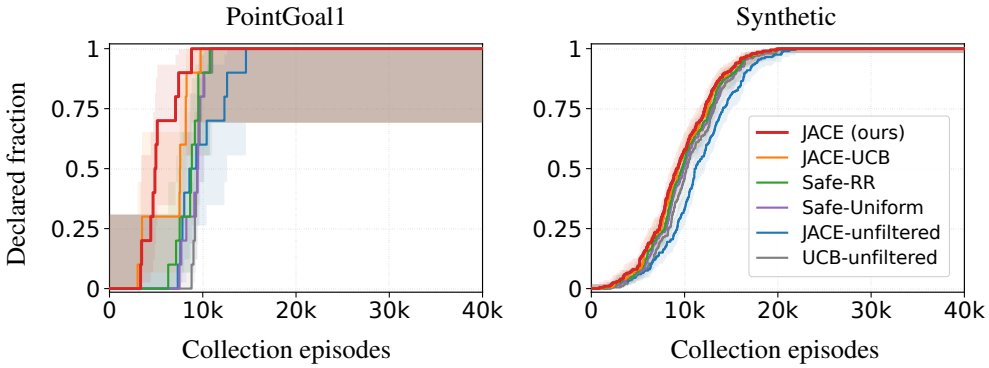


Figure 2: **All methods declare within the window, but JACE uses fewer episodes on average.** PointGoal1 uses 10 paired runs and synthetic data 200 seeds, each over 40,000 episodes. Methods share frozen utilities and local streams within each paired run. Shading gives pointwise exact 95% binomial intervals. Correctness is reported separately in Table 1.

utility does not improve deployment. Completion must instead resolve every potential margin-safe improver, so filtering can delay the first certificate while reducing the work needed for the final safe and useful deployment decision.

Early certification and final completion differ. On CarPush1, JACE reduces mean first-certification cost by 49.7% relative to JACE-UCB. The paired completion difference is 11 episodes, with a multiplicity-adjusted 95% interval of $[-123, 145]$; all four matched methods complete in the same seven of ten runs. JACE reduces mean capped completion cost by about 1.2% relative to rotating and uniform allocation. In the three unfinished JACE runs, the higher-utility competitor remains unresolved because its online lower risk bound is 0.170–0.172, below the exclusion threshold of 0.18 (Appendix B.2.2).

5.3 Scaling of Completion Cost

To test the bound's scaling, we start with a structural baseline of risk β_0 and utility zero, and M unobserved utility-one candidates. All candidate risks equal $\beta + \eta$; positive instances change one to $\beta - \eta$. Fixing $\beta = 0.15$, $\rho = 0.1$ and default $(M, s, \delta, \eta) = (8, 0.048, 0.05, 0.02)$ gives 22 configurations, each with four methods and 100 paired seeds. Figure 3 shows completion scaling above and separate first-certification scans below.

Default positive costs are 18,455, 21,092, 32,418 and 33,901 episodes for JACE, JACE-UCB, Safe-RR and Safe-Uniform. Without an improvement, costs are nearly equal: 32,753, 32,652, 32,716 and 32,723. Every equally difficult candidate must then be excluded, so ordering offers little benefit. Appendix B.4 reports the full results and parameter sensitivity.

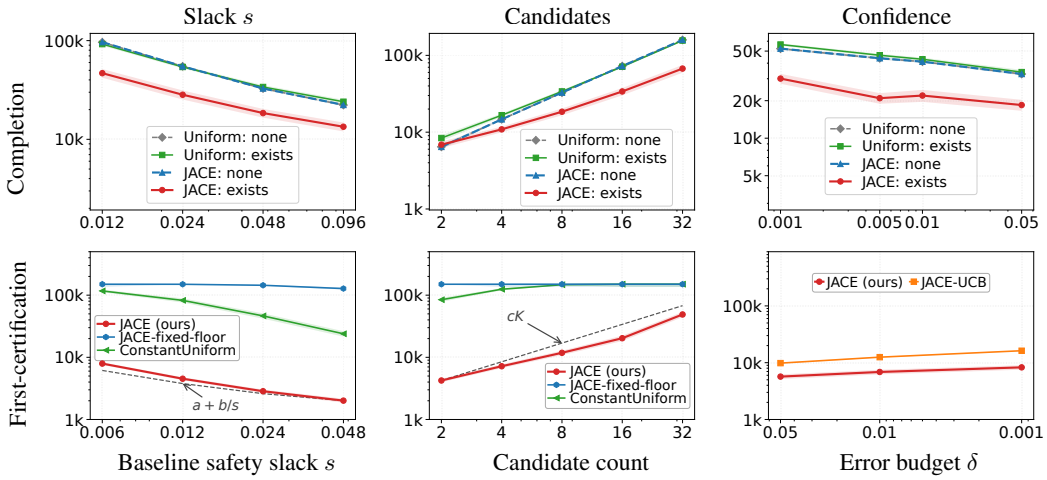


Figure 3: **Completion and first certification reveal how evidence demand and safety slack shape collection cost.** Top: baseline-only start, improvement exists/none, 100 paired seeds, 10^6 episodes. Bottom: baseline-only start, 200 seeds and 150,000 episodes. JACE-UCB uses 1,000 seeds in the confidence scan. Middle-column counts are M above and K below. Curves show capped means with pointwise 95% intervals, varying only the horizontal-axis parameter. Top blue/grey dashed curves mark no improvement. Bottom grey curves are single-point-normalized references.

5.4 Comparison with identification methods

These comparisons assess complete procedures; matched completion comparisons isolate allocation choices. We compare existing methods on strict identification: finding a safe nonbaseline candidate or excluding every candidate. All 58 instances use 1,000 seeds and the strict-identification continuation in Appendix B.6. Table 2 and Figure 4 summarize performance. JACE has lower mean capped cost than SEE on 41 instances and SCUCB on all 58. Complete per-instance costs, correct-identification rates and collection-risk excursions appear in Appendix Figures 7 and 8.

The exceptions are informative. In Unique, one candidate has risk 0.35 and the other $K - 1$ lie at $\beta = 0.5$. Equal utilities provide no initial screening advantage. Certifying the sole interior candidate is difficult, and JACE is slower than SEE and LUCB-G at $K = 50$. All-worse instead requires excluding every candidate while allocating at least $5/13$ of collection to the baseline. At $K = 150, 200$, all 1,000 runs correctly identify no safe candidate, averaging 43,493 and 59,278 episodes. Collection-risk excursions occur in at most 0.3% of JACE runs per instance.

6 Related Work

Safe identification and improvement. SafeBAI [31] uses controllable doses and known safe starting doses to identify the best attainable safe reward. Constrained identification learns feasibility and performance [3, 15, 2], or recommends feasible mixtures [4]. Safe policy improvement [29, 16] and constrained policy optimization [1] instead construct policies from data. Our setting fixes policies and scores, asking when risk evidence certifies an individually safe recommendation whose utility matches or exceeds that of every margin-safe alternative.

Safe data collection. Wan et al. [30], SaVeR [21] and SCOPE [5] design safe collection for policy evaluation. OPB already combines a baseline with cost-constrained mixtures to optimize cumulative reward [24, 23]. Conservative bandits and reinforcement learning constrain cumulative performance [34, 14, 6, 36]. JACE instead studies valid stopping under per-round collection safety. The contribution is the stopping-cost analysis, not the baseline mixing rule.

Sequential evidence and stopping. Good-arm search [10, 12], search/verification separation [17, 18] and anytime identification [9] study evidence for recommendations. BAEC separates finding a qualifying arm from excluding all arms using two thresholds [28]. These stopping costs count direct observations, whereas our acquisition rates depend on certificates. Building on off-policy confidence

Table 2: **JACE's identification-time advantage depends on the instance, despite low collection-risk excursions.** Each cell uses 1,000 seeds and 100,000 episodes. Time reports capped mean ± pointwise 95% interval half-width. Exc. reports the percentage of runs with a collection-risk excursion. Red/blue rank mean time, not safety or significance. Unrestricted methods are not required to meet the same per-round collection-safety constraints.

Method	Unique, $K = 50$		APGAI thr1		RealL		APGAI thr2	
	Time ↓	Exc. ↓	Time ↓	Exc. ↓	Time ↓	Exc. ↓	Time ↓	Exc. ↓
JACE (ours)	6,376 ± 757	0.0	381 ± 27	0.0	176 ± 15	0.0	1,398 ± 108	0.0
SEE [17]	3,826 ± 94	0.0	775 ± 18	100.0	267 ± 5	100.0	2,089 ± 60	100.0
APGAI [9]	53,850 ± 2,970	0.0	1,051 ± 203	100.0	569 ± 284	100.0	4,187 ± 434	100.0
HDoC [10]	8,123 ± 157	0.0	1,272 ± 26	100.0	426 ± 7	100.0	2,801 ± 63	100.0
LUCB-G [10]	5,756 ± 113	0.0	1,071 ± 23	100.0	333 ± 6	100.0	2,392 ± 55	100.0
Uniform [9]	24,311 ± 465	0.0	2,987 ± 56	100.0	1,528 ± 24	100.0	6,564 ± 137	100.0
SCUCB [20]	10,932 ± 189	0.0	1,750 ± 34	0.0	445 ± 16	0.0	3,452 ± 70	0.0
CUCB [34]	9,550 ± 158	0.0	1,488 ± 28	99.4	368 ± 13	88.7	2,985 ± 61	100.0

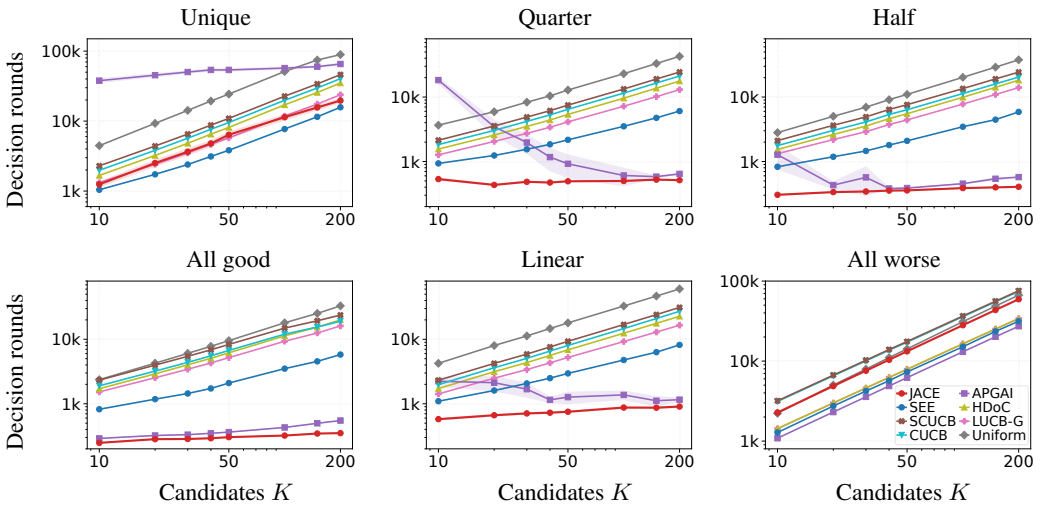


Figure 4: **JACE's identification gains depend on candidate composition.** Six SEE families use 1,000 seeds per point and 100,000 episodes. Curves show capped means with pointwise 95% intervals. All-worse requires negative declarations, the other families positive identification. For costs with zero empirical variance, we report intervals from the bounded Hoeffding construction.

sequences [11, 33], anytime inference [7, 25, 32] and stopped-information arguments [13, 26], we account for state-dependent waiting above and a shared safety constraint below.

7 Conclusion

JACE connects safe collection to a valid stopping decision through time-uniform risk certificates. Completion certifies an improvement or justifies returning the baseline, with a utility guarantee over the prescribed margin-safe set. Matching worst-case bounds separate evidence requirements from certificate-dependent waiting. PointGoal1 shows substantial mean completion savings, while CarPush1 separates fast initial certification from the evidence needed to resolve every remaining competitor. These results support treating search, exclusion and safe data acquisition as parts of the same stopping problem.

Limitations and open questions. Guarantees assume a fixed finite library and stationary episode laws, and concern frozen utilities over a prescribed margin-safe set. Control results measure online costs for two fixed libraries after preparation and baseline qualification. Utility uncertainty, historical risk-data reuse and variation across independently trained libraries require further evaluation. Instance-dependent priority analysis could explain when evidence ordering reduces completion cost.

AI USE STATEMENT

AI helps us polish our articles and correct grammatical errors. The authors take responsibility for all content and results.

REPRODUCIBILITY STATEMENT

The supplementary materials include the complete code for reproducing our experimental results.

REFERENCES

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.

[2] Ting Cai and Kirthevasan Kandasamy. Constrained best arm identification with tests for feasibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[3] Romain Camilleri, Andrew Wagenmaker, Jamie Morgenstern, Lalit Jain, and Kevin Jamieson. Active learning with safety constraints. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[4] Emil Carlsson, Debabrota Basu, Fredrik Johansson, and Devdatt Dubhashi. Pure exploration in bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 334–342. PMLR, 2024.

[5] Claire Chen, Shuze Liu, and Shangtong Zhang. Efficient policy evaluation with safety constraint for reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pages 34989–35010, 2025.

[6] Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Conservative exploration in reinforcement learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 1431–1441. PMLR, 2020.

[7] 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.

[8] Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. *Advances in Neural Information Processing Systems*, 36:18964–18993, 2023.

[9] Marc Jourdan, Andrée Delahaye-Duriez, and Clémence Réda. An anytime algorithm for good arm identification. *Journal of Machine Learning Research*, 27(19):1–90, 2026.

[10] Hideaki Kano, Junya Honda, Kentaro Sakamaki, Kentaro Matsuura, Atsuyoshi Nakamura, and Masashi Sugiyama. Good arm identification via bandit feedback. *Machine Learning*, 108(5):721–745, 2019.

[11] Nikos Karampatziakis, Paul Mineiro, and Aaditya Ramdas. Off-policy confidence sequences. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2021.

[12] Julian Katz-Samuels and Kevin Jamieson. The true sample complexity of identifying good arms. In *International Conference on Artificial Intelligence and Statistics*, pages 1781–1791. PMLR, 2020.

[13] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

[14] Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. *Advances in neural information processing systems*, 30, 2017.

[15] Tyron Lardy, Christina Katsimerou, and Wouter M. Koolen. Constrained best arm identification. In *Advances in Neural Information Processing Systems*, 2025.

[16] Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International conference on machine learning*, pages 3652–3661. PMLR, 2019.

[17] Zitian Li and Wang Chi Cheung. Near optimal non-asymptotic sample complexity of 1-identification. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 34144–34184. PMLR, 2025.

[18] Zitian Li and Wang Chi Cheung. Closing the gap on the sample complexity of 1-identification. *arXiv preprint arXiv:2601.15620*, 2026.

[19] Zuxin Liu, Zijian Guo, Haohong Lin, Yihang Yao, Jiacheng Zhu, Zhepeng Cen, Hanjiang Hu, Wenhao Yu, Tingnan Zhang, Jie Tan, and Ding Zhao. Datasets and benchmarks for offline safe reinforcement learning. *Journal of Data-centric Machine Learning Research*, 2024.

[20] Ahmadreza Moradipari, Christos Thrampoulidis, and Mahnoosh Alizadeh. Stage-wise conservative linear bandits. *Advances in neural information processing systems*, 33:11191–11201, 2020.

[21] Subhojyoti Mukherjee, Josiah P Hanna, and Robert Nowak. SaVeR: Optimal data collection strategy for safe policy evaluation in tabular MDP. *arXiv preprint arXiv:2406.02165*, 2024.

[22] Andrew Y. Ng and Michael I. Jordan. PEGASUS: A policy search method for large MDPs and POMDPs. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 406–415, 2000.

[23] Aldo Pacchiano, Mohammad Ghavamzadeh, and Peter Bartlett. Contextual bandits with stage-wise constraints. *Journal of machine learning research*, 26(170):1–57, 2025.

[24] Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International conference on artificial intelligence and statistics*, pages 2827–2835. PMLR, 2021.

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

[26] Venkatesh Saligrama. When can safe controllers adapt? information before commitment. *arXiv preprint arXiv:2607.16895*, 2026.

[27] Xuedong Shang, Igor Colin, Merwan Barlier, and Hamza Cherkaoui. Price of safety in linear best arm identification. *arXiv preprint arXiv:2309.08709*, 2023.

[28] Koji Tabata, Atsuyoshi Nakamura, Junya Honda, and Tamiki Komatsuzaki. A bad arm existence checking problem: How to utilize asymmetric problem structure? *Machine Learning*, 109(2):327–372, 2020.

[29] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, pages 2380–2388. PMLR, 2015.

[30] Runzhe Wan, Branislav Kveton, and Rui Song. Safe exploration for efficient policy evaluation and comparison. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 22491–22511. PMLR, 2022.

[31] Zhenlin Wang, Andrew J Wagenmaker, and Kevin Jamieson. Best arm identification with safety constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 9114–9146. PMLR, 2022.

