

CLARK Frontier: Adaptive Monte Carlo for Evaluating Superintelligence

Lloyd Clark, PhD

Federal AI · October 22, 2026

Abstract

A training organization must decide when a cohort is likely to graduate and whether a forecasting or planning system is worth deploying. Individual prediction accuracy misses resource externalities: advancing one student can delay others needing the same aircraft, instructor, or simulator. We formulate evaluation around resource-coupled completion outcomes and a declared operational contract. CLARK Frontier independently trains and tunes bounded control variates, then freezes a stratified allocation before fresh evaluation. Simultaneous empirical-Bernstein intervals cover every declared policy pair, metric, and stage; exact total-variation mixture optimization yields robust comparisons and conservative stability witnesses. A separate cohort-level conformal layer calibrates forecasts under exchangeability. These compose established tools rather than improve the Monte Carlo convergence rate. A 170-cohort synthetic benchmark fits three regressors and compares three procedural forecasts. A validation-selected 48-world simulator forecast has restricted-completion MAE 3.942 days and whole-cohort coverage 37/40 on matched tests, falling to 8/20 under shift. Its access to the true generating simulator limits the comparison. In a four-context learned-priority experiment, the proposed allocation fails to beat pilot-free stratified simulation: at 8,192 worlds it resolves 27/36 contrasts versus 34/36. A learned priority can improve one completion deadline while worsening mean delay and individual target misses. Superintelligence is an intended future target; no superintelligent system or language model is tested. Guarantees remain conditional on their simulator or cohort assumptions.

1 Introduction

Predicting a pilot’s graduation date is a decision problem with shared resources. A student may have only a few flight events remaining, yet a maintenance outage, persistent bad weather, an unavailable instructor, or a prerequisite still blocks completion. A policy that favors that student may consume the same scarce slots needed by several others. The organization therefore needs more than a point estimate of days remaining. It needs a distribution of completion outcomes, a description of the information and assumptions used, and a defensible way to compare alternative forecasting and planning systems.

This problem is relevant to operational AI evaluation because a model’s recommendation changes the system being forecast. A persuasive explanation, an accurate isolated answer, and a feasible resource allocation are different properties. A model may infer a plausible graduation date while implicitly assuming more aircraft or instructor capacity than exists. Conversely, a simple simulator

can outperform a more flexible learned predictor when the governing constraints are known. An evaluation should allow either outcome and should preserve the option to return an unresolved comparison.

We introduce CLARK Frontier as a research framework for this bounded question: under a specified resource-coupled training process, what evidence supports choosing one forecasting configuration or planning policy over another? The term *frontier* has two meanings. First, the method retains a set of configurations that cannot yet be ruled out by valid comparisons, rather than forcing a total ranking. Second, it reports how a comparison changes as the mixture of disruption regimes moves away from a nominal distribution. The output is an evidence record: estimated contrasts, interval bounds, resource checks, mixture witnesses, and the provenance needed to reproduce the calculation.

The title’s reference to superintelligence denotes an intended future use, not an empirical finding. An increasingly capable agent may propose forecasts, priority rules, or contingent policies that are difficult for a human to inspect informally. Its outputs can nevertheless be subjected to a bounded contract with externally executable constraints and measurable consequences. Our evaluation does not define a scalar measure of general intelligence, establish performance beyond humans, or infer global safety from task success. The current experiments evaluate ordinary learned regressors, procedural scheduling policies, and a frozen priority rule derived from an actually fitted random forest. This restricted scope is essential to the validity of the claims.

The paper makes three candidate contributions. First, it defines an operational evaluation contract that connects cohort graduation forecasting, resource externalities, and configuration choice. Second, it implements a pilot-frozen stratified evaluation procedure whose simultaneous interval boxes support all declared pairwise comparisons and all total-variation mixture queries. The statistical components are established; the contribution is the explicit composition and its executable decision semantics. Third, it supplies an independently generated synthetic cohort benchmark, real model fitting, a procedural simulation baseline, censoring-aware outputs, and ablations intended to expose rather than hide the limits of adaptive allocation.

