

CLARK Criterion: Proof-Assisted Paired Monte Carlo for Risk-Constrained Planning

Lloyd Clark, PhD

Federal AI · October 27, 2026 · version 0.6.1

Abstract

We present a reproducible framework for choosing among frozen interventions under shared resources and uncertain availability. CLARK Criterion, the mathematical evaluation engine within the CLARK Frontier workbench, combines resource-infeasibility tests, paired anytime-valid inference and risk-constrained finite-menu selection. A sound resource witness determines a binary failure label without a scheduling rollout. Conditional on a fixed simulator, independent pilot fitting and fresh iid regime-level world streams, simultaneous confidence bounds support adaptive allocation, count-based stopping, risk-plus-cost regret bounds, minimum externally priced menu-cost decisions, and abstention. The seven propositions specify this composition; they do not establish a new generic Monte Carlo convergence rate, an optimal acquisition rule or mathematical novelty of the ingredients. Corrected pseudorandom studies compare adaptive and static allocation, paired and marginal inference, and separate risk feasibility from relative ranking. The release fixes inconsistent completion events, calendar weather initialization, cross-replication seed reuse and candidate-controlled resource prices. Screening checks exercise 12,288 labels, including 6,342 replaced rollouts; selective computational savings do not imply universal speedup. Three synthetic profiles illustrate sport-pilot training, clinical placements and military training. A frozen-proposal tournament, operational adapters and source-linked knowledge contract support further studies. No real operational cohort, authenticated live language model or superintelligent system is evaluated. The contribution is a specified, inspectable implementation of simulator-conditional decision evidence, with negative results and reproducible adversarial checks.

1 Introduction

A training operator deciding whether to add an instructor needs more than a predicted graduation date. The useful question is whether that intervention improves the probability of meeting a declared target, at an acceptable cost, despite uncertain weather, staffing, retrials, and learner availability. Advancing one learner can delay another competing for the same resource. The same structure appears in flight schools, clinical placements, technical qualification, and maintenance certification.

This is also an operational AI evaluation problem. A model may propose a persuasive schedule that violates capacity, recommend an expensive intervention without material benefit, or promise a target that no feasible schedule can meet. An evaluator should distinguish a resource impossibility, a poor decision, and an unresolved comparison. Its output should identify a supported action, the remaining uncertainty, and the next experiment that could change that action.

CLARK Criterion evaluates frozen plans against the same simulated futures and adaptively selects the next disruption regime to sample. It bounds how much worse each plan could be than the best listed alternative and separately bounds absolute target-failure risk. Being close to the best plan does not make a target acceptable. When a resource inequality already determines failure, a checkable witness replaces a full scheduling rollout while retaining the failure in the sample.

The evaluation target is a finite menu of frozen plans under an explicit simulator. It is not general intelligence or AI safety. A resource witness can apply even to a planner with perfect foresight, but that establishes a physical limitation of a sampled future, not a measurement of intelligence. The built-in learned candidate is a fitted ridge model. Imported proposals retain provenance declarations without authenticating model authorship; the supplied experiments establish no live-language-model ranking.

The candidate contribution is the precise composition of resource proofs, paired sequential inference, and actionable regret semantics, supplied as reproducible software. Its mathematical ingredients have substantial prior art. Whether the composition and implementation are sufficiently original for a particular venue requires independent review. The experiments deliberately retain cases in which simpler evaluation methods perform as well or better.

2 Related work

2.1 Simulation and resource reasoning

Pilot training simulation predates this work. Okal models F-16 training resources, Erickson studies aircraft systems and student flow, Dyer and colleagues examine Undergraduate Pilot Training 2.5, and Case studies weather-sensitive sortie generation (Okal 2008; Erickson 2014; Dyer et al. 2023; Case 2025). A graduation simulator alone is not a new contribution. Language-model-generated scheduling priorities and simulation-based prompt selection are also established (Luo, Vanhoucke, and Coelho 2026; Zhang et al. 2024).

Energetic reasoning for cumulative scheduling already compares required work with available capacity (Vilim 2009). Bousquet uses monotonicity information to bound and accelerate Monte Carlo probability estimation (Bousquet 2012). We do not claim to invent capacity bounds or proof-based acceleration. Our implementation uses a weak aggregate relaxation, shares witnesses across policies with the same resource plan, retains every proved failure, and passes those exact labels into the same sequential inference process as rolled-out labels.

2.2 Adaptive evaluation and anytime inference

Active testing selects informative observations for risk estimation (Kossen et al. 2021). Prediction-powered inference corrects model predictions using measured residuals (Angelopoulos et al. 2023; Angelopoulos, Duchi, and Zrnic 2024). Prediction-Powered Active Testing combines active acquisition and prediction-powered control variates (Ashouritaklimi et al. 2026). Evaluating AI with fewer samples is already an established research objective; the present implementation does not reproduce or outperform those general frameworks.

Time-uniform inference is established (Howard et al. 2021). Lindon and Malek provide anytime inference for multinomial counts and Bernoulli contrasts (Lindon and Malek 2022). Our beta-mixture construction is an elementary instance of this literature. Decomposing a paired binary difference into disagreement probability and conditional sign makes ties informative, without creating a new

general convergence law. Our allocation heuristic inherits validity from the confidence event, not from an optimal acquisition theorem.

Ao and colleagues combine biased LLM proxies, selective human audits, propensity correction and anytime confidence sequences for best-arm identification (Ao et al. 2026). That work is directly relevant to model choice and limits broad claims that adaptive AI evaluation is new. Our exact outcome substitution uses a sound resource proof rather than a noisy judge. It does not estimate judge bias or selectively missing human labels.

Multi-fidelity best-arm identification already has instance-dependent cost lower bounds and asymptotically optimal allocation in its setting (Poiani et al. 2024). Our exact screening and full rollout are two routes to the same binary label, not an implementation of that general biased-fidelity model. The present controller is heuristic and has no new optimality result.

2.3 Persistent knowledge and operational contracts

Karpathy’s LLM Wiki pattern preserves raw sources while maintaining linked derived knowledge (Karpathy 2026). Wright compares compiled wiki and vector retrieval on 59 course questions, including a syllabus revision, and describes directional evidence requiring further testing (Wright 2026). Neither source establishes operational planning validity. We therefore treat source interpretation as a separate, testable input to the simulator, not as ground truth supplied by a fluent answer.

2.4 Predictive validity and contribution boundary

Conformal prediction addresses predictive coverage under exchangeability, which can deteriorate under shift (Lei et al. 2018; Barber et al. 2023). That is distinct from Monte Carlo uncertainty conditional on a simulator. Neither a valid conditional confidence sequence nor a correct capacity proof validates the simulator against an actual training organization.

The contribution map is deliberately narrow:

Component	Established foundation	Scope of this release
Capacity failure witness	Cumulative-resource reasoning (Vilim 2009)	Aggregate target-group cut, exact integer verifier and shared labels
Time-uniform paired evidence	Mixture confidence sequences and multinomial contrasts (Howard et al. 2021; Lindon and Malek 2022)	Disagreement/sign decomposition and one recorded error budget
Feasible-set selection	Constrained identification and confidence-set projection (Cai and Kandasamy 2026; Li and Zhu 2026)	Paired risk-plus-cost projection with mixture stress and abstention
Computational allocation	Active testing and multi-fidelity identification (Kossen et al. 2021; Poiani et al. 2024)	A tested heuristic; no optimality or lower-bound claim
Operational software	Existing training simulation	Frozen contracts, adversarial checks, explicit costs and reproducible exports

These distinctions specify an implementation contribution, not a claim that the integration is unprecedented. Exhaustive novelty and independent formal verification remain unestablished.

3 The operational contract

3.1 Milestones and target outcomes

A context contains N learners, ordered residual stages, release days, deadline D , horizon $H \geq D$, resources, and availability laws. Each stage requires successful work units. An attempt consumes a stated nonnegative integer amount of each required resource; failure can require repetition. A learner attempts at most one unit per working day. The generic engine uses a five-day week beginning on a Monday; it does not model real holidays or consecutive clinical rotation blocks.