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

[33] Ian Waudby-Smith, Lili Wu, Aaditya Ramdas, Nikos Karampatziakis, and Paul Mineiro. Anytime-valid off-policy inference for contextual bandits. *ACM/IMS Journal of Data Science*, 1(3):1–42, 2024.

[34] Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pages 1254–1262. PMLR, 2016.

[35] Sandarbh Yadav, Frederic J Maliakkal, Harshad Khadilkar, and Shivaram Kalyanakrishnan. Using common random numbers for simulation-based planning with rollouts. *arXiv preprint arXiv:2605.04732*, 2026.

[36] Yunchang Yang, Tianhao Wu, Han Zhong, Evrard Garcelon, Matteo Pirotta, Alessandro Lazaric, Liwei Wang, and Simon Shaolei Du. A reduction-based framework for conservative bandits and reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022.

A Proofs of the Main Results

We first specify the probability space and coverage event, then establish the certificate and allocation lemmas. These support the main safety theorem, completion bounds and their consequences. Proofs for JACE-MS remain separate in Appendix C.

A.1 Probability space and coverage event

For each fixed policy, pre-generate an independent infinite episode stream and reveal its next unused observation when selected. This construction preserves the joint law of adaptive collection under Assumption 2.1. On the local clock, $X_{j,n}$ are i.i.d. Bernoulli(p_j); their global reveal times are adaptive. Conditioning on the policy draw gives (1). This identity uses the protocol of executing one fixed policy per complete episode.

Let $\mathcal{F}_0$ contain the frozen library, independent development and calibration data, and online settings. Write $\widetilde{U}_j$ for the online upper bounds before the baseline cap and define $\mathcal{E}_{\mathrm{on}} = \{\forall j, n : L_j(n) \le p_j \le \widetilde{U}_j(n)\}$. If the baseline lower certificate is omitted, set $L_0 \equiv 0$ and $\varepsilon_0^L = 0$. Lemma A.1 and a union bound give

$$\mathbb{P}(\mathcal{E}_{\mathrm{on}}^c \mid \mathcal{F}_0) \le \sum_{j=0}^{K} (\varepsilon_j^U + \varepsilon_j^L) \le \delta_{\mathrm{on}} := \delta - \delta_0. \tag{EC.1}$$

The deterministic budget cap allows the allocation among certificates to depend on $\mathcal{F}_0$. Calibration supplies $\mathcal{E}_{\mathrm{cal}} = \{p_0 \le \beta_0\}$ with marginal probability at least $1 - \delta_0$. On $\mathcal{E}_{\mathrm{cov}} := \mathcal{E}_{\mathrm{cal}} \cap \mathcal{E}_{\mathrm{on}}$, the implemented bound $U_0 = \min\{\beta_0, \widetilde{U}_0\}$ and all other bounds cover their risks. Total probability gives $\mathbb{P}(\mathcal{E}_{\mathrm{cov}}) \ge 1 - \delta_0 - \delta_{\mathrm{on}} = 1 - \delta$. Expected-time bounds hold pointwise in admissible frozen parameters; averaging over them requires integrability of the stated cost bound.

Presampled rollout streams. Under Assumption 2.1, generating policy-local episodes in advance does not change an adaptive experiment, provided unused episodes remain hidden and each execution reveals the next unused episode of the selected policy. To see this, pre-generate independent streams $(X_{j,n})_{n\ge1}$ and allocation uniforms $(Z_t)_{t\ge1}$. After any revealed history and policy draw $J_t = j$, the unrevealed $X_{j,N_j(t-1)+1}$ is independent of that history and has law Bernoulli(p_j). Induction on t gives the same transition kernel and history law as on-demand simulation. The identity includes decisions and stopping times. For finite streams, it holds for the experiment stopped before the first request beyond a stored capacity; exhaustion is not a statistical stopping decision. Sharing streams across methods preserves each method's marginal law and induces the pairing used for comparisons. It does not require revealing another method's observations or guarantee variance reduction for every contrast [22, 35].

A.2 Proof of Lemma A.1

Write $\mathcal{G} = \{0 = g_0 < \cdots < g_{n_g-1} = 1\}$ and let $\mathcal{R}_n$ contain every point whose evidence has crossed its threshold by local count n. Define

$$r_n = \max\big(\{i : g_i \notin \mathcal{R}_n\} \cup \{-1\}\big), \qquad U_n = g_{\min(r_n+1, n_g-1)}. \tag{EC.2}$$

Rejections are permanent. The reflected construction on $1 - X$ gives $L_n = 1 - U_n^{\mathrm{ref}}$, with reflected grid $1 - \mathcal{G}$. Thus U_n is nonincreasing and L_n nondecreasing on every path. Implement reflection and the strict test $L_j > \theta$ using stored grid indices, so floating-point subtraction cannot turn equality into strict exceedance.

Lemma A.1 (Anytime coverage)**.** *Let $\varepsilon \in (0,1)$. Suppose stakes are predictable and nonnegative, with $\lambda_n(q) \le c/(1-q)$ for $q < 1$, $0 < c < 1$, and finite stakes at $q = 1$. For a fixed convex mixture of such capital processes and rejection level $1/\varepsilon$, (EC.2) satisfies $\mathbb{P}(\forall n : p \le U_n) \ge 1 - \varepsilon$.*

Proof. At a true null $q \le p$, each multiplier is nonnegative and has conditional expectation $1 + \lambda_n(q)(q - p) \le 1$. The capital and any fixed convex mixture are nonnegative supermartingales initially equal to one. For $q_* := \max\{g \in \mathcal{G} : g \le p\}$, Ville's inequality bounds the probability of ever rejecting q_* by ε. While q_* remains, its successor is at least p. At $p = 1$, the capital at $q = 1$

stays one. This single true-null argument requires neither monotonic evidence across q nor a grid union bound. □

Implemented upper hedge. After n local observations maintain a Welford mean and centered sum of squares $(\widehat{\mu}_n, M_{2,n})$. For choosing a stake while $n < 2$, use mean $1/2$ and variance one; otherwise use the empirical mean and $\widehat{v}_n = \max\{M_{2,n}/(n-1), 10^{-8}\}$. Set

$$\lambda^{\mathrm{ag}}_{n+1}(q) = \min\left\{\frac{(q-\widehat{\mu}_n)_+}{\widehat{v}_n + (q-\widehat{\mu}_n)^2}, \frac{1/2}{1-q}\right\},$$
$$K^{\mathrm{ag}}_{n+1}(q) = K^{\mathrm{ag}}_n(q)[1 + \lambda^{\mathrm{ag}}_{n+1}(q)(q - X_{n+1})]. \tag{EC.3}$$

At $q = 1$, omit the infinite cap. Moments update after the observation. The upper capital is $K^{\mathrm{hed}} = (K^{\mathrm{mix}} + K^{\mathrm{ag}})/2$. The lower certificate uses K^{mix} alone on the reflected stream. Both satisfy Lemma A.1. Pathwise,

$$U^{\mathrm{hed},\varepsilon}_n \le \min\{U^{\mathrm{mix},\varepsilon/2}_n, U^{\mathrm{ag},\varepsilon/2}_n\}. \tag{EC.4}$$

Indeed, either component crossing $2/\varepsilon$ forces the hedge across $1/\varepsilon$. Since adaptive stakes depend on q, the algorithm checks all grid points; it does not infer rejection above β merely from the full hedge crossing at β.

A.3 Proof of Lemma A.2

At all-floor allocation, with $b \le \beta_0$ and $U_j \le 1$,

$$b + w_{\min}\sum_{j\in A_t}(U_j - b) \le \beta_0 + Kw_{\min}(1-\beta_0) \le \beta_0 + \rho s < \beta. \tag{EC.5}$$

The mass constraint is feasible because $(K+1)w_{\min} \le \rho < 1$. Solving the affine risk constraint for a priority candidate with $v = U_h > \beta$ gives (4).

Lemma A.2 (Priority sampling probability)**.** *On every path, any priority candidate h satisfies* (5)*.*

Proof. Let $f = (|A_t| - 1)w_{\min}$ be the other candidates' floor mass. The risk endpoint in (4) is at least $\phi(b) = [\beta - b - f(1-b)]/(v-b)$. Its derivative has the sign of $(\beta - v) - f(1-v) \le 0$. Therefore

$$\phi(b) \ge \phi(\beta_0) \ge \frac{(1-\rho)s}{v - \beta_0}. \tag{EC.6}$$

The mass endpoint is at least $1-\rho$, which is no smaller than this last quantity since $v > \beta$. Feasibility also gives $w_{t,h} \ge w_{\min}$. □

A.4 Proof of Theorem 3.1

Proof of Theorem 3.1. On $\mathcal{E}_{\mathrm{cov}}$, the baseline cap is valid, every eligible deployment has $p_{d_t} \le \beta$, and (1) together with the allocation constraint gives $r_t \le \beta$ at every round. Safe candidates cannot be eliminated since $L_j \le p_j \le \beta$.

If Q_t holds and a candidate has $p_j \le \theta$, then $L_j \le \theta$. Such a candidate cannot have utility above u_{d_t} by (6). The baseline is always eligible, proving $u_{d_t} \ge u_\eta$.

At every decision from round one, Q_t is equivalent to $V_t = \emptyset$ on every path. One direction follows immediately from the definition of V_t. Conversely, a higher-utility candidate with $L_j \le \theta$ cannot have been eliminated, because elimination requires the permanently maintained inequality $L_j > \beta > \theta$. It cannot be certified and active, by maximality of u_{d_t}. It must therefore be in V_t.

On coverage, eligible sets expand among safe policies, deployment utility does not decrease, and both certificate bounds move monotonically. Hence V_t only shrinks and remains empty after its first empty state. When baseline fallback is used, (EC.5) gives $r_t \le \beta_0 + \rho s$. All statements use the same coverage event. □

A.5 CONTINUING TO STRICT IDENTIFICATION

Proposition A.3 (Safe continuation after completion)**.** *Run the strict-identification protocol of Appendix B.6 under the assumptions of Theorem 3.1. Up to its stopping time, it retains joint safety and valid margin declarations at level $1-\delta$, and any strict positive or negative answer is correct on the same event. With identical initial information and random streams, its history agrees with Algorithm 1 until the first margin completion or strict answer, whichever occurs first. If a strict answer cannot precede margin completion, the first completion costs are pathwise equal and Theorem 4.1 and its corollaries transfer under their stated conditions. This condition holds when all nonbaseline utilities are equal and greater than u_0, as in the identification benchmarks.*

Proof. Before the first margin declaration or strict answer, the two runs have identical working sets and transition rules. Induction with common allocation uniforms and policy-local observations yields identical histories up to that time, without conditioning on coverage. With equal nonbaseline utilities above u_0, a positive answer attains the largest library utility and a negative answer leaves no active candidate; either implies $V_t = \emptyset$ on every path. Thus strict stopping cannot precede margin completion in this case. After margin completion, every continuation priority has $U_j > \beta$ and uses (4), so the certificate risk constraint and sampling floors remain feasible. On joint coverage, $r_t \le \beta$ and every eligible deployment is safe. A positive identification has $p_j \le U_j \le \beta$; a negative identification has $p_j \ge L_j > \beta$ for every nonbaseline policy. The quality proof uses certificates and maximal-utility deployment, not the collection priority set. The smaller post-completion risk bound $\beta_0 + \rho s$ specifically requires baseline fallback. The pathwise equality transfers first-completion bounds, not bounds on the additional time to strict identification. □

Scope of early screening. At the first decision, no candidate with $u_j > u_0$ is filtered because $U_j = 1$ and $L_j = 0 < \theta$. More generally, on $\mathcal{E}_{\mathrm{cov}}$, any candidate with $p_j \le \theta$ and $u_j > u_{d_t}$ belongs to V_t. Indeed, $L_j \le p_j \le \theta$ prevents elimination and margin screening. If $U_j \le \beta$, maximality of the deployed utility would contradict $u_j > u_{d_t}$. Thus $U_j > \beta$ and every priority condition holds. This is preservation of a still-better margin-safe comparator, not a guarantee that every such policy must be sampled or deployed: a later incumbent can dominate its utility.

For $p_j \in (\theta, \beta]$, valid bounds can satisfy $\theta < L_j \le p_j \le \beta < U_j$. Margin screening then removes priority while retaining the sampling floor. Such a candidate can have arbitrarily high utility, so completion relative to u_η need not attain the best utility among all policies with $p_j \le \beta$. This limitation does not require a coverage failure or vanish merely because the margin is numerically small.

Permanent elimination is different: $L_j > \beta$ cannot occur for a safe candidate on coverage. Outside that event, erroneous elimination remains possible. Continued floor sampling does not undo a declaration already made, and a first-certification guarantee requires both continued collection and the resolution condition in Proposition A.9. Finally, all utility statements concern frozen u_j. A truly higher-return policy can receive $u_j \le u_0$ through development estimation error and lose priority immediately. Freezing utility estimates does not certify their ranking by true expected return.

Corollary A.4 (Transfer to expected returns)**.** *Under the conditions of Theorem 3.1, let μ_j be policy j's finite expected return. Suppose that, for some $\epsilon_u \ge 0$ and $0 < \alpha < 1-\delta$, the frozen utilities satisfy*

$$\mathbb{P}\left(\max_{0\le j\le K} |u_j - \mu_j| \le \epsilon_u\right) \ge 1-\alpha. \tag{EC.7}$$

With probability at least $1-\delta-\alpha$, every completion declaration returns a safe policy d satisfying

$$\mu_d \ge \max\big(\{\mu_0\} \cup \{\mu_j : p_j \le \theta,\ j \ge 1\}\big) - 2\epsilon_u. \tag{EC.8}$$

Proof. Intersect the simultaneous utility-accuracy event with the joint event in Theorem 3.1; a union bound gives probability at least $1-\delta-\alpha$, without requiring independence between these events. Let $j^\star$ maximize expected return over the baseline and margin-safe candidates. At every completion declaration, the existing utility guarantee gives $u_d \ge u_{j^\star}$, and therefore

$$\mu_d \ge u_d - \epsilon_u \ge u_{j^\star} - \epsilon_u \ge \mu_{j^\star} - 2\epsilon_u. \tag{EC.9}$$

Safety follows on the same event. The simultaneous utility-error bound is an additional assumption; the reported development means alone do not supply a numerical value for ϵ_u. □

Notation for the completion bounds. Throughout this appendix, the library, utilities, grid, stakes and positive budgets are fixed. Write $a = 1 - \rho$, $v_j = p_j(1 - p_j)$ and $r_{\mathrm{cfg}} = h + 5\lambda_{\min}$. The two-threshold results assume (7); the first-certification result in Appendix A.12 only needs an appropriate stake at β. No local-sample priority cap is enabled.

A.6 Proof of Lemma A.5

Lemma A.5 (Sample bound for risk testing). *Let $\varepsilon, \alpha \in (0,1)$. For a grid point $q > p$, set $d = q - p$, $v = p(1-p)$ and $\lambda^* = \min\{1/2, d/[4(v+d^2)]\}$. Suppose a geometric stake $\lambda \in [\lambda^*/2, \lambda^*]$ has weight $\omega > 0$. If $L \geq \max\{\log[2/(\varepsilon\omega)], \log(1/\alpha)\}$, then with probability at least $1-\alpha$, q and every higher grid point are permanently rejected by*

$$n_q = \left\lceil 128 \left(\frac{v}{d^2} + \frac{1}{d} \right) L \right\rceil. \tag{EC.10}$$

This holds for the upper hedge and, conservatively, for the pure geometric reflected certificate.

Proof. Let $Y = \log[1 + \lambda(q - X)]$. Since $\lambda \leq 1/2$ and $\log(1+x) \geq x - x^2$ for $x \geq -1/2$,

$$\mu := \mathbb{E}Y \geq \lambda d - \lambda^2(v + d^2) \geq 3\lambda d/4. \tag{EC.11}$$

The two possible increments differ by at most 2λ, giving $\mathrm{Var}(Y) \leq 4\lambda^2 v$ and $|Y - \mu| \leq 2\lambda$. Bernstein's inequality yields

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i < n\lambda d/2 \right) \leq \exp\left(-\frac{nd^2}{128(v+d)} \right). \tag{EC.12}$$

Also $1/(\lambda d) \leq 8v/d^2 + 12/d$. At n_q, the evidence exceeds $2/(\varepsilon\omega)$ except with probability α, so the hedge exceeds $1/\varepsilon$. A fixed nonnegative stake has capital nondecreasing in q on the same data. Thus the same component rejects all higher grid points simultaneously, regardless of whether the adaptive component is monotone. Permanent rejection preserves the conclusion. □

For Bernoulli observations $\lambda^* \geq d/5$, so $\lambda_{\min} \leq d/5$ is sufficient for the required stake to exist. Applying the lemma at $q = \beta$ proves certification without a mesh condition involving $\beta - p$.