The central empirical question is whether the allocation improves the measured cost or precision of an operationally relevant comparison over simpler alternatives after fitting and tuning costs are counted. The framework remains useful if the answer is negative in some regimes: a valid comparison should reveal when a learned predictor or an adaptive allocation is unhelpful. A claim of universal superiority would be stronger than either the theory or the experiment supports.

2 Related work and the boundary of originality

2.1 Pilot training, stochastic scheduling, and simulation

Simulation of resource-constrained pilot training has substantial precedent. Okal models syllabus-based F-16 training duration and interactions among student pilots, instructors, bandits, and aircraft (Okal 2008). Erickson studies the intersection of student flow and aircraft availability (Erickson 2014). Dyer and colleagues use discrete-event simulation to identify bottlenecks and resource requirements in Undergraduate Pilot Training 2.5 at Vance Air Force Base (Dyer et al. 2023). Case studies weather-sensitive sortie-generation timelines with adjustable aircraft, personnel, equipment, and process inputs (Case 2025). These works rule out a novelty claim based merely on simulating graduation outcomes under resource constraints.

Recent work also uses language models to design scheduling rules. Luo, Vanhoucke, and Coelho generate reusable RCPSP priority functions with language-model-guided population search and evaluate the selected rules on unseen projects (Luo, Vanhoucke, and Coelho 2026). Zhang and colleagues cast language-model prompt selection as simulation optimization, including surrogate-guided sequential evaluation (H. Zhang et al. 2024). Thus neither scoring a language-model-generated rule in a simulator nor applying simulation optimization to AI configuration selection is new by itself. The question here concerns a specific operational outcome contract, simultaneous uncertainty, and explicit resource externalities.

2.2 Active and prediction-powered evaluation

Active testing selects informative test observations while correcting selection bias (Kossen et al. 2021). Berrada and colleagues show how cheap frozen surrogates can guide evaluation of large language models (Berrada et al. 2025). Prediction-powered inference combines model predictions with measured residuals, and PPI++ adapts reliance on predictions to their quality (Angelopoulos et al. 2023; Angelopoulos, Duchi, and Zrnic 2024). These are direct precedents for the estimator used here.

The closest recent work is Prediction-Powered Active Testing, which combines an unbiased active-testing estimator with prediction-powered control variates and derives acquisition rules tailored to residual estimation (Ashouritaklimi et al. 2026). R-AutoEval+ uses adaptive prediction reliance and testing by betting for reliable AI configuration evaluation (Park, Zecchin, and Simeone 2025). Sfyraki and Wang show that uncertainty-directed querying can fail to improve a confidence-width objective, emphasizing the importance of comparing simple query schemes (Sfyraki and Wang 2026). Filtered PPI addresses heterogeneous prediction quality (Xu and Sun 2026). Accordingly, the present paper does not claim to invent adaptive residual correction, robust fallback, or the use of cheap model predictions in evaluation.

Our implemented design makes a deliberate statistical tradeoff. It learns the predictor, control coefficients, and allocation on independent pilot simulations, then freezes them. Fresh observations within each stratum are consequently conditionally independent and identically distributed. This permits an explicit finite-stage empirical-Bernstein analysis without asserting a general martingale central limit theorem for an evolving allocation. The design sacrifices some possible online adaptivity to make the implemented guarantee straightforward to audit.

2.3 Model selection and sequential uncertainty

SELECT-LLM identifies informative annotation queries for choosing among language models (Durmazkeser et al. 2026). Continuous-score adaptive ranking already includes pairwise confidence, stopping, and cost-aware allocation (Balkir et al. 2026). Prediction-powered confidence sequences and Bayes-assisted sequential procedures also precede this work (Kilian, Cortinovis, and Caron 2025, 2026), as does the wider confidence-sequence literature (Howard et al. 2021). A generic promise to select an AI model with fewer evaluations would therefore be insufficiently distinctive.