A frozen action a supplies a priority rule, a resource change, and nonnegative relative cost c_a . It obeys stage order and capacity. Future W contains environment, resource availability, attendance, and success draws. Fix an integer required-completion target $m \in \{1, \dots, N\}$. For completion time $T_i(a, W)$, define

$$C_a(W) = \sum_{i=1}^N \mathbf{1}\{T_i(a, W) \leq D\}, \quad Y_a(W) = \mathbf{1}\{C_a(W) < m\}.$$

The integer m is stored in the validated context and shared by simulation, screening, witness verification and exports. In exact mathematics a fraction η would give $m = \lceil \eta N \rceil$. The browser and Python interfaces instead define their machine input conversion explicitly: for accepted finite binary64 $\eta \in (0, 1]$, choose the least integer k with $\text{fl}(k/N) \geq \eta$, where fl is binary64 division. This grid rule preserves representable UI cohort fractions without a tolerance; the next representable value above a grid point raises the required count. Imported derived counts are recomputed. The schema permits $1 \leq N \leq 100$ and rejects boolean or nonfinite fractions. All downstream outcomes use only the integer comparison $C_a < m$. Thus $Y_a = 1$ is a cohort target miss. Practical-test readiness, certification, graduation, and licensure are different events; an adapter must define the actual milestone.

Stateless draws indexed by seed, channel, day, and entity couple actions. Adding a resource preserves existing entities' draws. Common random numbers pair plans without assuming their outcomes independent. Python and JavaScript share the integer hash generator. Theory assumes ideal independent world streams; the implementation uses a seeded pseudorandom approximation.

3.2 Regimes and decision losses

There are K regimes with nominal probabilities $p_h > 0$. Examples use $(.65, .15, .15, .05)$ for ordinary, environment, resource, and coupled disruptions. Conditions can persist within a world; worlds are independent conditional on regime. These probabilities are assumptions, not measured military, medical, or flight-school frequencies.

Let $\mu_{ha} = \mathbb{E}_h Y_a$ and $\Delta_{hab} = \mu_{ha} - \mu_{hb}$. For mixture q and cost preference $\lambda \geq 0$, set

$$L_a(q, \lambda) = \sum_h q_h \mu_{ha} + \lambda c_a, \quad \mathcal{Q}_\rho = \left\{ q \geq 0 : \sum_h q_h = 1, \frac{1}{2} \|q - p\|_1 \leq \rho \right\}.$$

Only mixture masses vary. Conditional laws, curriculum, cutoff, learners, deadline, and actions remain fixed. Toy costs use a declared dimensionless unit. In a model tournament, evaluator-owned prices, acquisition limits and a common budget are frozen before candidate generation; candidate fields cannot redefine them. These costs are not measured API bills or validated market prices. Robust regret is

$$R_a(\rho, \lambda) = \sup_{q \in \mathcal{Q}_\rho} \left[L_a(q, \lambda) - \min_b L_b(q, \lambda) \right] = \max_b \sup_{q \in \mathcal{Q}_\rho} \left[q^\top \Delta_{ab} + \lambda(c_a - c_b) \right].$$

The maximum includes $b = a$, making regret nonnegative. Tolerance .10 permits .10 extra risk-plus-cost loss against the best listed plan under the declared ambiguity. It does not prove baseline improvement or real-world safety.

4 Exact capacity screening

4.1 A policy-independent witness

Fix a future and resource plan. Let C_r be total available resource r through D . Let w_{ir} be learner i 's minimum residual successful-work demand, ignoring retries. Remove learners whose total minimum work exceeds their available attendance days after release; call the remaining set E . This removal is sound because at most one unit can be attempted daily.

For weights $y_r \geq 0$, set $v_i(y) = \sum_r y_r w_{ir}$ and sort these values over E as $v_{(1)}(y) \leq \dots$. Timing conflicts, failures, and prerequisites are relaxed. The resulting condition is necessary, not sufficient, for feasibility.

Proposition 1. If $|E| < m$, or if some $y \geq 0$ satisfies

$$\sum_{j=1}^m v_{(j)}(y) > \sum_r y_r C_r, \tag{1}$$

then $Y_a(W) = 1$ for every policy respecting the same work, attendance, daily attempt rule, and resource plan, including an omniscient scheduler.

Proof. Every completed learner belongs to E . If m finish, select a completed group S of size m . Successful work alone consumes at least $\sum_{i \in S} w_{ir}$ units of resource r . Hence $\sum_{j=1}^m v_{(j)}(y) \leq \sum_{i \in S} v_i(y) \leq \sum_r y_r C_r$, contradicting (1). Retries and other constraints cannot reduce this required work. $\square$

The browser searches nonempty zero-one weight subsets for its three resources. A separate verifier accepts nonnegative integer weights within exact safe-integer arithmetic. This is not a complete separation oracle: no witness does not mean feasible. Changed resources need a new capacity calculation; changed work requirements are outside the frozen contract.

An exported witness records resource weights, minimum target-group demand, available capacity, seed, and regime. It proves impossibility in that sampled future, not every future or the actual organization. It can refute an AI promise within an explicit contract without trusting the model's explanation.

4.2 Substitution preserves the experiment

Proposition 2. In an evaluator whose sampling and stopping choices depend only on the count, observed binary target outcomes, fixed settings and shared auxiliary randomization, substitute label one whenever a sound cut proves failure; otherwise run the original policy. With identical world streams and controller randomization, all binary labels, counts, adaptive sampling choices, and count/history-based stopping decisions are unchanged. Wall-clock stopping and time-dependent sampling are excluded.

Proof. Proposition 1 gives label equality in screened worlds. Other labels use the same simulator. Inductively, equal observed histories give equal sampling and stopping decisions, followed by equal next labels. $\square$

Proved failures remain in the sample; none are rejected or redrawn. Completion dates and completed fractions remain unknown when only a binary miss is proved. Screening reduces work to obtain a label, not variance or the number of worlds. Cut computation may cost more than an inexpensive rollout.

The former release used different tolerances for the rollout fraction and screening count. A retained importer accepted a two-learner profile with $\eta = 0.50000000000075$, producing incompatible labels. Version 0.6 replaces those conventions with the single stored integer target above; that input requires two completions in both paths. Boundary regression tests cover exact cohort grid points and adjacent representable inputs. This is a software correction, not a change to Proposition 1. Historical receipts remain identifiable as historical evidence.

5 Anytime paired inference

5.1 Primitive confidence sequences

For s successes in n Bernoulli observations, the Jeffreys beta-mixture process at hypothesized mean u is

$$M_n(u) = \frac{B(s + \frac{1}{2}, n - s + \frac{1}{2})}{B(\frac{1}{2}, \frac{1}{2})u^s(1 - u)^{n-s}}. \quad (2)$$

Invert $M_n(u) < 1/\delta$ and include boundary endpoints conservatively. At $n = 0$ the set is $[0, 1]$. This is a mixture likelihood-ratio martingale for interior means and a nonnegative supermartingale at boundary means. Ville’s inequality gives coverage over all counts. Running intersections preserve validity. Software uses floating-point log-beta evaluation and bisection; theoretical statements concern exact arithmetic. Independent root and cross-language checks assess numerical accuracy, but do not constitute validated interval arithmetic. An independent 80-digit Decimal root audit over 35 count pairs through $n = 4096$ finds maximum absolute endpoint differences 3.10×10^{-15} in Python and 1.15×10^{-14} in JavaScript, with 15 and six inward-endpoint cases respectively. These checks diagnose floating-point error; they do not quantify undercoverage or provide a uniform bound. The total-variation optimizer now processes positive residual budgets below the former 10^{-14} cutoff; its $\rho = 10^{-15}$ regression agrees with the analytic two-state optimum. Neither confidence-sequence endpoints nor total-variation optimization are claimed to be machine-certified outward enclosures. Formal numerical guarantees would require validated arithmetic and propagated optimization errors; the present qualification must accompany exported statistical bounds.

For pair (a, b) in regime h , $Z = Y_a - Y_b \in \{-1, 0, 1\}$. Write

$$d_{hab} = P_h(|Z| = 1), \quad s_{hab} = P_h(Z = 1 \mid |Z| = 1), \quad \Delta_{hab} = d_{hab}(2s_{hab} - 1).$$

When $d = 0$, the sign parameter is immaterial. Apply (2) to disagreement indicators and to positive signs in the disagreement subsequence. A rectangle $[d^-, d^+] \times [s^-, s^+]$ maps to a contrast interval by taking the extrema of $d(2s - 1)$ at its corners. Dependence between the two confidence sequences is allowed: coverage uses a union bound.

5.2 Adaptive sampling and a joint error budget

For A frozen actions and $J = \binom{A}{2}$ unordered pairs, allocate

$$\delta = \frac{\alpha}{K(2J + A)} \tag{3}$$

to each disagreement, sign, and action-risk process. Denote the intersected contrast intervals by $[\ell_{hab,t}, u_{hab,t}]$ and absolute-risk intervals by $[v_{ha,t}^-, v_{ha,t}^+]$.