A.7 Proof of Lemma A.6

Lemma A.6 (Time-uniform confidence widths). *For $\gamma \in (0,1)$ put $\ell = \log[16Kn_g/(\varepsilon_{\min}\omega_{\min}\gamma)]$ and*

$$b_j(n) = 512 \left[\sqrt{\frac{v_j \ell}{n}} + \frac{\ell}{n} \right], \qquad n \geq 1. \tag{EC.13}$$

With probability at least $1 - \gamma/2$, simultaneously over candidates and local counts,

$$U_j(n) \leq \min\{1, p_j + b_j(n) + r_{\mathrm{cfg}}\}, \quad L_j(n) \geq \max\{0, p_j - b_j(n) - r_{\mathrm{cfg}}\}. \tag{EC.14}$$

Proof. For each upper or reflected grid target at distance at least $5\lambda_{\min}$ from its local mean, apply Lemma A.5 with failure probability $\gamma/(4Kn_g)$. This is a finite union over permanent rejection deadlines. For a fixed n, choose the first grid point at least $p_j + \max\{5\lambda_{\min}, b_j(n)\}$. If it exists, its distance $d \geq b_j(n)$ gives $128v_j\ell/d^2 \leq n/2048$ and $128\ell/d \leq n/4$. Existence also implies $b_j(n) \leq 1$, hence $n \geq 512\ell$, which absorbs the ceiling in (EC.10). The successor upper bound is at most this target, and its distance is at most $b_j(n) + 5\lambda_{\min} + h$. If the target does not exist, the claimed upper bound is at least one. Reflecting the argument gives the lower bound. No union over random stopping times is used. □

For each candidate define its permanent local exit time

$$M_j = \inf\{n \geq 0 : U_j(n) \leq \beta \text{ or } L_j(n) > \theta\}. \tag{EC.15}$$

These local states are defined on the entire potential stream, whether or not the online run reveals them. If $g_j = \beta - p_j$, use the upper target β and budget $\varepsilon_j^\star = \varepsilon_j^U$; otherwise use the reflected target $1 - q_\theta$ and $\varepsilon_j^\star = \varepsilon_j^L$. Under (7), $g_j \geq \eta_{\text{eff}}/2$, so $\lambda_{\min} \leq g_j/5$. Choose a stake index $m(j)$ satisfying Lemma A.5 in this direction. Moreover,

$$r_{\text{cfg}} \leq 3g_j, \qquad (p_j - \beta)_+ \leq g_j, \qquad R_j \geq 1. \tag{EC.16}$$

The use of $q_\theta > \theta$ is essential: rejecting the reflected target yields $L_j \geq q_\theta > \theta$. A one-cell margin with $q_\theta = \beta$ would leave zero drift at $p_j = \beta$.

Let $L_0 = \log[4/(\varepsilon_{\min}\omega_{\min})]$ and $r_j = 128R_j$. The local tail argument gives

$$\mathbb{P}(M_j > n) \leq e^{-n/r_j}, \qquad n \geq N_j := \lceil r_j L_0 \rceil. \tag{EC.17}$$

For the selected component, let S_n be its log-capital and $\mu_j = \mathbb{E}(S_n - S_{n-1}) > 0$. Set $b_j^* = \log[2/(\varepsilon_j^\star \omega_{m(j)})]$ and $T_j = \inf\{n : S_n \geq b_j^*\}$. Then $M_j \leq T_j$, and positive overshoot is at most $1/2$. Bounded-stopping Wald gives $\mu_j \mathbb{E}(T_j \wedge n) = \mathbb{E}S_{T_j \wedge n} \leq b_j^* + 1/2$. Since $\mu_j^{-1} \leq 16R_j$, $\varepsilon_j^\star \geq \varepsilon_{\min}$ and $\omega_{m(j)} \geq \omega_{\min}$, monotone convergence yields

$$\mathbb{E}M_j \leq 16R_j \left[\log \frac{2}{\varepsilon_{\min}\omega_{\min}} + \frac{1}{2}\right]. \tag{EC.18}$$

Thus local evidence is finite even on coverage-failure paths.

A.8 Proof of Theorem 4.1

This proof analyzes Algorithm 1 with baseline fallback after completion. Let $A_\infty = \sum_{t \geq 1} \mathbf{1}\{V_t \neq \emptyset\}$ count all its nonempty-priority rounds from initialization. Every priority candidate belongs to I_+ and has local count $n < M_j$. Local counts never decrease, even if an incorrectly certified policy is later eliminated and a utility-filtered candidate regains priority.

Condition on all potential observation streams, fixing $U_j(n)$ and M_j. Selection randomization remains independent. For each candidate and count layer, pre-generate independent uniform random variables for its successive priority attempts. By Lemma A.2, an attempt succeeds with probability at least $as/[\max\{\beta, U_j(n)\} - \beta_0]$. A success leaves the layer permanently. Floor observations or other exits can only shorten its occupancy. Geometric comparison and Tonelli's theorem therefore give

$$\mathbb{E}A_\infty \leq \frac{1}{as} \sum_{j \in I_+} \mathbb{E} \sum_{n < M_j} [\max\{\beta, U_j(n)\} - \beta_0]. \tag{EC.19}$$

This conditions on observation streams, not on future coverage, and does not assert that the actual adaptive waits are independent.

To evaluate (EC.19), define

$$L_n = \log \frac{16 n_g (n+1)^2}{\varepsilon_{\min}\omega_{\min}}, \qquad q_j(n) = \begin{cases} 1, & n < N_j, \\ e^{-n/r_j}, & n \geq N_j, \end{cases} \tag{EC.20}$$

$$f_j(n) = s + (p_j - \beta)_+ + r_{\text{cfg}} + \min\left\{1, 512\left[\sqrt{\frac{v_j L_n}{n}} + \frac{L_n}{n}\right]\right\}. \tag{EC.21}$$

The one-candidate width argument with failure budget $(n+1)^{-2}$ gives an upper-width failure probability at most that value. On its complement, the cost is at most $f_j(n)$; on failure it is at most one. Combining with (EC.17), without independence, yields

$$\mathbb{E}A_\infty \leq \frac{1}{as} \sum_{j \in I_+} \left[1 - \beta_0 + \sum_{n \geq 1} \{f_j(n) q_j(n) + (n+1)^{-2}\}\right]. \tag{EC.22}$$

The series is finite. To expose its confidence dependence, write $\overline{L} = \log[16n_g(1 + \max_{j\in I_+} N_j)^2/(\varepsilon_{\min}\omega_{\min})]$. Below N_j, use $L_n \le \overline{L}$ and

$$\sum_{n=1}^{m} n^{-1/2} \le 2\sqrt{m},$$
$$\sum_{n=1}^{m} \min\{1, A/n\} \le A[1 + \log_+(m/A)] + 1 \quad (A \ge 1). \tag{EC.23}$$

Since $N_j \le 129R_jL_0$ and $\overline{L} \ge L_0$, the square-root contribution is at most $1024\sqrt{v_j\overline{L}N_j}$ and the harmonic contribution at most $C\overline{L}\log(eR_j)$. This truncation prevents an unnecessary factor $\log L_0$. Above N_j, L_n/n decreases, so $f_j(n) \le f_j(N_j)$. The exponential tail sum is at most $(r_j + 1)e^{-N_j/r_j}$, and its total contribution is bounded by

$$C\left[sN_j + g_jN_j + \sqrt{v_j\overline{L}N_j} + \overline{L}\right]. \tag{EC.24}$$

Using $\sqrt{v_jR_j} \le H_j/g_j$ and (EC.16), we obtain

$$\mathbb{E}A_\infty \le \frac{C\overline{L}}{a}\left[\sum_{j\in I_+}\frac{H_j}{g_j^2} + \frac{1}{s}\sum_{j\in I_+}\left(\frac{H_j}{g_j} + \log(eR_j)\right)\right]. \tag{EC.25}$$

Finally $1 + N_j \le 130R_{\max}L_0$ and $\log L_0 \le L_0$ imply $\overline{L} \le 8\Lambda$. Absorb the factor into C. Actual priority executions are separately bounded by $\mathbb{E}N_{\text{pri}} \le \sum_{j\in I_+}\mathbb{E}M_j$; they are not the same object as A_∞.

Proof of Theorem 4.1. At every decision, Q_t is equivalent to $V_t = \emptyset$ by the pathwise argument in Appendix A.4. Consequently every one of the $\tau_Q - 1$ rounds preceding the first declaration has nonempty priority. Thus $T_Q \le A_\infty$ on every path, even if erroneous elimination subsequently lowers deployment utility and makes priority nonempty again. Equation (EC.25) proves (8). Its finite right-hand side also proves $\mathbb{P}(\tau_Q = \infty) = 0$. On $\mathcal{E}_{\text{cov}}$, priority only shrinks, so $T_Q = A_\infty$, and Theorem 3.1 supplies quality and fallback from the first declaration.

If $I_+ = \emptyset$, baseline deployment already dominates every utility at decision one, and $V_1 = \emptyset$. Otherwise, the argument charges only candidates in I_+ because $u_{d_t} \ge u_0$ on every path, including coverage failures. This is the step that permits charging the whole run, rather than a segment following first certification. No assumption on τ_0 is used.

Finally, $H_j \le \beta + 2g_j$, $g_j \ge \eta_{\text{eff}}/2$ and $\eta_{\text{eff}} < \beta$ imply $H_j/g_j^2 \le 8\beta/\eta_{\text{eff}}^2$ and $H_j/g_j \le 4\beta/\eta_{\text{eff}}$. These inequalities give (10) with the stated logarithmic factors. All expectations average the original online randomness, not a selected successful history. □

A.9 Proof of Corollary A.7

Corollary A.7 (Logarithmic configuration overhead)**.** *Under Assumption 2.1, take a structural baseline and the parameter range of Theorem 4.3. Fix constants $\rho, \alpha \in (0, 1)$ independently of the problem parameters. Set*

$$D = \lceil 4/\eta \rceil, \qquad \mathcal{G} = \{i/D : 0 \le i \le D\} \cup \{\theta, \beta\}. \tag{EC.26}$$

Let $m_\star$ be the smallest nonnegative integer with $2^{-m_\star-1} \le \eta_{\text{eff}}/10$. Use $\lambda_m = 2^{-m-1}$ for $0 \le m \le m_\star$ and

$$\omega_m = \frac{m_\star + 2}{(m_\star + 1)(m + 1)(m + 2)}, \tag{EC.27}$$

$$\delta_0 = 0, \qquad \varepsilon_j^U = \frac{\alpha\delta}{K + 1}, \qquad \varepsilon_j^L = \frac{(1 - \alpha)\delta}{K + 1}, \quad 0 \le j \le K. \tag{EC.28}$$

Then (7) holds and, for a constant $C_{\rho,\alpha}$ depending only on the fixed fractions,

$$\mathbb{E}T_Q \le C_{\rho,\alpha} K \log \frac{e(K+1)}{\delta\eta} \left[\frac{\beta}{\eta^2} + \frac{\beta}{s\eta} \log \frac{e}{\eta} \right]. \tag{EC.29}$$

Together with Theorem 4.3, this gives matching worst-case scales up to logarithmic factors in its stated parameter range.

Proof. Adding θ, β cannot increase the uniform grid spacing, so $h \le \eta/4$, $n_g \le D+3$, and $\eta_{\text{eff}} \ge \eta - h \ge 3\eta/4$. Thus the resolution conditions hold. Minimality gives $m_\star + 1 = O(\log(e/\eta))$. The weights sum to one and $\omega_{\min} = (m_\star+1)^{-2}$, while $\varepsilon_{\min} = \min\{\alpha, 1-\alpha\}\delta/(K+1)$. Since $g_j \ge 3\eta/8$ and $R_j \le \eta_{\text{eff}}^{-2} + 2\eta_{\text{eff}}^{-1}$, we have $R_{\max} = O(\eta^{-2})$ and $\Lambda = O_\alpha(\log[e(K+1)/(\delta\eta)])$. Finally, $H_j \le \beta + 2g_j$ and $\eta \le \beta/2$ give $H_j/g_j^2 = O(\beta/\eta^2)$, $H_j/g_j = O(\beta/\eta)$ and $\log(eR_j) = O(\log(e/\eta))$. Substitute into (8), use $K_+ \le K$ and absorb the fixed fractions. If $K_+ = 0$, the claim follows from $T_Q = 0$. □

Initialization and comparison scope. Theorems 4.1 and 4.3 can be compared using only a structural safe baseline and fresh candidate streams. No auxiliary policy is certified in advance. In the lower-bound family, $K = M$, $u_0 = 0$ and all candidate utilities are one. Under P^0, every candidate is unsafe; declaring the baseline is correct only after collecting evidence sufficient to rule out the positive alternatives. Under P^j, candidate j is margin-safe and must be selected by any correct declaration. Baseline information and external randomization have the same law throughout the family.

The upper bound additionally permits independent baseline calibration, with the joint probability accounting described in Appendix A.1. The lower bound uses the stronger structural baseline information, so its cost remains necessary on that subclass. Frozen utilities simplify the reward-learning problem. These results do not provide an instance-optimal characteristic time, a comparison to arbitrary side information, or a uniform conditional remaining-time guarantee at every realized history. Replacing unknown risks by empirical risks in the bound does not create such a guarantee.

A.10 Completion with a certified auxiliary

Corollary A.8 (A common certified start)**.** *Take the parameter range of Theorem 4.3 and the resolution conditions (7). Add to its instance family one auxiliary policy a with $p_a = \beta_0$ and $u_a = u_0 = 0$, alongside the M fresh utility-one candidates. Supply the structural certificate $U_a(0) = \beta$, with initial capitals one and grid points at or above β rejected. The auxiliary's initialization and episode law are identical under every instance and reveal no candidate information. Use the floor (3) and certificate budgets for $K = M+1$. Then JACE satisfies Theorem 3.1 and (8) with I_+ consisting of the M fresh candidates. Every jointly valid completion method on this augmented family still satisfies (12). Costs start after the auxiliary certificate is supplied.*

Proof. Every initially rejected auxiliary grid point is strictly above p_a, since $\beta_0 < \beta$. The structural cap is valid and subsequent certificates use fresh observations, so the joint-coverage and allocation proofs remain valid with $K = M+1$. The auxiliary never belongs to V_t, because $u_a = u_0 \le u_{d_t}$. Every priority round therefore still charges a local count $n < M_j$ of a fresh candidate. Lemma A.2 accounts for the auxiliary's floor as part of the other active mass. The local deadlines and widths are unchanged, and the count-layer argument in Appendix A.8 gives the stated upper bound. With the configured budgets, the added policy enters through $M+2$ in the logarithmic budget factor, not as another evidence-acquisition term.

For the lower bound, let W_t be the total mass on the M fresh candidates. Under P^0, both baseline and auxiliary have risk β_0, hence

$$r_t = \beta_0 + (s+\eta)W_t, \qquad r_t \le \beta \implies W_t \le s/(s+\eta). \tag{EC.30}$$

Auxiliary observations and its supplied certificate contribute zero log-likelihood ratio between P^0 and P^j. A utility-zero declaration is correct under P^0 and incorrect under P^j, regardless of which zero-utility policy is returned. The stopped change-of-measure argument in Appendix A.14 thus gives exactly (12). This proves matching parameter scales for the common certified start without assigning zero cost to acquiring a certificate from data. □

A.11 Proof of Corollary 4.2

Proof of Corollary 4.2. Once every $j \in I_+$ has reached M_j, Q_t must hold at the next decision: a higher-utility candidate not screened by its lower bound is certified; if active it would participate in maximal-utility deployment, and if eliminated its lower bound exceeds $\beta > \theta$.

Reserve disjoint intervals of length $w_{\min}$ for the K candidates in each fresh uniform selection draw. Their total length is below one. Before a candidate exits, it remains active and its interval forces its execution. After exit, retain interval hits only as a comparison count. Thus each comparison count $B_j(t)$ is marginally $\mathrm{Bin}(t, w_{\min})$; cross-candidate independence is unnecessary.

Let $r_{\max} = 128\overline{R}$, $z = \log(2K_+) + x$, and

$$n = \lceil r_{\max}(L_0 + z) \rceil, \qquad T(x) = \left\lceil \frac{4r_{\max}}{w_{\min}} [L_0 + \log(2K_+) + x] \right\rceil. \tag{EC.31}$$

For $x \geq 0$, $T(x)w_{\min} \geq 2n$ and $T(x)w_{\min} \geq 8z$. Equation (EC.17) and a Chernoff lower tail give, for each candidate,

$$\mathbb{P}(\text{not exited by } T(x)) \leq \mathbb{P}(M_j > n) + \mathbb{P}(B_j(T(x)) < n) \leq 2e^{-z}. \tag{EC.32}$$

A union bound gives $\mathbb{P}(\tau_Q - 1 > T(x)) \leq e^{-x}$. No independence between exit times and selection counts is asserted. Integrating this tail, including the ceiling, proves (11). The quality statement uses the same event $\mathcal{E}_{\mathrm{cov}}$ as Theorem 3.1. □

Since $p_j(1-p_j) \leq 1/4$ and $g_j \geq \eta_{\mathrm{eff}}/2$, $R_j \leq \overline{R}$, so all parameters in this tail bound are known. This conservative bound applies to any allocation retaining the floors; it does not replace the refined waiting integral in Theorem 4.1. It does not require $\tau_0 < \infty$.