Here, the configurations are evaluated against completion outcomes induced by a shared resource system. The statistical comparison is tied to explicit losses and constraints, not to an assumed latent intelligence score. The tested forecasting configurations include three fitted regressors and three procedural baselines. The tested planning configurations are deterministic policies, including one whose priorities are computed by a fitted random forest and frozen at the decision cutoff.

Evaluation of actual language-model-generated policies remains a future experiment rather than a hidden interpretation of the current results.

2.4 Calibration, robustness, and increasingly difficult tests

Split conformal prediction supplies finite-sample marginal predictive coverage when its exchangeability assumptions hold (Lei et al. 2018). Grouped and partially observed data require more careful inference (Lee and Ren 2025; Mallick et al. 2026). Recent panel methods show that arbitrary cross-unit mismatch and feedback mechanisms cannot simply be ignored (Tu and Giesecke 2026), while Barber and colleagues quantify consequences of departures from exchangeability (Barber et al. 2023). We therefore use one calibration score per cohort and distinguish a simulator-conditional Monte Carlo interval from a predictive interval calibrated on historical or synthetic cohorts.

Distributional robustness is also established. Recent papers study Wasserstein-robust rare-event simulation and robust quantile regression (Ahn, Chen, and Zheng 2026; C. Zhang, Mao, and Wang 2026). Our uncertainty set is intentionally simpler: only the mixture masses of a finite set of declared disruption regimes may change. Within-regime dynamics remain fixed. This limited ambiguity set admits an exact linear optimization and an interpretable probability-mass-transfer witness; it is not protection against arbitrary simulator misspecification.

FrontierMath, Humanity’s Last Exam, MathArena, and MIT-linked MathNet motivate independently checkable answers, harder maintained evaluations, false-premise recognition, and tests of structural rather than superficial equivalence (Glazer et al. 2025; Center for AI Safety, Scale AI, and HLE Contributors Consortium 2026; Dekoninck et al. 2026; Alshammari et al. 2026). We transfer those design principles, not their difficulty level or leaderboard status. A constrained planning benchmark should maintain representative operational cases alongside harder challenge cases. Their failure frequencies have different interpretations and should not be merged into an operational risk estimate.

3 A resource-coupled operational contract

3.1 Contexts, policies, disturbances, and completion

Let a context c contain the decision cutoff, a cohort $I(c)$ of N_c pilots, remaining syllabus stages, deadlines, qualifications, and a resource snapshot. Let π be a policy whose action at day t depends only on observations available by that day. Future disturbance paths W contain weather, aircraft availability, instructor availability, pilot availability, and retrieval outcomes. For a declared law P_c , define the completion time

$$T_i(c, \pi, W) \in \{0, 1, \dots\} \cup \{\infty\}.$$

An evaluation horizon H restricts what can be observed or simulated. We use the restricted completion outcome and a completion indicator

$$R_i = \min(T_i, H), \quad G_i(d) = \mathbf{1}\{T_i \leq d\}, \quad 0 \leq d \leq H.$$

A completion beyond H is right-censored. The equality $R_i = H$ is an exact statement about the restricted target and is not a claim that the pilot graduates on day H . Completion probabilities for deadlines beyond the horizon and quantiles whose mass is unresolved at the horizon must remain

unavailable or open-ended. This distinction applies equally to forecasts displayed in a user interface and to benchmark labels.

Shared capacity is explicit. At each day, flight attempts consume both aircraft-derived sortie capacity and instructor capacity; simulator attempts consume simulator slots and the same instructor pool. A pilot makes at most one attempt per day, and events follow the declared ordered stages. If x_{it}^F and x_{it}^S indicate flight and simulator attempts, representative constraints are

$$\sum_i x_{it}^F \leq A_t, \quad \sum_i (x_{it}^F + x_{it}^S) \leq U_t, \quad \sum_i x_{it}^S \leq V_t, \quad x_{it}^F + x_{it}^S \leq 1.$$

The full contract also records stage eligibility and the information cutoff. Violating one of these rules produces a hard-check failure; a favorable average loss cannot offset it.