Proposition 3. Conditional on independent pilot fitting, suppose each regime supplies an iid stream of fresh world outcomes for a frozen action set. Regime selection may depend on observed history, without access to unseen outcomes. With probability at least $1 - \alpha$, every contrast and absolute-risk interval covers its target at every count, under adaptive allocation and any stopping time.

Proof. Reveal a fixed infinite iid stream for each regime. For each pair, disagreement indicators are Bernoulli and signs at disagreement occurrences form an iid Bernoulli subsequence. Each primitive confidence sequence fails at some local count with probability at most δ by (2) and Ville's inequality. A union bound over $K(2J + A)$ processes gives the joint event. Adaptive revealing and stopping select counts already included in it. Rectangle mapping and intersection preserve coverage. Conditioning on independent pilot data leaves this argument unchanged. $\square$

An entire world is the independence unit. Correlated students and paired plans within it are permitted. Reusing a seed as a fresh world or selecting a new policy after examining evaluation labels violates the experiment. Revised policies require fresh evaluation or a different selection-valid analysis.

5.3 Regret and absolute risk

For $a \neq b$, compute

$$U_{ab,t}(\rho, \lambda) = \sup_{q \in \mathcal{Q}_\rho} q^\top u_{ab,t} + \lambda(c_a - c_b), \quad U_{aa,t} = 0, \quad \bar{R}_{a,t} = \max_b U_{ab,t}. \tag{4}$$

Reverse intervals use $\ell_{ba} = -u_{ab}$ and $u_{ba} = -\ell_{ab}$. In exact arithmetic, finite-mixture optimization transfers probability mass from low-valued to high-valued coordinates within the total-variation budget. Independent linear programming checks it. A maximizing mixture witnesses the confidence envelope, not necessarily the true losses.

Proposition 4. On the event of Proposition 3, simultaneously over counts, actions, $\rho \in [0, 1]$, and finite $\lambda \geq 0$, $R_a(\rho, \lambda) \leq \bar{R}_{a,t}(\rho, \lambda)$. The absolute-risk interval

$$\left[\inf_{q \in \mathcal{Q}_\rho} q^\top v_{a,t}^-, \quad \sup_{q \in \mathcal{Q}_\rho} q^\top v_{a,t}^+ \right] \quad (5)$$

contains risk for every permitted mixture. Selecting an action and stopping when its upper regret is at most ϵ certifies that tolerance within this simulator contract.

Proof. Nonnegative weights give $q^\top \Delta_{ab} \leq q^\top u_{ab,t}$. Take the supremum and finite maximum over b . The same reasoning gives (5). Since primitive coordinates are bounded before queries over ρ or λ , all such queries hold on the same event. Selecting a query, action, or count requires no additional union bound. $\square$

Equation (4) describes an unconstrained comparison, retained to reproduce the earlier experiments. It must not by itself authorize a recommendation under a hard risk ceiling. Improvement over baseline requires a positive lower bound on baseline-minus-candidate loss. Relative qualification alone is insufficient.

5.4 What observed ties imply

Proposition 5. Along $n \geq 1$ observed ties for a pair and regime, the disagreement upper endpoint in (2) is

$$u_d = 1 - (\delta C_n)^{1/n}, \quad C_n = \frac{\binom{2n}{n}}{4^n}, \quad u_d \leq \frac{\log(2/\delta) + \frac{1}{2} \log n}{n}. \quad (6)$$

Before further valid intersection the contrast interval is $[-u_d, u_d]$.

Proof. Zero disagreements give $M_n(d) = C_n/(1-d)^n$. Solve $M_n(d) = 1/\delta$. The bound $C_n \geq 1/(2\sqrt{n})$ and $1 - e^{-x} \leq x$ give (6). With no observed sign, its interval is $[0, 1]$, so the product interval is $[-u_d, u_d]$. $\square$

This path property follows from established confidence sequences. It does not assume future ties, prove a zero true contrast, or improve a generic Monte Carlo rate. Frequent disagreement and small decision gaps may require many worlds. Cheap proved labels and faster confidence contraction are different computational effects.

5.5 Risk-constrained selection and abstention

A cheap plan can minimize risk-plus-cost loss while violating a mandatory risk limit. Ranking all plans and displaying a separate warning does not solve the constrained problem. Let

$$g_a(\rho) = \sup_{q \in \mathcal{Q}_\rho} q^\top \mu_a, \quad F = \{a : g_a(\rho) \leq \kappa\}, \quad \kappa \in [0, 1].$$

Here acceptability concerns worst-case target risk, while cost enters the objective only. Bound the worst-case risk itself, rather than the wider interval for every mixture in (5):

$$g_{a,t}^- = \sup_{q \in \mathcal{Q}_\rho} q^\top v_{a,t}^-, \quad g_{a,t}^+ = \sup_{q \in \mathcal{Q}_\rho} q^\top v_{a,t}^+. \quad (7)$$

The supremum in both endpoints is intentional. Define certified and possible acceptable sets

$$C_t = \{a : g_{a,t}^+ \leq \kappa\}, \quad P_t = \{a : g_{a,t}^- \leq \kappa\}.$$

For $a \in C_t$, use the constrained upper regret

$$B_{a,t} = \max_{b \in P_t} U_{ab,t}, \quad a_t \in \operatorname{argmin}_{a \in C_t} B_{a,t}. \quad (8)$$

Proposition 6 (conditional constrained decision). On the event of Proposition 3, $C_t \subseteq F \subseteq P_t$ simultaneously over counts, ρ , λ and κ . Whenever C_t is nonempty, a selected a_t has $g_{a_t} \leq \kappa$ and

$$\sup_{q \in \mathcal{Q}_\rho} \left[L_{a_t}(q, \lambda) - \min_{b \in F} L_b(q, \lambda) \right] \leq B_{a_t,t}. \quad (9)$$

Thus recommending only when $B_{a_t,t} \leq \epsilon$ controls both requirements on that same event. If P_t is empty, no listed action satisfies the modeled risk requirement. Otherwise an empty C_t means unresolved risk; a bound above tolerance means unresolved comparison. No action is recommended in these three cases.

Proof. Coordinate coverage and nonnegative mixture weights imply $g_{a,t}^- \leq g_a \leq g_{a,t}^+$, giving the set inclusions. Every member of C_t is truly acceptable. For such an action, take the supremum and finite maximum of the covered paired losses over $F \subseteq P_t$ to obtain (9). The same joint coordinate event holds before preferences or stopping times are selected, so these post-query projections spend no additional error budget. If P_t is empty then F is empty. $\square$

This is a composition result, not a claim to invent constrained selection. Cai and Kandasamy study fixed-confidence identification with feasibility tests (Cai and Kandasamy 2026). Li and Zhu explicitly project certified and possible feasibility sets for conformal model selection (Li and Zhu 2026). Our target differs: fixed-regime paired simulation, an unknown paired loss objective, total-variation mixture stress, deterministic costs, and exact resource-proof labels. No new lower bound, optimal allocation, or universal efficiency theorem follows from Proposition 6. At a binding risk boundary, abstention may persist indefinitely without a separation assumption.

5.6 Minimum-cost selection over a frozen menu

A user asking for the cheapest adequate resource change has a different objective from minimizing risk-plus-cost regret. Freeze a finite menu $\mathcal{M}$ of admissible actions before revealing its evaluation outcomes. Each member has a deterministic externally specified cost c_a , in one explicit unit, and worst-case target risk g_a as above. Use the joint absolute-risk bounds to construct C_t and P_t for this menu.

Proposition 7 (conditional minimum menu cost). On the event of Proposition 3, suppose C_t is nonempty and choose $a_t \in \operatorname{argmin}_{a \in C_t} c_a$. If

$$\{b \in P_t : c_b < c_{a_t}\} = \emptyset, \quad (10)$$

then a_t is acceptable and $c_{a_t} = \min_{b \in F} c_b$. If cheaper possibly acceptable actions remain, the minimum-cost conclusion is unresolved. An empty P_t excludes only the listed menu.

Proof. Proposition 6 gives $C_t \subseteq F \subseteq P_t$. Thus $a_t \in F$. Condition (10) implies that no member of F has smaller cost than a_t . The reverse inequality follows because a_t itself is a member of F . Tied costs are allowed. All conclusions use the same joint event and remain valid when the count or risk query is selected from the observed history. $\square$

This elementary projection does not prove the globally cheapest intervention, infer dollar costs, or search an unrestricted policy space. External accounting is an assumption: candidates cannot earn a meaningful economic comparison by declaring additional resources free. The menu must include every alternative covered by the claim and be frozen before evaluation. Adding a plan after inspecting its outcomes requires a new evidence family. The minimum-cost status and the paired risk-plus-cost-regret status are reported separately because their objectives differ. Admissibility, including approval scope and a valid frozen output, is fixed before the evaluation stream; malformed workflows remain in benchmark accounting without becoming eligible interventions.