A.12 Proof of Proposition A.9

Proposition A.9 (First candidate certification)**.** *Suppose a candidate has $p_* = \beta - \Delta_* < \beta$, $\beta \in \mathcal{G}$ and $\lambda_{\min} \leq \Delta_*/5$. For $\gamma, \zeta \in (0,1)$ let*

$$n_* = \left\lceil 128 \left[\frac{p_*(1-p_*)}{\Delta_*^2} + \frac{1}{\Delta_*} \right] \log \frac{2}{\varepsilon_*^U \omega_{\min} \gamma} \right\rceil, \tag{EC.33}$$

$$T_* = \left\lceil \frac{2n_* + 8\log(1/\zeta)}{w_{\min}} \right\rceil. \tag{EC.34}$$

If collection continues as specified in the algorithm rather than being externally terminated at declaration, then with probability at least $1 - \delta - \gamma - \zeta$, joint safety holds and a nonbaseline certificate is obtained by T_*. Other candidate risks are arbitrary.

Proof. Lemma A.5 at $q = \beta$ supplies the target's local deadline. On coverage it cannot be eliminated and its pre-certification probability is at least $w_{\min}$. Pre-generate floor comparison counts before intersecting with coverage. Their binomial lower tail, by the same Chernoff argument as above, contributes at most ζ. Combining the events proves the claim. The target deadline needs neither a target-gap mesh restriction nor a grid-size union bound. □

For a high-probability bound from initialization, use ℓ from Lemma A.6 and $m_j = \lceil 128 R_j \ell \rceil$. With probability at least $1 - \gamma$, all width bounds hold and each candidate exits by m_j. Define

$$d_j(0) = 1 - \beta_0, \quad d_j(n) = \max\{\beta, \min\{1, p_j + b_j(n) + r_{\mathrm{cfg}}\}\} - \beta_0, \tag{EC.35}$$

$$\nu_j(n) = \max\{w_{\min}, as/d_j(n)\}, \quad \nu_* = \max\{w_{\min}, as/(1-\beta_0)\}, \tag{EC.36}$$

$$S = \sum_{j \in I_+} \sum_{n < m_j} \nu_j(n)^{-1}, \quad B = \left\lceil 2S + \frac{2}{\nu_*} \log \frac{1}{\zeta} \right\rceil. \tag{EC.37}$$

The count-layer coupling bounds priority attempts by independent geometric comparison variables with parameters $\nu_j(n)$. For $\lambda = \nu_*/2$, their log moment-generating functions are at most $2\lambda/\nu_j(n)$. Chernoff's inequality gives $A_\infty \leq B$ except with additional probability ζ. Thus, with probability

at least $1-\delta-\gamma-\zeta$, $\tau_Q \le 1+B$, all subsequent deployments attain u_η, and baseline-fallback collection risk is at most $\beta_0+\rho s$. The comparison variables are generated independently of local streams before event intersection. This is a joint-probability statement from initialization, without conditioning on a selected history or future coverage.

A.13 Actual catastrophe counts

For continued collection with baseline fallback after completion, let $C_n = \sum_{t=1}^{n} X_t$ and $V_n = \sum_{t=1}^{n} r_t$, both measured from initialization. On the joint success event of the preceding high-probability bound, completion occurs after at most B collection episodes. Theorem 3.1 therefore gives

$$V_n \le b_n := (\beta_0+\rho s)n + (1-\rho)s\min\{n, B\}. \tag{EC.38}$$

For $\kappa \in (0,1)$ set $x_n = \log[\pi^2 n^2/(6\kappa)]$. With joint probability at least $1-\delta-\gamma-\zeta-\kappa$, simultaneously for every $n \ge 1$,

$$C_n \le b_n + \sqrt{2b_n x_n} + x_n/3. \tag{EC.39}$$

Indeed, for Bernoulli X with conditional mean r, $\mathbb{E}[e^{\lambda X} \mid \mathcal{F}] \le \exp[(e^\lambda-1)r]$. Thus $\exp[\lambda C_n - (e^\lambda-1)V_n]$ is a nonnegative supermartingale from initialization. The inequality $e^\lambda - 1 - \lambda \le \lambda^2/[2(1-\lambda/3)]$ and the choice $\lambda = t/(1+t/3)$, $t = \sqrt{2x/b_n}$, give a tail e^{-x} for exceeding $b_n + \sqrt{2b_n x} + x/3$ while $V_n \le b_n$. A union bound over n proves the result without conditioning on future coverage. This is an absolute risk-envelope bound, not a predicted percentage improvement over a comparator. Empirical catastrophe comparisons require the same episode horizon across methods.

Limits of finite expected completion. A declaration may be wrong on a probability-δ event. Finite expected declaration time therefore does not imply a finite unconditional expected hitting time of a truly optimal safe deployment. For example, choose a grid containing $q^+ \in (\beta, 1)$, one unique high-utility candidate with $0 < p \le \theta$, and otherwise only low-utility certified policies. Along an all-catastrophe local prefix, a reflected geometric component at $1-q^+ > 0$ grows without bound. It eventually rejects that point and all higher reflected grid points, giving $L \ge q^+ > \beta$ and permanent erroneous elimination. Upper certification does not occur on that prefix. The event has positive probability because the candidate retains a positive floor until elimination. Once removed, the unique correct target is never deployed, although the algorithm can declare completion. This distinction is intrinsic to permanent decisions with a nonzero allowed error probability.

A.14 Proof of Theorem 4.3

Proof of Theorem 4.3. Use the instance family and structural common initialization specified in the main text and Appendix A.9. Set $H = 1$ with one action per policy and a Bernoulli catastrophe observation. Any remaining observation has the same conditional law given that indicator across instances. The full-feedback KL is therefore exactly the Bernoulli KL; no trajectory field additionally reveals the instance or risk parameter. Write $p_+ = \beta+\eta$, $p_- = \beta-\eta$, $w_+ = s/(s+\eta)$ and $d_+ = \mathrm{kl}(p_+, p_-)$. If $\mathbb{E}^0(\tau_{\mathrm{dec}} - 1) = \infty$, the conclusion is immediate; otherwise the declaration is almost surely finite under P^0.

Pre-generate the algorithm's external decision randomization, identically distributed under all instances. Do not include fresh policy-draw randomization in the past: conditional execution probabilities must remain $\mathbb{P}(J_t = j \mid \mathcal{F}_{t-1}) = w_{t,j}$. Decisions and declarations are measurable before the current observation. Let W_t be total allocation mass on the M high-utility candidates. Under P^0, mixture risk is $\beta_0 + (s+\eta)W_t$, so $W_t > w_+$ violates collection safety.

Let τ_{cap} be the first decision with $W_t > w_+$ and stop the observed experiment at

$$S_n = \min\{\tau_{\mathrm{dec}} - 1, \tau_{\mathrm{cap}} - 1, n\}. \tag{EC.40}$$

If collection ends at declaration, extend the unobserved allocation sequence by baseline-only decisions. The stopping time S_n is bounded, and every observation before it has $W_t \le w_+$ on every path, including safety failures. Define

$$E_n = \{\tau_{\mathrm{dec}} \le n+1,\ \tau_{\mathrm{dec}} \le \tau_{\mathrm{cap}},\ u_d = 0\}. \tag{EC.41}$$

This event is measurable in the stopped history: declaration is known before the next observation; if an exposure violation stops first it is false; if the cap n is reached, the next decision is already measurable. The weak inequality allows a declaration at the same decision as a proposed unsafe allocation, since that allocation has not yet been executed.

Under P^0, joint correctness requires a utility-zero declaration without a preceding collection violation. Hence $a_n := P^0(E_n)$ increases to at least $1-\delta$. Under P^j, $u_\eta = 1$, so $P^j(E_n) \le \delta$. Only candidate j has a different observation law. Bounded stopping and the likelihood-ratio chain rule [13] yield, for sufficiently large n with $a_n > \delta$,

$$d_+ \mathbb{E}^0 \sum_{t \le S_n} w_{t,j} = \mathrm{KL}(P^0|_{\mathcal{F}_{S_n}} \| P^j|_{\mathcal{F}_{S_n}}) \tag{EC.42}$$

$$\ge \mathrm{kl}(a_n, P^j(E_n)) \ge \mathrm{kl}(a_n, \delta). \tag{EC.43}$$

The equality uses the fact that allocation rules agree at the same observed history and that expected execution counts equal expected predictable allocation mass. Sum over j and apply the pathwise exposure cap:

$$M\,\mathrm{kl}(a_n, \delta) \le d_+ \mathbb{E}^0 \sum_{t \le S_n} W_t \le d_+ w_+ \mathbb{E}^0(\tau_{\mathrm{dec}} - 1). \tag{EC.44}$$

Letting $n \to \infty$ gives the first inequality in (12). We used only bounded stopped experiments; no finite expected stopping time under the positive alternatives is needed.

For the explicit scale, Bernoulli KL satisfies $\mathrm{kl}(x, y) \le (x-y)^2/[y(1-y)]$. Thus

$$d_+ \le \frac{4\eta^2}{(\beta-\eta)(1-\beta-\eta)}, \qquad (\beta-\eta)(1-\beta-\eta) \ge \beta/8, \tag{EC.45}$$

$$\frac{1}{w_+ d_+} \ge \frac{1}{32}\left(\frac{\beta}{\eta^2} + \frac{\beta}{s\eta}\right). \tag{EC.46}$$

For $\delta < 1/4$, $\mathrm{kl}(1-\delta, \delta) = (1-2\delta)\log[(1-\delta)/\delta] \ge \frac{1}{4}\log(1/\delta)$. Combining proves the factor $1/128$. □

The lower bound applies to any valid declaration rule on a class containing both improvement and no-improvement instances with identical initial risk information. Its $M\log(1/\delta)$ term captures the cost of reliably excluding every possible improvement. A promise that an improvement exists, or an objective requiring only one acceptable policy, changes the relevant class of instances.

Observation kernel and information cost. In SafeBAI's linear Gaussian submodel [31], changing a safety coefficient from μ to μ' at dose a gives query KL $a^2(\mu-\mu')^2/(2\sigma^2)$. Changing only candidate j here gives labelled-episode KL $w_j \mathrm{kl}(p_j, p'_j)$. Dose controls signal amplitude; safe mixtures control the frequency of full observations, so their information bounds are not interchangeable. Our lower bound excludes each positive alternative on the no-improvement instance under a shared exposure cap. The upper bound sums local-state waiting in (9), instead of charging all samples at the initial conservative rate. Both use standard stopped-KL and anytime-inference tools to analyze completion from the same baseline-only start.

OPB's multi-armed specialization already uses selected-arm feedback and mixtures constrained by cost upper bounds [23]. This shared collection mechanism does not transfer its regret guarantees to a stopping-time bound for an individually safe recommendation. BAEC [28] addresses two-threshold existence testing without a collection-safety constraint or utility ordering. Its original definition imposes no requirement when the largest mean lies in the neutral interval. JACE instead guarantees finite expected observable completion for every risk vector under Theorem 4.1, with an individually safe output dominating all margin-safe candidates on the coverage event. Local discrimination bounds alone do not account for the certificate-dependent waiting needed to collect those observations.

B SUPPLEMENTARY EXPERIMENTS

B.1 SHARED PROTOCOL

Costs and pairing. A declaration after n collected episodes costs n, including baseline episodes. Capped costs include all planned seeds; nondeclarations contribute the full window. Methods share policy-local observation streams and allocation random numbers, yielding identical selections whenever their probability vectors coincide. Pool exhaustion is a technical error, not censoring. First certification, strict identification and common-window utility are evaluated at their respective endpoints.

Validity and uncertainty. A valid completion satisfies $p_d \leq \beta$ and $u_d \geq u_\eta$. Synthetic risks are known. Control outputs are assessment-supported when the independent upper endpoint for the output is at most β and every higher-utility policy has a lower endpoint above θ. These assessment intervals are withheld from online candidate certification. Means and paired differences use seed-level Student intervals, rates use Clopper–Pearson intervals, and the three allocation contrasts additionally use within-library, within-metric Bonferroni adjustment. Zero-variance bounded costs, event counts and losses use Hoeffding intervals; utility-value summaries retain Student intervals without implying equality of population means.

Shared completion settings. Unless varied explicitly, certificates use 501 base-grid points, 21 geometric stakes and a 90/10 upper/lower budget split. Parameters are $\delta = 0.05$, $\rho = 0.1$, $\eta = 0.02$, $T = 40{,}000$ and $\beta = 0.15$ (0.20 on CarPush1). Allocation controls share certificates, filters, floors and stopping. JACE-UCB prioritizes the smallest risk lower bound, Safe-RR rotates through priority candidates, and Safe-Uniform maximizes feasible uniform excess probability.

Synthetic library. Baseline risk is 0.051; candidate risks are $(0.032, 0.103, 0.151, 0.195, 0.674)$. Each of 200 paired seeds freezes utilities from 1,024 independent Bernoulli development rewards per policy, with respective means $(0.05, 0.02, 0.60, 0.95, 0.30, 0.99)$. Rewards are independent of risk observations. Thus synthetic averages include development and collection randomness, whereas control averages condition on one fixed library and utility vector per task.

B.2 CONTROL POLICY LIBRARIES

Both tasks use 16 training-seed-10 OSRL checkpoints [19]: BC-safe, BC-all, BCQL, BEARL, COptiDICE and CPQ at target costs $\{20, 40, 80\}$, excluding CDT and retaining duplicates once. Released-policy evaluation conventions are used in `OfflinePointGoal1Gymnasium-v0` and `OfflineCarPush1Gymnasium-v0`; the former wraps `SafetyPointGoal1-v0` [8]. Episodes last at most 1,000 steps, and catastrophe means total episode cost strictly above 20.

Preparation is separated from online search. A 2,000-episode pilot per policy determines β and the baseline, after which settings are frozen. Independent development and assessment use 10,000 and 20,000 episodes per policy. Two-sided Clopper–Pearson intervals allocate error $0.05/16$ per policy, while baseline qualification uses a separate one-sided budget $\delta_0 = 0.001$; online certificates use the remaining confidence budget. Each task therefore uses 32,000 pilot, 160,000 development and 320,000 assessment episodes, shared across methods and excluded from online completion costs. Online candidate certificates start without pre-online candidate risk evidence, providing a common baseline-only information state for the comparison. The reported savings concern additional online collection after preparation. Pilot and development episodes support benchmark setup and utility estimation. Assessment episodes support output evaluation and baseline qualification. The experiments do not quantify the benefit of reusing historical candidate risk observations.

Each task uses one fixed policy library and utility vector across ten independent collection repetitions. The resulting intervals quantify online collection variability conditional on that library and utility vector. They do not measure variation across independently trained libraries. Within a repetition, methods reveal only the selected policy's next unused episode from common policy-local streams [22, 35]. These streams are independent of preparation data; Appendix A.1 justifies adaptive revelation. PointGoal1 presamples 40,000 episodes for the baseline, `c40_CPQ`, `c20_CPQ` and `c20_BC-safe`, and 1,000 for each other policy per repetition. CarPush1 generates streams on demand, with a fresh environment instance per episode.

B.2.1 POINTGOAL1

PointGoal1 separates the preferred safe policy from unsafe competitors. At $\beta = 0.15$, simultaneous assessment supports three safe policies and excludes thirteen (Table 3). Baseline `c80_CPQ` has 1,171 catastrophes in 20,000 episodes, giving qualification bound $\beta_0 \approx 0.0639$. Policy `c40_CPQ` has the highest frozen utility among the three, and its risk upper endpoint is below $\theta = 0.13$.

Table 3: **Simultaneous risk intervals lie below the safety limit for three policies and above it for thirteen.** Each policy has 20,000 assessment episodes and a simultaneous 95% risk interval; $\beta = 0.15$. Utilities average 10,000 separate development episodes, and target selection uses their unrounded frozen values.

Policy	Events	Risk interval	Verdict	Utility ↑
`c20_CPQ`	477	[0.0208,0.0272]	safe	-13.334
`c80_CPQ`	1,171	[0.0538,0.0636]	safe	-2.097
`c40_CPQ`	2,087	[0.0981,0.1109]	safe	16.613
`c20_BC-safe`	3,684	[0.1762,0.1924]	unsafe	7.836
`c40_BC-safe`	6,607	[0.3206,0.3403]	unsafe	11.769
`c80_BEARL`	8,792	[0.4292,0.4500]	unsafe	14.697
`c40_BEARL`	8,931	[0.4362,0.4570]	unsafe	14.745
`c80_BC-safe`	11,361	[0.5577,0.5784]	unsafe	17.055
`c40_COptiDICE`	11,549	[0.5671,0.5878]	unsafe	10.680
`c40_BC-all`	12,311	[0.6053,0.6257]	unsafe	18.234
`c20_COptiDICE`	12,376	[0.6086,0.6289]	unsafe	11.167
`c80_COptiDICE`	12,562	[0.6179,0.6382]	unsafe	11.121
`c80_BCQL`	12,596	[0.6196,0.6399]	unsafe	19.788
`c20_BCQL`	13,619	[0.6711,0.6907]	unsafe	21.117
`c40_BCQL`	13,712	[0.6758,0.6953]	unsafe	21.186
`c20_BEARL`	13,777	[0.6791,0.6985]	unsafe	20.946

B.2.2 CARPUSH1

CarPush1 exposes the cost of resolving a near-threshold competitor. The pilot fixes $\beta = 0.20$ and baseline `c80_CPQ`, whose 1,242 assessment catastrophes give $\beta_0 \approx 0.0675$. The four allocation methods obtain a nonbaseline certificate in ten runs and complete in the same seven, returning assessment-supported `c40_BC-safe`. JACE certifies this policy in every run, but the slightly higher-utility `c80_BC-safe` remains unresolved in all three capped runs: its online lower bound is 0.170–0.172, below $\theta = 0.18$. Its independent risk estimate of 0.2008 (Table 4) cannot substitute for the online certificate. Completion requires resolving every higher-utility candidate; first certification requires only one safe candidate. The frozen utility difference between `c80_BC-safe` and `c40_BC-safe` is 0.0087. The completion requirement follows this estimated ordering; the reported utility means alone do not establish an ordering of their true expected returns.