3.2 Forecasting and planning are different interfaces

A forecasting configuration maps the current context and a pilot’s state to an estimate or predictive interval for R_i . Its predictive accuracy can be measured against a later realized cohort. A planning configuration maps an available state to a feasible allocation of attempts and thereby changes the completion distribution. A forecast conditional on one policy cannot be silently treated as a forecast under another policy. The implementation exposes both interfaces but evaluates them in separate experiments.

For a planning policy, useful bounded losses include mean restricted completion and deadline-miss fraction,

$$Y_{\pi,0}(W) = \frac{1}{N_c H} \sum_i R_i(c, \pi, W), \quad Y_{\pi,d}(W) = 1 - \frac{1}{N_c} \sum_i G_i(c, \pi, W; d).$$

All lie in $[0, 1]$, with smaller values preferred. Prespecified individual or subgroup losses can also be included, but adding metrics increases the simultaneous-inference burden. A cohort average alone may conceal concentrated harm, so the loss contract should identify which distributional summaries and subgroups matter to the decision maker.

For policies a and b and metric k , the target contrast is

$$\Delta_{abk}(q) = \sum_{h=1}^S q_h \mathbb{E}_h[Y_{a,k}(W) - Y_{b,k}(W)].$$

The same disturbance path is used for both policies. This coupling changes the variance of the contrast without changing either policy’s marginal law. It does not give a policy knowledge of the future. Every unordered policy pair is included, rather than comparing alternatives only with a reference that may be far from the decisive boundary.

3.3 Simulator implementation

The daily simulator uses persistent weather and aircraft availability, shared instructors, shared simulator slots, pilot availability, and retrievals. Its default horizon is 120 calendar days from November 2, 2026, with no training attempts on weekends. Default resources are 18 aircraft, 1.2 sortie slots

per available aircraft per day, 12 instructors, and eight simulator slots. Nominal aircraft availability is 0.65 and nominal good-weather probability is 0.8. Weather and aircraft persistence parameters are 0.65 and 0.75. Instructor and pilot availability probabilities are 0.94 and 0.96, and the default attempt failure probability is 0.08.

These are synthetic demonstration parameters, not estimates of a particular unit. Default remaining stages contain an ordered sequence of simulator and flight phases. A supplied scenario may provide explicit remaining stages instead. Flight capacity is stochastically rounded from available aircraft, sortie rate, and a weather factor; bad weather retains a factor of 0.20. Instructor capacity is shared by both event types. A failed event is retried on a later day.

Four declared disruption strata have nominal masses (0.70, 0.15, 0.10, 0.05): nominal, weather, maintenance, and coupled disruption. Weather stress scales the nominal good-weather probability by 0.60; maintenance stress scales aircraft availability by 0.70; the coupled regime applies both. These regime definitions are part of the supplied simulator law, not a learned claim about the frequency of actual Air Force disruptions.

Future randomness is indexed by resource, pilot, and day independently of which policy consumes it. This prevents a policy’s different scheduling order from accidentally receiving a different random-number mechanism. Baseline policies include immutable-order FIFO, earliest deadline, and a balance rule that prioritizes current deadline slack with waiting-time and priority tie breaks. A frozen external priority mapping supplies the adapter boundary for model-generated planning proposals. The tested policies remain explicitly identified in each result.