6 Implementation and experimental protocol

6.1 Actual local model fitting

The live application fits weighted ridge regression to learner features and restricted completion labels from 32 independent baseline-policy pilot worlds, eight per regime. Features include remaining work, stage count, a resource-demand summary, target day, and an intercept. The response is $\min(T, H)/H$. Priority is predicted target lateness, rounded identically across Python and JavaScript. Coefficients and training cost are exported. Restricted labels are not known eventual graduation dates.

Four actions are FIFO, frozen learned priorities, one added mentor, and one added gate resource. Costs are (0, .02, .12, .18). The browser additionally accepts one to four frozen priority/resource proposals in a bounded JSON schema. Provenance is self-declared; import does not authenticate a model. No arbitrary imported code executes. These are static proposals, not interactive agents or the complete space of feasible schedules.

Evaluation starts with eight observations per regime. The current heuristic targets unresolved risk classifications and the active paired competitor. It favors a regime with a large interval-width contribution, using adverse-mixture weights, nominal mass, and an approximate batch contraction factor. Every fifth batch explores the least-sampled regime; batch size is 16. Future allocation depends on current evidence, making this actual online adaptive Monte Carlo. No optimal acquisition result is claimed.

The local browser shows actions, regret, absolute risk, next sampling regime, disagreement examples, capacity witnesses, fitted coefficients, and rollout counts. Changing mixture stress or cost preference reuses the joint evidence. Changing context or action set invalidates it. The current browser stops at risk-constrained qualification or exclusion of the listed menu. The unconstrained studies below stop at relative qualification and remain separately labeled; they must not be reinterpreted as

risk-constrained recommendations. A separate legacy tool retains detailed date forecasts and earlier recorded studies. Calculations need no provider API, key, telemetry, or network service.

6.2 A standalone tournament and knowledge contract

The CLARK Frontier workbench organizes decision evidence, while its CLARK Criterion engine evaluates frozen decisions. BOLT Wiki holds source-linked knowledge; Ask Bolt is the intended question and tool interface; Optimize supplies the future operational review workflow. The released prototype runs independently. It does not connect to, modify, or deploy the production Ask Bolt service.

A candidate registry binds a workflow identifier, model version, prompt and tool versions, proposal context, and declared approval scope. The default allowlist permits local synthetic demonstrations; it does not attest Air Force or organizational approval. Built-in candidates include an actually fitted priority model and deterministic policies. Imported outputs remain static workflow proposals with self-declared authorship. Failed or malformed outputs are retained as declared failure outcomes rather than removed from the denominator, and are structurally excluded from recommendation even at a permissive risk limit. The current event contract remains one frozen plan per candidate; it is not a general interactive-agent evaluation.

Before proposals are generated, the evaluator freezes an acquisition contract: positive unit prices, per-resource addition limits, a total-addition limit, a common budget, units and source provenance. Version 2 proposal objects contain priorities and resource changes, not candidate-declared costs. The evaluator derives $c_a = \sum_r \pi_r \Delta r_a$ and rejects changes outside that contract; imported registries cannot replace the caller’s contract. The synthetic default allows at most one added resource, prices an added mentor at .12 and each other resource at .18, and sets a common .18 budget. FIFO, ridge and scheduling-search policies with no additions consequently have resource cost zero; training and generation compute are reported separately. Equal-resource scheduling comparisons and externally priced resource-acquisition comparisons have different interpretations. The cost contract prevents the former 199-resource, zero-price adversarial proposal from being an admissible intervention, without implying market-price validity.

Every participating candidate sees the same sampled future. The recommendation gate is projected onto the predeclared admissible menu; penalty-only invalid candidates retain their confidence-family allocation. The constrained gate returns an acceptable action only with sufficient evidence; an estimated leader or Pareto point is not itself a certificate. Training, proposal work when recorded, policy evaluations, proof work, rollouts and measured local timing are reported separately. Relative resource costs are not silently converted into API bills. A local executable adapter is available for separately authorized, allowlisted model runners, but its existence is not evidence of a live model result.

A knowledge record distinguishes preserved source material, generated interpretation, proposed rule and approved operational rule. Compilation carries source passages, units, effective dates, review state and dependencies into a typed scenario. The supplied demonstration links three required fields per profile; remaining profile inputs are explicitly declared assumptions. Missing or conflicting required rules block promotion. A changed source identifies affected scenarios and tests for fresh review; a change of within-regime laws also requires fresh simulation evidence. The included ingest, query, lint and drafting operations are deterministic local routines, not a concealed language-model service. This separation prevents an AI-written explanation from becoming its own corroborating source.

The intervention workbench enumerates a finite resource menu before sampling and applies Proposition 7. Within-regime stress changes attendance, resource availability or work assumptions in separately frozen contexts. Its confidence allocation is divided across the predeclared stress family when reporting joint coverage. A timing-based controller can choose exact screening or full rollout, but always obtains the original binary label. Timing selects the route to a label, not the next regime or a wall-clock stopping boundary; only that restricted implementation falls under Proposition 2. A tiny exact deterministic scheduler is supplied as a restricted computational comparator, not a deployable stochastic policy or a substitute for a strong operational baseline. It does not substitute an uncorrected approximate probability or generate dates for a proof-only label.

Historical-replay and outcome-ledger adapters preserve cutoff dates, cohort units and censoring; supplied examples remain synthetic. Replay date error and interval coverage are computed only for observed completions, while target-risk scoring uses only cohorts with a resolved target observation window. These denominators are exported; they are not a survival-analysis correction or population coverage guarantee.

An optional operational adapter adds an ISO start date, closure dates, learner/resource unavailability, pooled resource qualifications and required stage qualifications while retaining ordered stages. Its weather process is indexed by open training days: initialize from the stated good-weather probability on the first open day, then update on subsequent open days; closed dates do not advance the process. This explicitly fixes a former closed-start bug in which an uninitialized good-weather state could persist when good-weather probability was zero. It is a bounded schema, not a full institutional prerequisite graph or individually matched instructor scheduler. Because an arbitrary calendar can differ from the canonical Monday-start tape, this mode uses full rollouts and disables the original capacity screening. Reusing that tape could otherwise give an unsound cut. The original ridge pilot remains fitted under the canonical work model; its transfer to an operational calendar is not validated by this release.

A next-evidence heuristic distinguishes more simulation from a missing measurement or rule clarification. No optimal value-of-information policy is claimed.

6.3 Known-truth validation and stronger baselines

The protocol fixes $\alpha = .05$, $\rho = .05$, $\lambda = .25$, $\epsilon = .10$, and a 2,048-world cap. Three original categorical joint-outcome laws have known exact risks: all ties, disagreement concentrated in stressed regimes, and homogeneous disagreement. After the original comparison, uniform and static square-root allocation proportional to $\sqrt{p_h + \rho}$ were added. Those design choices were not externally preregistered.

The referee identified collisions in the original XOR-derived replication seeds. Version 0.6 replaces the affected runs rather than treating them as independent validation. A bounded, injective key encodes study/law, replication, regime and sample; an invertible 32-bit mixer maps it to the engine seed. Comparison-method identifiers are deliberately omitted to preserve common random numbers within a replication. Separate namespaces distinguish studies and laws. The finite key-space audit establishes absence of accidental seed reuse within its declared ranges, not mathematical independence or cryptographic randomness. The conditional theorems retain their ideal iid assumption.

The corrected confirmation study uses 100 replications per law and method, five laws and six methods, totaling 3,000 runs. In addition to the original three laws, an anticorrelated law places

equal mass on $(0, 1, 0, 1)$ and $(1, 0, 1, 0)$, so opposing candidates always disagree despite identical marginal risks; a rare-adverse law places .49 mass on all-zero outcomes, .49 on all-one outcomes and .01 each on two opposing paired outcomes. Their protocols were fixed before corrected runs, with no subsequent adaptive-policy tuning. These unfavorable or mixed coupling laws limit conclusions drawn only from frequent ties.