Table 4: **Selected CarPush1 policies explain the certification–completion gap.** The baseline and three candidates from the full 16-policy library. Risk intervals have simultaneous 95% coverage across the library; utilities use independent development episodes.

Policy	Risk estimate	Risk interval	Frozen utility
`c80_CPQ`	0.0621	[0.0572, 0.0673]	−0.0950
`c20_BC-safe`	0.1567	[0.1491, 0.1644]	2.5091
`c40_BC-safe`	0.1872	[0.1791, 0.1955]	2.8004
`c80_BC-safe`	0.2008	[0.1925, 0.2093]	2.8091

B.3 PAIRED COMPARISONS AND ABLATIONS

Allocation gains depend on library structure. For JACE minus JACE-UCB, Safe-RR and Safe-Uniform, paired completion differences are $-1{,}278 \pm 1{,}294$, $-3{,}203 \pm 940$ and $-3{,}734 \pm 946$

episodes on PointGoal1, and -178 ± 53, -517 ± 57 and -732 ± 63 on synthetic data. The corresponding CarPush1 differences are 11 ± 104, -444 ± 269 and -427 ± 263. Here $\pm$ denotes nominal 95% Student half-widths. Three-contrast Bonferroni half-widths are $(1{,}677, 1{,}219, 1{,}226)$, $(65, 70, 77)$ and $(134, 349, 341)$, respectively. JACE's first-certification advantage on CarPush1 is larger: paired differences are $-3{,}950 \pm 1{,}727$, $-9{,}785 \pm 2{,}620$ and $-11{,}643 \pm 2{,}378$ episodes. All methods retain two synthetic false certifications of the risk-0.151 candidate in their reported statistics.

Screening reduces completion cost and catastrophes. Removing utility and margin screening increases both means under either priority rule in all three libraries (Table 1). On CarPush1, paired JACE-minus-unfiltered differences are $-4{,}649 \pm 3{,}233$ episodes and -502 ± 363 catastrophes, using nominal 95% Student half-widths. The ablations retain certificates and safe allocation.

Direct sampling sacrifices collection safety. JACE-direct removes mixture allocation and executes the priority candidate with probability one. Mean identification cost falls from 381 to 267 episodes on thr1 and from 176 to 103 on RealL, but collection-risk excursions occur in 93.8% and 94.7% of runs. On Unique, cost instead rises from 6,376 to 11,451 and identification falls from 996/1,000 to 970/1,000.

B.4 Completion scaling and sensitivity

Prioritization helps locate an improvement, but offers little gain when every candidate must be excluded. At the positive baseline-only default, JACE's paired differences from JACE-UCB, Safe-RR and Safe-Uniform are $-2{,}637 \pm 1{,}441$, $-13{,}963 \pm 1{,}605$ and $-15{,}446 \pm 1{,}604$ episodes, using three-contrast Bonferroni half-widths. Across no-improvement settings, JACE is 0.21%–0.56% slower than JACE-UCB: equal exclusion gaps leave little scope for prioritization.

Baseline-only initialization. The baseline has structural bound $\beta_0 = \beta - s$ and utility zero. Each of M utility-one candidates starts with $U_j = 1$, $L_j = 0$ and risk $\beta + \eta$; positive instances change one risk to $\beta - \eta$. We fix $\beta = 0.15$, $\rho = 0.1$, $\eta = 0.02$ and vary one coordinate of $(M, s, \delta) = (8, 0.048, 0.05)$: $M \in \{2, 4, 8, 16, 32\}$, $s \in \{0.012, 0.024, 0.048, 0.096\}$ or $\delta \in \{0.05, 0.01, 0.005, 0.001\}$. The 22 distinct positive/negative configurations use four methods, 100 paired seeds and a 10^6-episode window. All 8,800 runs declare; Table 5 retains incorrect declarations. Collection-risk excursions occur in 0–3/100 runs per method and setting.

Table 5: **Baseline-only completion separates finding an improvement from excluding all improvements.** Mean completion episodes ↓, 100 paired seeds per setting, 10^6-episode window. Each column varies one parameter from $(M, s, \delta, \eta) = (8, 0.048, 0.05, 0.02)$. Red/blue mark the smallest/second-smallest means. All runs declare; each method is correct in 99/100 runs at positive $M = 2$ and 100/100 elsewhere.

Improvement	Method	Candidates M					Slack s			Confidence δ		
		2	**4**	**8**	**16**	**32**	**0.012**	**0.024**	**0.096**	**0.01**	**0.005**	**0.001**
No	**JACE**	6,433	14,695	32,753	72,652	159,226	96,628	54,899	22,311	41,128	43,801	52,370
	JACE-UCB	6,418	14,665	32,652	72,407	158,529	96,111	54,732	22,187	40,986	43,667	52,209
	Safe-RR	6,432	14,689	32,716	72,568	158,903	96,375	54,840	22,260	41,079	43,710	52,262
	Safe-Uniform	6,436	14,707	32,723	72,585	158,955	96,314	54,891	22,251	41,115	43,743	52,258
Yes	**JACE**	6,880	10,880	18,455	33,886	67,273	46,801	28,293	13,413	21,984	20,942	30,082
	JACE-UCB	7,136	11,744	21,092	38,890	81,555	58,714	33,111	14,568	25,800	27,936	33,965
	Safe-RR	8,247	16,045	32,418	66,982	145,982	80,241	49,237	23,456	40,621	44,448	54,858
	Safe-Uniform	8,388	16,625	33,901	71,171	156,091	92,210	54,086	24,042	42,989	46,298	56,441

Supplied certificates give the same qualitative comparison. Table 6 adds a zero-utility auxiliary of risk β_0 to the same families. Its grid points at and above β are excluded structurally, and remaining capitals start at one. These $M + 2$-policy instances begin after the auxiliary certificate is supplied, excluding its acquisition cost, as analyzed by Corollary A.8. We use the preceding scans plus $\eta \in \{0.01, 0.02, 0.04\}$, with 100 paired seeds and the same window. Without an improvement, JACE costs 0.09%–0.47% more than JACE-UCB.

Grid resolution and lower-certificate budget affect exclusion cost. Figure 5 uses the supplied-certificate start, 100 paired seeds and a 300,000-episode window, varying the floor, grid size or budget split one at a time. Without an improvement, refining the grid from 101 to 1,001 points lowers JACE's cost from 52,616 to 32,693; raising the upper-budget fraction from 0.5 to 0.95 increases cost from 29,489 to 36,064. Every method completes correctly in all sensitivity runs.

Table 6: **JACE reduces mean cost relative to rotating and uniform allocation on improvement instances, while no-improvement costs are closely matched.** Mean completion episodes ↓. Common certified start, 100 paired seeds, 10^6 episodes. Each column varies one parameter from $(M, s, \delta, \eta) = (8, 0.048, 0.05, 0.02)$. Red/blue rank mean costs within each column and improvement condition. All runs declare; each method is correct in 99/100 runs at positive $M = 2$ and 100/100 elsewhere.

Improvement	Method	Candidates M					Slack s			Confidence δ			Margin η	
		2	**4**	**8**	**16**	**32**	**0.012**	**0.024**	**0.096**	**0.01**	**0.005**	**0.001**	**0.01**	**0.04**
No	**JACE**	6,830	15,197	33,288	73,260	159,794	98,218	55,990	22,668	41,734	44,399	52,980	108,754	14,848
	JACE-UCB	6,824	15,171	33,185	72,950	159,165	97,760	55,808	22,562	41,622	44,252	52,843	108,578	14,818
	Safe-RR	6,827	15,186	33,240	73,168	159,514	97,966	55,925	22,626	41,692	44,306	52,888	108,694	14,797
	Safe-Uniform	6,834	15,207	33,246	73,187	159,543	97,930	55,955	22,623	41,711	44,334	52,879	108,689	14,852
Yes	**JACE**	7,231	11,290	18,988	34,282	66,247	47,044	28,614	13,553	21,994	21,298	30,175	61,539	6,294
	JACE-UCB	7,535	12,231	21,662	39,163	81,585	59,814	33,737	14,769	26,109	28,454	34,267	60,711	10,514
	Safe-RR	8,741	16,749	33,242	67,519	146,631	81,896	50,271	23,821	41,173	45,150	55,556	108,575	12,964
	Safe-Uniform	8,928	17,326	34,673	71,730	156,736	93,901	55,184	24,420	43,580	46,987	57,078	112,355	14,433

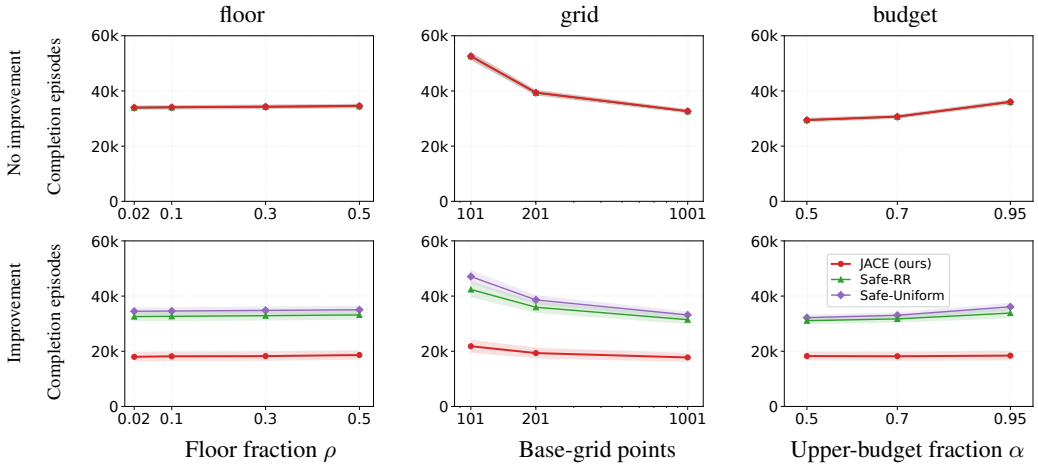


Figure 5: **Finer grids lower mean completion cost, while smaller lower-certificate budgets mainly delay no-improvement decisions.** Top: no improvement; bottom: one improvement. Common certified start, 100 paired seeds per setting, 300,000-episode window. Each column varies one parameter from $\rho = 0.1$, 501 base-grid points and upper-budget fraction $\alpha = 0.9$; the grid axis is logarithmic. Curves show means with pointwise 95% Student intervals; the legend applies to all panels. Every method completes correctly in 100/100 runs at each setting.

B.5 Margin and utility tradeoff

A larger margin can accelerate completion while sacrificing full-safe optimality. We add a risk-0.145, utility-one candidate to the synthetic library, fix all utilities at $(0.05, 0.02, 0.60, 0.95, 0.30, 0.99, 1)$, and scan $\eta \in \{0.004, 0.006, 0.01, 0.02, 0.03\}$ with 200 paired seeds and $T = 40{,}000$. The added candidate leaves the margin comparator from $\eta = 0.006$ onward. For $\eta \leq 0.01$, JACE completes in only 18–19/200 runs, all attaining full-safe optimal utility. At $\eta = 0.02$, all 200 runs complete and the paired cost difference from unfiltered JACE is $-10{,}685 \pm 941$ episodes. The mean loss at declaration is 0.368 relative to the full-safe optimum, whereas the paired common-window utility difference is 0.006 ± 0.015. Both intervals are nominal 95%. At $\eta = 0.03$, declaration loss is 0.390. Each of these two margins retains one incorrect

declaration. Figure 6 separates declaration-conditional loss from all-seed common-window performance.

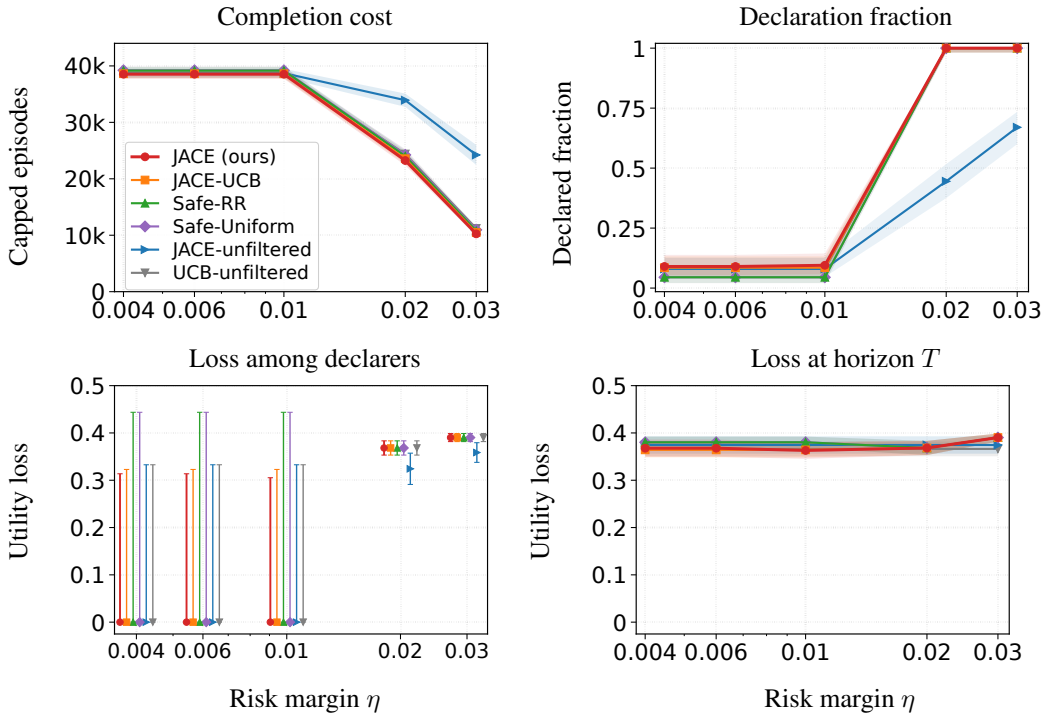


Figure 6: **Larger margins can accelerate declarations, but declaration-only loss can conceal the common-horizon utility gap.** Each method uses 200 paired seeds and $T = 40{,}000$ episodes; horizontal axes are logarithmic. Top: capped completion cost and declaration fraction use all seeds. Bottom left: mean loss among declared runs, shown as points with error bars; small horizontal offsets separate methods. For $\eta \leq 0.01$, only 9–19 declared runs per method enter this panel. Bottom right: loss of the deployment recommendation at the common horizon averages all seeds. Both losses are relative to the full-safe optimum and retain erroneous-declaration paths. Error bars at bottom left and shading elsewhere give pointwise 95% intervals: Clopper–Pearson for fractions, Student for means, and the audited Hoeffding fallback for bounded loss with zero empirical variance. The legend is shared; margins change the comparator, and conditional declaration loss is not the all-seed horizon performance.

B.6 Identification and first certification

Benchmark protocol. The 58 instances comprise six SEE families at eight candidate counts [17] and ten APGAI instances [9], each with 1,000 paired seeds and a 100,000-episode window. Bernoulli successes are converted to catastrophes; Gaussian instances supply binary events $\mathbf{1}\{\text{raw draw} \leq \text{threshold}\}$ to every method. RealL generates Bernoulli observations from public scores. JACE uses a 501-point beta-only grid, $\rho = 0.1$, $\eta = 0.02$ and $\delta = 0.01$.

Strict-identification continuation. Positive identification requires an active nonbaseline $U_j \leq \beta$; negative identification requires $A_t = \emptyset$. If neither holds at first completion, JACE changes priority from V_t to $S_t = \{j \in A_t : U_j > \beta\}$, retaining rotation, evidence ordering, certificates, floors and allocation (4). Proposition A.3 gives joint safety and unchanged first-completion costs on these equal-utility instances. Continuation retains the general risk bound β, not the tighter baseline-fallback bound $\beta_0 + \rho s$. Completion and first-certification experiments retain Algorithm 1.

Baseline implementations. SEE uses the authors' code; HDoC and LUCB-G use its comparison implementations [10]. APGAI uses the authors' sampling rule with the Bernoulli TVT GLRT boundary. SCUCB [20] and CUCB [34] use per-policy Hoeffding certificates and minimum-lower-bound priority. SCUCB applies feasible mixture allocation with floors; CUCB directly samples

a candidate when its cumulative risk budget allows, otherwise the baseline. Uniform rotates candidates with a Hoeffding stopping rule. SEE, APGAI, HDoC, LUCB-G and Uniform impose no collection constraint.