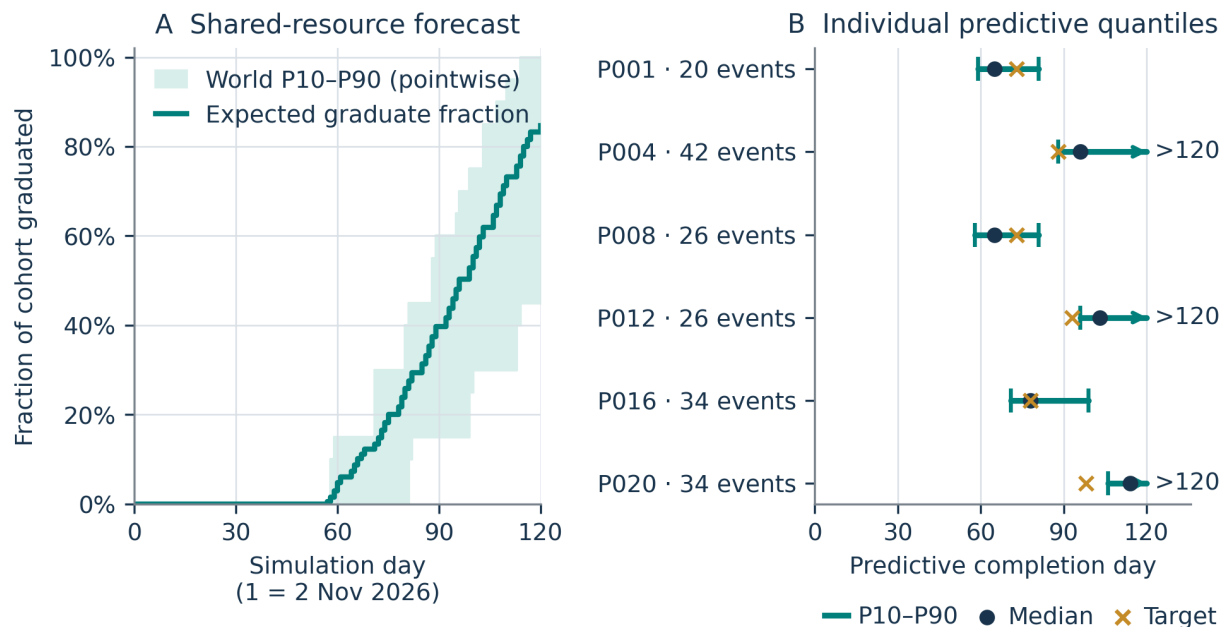


Figure 1: Configured completion forecasts for a synthetic twenty-pilot cohort, starting November 2, 2026, under the balanced policy and 300 simulated worlds. Panel A shows the expected graduate fraction and pointwise world P10–P90 range, which is predictive under the configured simulator rather than a simultaneous confidence band. Panel B shows selected pilots’ predictive P10/P50/P90 completion times and requested targets; arrows preserve quantiles beyond the 120-day horizon.

4 Pilot-frozen adaptive Monte Carlo

4.1 Independent learning, tuning, and evaluation

Let $j = (a, b, k)$ index an unordered policy pair and a bounded loss metric; there are J such contrasts. Within each stratum h , independent training simulations fit a ridge predictor $m_{a,k}(X)$, clipped to $[0, 1]$. Cheap features X must be computable without executing the expensive outcome evaluation. Independent tuning simulations choose a contrast-specific coefficient $\beta_j \in [0, 1]$ and the evaluation allocation. The coefficient estimates a weighted within-stratum covariance-to-variance ratio and is clipped; it is not fixed to a favorable algebraic transform of the true outcome.

These evaluation controls and the deployed forecasting models have different information sets. The control's six features summarize a sampled future disturbance tape: good-weather fraction, flight and instructor capacity per pilot, simulator capacity per pilot, pilot availability fraction, and the longest zero-flight-capacity run. Generating that tape does not execute a scheduling policy, but it still has a computational cost. Such future summaries are legitimate internal Monte Carlo controls because the evaluator has drawn the simulated future; they are unavailable to a real forecaster at the decision cutoff. Neither the forecasting regressors nor the tested priority policies receive them. Their separate 27-feature input contains only cutoff information.

After tuning, the predictor, coefficients, mixture radius, allocation, metrics, policy set, and evaluation stages are frozen. Neither training nor tuning outcomes enter the final estimation sample. For a fresh expensive observation define

$$D_{hj} = Y_{a,k} - Y_{b,k}, \quad M_{hj} = \beta_j(m_{a,k}(X) - m_{b,k}(X)), \quad Z_{hj} = D_{hj} - M_{hj}.$$

An independent cheap feature pool of size N_h^C estimates $\mathbb{E}_h M_{hj}$. At declared stage ℓ , the fresh expensive pool contains $n_{h\ell}$ observations. The stratum estimator is

$$\hat{\delta}_{hj\ell} = \overline{M}_{hj}^C + \overline{Z}_{hj\ell}^E.$$

The cheap mean is measured and carries uncertainty. It is not treated as an exact expectation merely because predictions are inexpensive. Sampling with $\beta_j = 0$ reduces the estimator to an ordinary stratified paired mean. Conditional on the independent pilot design, linearity of expectation gives unbiasedness at each fixed declared stage. No optimal-stopping unbiasedness claim is made.