Four methods use disagreement/sign inference with adaptive, target, uniform or square-root allocation. The target-proportional marginal ablation discards pair dependence and uses its $\alpha/(KA)$ absolute-risk budget. A direct paired predictable plug-in empirical-Bernstein comparator applies Theorem 2 and Equation 15 of Waudby-Smith and Ramdas (Waudby-Smith and Ramdas 2023) to $X = (Y_a - Y_b + 1)/2$, with fixed tuning cap .5 and square-root allocation. It allocates $\alpha/[K(J + A)]$ to each contrast and absolute-risk process; the disagreement/sign method splits $\alpha/[K(2J + A)]$. Both control the same whole-family error budget in their ideal arithmetic models. Holding square-root allocation fixed isolates this inference comparison. It is one prior-work comparator, not an exhaustive evaluation of multinomial or betting confidence sequences.

Methods share world streams within replications; they are not independent cross-method replications. Coverage is checked at reported batches, while theory concerns all counts. Exact categorical risks supply true regret. Independent linear programs check those regrets, and Python replays verify selected JavaScript paths. Binomial frequency intervals are descriptive Clopper–Pearson summaries under the independent-replication idealization; the deterministic pseudorandom implementation does not itself prove that model.

6.4 Separate risk-constrained follow-up

A protocol saved before the follow-up runs fixes four Bernoulli laws, 100 replications per law and method, and the same 2,048-world cap. The conflict law deliberately uses cost weight 6, while the other laws use .25. Common uniforms couple all four actions; for given Bernoulli marginals this minimizes pairwise disagreement and favors tie-exploiting inference. The separate anticorrelated confirmation law tests a less favorable dependence structure. One law deliberately makes the unconstrained cost-preferred plan risk-unacceptable; another makes every plan unacceptable. A third places risks at or near the .20 ceiling; a fourth has comfortably acceptable alternatives. Both the new risk-aware heuristic and the square-root allocator use the same gate, error budget and within-replication streams. This is a subsequent study, not a retrospective change to the original experiments. It contains 800 runs and no learned pilot or screening costs.

An independent SciPy linear program recomputes true robust risks and constrained regrets. Every recorded decision is checked against that truth; the first and last replication per law and method are independently replayed in Python, including allocation, counts, confidence coverage and all 718 reported stages across 16 runs. Source hashes bind the protocol and original engines to the results.

6.5 Domain and proof studies

The synthetic profiles contain six sport-pilot learners, ten clinical learners, and twenty military-training learners. Four fresh-seed replications per profile under adaptive and target-proportional paired sampling give 24 runs. Profiles were constructed using separate design and pilot seeds. Four seeds do not represent variation across actual institutions or curricula.

For proof validation, each profile uses default deadline 100 and tight deadline 55, with 128 worlds per regime and four plans: $3 \times 2 \times 4 \times 128 \times 4 = 12,288$ comparisons. Every screened target label

is checked against the original full-world scheduling implementation. Separately selected positive witnesses are checked by an independent mixed-integer relaxation maximizing completed learners under aggregate capacity and individual-workday constraints. Timing paths share prepared tapes and charge tape generation to both. One-host runtime measurements are not portable performance guarantees.

7 Results

7.1 Paired inference and static allocation

Table 1 reports the corrected 3,000-run study. Each cell contains 100 replications. Counts include budget-capped unresolved runs; fewer worlds are not useful if the requested decision remains unresolved.

Law	Adaptive	Target	Uniform	Square root	Marginal	Paired EB
All ties	320.00	384.00	368.00	320.00	2,048.00	992.00
Localized	831.84	1,006.24	924.48	823.52	2,048.00	1,250.56
Homogeneous	537.60	685.76	639.52	546.08	1,375.04	656.00
Anticorrelated	2,048.00	2,048.00	2,048.00	2,048.00	2,048.00	2,048.00
Rare adverse	553.60	697.92	661.76	564.32	2,048.00	1,021.28

Table 1. Mean worlds to stopping or budget. Paired EB uses static square-root allocation. Every method is unresolved in all 100 anticorrelated replications. On the other four laws all paired methods qualify; marginal inference qualifies 99/100 homogeneous cases and none of the other cases. Analytic policy-outcome cost is four times worlds; pilot fitting and screening are absent here.

Adaptive and square-root allocation tie on all ties, and each modestly improves the other on different laws. There is no general adaptive-acquisition advantage. At fixed square-root allocation, disagreement/sign inference uses fewer worlds than this direct predictable empirical-Bernstein comparator in four laws, while neither resolves the anticorrelated law within budget. This is a finite comparison of specified constructions and tuning, not superiority over all paired inference, multinomial confidence sets or optimal betting methods.

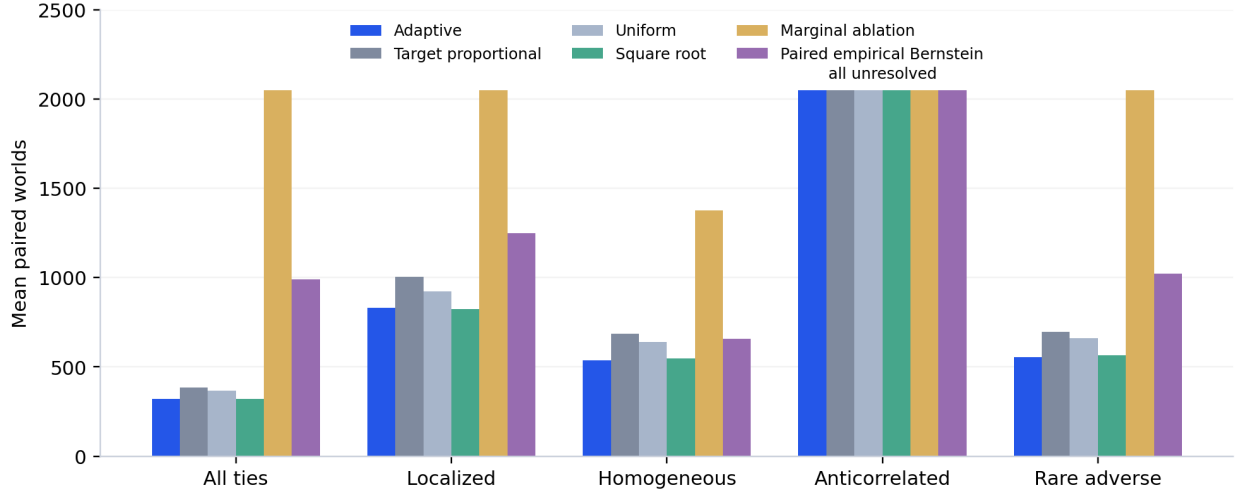


Figure 1: Corrected mean paired worlds across five known-truth laws. Every method reaches the cap without qualification in the anticorrelated law. The direct paired empirical-Bernstein comparator holds square-root allocation fixed.

One marginal/localized run misses joint coverage at a reported batch; that cell has 99/100 covered runs. Every other cell has 100/100. No false relative certificate is observed. Nominal two-sided 95% Clopper–Pearson intervals are $[\text{.9455}, \text{.9997}]$ for 99/100 coverage, $[\text{.9638}, 1]$ for 100/100 coverage, and $[0, \text{.0362}]$ for zero errors in 100 runs. These intervals assume independent replications and are descriptive for the keyed pseudorandom experiment. They are not pooled across laws or correlated methods, nor simultaneous over all cells. Observed coverage does not replace the conditional $1 - \alpha$ theorem.

7.2 Risk limits change the decision

Known-truth law	Adaptive worlds	Square-root worlds	Outcome in each 100-run cell
Cost versus risk conflict	339.52	345.92	Qualified acceptable plan
Every listed plan fails	67.68	73.60	Listed menu excluded
At the risk boundary	2,048.00	2,048.00	Risk unresolved at cap
Comfortable acceptable choices	476.80	480.32	Qualified acceptable plan

Table 2. Separate 800-run constrained follow-up. Mean worlds include budget-capped unresolved runs. Both methods enforce the same hard risk constraint. Every final unconstrained candidate in the first two laws violates the risk ceiling; it is diagnostic, not a recommendation.

There are no observed false recommendations or incorrect menu exclusions. Each 100-run cell has the nominal two-sided 95% Clopper–Pearson error-frequency interval $[0, \text{.0362}]$, under the independent-replication model; heterogeneous cells are not pooled. Independent replay supports implementation consistency. Modest allocation differences do not establish a new sample-complexity advantage. Persistent boundary abstention is the correct uncertainty output, not successful selection.