A single interior candidate limits identification speed. On Unique at $K = 50$, JACE takes 6,376 episodes versus 3,826 for SEE and 5,756 for LUCB-G. Only one candidate has risk 0.35; the others sit at $p_j = \beta = 0.5$, with no positive expected log-growth of threshold capital. Equal utilities cannot distinguish candidates before certification, while safe allocation limits exposure to the interior candidate. JACE and APGAI identify in 996/1,000 and 500/1,000 runs, respectively; other methods identify in all runs. APGAI identifies in 998/1,000 RealL runs; all remaining Table 2 cells identify in every run.

Excluding all candidates entails unavoidable baseline exposure. On All-worse, $p_+ = 0.75$, $p_0 = 0.1$ and $\beta = 0.5$. Under the safety constraint, total candidate probability is at most $(\beta - p_0)/(p_+ - p_0) = 8/13$, leaving at least $5/13$ for uninformative baseline episodes. Strict continuation gives correct negative answers in all 1,000 runs at $K = 50, 150, 200$, averaging 13,228, 43,493 and 59,278 episodes with no censoring or observed collection excursions. Across all 58 instances, JACE has nine collection excursions in 58,000 runs (at most 0.3% per instance) and one incorrect positive answer on APGAI thr3; SCUCB has no observed excursions.

Complete results. Figures 7 and 8 cover all 58 instances. SEE uses $K \in \{10, 20, 30, 40, 50, 100, 150, 200\}$. Unique, Quarter, Half and All good contain 1, $\lfloor K/4 \rfloor$, $\lfloor K/2 \rfloor$ and K candidates of risk 0.35, respectively, with others at 0.5. Linear risks span $[0.35, 0.65]$ evenly, and All worse risks equal 0.75. All instances use baseline risk 0.1, baseline utility zero, candidate utilities one and safety threshold 0.5. Gaussian APGAI observations have unit standard deviation. Source data provide full inputs and 95% intervals.

Identification cost

Unique	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	**1,245**	**1,044**	37,839	1,673	1,321	4,455	2,292	1,979
K = 20	2,514	**1,733**	45,310	3,227	**2,412**	9,283	4,400	3,821
K = 30	3,651	**2,394**	50,266	4,811	**3,503**	14,215	6,525	5,689
K = 40	4,789	**3,123**	53,870	6,498	**4,651**	19,388	8,754	7,635
K = 50	6,376	**3,826**	53,850	8,123	**5,756**	24,311	10,932	9,550
K = 100	**11,360**	**7,703**	57,223	17,069	11,798	51,179	22,676	19,823
K = 150	**15,669**	**11,517**	59,990	25,647	17,510	74,649	33,981	29,743
K = 200	**19,662**	**15,758**	65,450	34,988	23,784	89,252	46,209	40,436

Quarter	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	**534**	**934**	18,300	1,557	1,267	3,661	2,139	1,830
K = 20	**434**	**1,240**	3,462	2,566	2,049	5,920	3,550	3,036
K = 30	**484**	**1,548**	1,970	3,502	2,733	8,327	4,881	4,192
K = 40	**472**	1,847	**1,171**	4,397	3,401	10,406	6,141	5,281
K = 50	**494**	2,170	**921**	5,376	4,118	12,845	7,492	6,451
K = 100	**499**	3,511	**607**	9,598	7,199	23,011	13,320	11,514
K = 150	**524**	4,757	**582**	13,721	10,143	33,121	19,036	16,517
K = 200	**513**	6,059	**643**	17,715	13,011	42,763	24,406	21,167

Half	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	**304**	**829**	1,277	1,536	1,322	2,806	2,110	1,764
K = 20	**335**	1,185	**430**	2,624	2,186	5,024	3,639	3,034
K = 30	**343**	1,458	**568**	3,551	2,899	6,959	4,925	4,107
K = 40	**354**	1,801	**381**	4,620	3,741	9,106	6,415	5,361
K = 50	**356**	2,092	**388**	5,510	4,417	10,978	7,608	6,353
K = 100	**387**	3,461	**453**	9,920	7,772	20,218	13,518	11,324
K = 150	**397**	4,461	**540**	14,141	10,954	29,096	19,095	15,996
K = 200	**405**	5,879	**573**	18,233	13,997	37,780	24,336	20,388

All good	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	**248**	822	**290**	1,710	1,528	2,365	2,326	1,898
K = 20	**282**	1,179	**322**	2,912	2,528	4,259	3,973	3,210
K = 30	**284**	1,443	**332**	4,016	3,433	6,048	5,443	4,384
K = 40	**292**	1,729	**348**	5,070	4,288	7,780	6,868	5,536
K = 50	**304**	2,095	**363**	6,167	5,193	9,525	8,327	6,716
K = 100	**321**	3,506	**429**	11,306	9,309	18,016	14,894	11,998
K = 150	**345**	4,544	**498**	15,415	12,520	25,205	19,370	15,488
K = 200	**350**	5,797	**549**	19,992	16,105	32,985	23,767	19,080

Linear	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	**577**	**1,095**	2,239	1,720	1,421	4,237	2,316	1,982
K = 20	**663**	**1,605**	2,123	3,142	2,493	8,008	4,226	3,621
K = 30	**709**	2,066	**1,684**	4,348	3,388	11,274	5,891	5,061
K = 40	**728**	2,495	**1,151**	5,614	4,311	14,681	7,604	6,557
K = 50	**752**	2,957	**1,268**	6,861	5,226	17,970	9,309	8,024
K = 100	**873**	4,772	**1,372**	12,414	9,226	32,790	16,917	14,666
K = 150	**869**	6,314	**1,114**	17,667	12,942	46,838	24,138	20,993
K = 200	**905**	8,185	**1,157**	22,748	16,533	60,442	31,201	27,187

All worse	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	2,277	**1,279**	**1,086**	1,430	1,430	2,223	3,186	3,128
K = 20	4,870	**2,742**	**2,305**	2,994	2,994	5,015	6,675	6,537
K = 30	7,575	**4,220**	**3,571**	4,601	4,601	7,927	10,246	10,046
K = 40	10,388	**5,684**	**4,875**	6,253	6,253	11,058	13,946	13,655
K = 50	13,228	**7,312**	**6,205**	7,892	7,892	14,184	17,579	17,222
K = 100	28,069	**15,109**	**13,051**	16,430	16,430	31,070	36,581	35,850
K = 150	43,493	**23,479**	**20,102**	25,159	25,159	48,647	56,056	54,943
K = 200	59,278	**31,491**	**27,271**	33,998	33,998	66,757	75,769	74,290

APGAI	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
thr1, K=10	**381**	**775**	1,051	1,272	1,071	2,987	1,750	1,488
thr2, K=6	**1,398**	**2,089**	4,187	2,801	2,392	6,564	3,452	2,985
thr3, K=10	**19,540**	76,351	**18,790**	89,405	80,803	99,944	87,859	84,583
med1, K=5	**29,553**	94,021	**36,249**	87,133	80,697	100,000	86,304	83,407
med2, K=7	**221**	348	**202**	430	365	1,192	605	511
isa1, K=10	**387**	**669**	990	1,009	833	2,487	1,419	1,206
isa2, K=7	**129**	**153**	337	185	155	482	303	247
noa1, K=5	1,033	**586**	**495**	670	670	1,292	1,480	1,430
noa2, K=4	**5,231**	5,555	**4,622**	8,461	8,461	30,625	11,519	10,820
RealL, K=18	**176**	**267**	569	426	333	1,528	445	368

Figure 7: **Identification cost across the complete benchmark.** Each cell gives mean capped episodes over 1,000 runs with a 100,000-episode window. Bold red and blue numbers mark the lowest and second-lowest means within each row, respectively. Ranks use unrounded means and do not indicate statistical significance. Blocks group instance families, with repeated method headers. All runs are retained.

Correct identification (%)

Unique	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	100	100	**64.9**	100	100	100	100	100
K = 20	100	100	**57.6**	100	100	100	100	100
K = 30	100	100	**52.8**	100	100	100	100	100
K = 40	**99.9**	100	**48.9**	100	100	100	100	100
K = 50	**99.6**	100	**50**	100	100	100	100	100
K = 100	**98.4**	100	**46.4**	100	100	**99.3**	100	100
K = 150	**96**	100	**44**	100	100	**83.5**	100	100
K = 200	**93.5**	100	**39**	100	100	**47.9**	100	100

Quarter	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	100	100	**84**	100	100	100	100	100
K = 20	100	100	**97.4**	100	100	100	100	100
K = 30	100	100	**99**	100	100	100	100	100
K = 40	100	100	**99.5**	100	100	100	100	100
K = 50	100	100	**99.7**	100	100	100	100	100
K = 100	100	100	**99.9**	100	100	100	100	100
K = 150	100	100	100	100	100	100	100	100
K = 200	100	100	100	100	100	100	100	100

Half	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	100	100	**99.4**	100	100	100	100	100
K = 20	100	100	100	100	100	100	100	100
K = 30	100	100	**99.8**	100	100	100	100	100
K = 40	100	100	100	100	100	100	100	100
K = 50	100	100	100	100	100	100	100	100
K = 100	100	100	100	100	100	100	100	100
K = 150	100	100	100	100	100	100	100	100
K = 200	100	100	100	100	100	100	100	100

All good	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	100	100	100	100	100	100	100	100
K = 20	100	100	100	100	100	100	100	100
K = 30	100	100	100	100	100	100	100	100
K = 40	100	100	100	100	100	100	100	100
K = 50	100	100	100	100	100	100	100	100
K = 100	100	100	100	100	100	100	100	100
K = 150	100	100	100	100	100	100	100	100
K = 200	100	100	100	100	100	100	100	100

Linear	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	100	100	100	100	100	100	100	100
K = 20	100	100	**99.5**	100	100	100	100	100
K = 30	100	100	**99.7**	100	100	100	100	100
K = 40	100	100	100	100	100	100	100	100
K = 50	100	100	**99.9**	100	100	100	100	100
K = 100	100	100	**99.9**	100	100	100	100	100
K = 150	100	100	**99.9**	100	100	100	100	100
K = 200	100	100	100	100	100	**99.9**	100	100

All worse	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	100	100	100	100	100	100	100	100
K = 20	100	100	100	100	100	100	100	100
K = 30	100	100	100	100	100	100	100	100
K = 40	100	100	100	100	100	100	100	100
K = 50	100	100	100	100	100	100	100	100
K = 100	100	100	100	100	100	100	100	100
K = 150	100	100	100	100	100	100	100	100
K = 200	100	100	100	100	100	100	100	100

APGAI	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
thr1 (10)	100	100	100	100	100	100	100	100
thr2 (6)	100	100	100	100	100	100	100	100
thr3 (10)	**99.9**	**81.6**	100	**50.2**	**74.1**	**0.3**	**56.1**	**66.4**
med1 (5)	100	**40.8**	**99.9**	**62.6**	**76.4**	**0**	**65.5**	**71.6**
med2 (7)	100	100	100	100	100	100	100	100
isa1 (10)	100	100	100	100	100	100	100	100
isa2 (7)	100	100	100	100	100	100	100	100
noa1 (5)	100	100	100	100	100	100	100	100
noa2 (4)	100	100	100	100	100	100	100	100
RealL (18)	100	100	**99.8**	100	100	100	100	100

Collection-risk excursions (%)

Unique	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	0	0	0	0	0	0	0	0
K = 20	0	0	0	0	0	0	0	0
K = 30	0	0	0	0	0	0	0	0
K = 40	0	0	0	0	0	0	0	0
K = 50	0	0	0	0	0	0	0	0
K = 100	0	0	0	0	0	0	0	0
K = 150	0	0	0	0	0	0	0	0
K = 200	0	0	0	0	0	0	0	0

Quarter	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	0	0	0	0	0	0	0	0
K = 20	0	0	0	0	0	0	0	0
K = 30	0	0	0	0	0	0	0	0
K = 40	0	0	0	0	0	0	0	0
K = 50	0	0	0	0	0	0	0	0
K = 100	0	0	0	0	0	0	0	0
K = 150	0	0	0	0	0	0	0	0
K = 200	0	0	0	0	0	0	0	0

Half	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	0	0	0	0	0	0	0	0
K = 20	0	0	0	0	0	0	0	0
K = 30	0	0	0	0	0	0	0	0
K = 40	0	0	0	0	0	0	0	0
K = 50	0	0	0	0	0	0	0	0
K = 100	0	0	0	0	0	0	0	0
K = 150	0	0	0	0	0	0	0	0
K = 200	0	0	0	0	0	0	0	0

All good	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	0	0	0	0	0	0	0	0
K = 20	0	0	0	0	0	0	0	0
K = 30	0	0	0	0	0	0	0	0
K = 40	0	0	0	0	0	0	0	0
K = 50	0	0	0	0	0	0	0	0
K = 100	0	0	0	0	0	0	0	0
K = 150	0	0	0	0	0	0	0	0
K = 200	0	0	0	0	0	0	0	0

Linear	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	0	**100**	**100**	**100**	**100**	**100**	0	**99.9**
K = 20	0	**100**	**100**	**100**	**100**	**100**	0	**99.7**
K = 30	0	**100**	**100**	**100**	**100**	**100**	0	**99.9**
K = 40	0	**100**	**100**	**100**	**100**	**100**	0	**100**
K = 50	0	**100**	**100**	**100**	**100**	**100**	0	**100**
K = 100	0	**100**	**100**	**100**	**100**	**100**	0	**100**
K = 150	0	**100**	**100**	**100**	**100**	**100**	0	**100**
K = 200	0	**100**	**100**	**100**	**100**	**100**	0	**99.9**

All worse	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
K = 10	0	**100**	**100**	**100**	**100**	**100**	0	**100**
K = 20	0	**100**	**100**	**100**	**100**	**100**	0	**100**
K = 30	**0.2**	**100**	**100**	**100**	**100**	**100**	0	**100**
K = 40	0	**100**	**100**	**100**	**100**	**100**	0	**100**
K = 50	0	**100**	**100**	**100**	**100**	**100**	0	**100**
K = 100	0	**100**	**100**	**100**	**100**	**100**	0	**100**
K = 150	0	**100**	**100**	**100**	**100**	**100**	0	**100**
K = 200	0	**100**	**100**	**100**	**100**	**100**	0	**100**

APGAI	JACE	SEE	APGAI	HDoC	LUCB-G	Uniform	SCUCB	CUCB
thr1 (10)	0	**100**	**100**	**100**	**100**	**100**	0	**99.4**
thr2 (6)	0	**100**	**100**	**100**	**100**	**100**	0	**100**
thr3 (10)	**0.1**	**100**	**100**	**100**	**100**	**100**	0	**100**
med1 (5)	**0.3**	**100**	**100**	**100**	**100**	**100**	0	**100**
med2 (7)	0	**100**	**100**	**100**	**100**	**100**	0	**96.2**
isa1 (10)	0	**100**	**100**	**100**	**100**	**100**	0	**98.5**
isa2 (7)	0	**100**	**100**	**100**	**100**	**100**	0	**78.5**
noa1 (5)	**0.3**	**100**	**100**	**100**	**100**	**100**	0	**100**
noa2 (4)	0	**100**	**100**	**100**	**100**	**100**	0	**100**
RealL (18)	0	**100**	**100**	**100**	**100**	**100**	0	**88.7**

Figure 8: **Correct identification and collection safety measure different properties.** All values are percentages of the 1,000 runs per cell. Left: valid positive or negative answers, with unresolved and incorrect answers counted as failures. Right: runs with at least one pre-draw mixture risk exceeding β. Red boldface highlights identification rates below 100% and nonzero excursion rates. Ordering follows Figure 7. Parentheses give K for APGAI instances.

First-certification scans. These use 200 seeds and a 150,000-episode window, except JACE-UCB's confidence scan uses 1,000 seeds. The bottom row of Figure 3 includes 3,600 fixed-floor and constant-uniform control runs. Many reach the window, so the comparison concerns capped costs.

C JACE-MS: Finding a Safe Candidate

C.1 Objective and guarantees

JACE-MS studies first-safe-policy search from initialization. It uses a supplied margin Δ and a structural baseline bound, and returns any certified nonbaseline candidate when one has risk at most $\beta - \Delta$. It does not optimize frozen utility or require a negative declaration when no safe candidate exists. Its search–confidence separation therefore concerns a different objective from the completion results in Section 4.

Use baseline index 0, candidates $1, \dots, K$, and $s = \beta - \beta_0$. Let τ_{dep} be the first nonbaseline deployment round; JACE-MS adopts its returned policy in the round following certification. Define

$$D(s, \Delta) = \frac{\beta}{\Delta^2} + \frac{\beta}{s\Delta}. \tag{EC.47}$$

Write $R = \lceil \log_2((1 - \beta)/\Delta) \rceil$ for the number of scales. Grid points, geometric weights and deterministic timeouts are specified below.