4.2 Allocation and ablations

The allocation is adaptive to the independent pilot, not updated online during the final evaluation. Let v_{hj} be a pilot estimate of residual variance, and let d_{hj} denote a pilot contrast mean. A target allocation uses nominal masses p_h . A residual-variance allocation pools $p_h^2 v_{hj}$ across contrasts. The proposed frontier allocation additionally emphasizes contrasts close to a mixture-robust decision boundary.

For a declared total-variation radius ρ , let q_j^+ and q_j^- maximize and minimize $q^T d_j$ over the ambiguity set defined below. Define

$$w_j = \left[\max\{\min(|(q_j^+)^T d_j|, |(q_j^-)^T d_j|), \tau\} \right]^{-2},$$

$$s_h^2 = \sum_j w_j \frac{(q_{jh}^+)^2 + (q_{jh}^-)^2}{2} \max(v_{hj}, \eta), \quad \lambda_h = \epsilon p_h + (1 - \epsilon) \frac{s_h}{\sum_g s_g}.$$

Here $\tau > 0$ is a tolerance floor, $\eta > 0$ is a numerical variance floor, and $\epsilon \in (0, 1]$ preserves exploration. The formula is a heuristic design choice. Unlike a one-step variance-minimizing allocation for a fixed scalar mean, it does not prove globally optimal selection or minimum time to certification. An allocation that reduces point error may still worsen conservative interval width.

Four ablations isolate the effects: target allocation without a predictor; target allocation with a tuned predictor; residual-variance allocation with the same predictor; and frontier allocation with the same predictor. The target baseline requires no training, tuning, or cheap integration. The three predictor methods receive the same independent training and tuning protocol. Comparisons match total expensive simulation budgets, so the baseline can spend the predictor methods' pilot budget on additional fresh outcomes. Integer sample counts use a deterministic nested schedule with at least two observations per stratum. Later declared stages add observations without removing previous ones. Conditional on the pilot, stage counts are fixed before final outcomes are observed.

5 Simultaneous inference and mixture robustness

5.1 Finite-stage empirical-Bernstein intervals

For iid observations $X_1, \dots, X_n$ in an interval of deterministic width B , write s_n^2 for unbiased sample variance and define

$$e(X_{1:n}; B, \delta) = \sqrt{\frac{2s_n^2 \log(4/\delta)}{n}} + \frac{7B \log(4/\delta)}{3(n-1)}, \quad n \geq 2.$$

Applying the empirical-Bernstein inequality to both signs yields a two-sided interval with failure probability at most δ (Maurer and Pontil 2009). The width B is a support assertion, not the observed sample range. In particular, a constant observed sample need not justify a zero uncertainty radius.

The cheap contrast has width $B_j^C = 2\beta_j$, and the expensive residual has width $B_j^E = 2(1 + \beta_j)$. With L prespecified stages and total error level α , allocate

$$\delta_C = \frac{\alpha}{2SJ}, \quad \delta_E = \frac{\alpha}{2SJL}.$$

Let $r_{hj\ell} = e(M^C; B_j^C, \delta_C) + e(Z^E; B_j^E, \delta_E)$. Define interval endpoints

$$L_{hj\ell} = \max(-1, \hat{\delta}_{hj\ell} - r_{hj\ell}), \quad U_{hj\ell} = \min(1, \hat{\delta}_{hj\ell} + r_{hj\ell}).$$

The truncation uses the known range of a difference of bounded losses. If numerical endpoints become reversed outside the coverage event, the implementation conservatively widens the interval to the full range rather than returning an empty interval. A zero coefficient has an identically zero cheap term and zero cheap uncertainty, rather than a simulated confidence penalty for a term that does not exist.