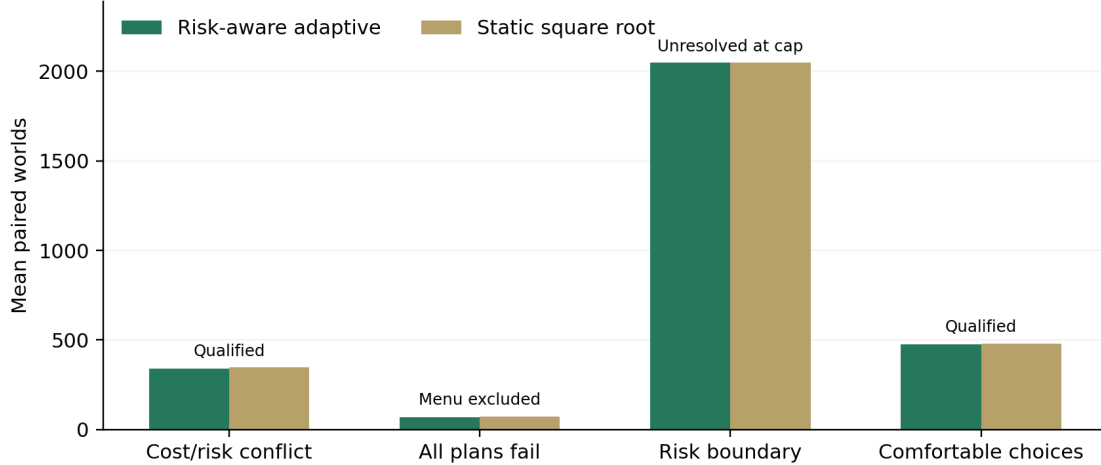


Figure 2: The risk-constrained gate returns different actionable states. Bars show mean worlds, not real-world performance. Boundary cases consume the entire budget without a recommendation.

7.3 Corrected unconstrained choices in synthetic domains

Synthetic profile	Adaptive worlds	Target worlds	Choice in all four seeds
Sport pilot	424	464	Learned priorities
Clinical placements	404	440	Add one preceptor
Military pilot training	280	348	Add one instructor

Table 3. Mean paired worlds. All runs meet .10 relative-regret tolerance. Each additionally charges 32 baseline-policy pilot evaluations; each world charges four candidate outcomes.

These unconstrained choices compare four plans under assumed costs. They do not establish the best action at a real school, a general preference for more staff, or superiority over language models. No independent exact risk oracle is available for the complete domain simulator here. The theorem supplies conditional inference; the small domain study demonstrates the workflow. Absolute risk must still be checked before acceptance.

7.4 Proof screening preserves labels with selective cost benefits

All 12,288 target outcomes match the original simulator; 6,342 full rollouts are replaced by proved failures. Each of 200 separately selected witnesses is confirmed by the independent mixed-integer relaxation. Proposition 2 is deterministic; the tests exercise the implementation rather than prove a population error rate.

Profile	Default avoided	Tight avoided	Default time ratio	Tight time ratio
Sport pilot	46.68%	99.22%	1.30	2.74
Clinical	0.00%	87.55%	0.90	2.73
Military	8.11%	68.12%	1.01	1.79

Table 4. Fraction of 2,048 policy outcomes per context obtained without a full rollout. Ratio is unscreened time divided by screened time, including tape generation; below one means slower screening. Fractions are equally stratified test frequencies, not nominal-mixture failure probabilities.

Tight targets provide many inexpensive witnesses. Default cases may be cheaper to simulate: the clinical default is approximately 11% slower with screening, while the military default is essentially tied. The corrected rerun improves sport-pilot default timing relative to the historical host run; this variation illustrates why one-run timing ratios are descriptive. Avoided rollouts must not be advertised as equal time savings. Screening may be more attractive for expensive downstream policies or frequent infeasibility; no expensive deployed AI policy is tested here.

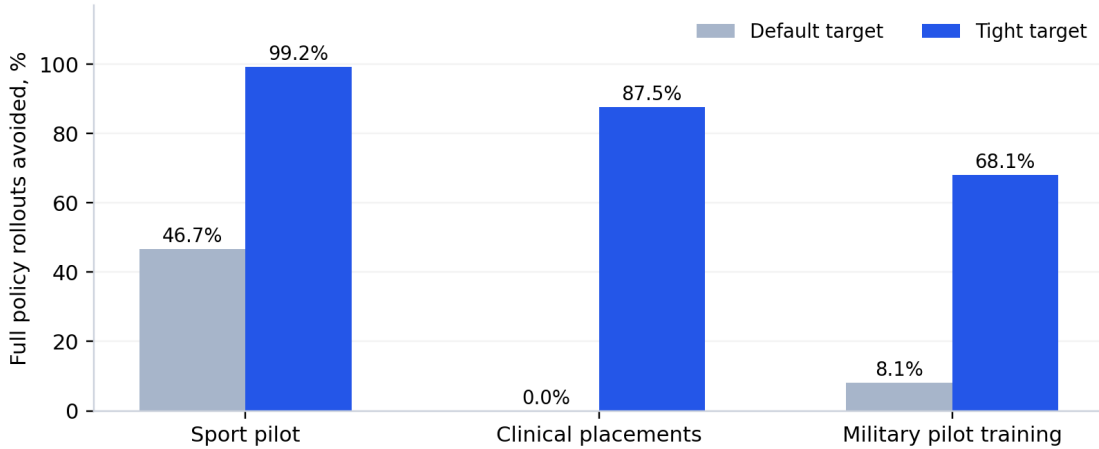


Figure 3: Full rollouts avoided in equally stratified test futures. Every binary target outcome is retained. Default and tight-target cases are different experiments, not operational risk estimates.

7.5 Preserved negative controls

The release retains the earlier manuscript and raw studies. Its pilot-frozen learned-control-variate allocator resolves 27/36 policy contrasts at an 8,192-world budget, versus 34/36 for pilot-free stratified evaluation. Training overhead and weak allocation gains can erase a benefit. The new anytime method is a separate mode; the old finite-stage empirical-Bernstein analysis is not unrestricted anytime inference.

The earlier 170-cohort study fits three regressors and compares three procedural forecasts. Its validation-selected 48-world method is an **oracle-assisted simulation reference**: it receives the full cohort and true generating simulator, whereas regressors receive restricted learner-level features that omit some competing learners’ information. Its matched-test restricted-completion MAE is 3.942 days; whole-cohort coverage is 37/40 and falls to 8/20 under shift, with matched median interval width approximately 35.79 days. The response is $\min(T, H)$, not eventual graduation. No

train/calibration/test leakage was identified in the inspected pipeline, but unequal information prevents a fair predictive-superiority claim. Equal-information features and simulator parameters estimated using training data only would be needed for that comparison. These historical results expose failure under shift and remain separate from operational validation.

7.6 Subsequent workbench validation

The intervention study is a subsequent engineering check with nine runs: three synthetic profiles under always-screen, never-screen and adaptive-screen modes. Each uses the same frozen four-action intervention menu and a 128-world cap. All three proof modes produce the same labels, evidence counts and stopping traces within each profile. Separately, 192 proof-versus-rollout comparisons at target days 55 and 100 include 84 proved failures and zero label mismatches. Python and browser counts agree for canonical search and a separate operational-calendar check; the latter uses 32 worlds and zero capacity proofs.

Adaptive screening reduces proof attempts relative to always screening. Measured elapsed times vary by profile and host load; neither always screening nor adaptive screening has a uniform speed advantage over full rollouts. Timing is a descriptive implementation result, not a portable efficiency guarantee. At a deliberately permissive failure-risk ceiling $\kappa = .80$, the clinical example qualifies the zero-addition minimum-cost menu action at 32 worlds with risk upper bound approximately .7498. It does not qualify at a .20 ceiling. Sport-pilot and military examples remain risk-unresolved at the 128-world cap. Unresolved cases are not successful recommendations.

Four predeclared within-regime stress contexts use fresh evidence and an error allocation of $\alpha/4$. Their results are conditional on each named changed law; they do not estimate an unknown deployment discrepancy. The replay fixture has three held-out synthetic cohorts: one observed completion and two right-censored cases, with two resolved target outcomes and one unresolved target. It validates accounting and denominator disclosure rather than real-world predictive accuracy.