Theorem C.1 (Safety and search time)**.** *Under Assumption 2.1, let $K \geq 1$, $0 < \beta < 1/2$, $0 \leq \beta_0 < \beta$, $0 < \Delta \leq \beta/2$, and $0 < \delta < 1/4$. Assume the baseline bound holds structurally. Algorithm 2, with the scales, screening probabilities and budgets and the precommitted grid, stakes, screening quotas and timeouts specified below, satisfies Definition 2.2 on every instance. If $\min_{j \geq 1} p_j \leq \beta - \Delta$, it terminates almost surely and, for a universal C,*

$$\mathbb{E}[\tau_{\mathrm{dep}}] \leq C \mathcal{D}_R \left[K \log(32KR) + \Lambda_{\mathrm{MS}} \log \frac{16e\Lambda_{\mathrm{MS}}}{\Delta^2} \right], \tag{EC.48}$$

where $\mathcal{D}_R := D(s, \Delta) + R/s$, $s = \beta - \beta_0$, and $\Lambda_{\mathrm{MS}} := \log[16R^2 n_g/(\omega_{\min}\delta)]$. Here n_g is the number of risk-grid points and $\omega_{\min}$ the smallest stake weight. The remaining candidate risks are arbitrary in $[0, 1]$.

The proof is in Appendix C.4.

With $n_g = O(1/\Delta + R)$ and $O(\log(1/\Delta))$ geometric stakes with uniform or harmonic weights, this is $\widetilde{O}(D(s, \Delta)[K + \log(1/\delta)])$. The companion lower bound concerns the same first-deployment objective and information access.

Theorem C.2 (Search-time lower bound)**.** *Let $K \geq 1$, $0 < \beta < 1/2$, $0 \leq \beta_0 < \beta$, $0 < \Delta \leq \beta/2$ and $0 < \delta < 1/4$. Consider the Bernoulli family with baseline risk β_0: one instance has all candidate risks $\beta + \Delta$; for each candidate, another instance changes only its risk to $\beta - \Delta$. The procedure observes only selected-policy outcomes, and its initial information has the same distribution across instances. It must satisfy Definition 2.2 on every instance. If a common deterministic deadline T satisfies $\mathbb{P}(\tau_{\mathrm{dep}} \leq T) \geq 1/2$ on every instance containing a safe candidate, then, for a universal $c > 0$,*

$$T - 1 \geq c D(s, \Delta) \max\{K, \log(1/\delta)\}. \tag{EC.49}$$

The proof is in Appendix C.5.

C.2 Algorithm and implementation details

For JACE-MS, the sampled rule ν may be a single candidate or a uniform mixture over a fixed candidate set. If U bounds its risk, selecting ν with probability a and the baseline otherwise has collection risk at most β whenever

$$(1 - a)\beta_0 + aU \leq \beta. \tag{EC.50}$$

This condition uses the group's average risk; it does not require every group member to be individually certified.

Fresh certification. The operation $\textsc{Certify}(\nu, q, \varepsilon, T)$ fixes ν at call start and initializes a new certificate state E with bound $U(E) = 1$. In each subsequent collection episode, it executes ν with probability

$$a(E) = \frac{s}{\max\{2s, U(E) - \beta_0\}}, \qquad s = \beta - \beta_0, \tag{EC.51}$$

and the baseline otherwise. Only an episode executed under ν updates E through (2) and (EC.2); every episode counts toward T. The call returns E as soon as $U(E) \le q$ or after T episodes. All its observations are fresh relative to screening and earlier calls. Thus ν may be selected using the preceding history without compromising conditional coverage. When the bound is valid, (EC.51) satisfies the collection constraint (EC.50).

Scales and certificates. Take $0 < \beta < 1/2$, $0 \le \beta_0 < \beta$, $0 < \Delta \le \beta/2$, and $0 < \delta < 1/4$. Write $N = K + 1$ and $s = \beta - \beta_0$, and set

$$R = \left\lceil \log_2 \frac{1-\beta}{\Delta} \right\rceil, \qquad h_\ell = \frac{1-\beta}{2^{\ell+1}}, \qquad q_\ell = \beta + 2h_\ell \quad (0 \le \ell \le R). \tag{EC.52}$$

Then $q_0 = 1$ and $\Delta/4 < h_R \le \Delta/2$. For nonempty $A \subseteq \{1, \dots, K\}$, let v_A draw a candidate uniformly from A and execute it for an entire episode. Precommit the grid of Section 3.1 to contain $\beta, q_1, \dots, q_R$, with maximum spacing $\varrho \le h_R/4$ on $[\beta, 1]$. Use the pure geometric capital (2) with the smallest nonnegative integer $m_{\max}$ such that $\lambda_{m_{\max}} \le h_R/64$ and fix positive weights ω_m summing to one. Partitioning $[\beta, 1]$ into intervals of length at most $h_R/4$ and adding the required thresholds and zero gives $n_g = O(1/\Delta + R)$. Uniform weights, or normalized weights proportional to $1/((m+1)(m+2))$, give logarithmic mixture overhead.

Budgets and screening. Let $\omega_{\min} = \min_m \omega_m$. For attempt $r \ge 1$, define

$$\xi = \frac{1}{16R}, \qquad B = \log(32KR), \qquad \varepsilon_r = \frac{\delta}{Rr(r+1)},$$
$$L_r = \log \frac{n_g}{\omega_{\min}\varepsilon_r \xi} = \Lambda_{\mathrm{MS}} + \log(r(r+1)), \qquad \Lambda_{\mathrm{MS}} = \log \frac{16R^2 n_g}{\omega_{\min}\delta}. \tag{EC.53}$$

At scale ℓ, a group of size b is screened using

$$a_{\ell-1} = \frac{s}{q_{\ell-1} - \beta_0}, \qquad m_\ell = \left\lceil 9 \frac{q_{\ell-1}}{h_\ell^2} B \right\rceil,$$
$$T_\ell^{\mathrm{scr}}(b) = \left\lceil \frac{8 b m_\ell}{a_{\ell-1}} \right\rceil. \tag{EC.54}$$

The operation $\textsc{Screen}(A, a, m, T)$ fixes A and a, executes v_A with probability a and π^0 otherwise, and stops once every candidate has m new observations or after T collection episodes. It returns the vector of empirical risks from each candidate's first m observations, or $\perp$ if any quota is unmet. Extra observations are excluded. Screening counts and certificate states are reset between phases; no observations are reused.

Certification timeout. For a certificate target (q, h) in attempt r, set

$$C_1 = 259, \qquad n_\star = n_\star(q, h, r) = \left\lceil 1.25 \times 10^5 \frac{q L_r}{h^2} \right\rceil. \tag{EC.55}$$

The default timeout, counting all collection episodes, is

$$T_{\mathrm{expl}}^{\mathrm{cert}}(q, h, r) = \left\lceil \frac{4}{s} \left[(1-\beta_0) \left(1 + \log \frac{1}{\xi} \right) + (n_\star - 1)(q - \beta_0) \right.\right.$$
$$\left.\left. + C_1 \left(2\sqrt{q L_r n_\star} + L_r(1 + \log n_\star) \right) + (n_\star - 1)\varrho \right] \right\rceil. \tag{EC.56}$$

The fresh-certification operation is defined above (Appendix C.2). A group call uses $(\nu, q, h) = (v_{A_\ell}, q_\ell, h_\ell)$; the final call uses $(\nu, q, h) = (\pi^{j^\star}, \beta, \Delta/2)$. Each call stops immediately on certification.

After return, both deployment and collection execute $\pi^{j^\star}$ from the next round onward, retaining its final certificate. A restart advances r, restores the full library, and discards the preceding attempt's screening counts and certificate states.

Algorithm 2 JACE-MS: Multiscale Safe Policy Search

Input: library Π, risk limits $\beta_0 < \beta$, margin Δ, error budget δ, precommitted grid and weights
Initialize scales, budgets and timeouts by (EC.52)–(EC.56)
Deploy π^0 until return
for $r = 1, 2, \ldots$ **do**
  $A_0 \leftarrow \{1, \ldots, K\}$
  for $\ell = 1, \ldots, R$ **do**
    // Screen the current group
    $\widehat{\boldsymbol{p}}_\ell \leftarrow \textsc{Screen}(A_{\ell-1}, a_{\ell-1}, m_\ell, T_\ell^{\mathrm{scr}}(|A_{\ell-1}|))$
    if $\widehat{\boldsymbol{p}}_\ell = \perp$ **then**
      Restart the outer attempt
    end if
    // Select a group or a final candidate
    if $\ell < R$ **then**
      $A_\ell \leftarrow \{j \in A_{\ell-1} : \widehat{p}_{\ell,j} \le \beta + h_\ell/2\}$
      if $A_\ell = \emptyset$ **then**
        Restart the outer attempt
      end if
      $(\nu, q, h) \leftarrow (v_{A_\ell}, q_\ell, h_\ell)$
    else
      $j^\star \leftarrow \min \arg\min_{j \in A_{R-1}} \widehat{p}_{R,j}$
      $(\nu, q, h) \leftarrow (\pi^{j^\star}, \beta, \Delta/2)$
    end if
    // Obtain a fresh certificate before proceeding
    $E \leftarrow \textsc{Certify}(\nu, q, \varepsilon_r, T_{\mathrm{expl}}^{\mathrm{cert}}(q, h, r))$
    if $U(E) > q$ **then**
      Restart the outer attempt
    end if
    if $\ell = R$ **then**
      Retain E and **return** $j^\star$
    end if
  end for
end for

Corollary C.3 (Worst-case optimality up to logarithmic factors)**.** *Fix the numerical parameters of Theorem C.1. Let $\mathcal{C}$ be the class of libraries satisfying Assumption 2.1 with a structural baseline of risk at most β_0, and let $\mathcal{C}_\Delta \subseteq \mathcal{C}$ consist of the libraries whose candidate risk vector $(p_j)_{j\ge1}$ lies in $\mathcal{P}_\Delta = \{p \in [0,1]^K : \min_j p_j \le \beta - \Delta\}$. Consider procedures that observe only the selected policy's trajectory, receive the same Δ, are jointly safe at level $1-\delta$ on every instance of $\mathcal{C}$, including instances with no safe candidate, and whose initial information has the same law across the Bernoulli instances of Theorem C.2. Then*

$$cD(s,\Delta)\max\{K, \log(1/\delta)\} \le \inf_{\mathsf{A}} \sup_{\mathcal{I}\in\mathcal{C}_\Delta} \mathbb{E}_{\mathcal{I}}^{\mathsf{A}} \tau_{\mathrm{dep}} \le C\mathcal{D}_R \left[K \log(32KR) + \Lambda_{\mathrm{MS}} \log \frac{16e\Lambda_{\mathrm{MS}}}{\Delta^2} \right]. \tag{EC.57}$$

For the explicit grid and weight choices above, the bounds match up to logarithmic factors. This is worst-case optimality over $\mathcal{C}_\Delta$, not instance-optimality.

The proof is in Appendix C.6.

C.3 Proof of Lemma C.4

We prove Theorem C.1 under its stated structural baseline assumption. Candidates are indexed $1, \ldots, K$, with baseline 0. Every time below counts complete collection episodes; deployment starts one round after the successful update. All algorithm parameters depend only on the supplied inputs and precommitted grids.

Use the grids, budgets, screening quotas and default timeout specified in Appendix C. The auxiliary envelope

$$T^{\mathrm{cert}}(q,h,r) = \left\lceil c_{\mathrm{cert}} \left(\frac{q}{h^2} + \frac{q}{sh} \right) L_r \log \frac{eqL_r}{h^2} \right\rceil, \tag{EC.58}$$

with $c_{\mathrm{cert}} = 2.2 \times 10^6$ bounds the default timeout (EC.56). Lemma C.4 establishes the required completion probability for the explicit timeout as well. The proof uses the local notation $L = L_r$ and $N = n_\star(q,h,r)$; this N is a certificate observation quota, not the policy-library size.

Lemma C.4 (Certification with fresh data)**.** *Fix the information at any finite call-start stopping time σ. Let v be a nonbaseline candidate or a uniform mixture of nonbaseline candidates, fixed by that information, with risk $p \in [0,1]$. With fresh observations and budget ε_r, the certificate covers p at all local times with conditional probability at least $1-\varepsilon_r$. For the time bound, suppose $\beta \le q < 1$, $0 < h \le q$, $q \le \beta + 2h$, and $p \le q - h$. If the risk grid contains q, its mesh above β is at most $h/4$, and its smallest geometric stake is at most $h/64$, the call succeeds within (EC.56), and hence within (EC.58), with conditional probability at least $1 - 2\xi$.*

Proof. Condition on $\mathcal{F}_\sigma$. For a group A, pre-generate independent pairs of a uniform index in A and its fresh rollout. Their catastrophe indicators are i.i.d. Bernoulli(p_A), where $p_A = |A|^{-1} \sum_{j \in A} p_j$. Reveal the next pair whenever the group is drawn. The group may depend on all data preceding σ; its new potential stream remains independent of that history conditional on the fixed group.

Coverage. Lemma A.1 applies conditionally on the call-start history to the fresh Bernoulli group stream. It spends one budget per call.

A local rejection deadline. Put $L = L_r$. For any grid point $u \ge q$, write $d = u - p \ge h$ and

$$\lambda^\star = \min\left\{ \frac{1}{2}, \frac{d}{8(d^2 + 2p)} \right\}. \tag{EC.59}$$

Since $d \le 1$ and $p \le 1$, $\lambda^\star \ge h/24$. The geometric grid therefore supplies $\lambda \in [\lambda^\star/2, \lambda^\star]$. For $X_i = \log[1 + \lambda(u - Z_i)]$, the inequality $\log(1+x) \ge x - x^2$ on $[-1/2, 1/2]$ gives

$$\begin{aligned} \mu_X := \mathbb{E}X_i &\ge \lambda d - \lambda^2(p + d^2) \ge \tfrac{3}{4}\lambda d, \\ \mathrm{Var}(X_i) &\le 4\lambda^2 p, \quad |X_i - \mu_X| \le 2\lambda. \end{aligned} \tag{EC.60}$$

The variance and range bounds follow because the two possible log factors differ by at most 2λ. Bernstein's inequality, with downward deviation $n\lambda d/4$, yields the explicit step

$$\mathbb{P}\left(\sum_{i=1}^n X_i < \tfrac{1}{2} n\lambda d \right) \le \exp\left[-\frac{nd^2}{128p + (16/3)d} \right] \le \exp\left[-\frac{nd^2}{128(p+d)} \right]. \tag{EC.61}$$

Also $1/(\lambda d) \le 32(1 + p/d^2 + 1/d)$. Thus $n \ge 128L(1 + p/d^2 + 1/d)$ ensures both $n\lambda d/2 \ge \log(1/(\varepsilon_r \omega_m))$ for the selected component and a failure probability at most $e^{-L} \le \xi/n_g$. On success, $K_n(u) \ge 1/\varepsilon_r$. Since $L(1 + p/d^2 + 1/d) \ge 1$, the integer deadline $\lceil 128L(1 + p/d^2 + 1/d) \rceil$ is at most $C_0 L(1 + p/d^2 + 1/d)$ with $C_0 = 129$. A union bound over the at most n_g thresholds $u \ge q$ gives an event $\mathcal{E}_{\mathrm{w}}$ of probability at least $1 - \xi$ on which all deadlines hold. Permanent rejections preserve them at later times.

Width and accumulated uncertainty. Let

$$w(n) = C_1 \left(\sqrt{\frac{pL}{n}} + \frac{L}{n} \right) + \varrho, \qquad N = \left\lceil C_2 \frac{qL}{h^2} \right\rceil. \tag{EC.62}$$

With $C_1 = 259$ and $C_2 = 1.25 \times 10^5$, on $\mathcal{E}_{\mathrm{w}}$,

$$\begin{aligned} U_n \le \max\{q, p + w(n)\} \quad (n \ge 1), \qquad & w(N) \le h, \\ \sum_{n=1}^{N-1} w(n) &\le C \left[\frac{qL}{h} + L \log \frac{eqL}{h^2} \right]. \end{aligned} \tag{EC.63}$$

For completeness, set $b = C_1(\sqrt{pL/n}+L/n)$. If $\max\{q, p+b+\varrho\} \geq 1$, the envelope follows from $U_n \leq 1$. Otherwise $b < 1$, and every $d \geq b$ has deadline at most $C_0 n(2/C_1 + 1/C_1^2) \leq n$: here $d \leq 1$ gives $L \leq n/C_1$, $p/d^2 \leq n/(C_1^2 L)$ and $1/d \leq n/(C_1 L)$, and $129(2/259 + 1/259^2) < 1$. The smallest grid point at least $\max\{q, p+b\}$ and all larger points have been rejected, proving the envelope by the successor rule and the mesh bound. The definition of N, $p \leq q$, $h \leq q$ and $\varrho \leq h/4$ imply $w(N) \leq h$, because $C_1(C_2^{-1/2} + C_2^{-1}) \leq 0.74 < 3/4$, hence $U_N \leq q$. Finally sum $n^{-1/2}$ and n^{-1}, and use $N\varrho = O(qL/h)$ to obtain the last inequality.

An auxiliary continuation and geometric waiting. For analysis, ignore both early success and the timeout and continue the same certificate and sampling rule indefinitely. Every sampling probability is bounded away from zero. The potential local stream and independent random numbers for the waits realize this continuation; the actual call coincides with it until success or timeout. If the continuation obtains N observations within the prescribed timeout on $\mathcal{E}_{\mathrm{w}}$, the actual call must have succeeded by then.