The frozen-workflow tournament has 24 synthetic runs: three profiles, two risk ceilings, two allocation schemes and two seeds, with a 128-world cap. Across 2,560 paired worlds and 12,800 full rollouts it returns eight menu exclusions, six risk-unresolved states, ten comparison-unresolved states and zero qualified recommendations. These outcomes under the revised external cost contract establish no allocation-efficiency advantage. Per profile, fitting and four-rule scheduling search consume 64 distinct pilot futures and 160 policy rollouts before evaluation. A separate 32-world negative fixture retains 64 missing or malformed output penalties. Three predetermined Python replays match browser counts and decisions. These are functional checks, not known-truth coverage or real-model ranking results. The knowledge-contract receipt checks 12 sources and 12 typed rules, 18 review or freshness transitions, nine rejected malformed inputs and nine complete Python/browser output comparisons. Three promoted synthetic scenarios execute 96 paired worlds in total; none yields a qualified recommendation at that small budget. Its 24 public question fixtures generate 72 deterministic representation-proxy records. These inspectable fixtures and keyword routines test schema behavior and provenance; they are not secret benchmark tasks or an LLM-versus-RAG quality comparison. One held-out outcome is excluded from development summaries and one development miss becomes an adjudication-required regression case.

These engineering checks are separate from the corrected analytic, domain and 800-run risk-gate experiments. No supplied test authenticates a live language model, evaluates an operational Air Force cohort, or compares actual LLM answer quality under vector retrieval and wiki retrieval.

8 Applications and deployment

8.1 Sport pilot planning under MOSAIC

A flight school can use this contract to compare instructor availability, usable aircraft, weather windows, learner attendance, repeated training, and examiner access. The output can identify a supported intervention, target-risk interval, or future-specific impossibility witness.

MOSAIC expands the planning setting but does not equate residual work units with FAA eligibility. The FAA identifies separate effective dates for pilot privileges and aircraft certification provisions (Hartman 2026). Current airplane sport-pilot experience requirements include total, dual, and solo flight time plus specific cross-country and practical-test preparation requirements (Electronic Code of Federal Regulations 2026a). Aircraft eligibility and endorsements have additional rules (Electronic Code of Federal Regulations 2026b). A deployable adapter needs actual logged hours, event types, endorsements, aircraft and instructor qualifications, and current rule versions. The example has residual sessions and one practical-test attempt with possible repetition. It is not a complete regulatory engine or a minimum-hours certification promise.

8.2 Clinical and other milestones

The clinical example models placement workdays, preceptor capacity, release days, assessment repetition, and assessor access. It does not implement consecutive rotation blocks, specialty prerequisites, accreditation, licensure, competency judgments, or a medical school’s graduation rules. Those require an institution-specific adapter and independent validation. This is scheduling research, not clinical decision support.

The same event contract could cover maintenance signoffs, industrial apprenticeships, professional recertification, and technical onboarding. Correct work semantics and measured disturbances are prerequisites; relabeling the interface is insufficient. Commercial value should be tested prospectively through reduced delay at matched resource cost or fewer unnecessary resource additions, not inferred from a synthetic leaderboard.

8.3 Evaluating actual AI proposals

To choose among AI models, freeze each complete workflow at the same cutoff and record version, prompt, tools, failures, cost, and resource assumptions. Execute its proposed plans on hidden futures. Compare with FIFO, a strong operations-research policy, a local learner, and strong static evaluators; count proposal generation and verification work. Use executable feasibility and outcomes rather than self-reported confidence as ground truth.

The tournament and import interfaces accept frozen proposals from a model or human. They do not authenticate imported model authorship, and the supplied experiments do not call a live provider. Interactive agents need a causal action interface, tool budgets, nonanticipation checks, repeated-run accounting, and hidden contexts. Selecting new proposals from the same evaluation stream needs a different analysis. General or superintelligent capability claims additionally require an explicit capability target, strong human and algorithmic baselines, and held-out tasks beyond this family.

8.4 A hands-on decision workflow

Start with the browser’s **Snapshot** download. It records ordered residual work, learner release days, named resource counts and availability, four regime probabilities and factors, deadline, and

cohort target. Edit a local copy, retain the built-in action definitions, and **Import snapshot**. Use **Import JSON** separately for frozen AI proposals. The adapter rejects unsupported templates rather than silently inventing a syllabus. Its practical scope remains residual sequential work, not full medical or aviation eligibility.

For an explicit example, retain the sport profile’s day-100 target, 80% cohort completion and 20% maximum failure risk. The fixed browser stream excludes all four listed plans after 48 paired worlds; this is conditional statistical exclusion of that menu, not proof that every conceivable intervention fails. A capacity witness is a different object: a deterministic proof for one sampled future and one resource plan. To inspect a looser requirement, move the risk limit to 80% and resume. This reuses the evidence because only the query changed. Increasing a resource or changing work invalidates the old evidence and starts a fresh evaluation. A larger compute budget cannot repair a wrong simulator.

Export **Decision brief** for the operational discussion and **Evidence JSON** for an audit. The authoritative **decision** object has **status**, nullable **recommended_action**, certified and possible acceptable sets, worst-case risk bounds, constrained upper regret and the active sampling question. The unconstrained comparison is explicitly diagnostic. For example, **status: risk_unresolved** with **recommended_action: null** means collect more evidence or revise assumptions; an integration must not substitute **candidate_action** as advice. Preserve the context, frozen model, training cost, seed, counts and preferences with every record.

The next operational milestone is a blinded replay of completed cohorts using only information available at the historical cutoff. Predeclare acceptable date error, target-risk calibration, abstention burden, compute cost, and intervention value against the existing workflow. Measure the time saved by a decision brief and the cost of a wrong intervention. Only after those checks should a recommendation affect real scheduling.

9 Limitations and research priorities

The strongest result is a conditional composition: sound capacity cuts preserve exact labels, established confidence sequences survive adaptive revealing, and simultaneous intervals support post-query regret and absolute-risk bounds. These are auditable guarantees, not a verified foundational breakthrough. Equation (6) is a special path property. Screening does not solve scheduling, and adaptive allocation has not outperformed the strongest simple static comparator.

Simulator validity is the central evidence gap. Historical replay must preserve the cutoff, split by whole cohorts, handle censoring, and test fleet, curriculum, staffing, and population shifts. Counterfactual resource advice needs prospective or otherwise causally justified validation. A total-variation ball over four weights cannot absorb arbitrary errors in within-regime laws, missing rules, or incorrect work demands. More simulation only reduces conditional Monte Carlo uncertainty.

A prospective extension is certificate-aware active evaluation of frozen AI workflows: model, prompt, retrieval, tools, and budget. The outer evaluation unit would be a new case and bounded agent execution; inner simulator worlds would remain separate evidence. Exact task-specific verifiers could settle multiple workflow comparisons, while biased LLM scores would require independently sampled audits. Active model selection (Durmazkeser et al. 2025), propensity-corrected anytime inference with LLM judges (Ao et al. 2026), graph-feedback pure exploration (Russo, Song, and Pacchiano 2025) and sequential constrained selection (Li and Zhu 2026) already establish central ingredients. Novelty would therefore require a new result on shared certificate acquisition cost or a

reproducible real-model advantage beyond these baselines. No such theorem, live-model experiment or cost reduction is established here.

A substantive mathematical program would derive instance-dependent complexity in disagreement, decision gaps, proof cost, and shared resource structure, with matching lower bounds or an independently verified new separation algorithm. Any practical advantage must survive proposal, pilot, proof-search, and verification costs on hidden, independently validated tasks. The local tournament and adapters are starting points; authenticated workflow execution, outer-case inference and selective audits remain unimplemented and unevaluated in this release.

10 Reproducibility and disclosure

The release includes Python and JavaScript engines, frozen profiles, protocols, raw outcomes and traces, witnesses and source hashes. The final Python suite passes 144 tests; separate Node checks exercise the live interface bindings and stored model fixtures. The v0.6 regression checks cover the former failures directly: 5,050 exact cohort-fraction grid targets and 5,050 next-float targets, 480 deterministic capacity/target combinations across both languages and both screening routes, and first-open-day weather endpoints. A separately written re-audit checks 891 boundary profiles, 168 calendar worlds and six cost/accounting attacks. It reports no tested label mismatch or accepted budget bypass. These are finite adversarial checks, not formal software verification.

Independent numerical roots reveal small inward endpoint errors, as disclosed in Section 5.1. Cross-language agreement is a consistency check, not an independent correctness oracle. The source-bound audit checks 80 linear-programming truth quantities, 50 Python stream replays (maximum bound discrepancy 8.33×10^{-15}), and 200 mixed-integer capacity witnesses. The risk-gate audit replays 16 complete paths and 718 reported stages. Fifteen independent empirical-Bernstein formula replays differ by at most 2.78×10^{-17} . A stream audit checks 2,390,928 distinct evaluation keys and seeds with zero accidental collisions; 384 domain pilot seeds have no overlap with that evaluated union. These checks test the implementation against separately calculated quantities; seed uniqueness does not establish random independence. All checks are conditional on synthetic contracts; none validates an actual organization or supports a general intelligence claim.

The canonical reproduction commands are:

```
node experiments/run_anytime.cjs analytic
node experiments/run_anytime.cjs domains
node experiments/confirm_anytime.cjs
PYTHONPATH=src python3 experiments/audit_eb_v6.py
PYTHONPATH=src python3 experiments/audit_streams_v6.py
node experiments/validate_screening.cjs
PYTHONPATH=src python3 experiments/audit_v3.py
node experiments/run_risk_gate.cjs
PYTHONPATH=src python3 experiments/audit_risk_gate.py
PYTHONPATH=src python3 -m unittest discover -s tests -v
PYTHONPATH=src python3 experiments/run_tournament_v5.py
PYTHONPATH=src python3 experiments/validate_knowledge_v5.py
PYTHONPATH=src python3 experiments/run_lab_v5.py
PYTHONPATH=src python3 experiments/audit_referee_fixes_v6.py
PYTHONPATH=src python3 experiments/audit_numerics_v6.py
```

Protocols specify seed keys, budgets, subsequent analyses and source hashes. Version 0.6.1 introduces the workbench and engine naming structure without changing the reported experiments. Version 0.6 superseded the affected analytic and risk-gate empirical results after repairing cross-replication seed reuse; archived receipts are provenance, not the current evidence. Unaffected historical forecast and finite-stage studies remain explicitly labeled. Implementation coverage, the referee response and open validation requirements accompany the release.

All cohorts are synthetic; no operational Air Force validation, external endorsement or MIT affiliation is claimed. AI assisted drafting, implementation and analysis. The author is responsible for claims, proofs, sources, credentials and submission. This is a research preprint and prototype, not a peer-reviewed result or intelligence evaluation.

References

- Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023. “Prediction-Powered Inference.” *Science* 382: 669–74. <https://doi.org/10.1126/science.ad i6000>.
- Angelopoulos, Anastasios N., John C. Duchi, and Tijana Zrnic. 2024. “PPI++: Efficient Prediction-Powered Inference.” <https://arxiv.org/abs/2311.01453v2>.
- Ao, Ruicheng, Hongyu Chen, Siyang Gao, Hanwei Li, and David Simchi-Levi. 2026. “Best Arm Identification with LLM Judges and Limited Human.” <https://arxiv.org/abs/2601.21471>.
- Ashouritaklimi, Kianoosh, Valentin Kilian, Daolang Huang, Tom Rainforth, and Francois Caron. 2026. “Prediction-Powered Active Testing.” <https://arxiv.org/abs/2607.08347v1>.
- Barber, Rina Foygel, Emmanuel J. Candes, Aaditya Ramdas, and Ryan J. Tibshirani. 2023. “Conformal Prediction Beyond Exchangeability.” *The Annals of Statistics* 51 (2): 816–45. <https://doi.org/10.1214/23-AOS2276>.
- Bousquet, Nicolas. 2012. “Accelerated Monte Carlo Estimation of Exceedance Probabilities Under Monotonicity Constraints.” *Annales de La Faculte Des Sciences de Toulouse Mathematiques* 21 (3): 557–91. <https://doi.org/10.5802/afst.1345>.
- Cai, Ting, and Kirthevasan Kandasamy. 2026. “Constrained Best Arm Identification with Tests for Feasibility.” <https://arxiv.org/abs/2511.09808>.
- Case, Markus. 2025. “Evaluating Weather Effects on Sortie Generation Using Discrete Event Simulation.” Master’s thesis, Air Force Institute of Technology. <https://scholar.afit.edu/etd/8235/>.
- Durmazkeser, Yavuz, Patrik Okanovic, Andreas Kirsch, Torsten Hoefler, and Nezihe Merve Gürel. 2025. “Large Language Model Selection with Limited Annotations.” <https://arxiv.org/abs/2510.09418v2>.
- Dyer, Allie, Laurel Klawes, Skyler Royse, Keane Singleton, John Miller, and Colby Shufeldt. 2023. “Optimizing Undergraduate Pilot Training 2.5 at Vance Air Force Base.” *Industrial and Systems Engineering Review* 11 (1-2): 15–20. <https://doi.org/10.37266/ISER.2023v11i1-2.pp15-20>.
- Electronic Code of Federal Regulations. 2026a. “14 CFR 61.313 What Aeronautical Experience Must i Have to Apply for a Sport Pilot Certificate.” <https://www.ecfr.gov/current/title-14/chapter-I/subchapter-D/part-61/subpart-J/section-61.313>.
- . 2026b. “14 CFR 61.316 What Are the Performance Limits and Design Requirements for the Aircraft That a Sport Pilot May Operate.” <https://www.ecfr.gov/current/title-14/chapter-I/subchapter-D/part-61/subpart-J/section-61.316>.

- Erickson, Bryngel J. 2014. “Simulation Modeling of Advanced Pilot Training: The Effects of a New Aircraft Family of Systems.” Master’s thesis, Air Force Institute of Technology. <https://scholar.afit.edu/etd/671/>.
- Hartman, Nicole. 2026. “MOSAIC Fact Sheet.” FAA Safety Briefing, March/April 2026, page 25. <https://www.faa.gov/sites/faa.gov/files/MarApr2026.pdf>.
- Howard, Steven R., Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. 2021. “Time-Uniform, Nonparametric, Nonasymptotic Confidence Sequences.” *The Annals of Statistics* 49 (2): 1055–80. <https://doi.org/10.1214/20-AOS1991>.
- Karpathy, Andrej. 2026. “LLM Wiki.” GitHub Gist idea file. <https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f>.
- Kossen, Jannik, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. 2021. “Active Testing: Sample-Efficient Model Evaluation.” In *Proceedings of the 38th International Conference on Machine Learning*, 139:5753–63. Proceedings of Machine Learning Research. <https://proceedings.mlr.press/v139/kossen21a.html>.
- Lei, Jing, Max G’Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. 2018. “Distribution-Free Predictive Inference for Regression.” *Journal of the American Statistical Association* 113 (523): 1094–1111. <https://doi.org/10.1080/01621459.2017.1307116>.
- Li, Jing, and Haibin Zhu. 2026. “Sequential Confidence Sets for Coverage-Constrained Conformal Model Selection.” <https://arxiv.org/html/2609.28522v1>.
- Lindon, Michael, and Alan Malek. 2022. “Anytime-Valid Inference for Multinomial Count Data.” In *Advances in Neural Information Processing Systems*. Vol. 35. https://papers.nips.cc/paper_files/paper/2022/hash/12f3bd5d2b7d93eadc1bf508a0872dc2-Abstract-Conference.html.
- Luo, Jingyu, Mario Vanhoucke, and Jose Coelho. 2026. “Automated Priority Rule Design for the Resource-Constrained Project Scheduling Problem: A Large Language Model-Guided Population-Based Search.” <https://arxiv.org/abs/2609.03754v1>.
- Okal, Adem. 2008. “Simulation Modeling and Analysis of f-16 Pilot Training Squadron.” Master’s thesis, Air Force Institute of Technology. <https://scholar.afit.edu/etd/2814/>.
- Poiani, Riccardo, Remy Degenne, Emilie Kaufmann, Alberto Maria Metelli, and Marcello Restelli. 2024. “Optimal Multi-Fidelity Best-Arm Identification.” <https://arxiv.org/abs/2406.03033v2>.
- Russo, Alessio, Yichen Song, and Aldo Pacchiano. 2025. “Pure Exploration with Feedback Graphs.” <https://arxiv.org/abs/2503.07824>.
- Vilim, Petr. 2009. “Max Energy Filtering Algorithm for Discrete Cumulative Resources.” In *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*. <https://vilim.eu/petr/cpaior2009.pdf>.
- Waudby-Smith, Ian, and Aaditya Ramdas. 2023. “Estimating Means of Bounded Random Variables by Betting.” Royal Statistical Society discussion paper, May 23. <https://doi.org/10.1093/jrsssb/qkad009>.
- Wright, Brian. 2026. “Potential for Enhanced Learning in Machine Learning Classes by Using Wiki LLM Indexing.” <https://arxiv.org/html/2609.25303v1>.
- Zhang, Haoting, Jinghai He, Rhonda Righter, and Zeyu Zheng. 2024. “Language Model Prompt Selection via Simulation Optimization.” <https://arxiv.org/abs/2404.08164v2>.