Define deterministic quantities, conditional on the call-start information,

$$b_0 = 1 - \beta_0, \qquad \nu_n = \frac{s}{2b_n}, \tag{EC.64}$$

$$b_n = \min\{1 - \beta_0, q - \beta_0 + w(n)\} \quad (1 \leq n < N). \tag{EC.65}$$

On $\mathcal{E}_{\mathrm{w}}$, $U_n - \beta_0 \leq b_n$ and $b_n \geq s$, so the continuation's probability of its next local observation is at least ν_n. Using a separate independent uniform sequence for each waiting interval couples its waits below independent $G_n \sim \mathrm{Geom}(\nu_n)$, supported on $\{1, 2, \ldots\}$, on $\mathcal{E}_{\mathrm{w}}$. The quantities involving p are only proof devices.

Put $\mu_G = \sum_{n=0}^{N-1} \nu_n^{-1}$ and $\nu_* = \min_n \nu_n = s/[2(1-\beta_0)]$. For $0 \leq \theta \leq \nu/2$, the geometric log moment-generating function satisfies

$$\frac{d}{d\theta} \log \mathbb{E}e^{\theta G} = \frac{1}{1-(1-\nu)e^{\theta}} \leq \frac{2}{\nu}, \qquad \log \mathbb{E}e^{\theta G} \leq \frac{2\theta}{\nu}. \tag{EC.66}$$

Indeed $e^{\theta} \leq (1-\theta)^{-1}$ gives a denominator at least $\nu/2$. Taking $\theta = \nu_*/2$ and applying Markov's inequality yields

$$\mathbb{P}\left(\sum_{n=0}^{N-1} G_n > 2\mu_G + \frac{2}{\nu_*}\log\frac{1}{\xi}\right) \leq \xi. \tag{EC.67}$$

The deterministic time threshold $2\mu_G + (2/\nu_*)\log(1/\xi)$ is bounded by $T^{\mathrm{cert}}_{\mathrm{expl}}(q,h,r)$ in (EC.56), with $L = L_r$ and $N = n_\star(q,h,r)$. This bound is obtained from $2\mu_G + (2/\nu_*)\log(1/\xi)$ by bounding $b_n \leq q - \beta_0 + w(n)$, replacing p by q in w, and summing $n^{-1/2}$ and n^{-1}. Write $\Phi := (q/h^2 + q/(sh))L\log(eqL/h^2)$; since $q/h \geq 1$, $h^2 \leq q$ and $\log(eqL/h^2) \geq 1$, we have $\Phi \geq \max\{qL/h^2, qL/(sh), L/s, 1/s\}$. Using $q - \beta_0 \leq s + 2h$, $N - 1 \leq C_2 qL/h^2$, $N \leq 2C_2 qL/h^2$, $\varrho \leq h/4$, $\log(1/\xi) \leq L$ and $1 + \log N \leq (1 + \log(2C_2))\log(eqL/h^2)$, the entire pre-ceiling expression, including the factor $4/s$, is at most $[8 + 9C_2 + 8C_1\sqrt{2C_2} + 4C_1(1 + \log(2C_2))]\Phi \leq 2.2 \times 10^6\,\Phi$. Hence $T^{\mathrm{cert}}_{\mathrm{expl}} \leq T^{\mathrm{cert}}$ in (EC.58) with $c_{\mathrm{cert}} = 2.2 \times 10^6$, and either timeout may be used. The width and waiting failure probabilities sum to 2ξ. This argument uses a coupling before conditioning on the width event; it does not assume that actual adaptive waits are independent after such conditioning. □

C.4 Proof of Theorem C.1

Proof of Theorem C.1. Joint safety on every instance. Each call starts at a stopping time, fixes its group or candidate using the available history, and uses a fresh stream. Its conditional coverage failure probability is at most ε_r by Lemma C.4; the coverage part does not require $p \leq q - h$. Attempt r launches at most R calls. For a potentially unlaunched call, multiply its conditional failure bound by the indicator that it is launched and apply the tower property. The countable union bound gives

$$\mathbb{P}(\text{some launched certificate fails}) \leq \sum_{r \geq 1} R\varepsilon_r = \delta \sum_{r \geq 1} \frac{1}{r(r+1)} = \delta. \tag{EC.68}$$

On the complementary event, the initial group has deterministic bound $q_0 = 1$. During any completed group's subsequent screening phase,

$$p_c(\mu_t \mid \mathcal{F}_{t-1}) \le \beta_0 + a_{\ell-1}(q_{\ell-1} - \beta_0) = \beta. \tag{EC.69}$$

Within a new certificate call, if $U - \beta_0 \le 2s$, its mixing probability is $1/2$ and risk is at most $\beta_0 + (U - \beta_0)/2 \le \beta$; otherwise risk is at most $\beta_0 + s = \beta$. The baseline is deployed until return, and the final valid certificate permits the returned candidate from the next round. Retaining that certificate permits its subsequent deployment and collection. This proves joint safety. No screening-accuracy event enters this argument.

One screening stage. Fix a candidate j_s with $p_{j_s} \le \beta - \Delta$. On the successful prefix of an attempt, suppose $j_s \in A_{\ell-1}$ and every $j \in A_{\ell-1}$ has $p_j \le q_{\ell-1}$. This holds initially because $q_0 = 1$. Given the current history, define the first m_ℓ fresh potential observations for each remaining candidate. Bernstein's inequality gives

$$\mathbb{P}\left(|\widehat{p}_{\ell,j} - p_j| > h_\ell/2 \mid \text{current history}\right) \le 2\exp\left[-\frac{m_\ell h_\ell^2}{8(p_j + h_\ell/6)}\right] \le \frac{\xi}{K}. \tag{EC.70}$$

The last step follows from $p_j \le q_{\ell-1}$ and $h_\ell \le q_{\ell-1}/4$:

$$8(p_j + h_\ell/6) \le \tfrac{25}{3} q_{\ell-1}, \qquad \frac{m_\ell h_\ell^2}{8(p_j + h_\ell/6)} \ge \frac{27}{25} B, \tag{EC.71}$$

$$2e^{-27B/25} \le 2e^{-B} = \xi/K. \tag{EC.72}$$

Thus the screening coefficient 9 suffices; $B = \log(32KR)$ retains the prescribed error allocation. A union bound makes all estimates accurate with probability at least $1 - \xi$.

Extend the fixed screening allocation to its full timeout using potential draws. Each candidate count is binomial with mean at least $8m_\ell$, so

$$\mathbb{P}(N_j < m_\ell) \le e^{-49m_\ell/16} \le e^{-3m_\ell}, \qquad \mathbb{P}(\text{some count is insufficient}) \le Ke^{-3m_\ell} \le \xi. \tag{EC.73}$$

Here $m_\ell \ge B = \log(2K/\xi)$. Combining these two events, rather than conditioning concentration on successful completion, gives stage completion and accuracy with probability at least $1 - 2\xi$.

For $\ell < R$, on this event $\widehat{p}_{\ell,j_s} \le \beta - \Delta + h_\ell/2 < \beta + h_\ell/2$, so j_s survives. Every survivor satisfies

$$p_j \le \widehat{p}_{\ell,j} + h_\ell/2 \le \beta + h_\ell, \qquad p_{A_\ell} \le \beta + h_\ell = q_\ell - h_\ell. \tag{EC.74}$$

Thus the fresh group call satisfies Lemma C.4, succeeds by its timeout with probability at least $1 - 2\xi$, and permits the next scale. The per-candidate induction follows from accurate screening, not from a bound on the group average alone.

Final selection. On the successful prefix through scale R,

$$p_{j^\star} \le \widehat{p}_{R,j^\star} + h_R/2 \le \widehat{p}_{R,j_s} + h_R/2 \le p_{j_s} + h_R \le \beta - \Delta/2. \tag{EC.75}$$

The final call therefore uses $q = \beta, h = \Delta/2$. Every call satisfies the lemma's resolution conditions because its $h \ge h_R$, and its target is on the precommitted grid. Group calls also satisfy $q_\ell = \beta + 2h_\ell < 1$ and $h_\ell \le q_\ell$.

Conditional attempt success. There are R screening phases, each with conditional failure at most 2ξ, and at most R calls, each with conditional timeout probability at most 2ξ along the successful prefix. Sequential union bounds give failure probability at most $4R\xi = 1/4$, conditional on any attempt-start history. The prefix events concern completed phases, so subsequent applications condition only on already available information. Resetting the full library and taking fresh observations makes the same bound valid after any preceding failure. If A is the number of attempts,

$$\mathbb{P}(A \ge r) \le 4^{1-r}, \qquad \mathbb{P}(A = \infty) = 0. \tag{EC.76}$$

A deterministic cap on every attempt. Write $\mathcal{D}_R = D(s, \Delta) + R/s$. The geometric scales imply

$$\begin{aligned} S &:= \sum_{\ell=1}^{R} \left(\frac{\beta}{h_\ell^2} + \frac{1}{h_\ell} + \frac{\beta}{s h_\ell} + \frac{1}{s} \right) \\ &\le \frac{64\beta}{3\Delta^2} + \frac{8}{\Delta} + \frac{8\beta}{s\Delta} + \frac{R}{s} \le 26\mathcal{D}_R. \end{aligned} \tag{EC.77}$$

Indeed $\sum h_\ell^{-2} \le 4/(3h_R^2) < 64/(3\Delta^2)$, $\sum h_\ell^{-1} \le 2/h_R < 8/\Delta$, and $8/\Delta \le 4\beta/\Delta^2$. As $b \le K$ and $a_{\ell-1}^{-1} = 1 + 4h_\ell/s$, integer rounding in (EC.54) gives

$$T_\ell^{\mathrm{scr}}(b) \le CKB\frac{\beta + 4h_\ell}{h_\ell^2}\left(1 + \frac{4h_\ell}{s}\right) + 1. \tag{EC.78}$$

Consequently all screening timeouts sum to at most $CKB\mathcal{D}_R$. For group certification, $q_\ell/h_\ell^2 + q_\ell/(sh_\ell) = \beta/h_\ell^2 + 2/h_\ell + \beta/(sh_\ell) + 2/s$; their sum is at most $2S$. The final coefficient is $4\beta/\Delta^2 + 2\beta/(s\Delta) \le 4D(s,\Delta)$. All certificate margins satisfy $h \ge h_R > \Delta/4$ and targets satisfy $q \le 1$, so $\log(eqL_r/h^2) \le \log(16eL_r/\Delta^2)$. Thus, on every path, attempt r lasts at most

$$B_r \le C\mathcal{D}_R\left[KB + L_r \log\frac{16eL_r}{\Delta^2}\right]. \tag{EC.79}$$

This deterministic bound includes inaccurate screening and failed certificates. Rounding costs are absorbed using $\mathcal{D}_R \ge R/s \ge R$.

Unconditional expectation. Equations (EC.76) and (EC.79) imply

$$\mathbb{E}\tau_{\mathrm{dep}} \le 1 + \sum_{r\ge 1} 4^{1-r} B_r. \tag{EC.80}$$

Put $u_r = \log(r(r+1))$ and $V = \log(16e\Lambda_{\mathrm{MS}}/\Delta^2)$. Since $L_r = \Lambda_{\mathrm{MS}} + u_r$, $\Lambda_{\mathrm{MS}}, V \ge 1$, and $\log(1+x) \le x$,

$$L_r \log\frac{16eL_r}{\Delta^2} \le \Lambda_{\mathrm{MS}}V + u_rV + u_r + u_r^2/\Lambda_{\mathrm{MS}}. \tag{EC.81}$$

The sums of 4^{1-r}, $4^{1-r}u_r$, and $4^{1-r}u_r^2$ are universal finite constants. Substitution gives (EC.48), including the single extra deployment round. The expectation is not conditioned on the coverage event. □

C.5 Proof of Theorem C.2

Proof of Theorem C.2. Use $H = 1$ and one action per policy. The baseline has Bernoulli risk β_0 on every instance. On $\mathcal{M}^0$, every candidate has risk $\beta + \Delta$; on $\mathcal{M}^j$, only j changes to $\beta - \Delta$. All other observations and the law of $\mathcal{F}_0$ are common. Set

$$w_+ = \frac{s}{s+\Delta}, \qquad B_\Delta = \max\{\mathrm{kl}(\beta-\Delta, \beta+\Delta), \mathrm{kl}(\beta+\Delta, \beta-\Delta)\} \le \frac{4\Delta^2}{(\beta-\Delta)(1-\beta-\Delta)}. \tag{EC.82}$$

On $\mathcal{M}^0$, any nonbaseline deployment or candidate mass above w_+ violates joint safety. The per-round KL divergence in either direction between $\mathcal{M}^0$ and $\mathcal{M}^j$ is at most $w_{t,j}B_\Delta$, since only outcomes from j differ.

Search. Let $\tau_{\mathrm{over}} = \inf\{t : \sum_{i\ge 1} w_{t,i} > w_+\}$, $\tau = \tau_{\mathrm{dep}} \wedge \tau_{\mathrm{over}}$, $S = (\tau - 1) \wedge (T-1)$ and $E_T = \{\tau \le T\}$. Decisions are predictable, so S is a bounded observation stopping time and $E_T \in \mathcal{F}_S$: for $t < T-1$, $E_T \cap \{S \le t\} = \{\tau \le t+1\} \in \mathcal{F}_t$, and for $t \ge T-1$ use $E_T \in \mathcal{F}_{T-1}$. The stopped chain rule [13] gives

$$\mathrm{KL}(\mathbb{P}^0|_{\mathcal{F}_S} \| \mathbb{P}^j|_{\mathcal{F}_S}) \le B_\Delta \mathbb{E}^0 \sum_{t\le S} w_{t,j}. \tag{EC.83}$$

Summing candidate exposures gives at most $(T-1)w_+$, so some j has divergence at most $(T-1)w_+B_\Delta/K$. Safety yields $\mathbb{P}^0(E_T) \le \delta$, while the deadline assumption gives $\mathbb{P}^j(E_T) \ge 1/2$. Data processing therefore implies

$$(T-1)w_+B_\Delta \ge K\,\mathrm{kl}(\delta, 1/2) \ge K/8. \tag{EC.84}$$

Confidence. For any fixed j, instead stop before $w_{t,j} > w_+$ or a nonbaseline deployment. The same measurability argument applies. Both events violate safety on $\mathcal{M}^0$, and reversing the KL direction gives

$$(T-1)w_+B_\Delta \ge \mathrm{kl}(1/2, \delta) = \frac{1}{2}\log\frac{1}{4\delta(1-\delta)}. \tag{EC.85}$$

For $0 < \delta < 1/4$, the last expression is bounded below by a universal positive multiple of $\log(1/\delta)$. Also $(\beta - \Delta)(1 - \beta - \Delta) \geq \beta/8$ and

$$\frac{1}{w_+ B_\Delta} \geq \frac{(\beta - \Delta)(1 - \beta - \Delta)}{4\Delta^2} \left(1 + \frac{\Delta}{s}\right) \geq \frac{1}{32} D(s, \Delta). \tag{EC.86}$$

Combining (EC.84) and (EC.85) proves the claim. The horizon is $T - 1$ because round-T deployment uses only earlier observations. □

Fixed allocations. The same argument explains the fixed-allocation controls. With one candidate, additionally require safety when its risk is one. Any deterministic, observation-independent weight sequence must then obey $w_{t,1} \leq s/(1 - \beta_0)$ at every round. Comparing candidate risks $\beta - \Delta$ and $\beta + \Delta$ gives

$$T - 1 \geq \frac{1 - \beta_0}{s} \frac{(\beta + \Delta)(1 - \beta - \Delta)}{4\Delta^2} \operatorname{kl}(1/2, \delta). \tag{EC.87}$$

Indeed, the fixed exposure cap holds on all instances; apply the stopped KL argument with $S = (\tau_{\mathrm{dep}} - 1) \wedge (T - 1)$. Relative to adaptive exposure $s/(s + \Delta)$, fixed collection pays an additional factor $(1 - \beta_0)/(s + \Delta)$ in this lower bound.

C.6 Proof of Corollary C.3

Proof. The hard positive instances of Theorem C.2 belong to $\mathcal{C}_\Delta$, and its all-unsafe instance belongs to $\mathcal{C}$. For any admissible procedure, let $M_* = \sup_{\mathcal{I} \in \mathcal{C}_\Delta} \mathbb{E}_{\mathcal{I}} \tau_{\mathrm{dep}}$. If finite, Markov's inequality gives success probability at least $1/2$ by $T = \lceil 2M_* \rceil$ on every hard positive instance. Since $T - 1 < 2M_*$, Theorem C.2 gives the lower bound; the infinite case is immediate. Algorithm 2 is jointly safe on $\mathcal{C}$, and Theorem C.1 supplies the uniform upper bound on $\mathcal{C}_\Delta$. Finally, $D(s, \Delta) \geq 2/s$ gives $\mathcal{D}_R \leq (1 + R/2) D(s, \Delta)$; the specified grid and weights incur only logarithmic overhead. □