Proposition 1 (simultaneous interval boxes). Conditional on a frozen independent pilot design, suppose fresh expensive and cheap observations are iid within each stratum, their laws match the declared target conditional laws, and the deterministic support widths hold. Then, with probability at least $1 - \alpha$, every stratum contrast belongs to its interval at every declared stage:

$$\mathcal{E} = \{L_{hj\ell} \leq \delta_{hj} \leq U_{hj\ell} \text{ for all } h, j, \ell\}, \quad \Pr(\mathcal{E}) \geq 1 - \alpha.$$

The proof is an explicit union bound and appears in Appendix A. Dependence between different policy contrasts, reuse of the cheap pool across stages, and nesting of expensive pools do not invalidate that union bound. The claim covers the declared finite set of stages for one frozen evaluation method. Simultaneous guarantees across additional method-selection experiments require an additional error allocation. It is not an unbounded-horizon anytime confidence sequence. Adding stages or metrics after inspecting results requires a new error allocation or a different inference method.

5.2 Exact mixture optimization

The nominal stratum masses are $p \in \mathbb{R}^S$, with $p_h > 0$ and $\sum_h p_h = 1$. For a declared radius ρ , define

$$\mathcal{Q}_\rho(p) = \{q \geq 0 : \mathbf{1}^T q = 1, \text{TV}(q, p) = \frac{1}{2} \|q - p\|_1 \leq \rho\}.$$

On $\mathcal{E}$, every mixture contrast is bounded by

$$\underline{\Delta}_{j\ell}(\rho) = \min_{q \in \mathcal{Q}_\rho(p)} q^T L_{j\ell}, \quad \overline{\Delta}_{j\ell}(\rho) = \max_{q \in \mathcal{Q}_\rho(p)} q^T U_{j\ell}.$$

Proposition 2 (robust propagation). Under Proposition 1’s assumptions, the bounds simultaneously cover all declared contrasts, all declared stages, and every radius $\rho \in [0, 1]$. No discretization penalty for the radius is required because the same stratum event $\mathcal{E}$ implies the inequalities for every simplex mixture.

The optimization is a finite linear program. Its implementation greedily transfers at most ρ probability mass from low-valued donor strata to high-valued receiver strata, respecting donor availability and receiver capacity. Sorting costs $O(S \log S)$ and the transfer loop is linear in S . Independently enumerated simplex-grid cases and a separate general linear-program solver check the implementation in addition to the exchange argument below. The latter compares both extrema for 36 random problems with two to seven strata and checks the returned witnesses’ feasibility; the optimization is not verified solely by rerunning its own formula.

This robustness claim has a narrow scope. It allows probabilities of encoded regimes to change while preserving each regime’s conditional simulator law. It cannot represent a missing syllabus rule, an incorrect repair-time distribution inside a regime, unmodeled maintenance dependence, biased source records, or an adversary with capabilities absent from the simulator. The radius is a sensitivity parameter unless estimated and validated from independent operational data.

5.3 Possible frontiers and fragility witnesses

A policy a is certified to weakly improve on b across the chosen losses if every robust upper contrast is nonpositive and at least one is strictly negative. Removing only policies with certified dominance

yields a *possible frontier*. Remaining policies need not be truly Pareto-optimal; limited data may simply fail to separate them. Hard feasibility checks remain separate and precede this stochastic comparison.

For a scalar improvement threshold t , define the confidence-envelope crossing radius

$$\hat{\rho}_{j\ell}^{\text{cert}} = \inf\{\rho \in [0, 1] : \bar{\Delta}_{j\ell}(\rho) \geq t\}.$$

If the upper contrast already reaches t at the nominal mixture, the reported radius is zero. If no simplex mixture reaches t , the implementation returns an explicit all-mixtures status. Otherwise a monotone bisection locates the crossing and returns an optimizing mixture. This is a conservative radius of *certified stability*. It is not an estimate of the true policy breakdown radius. The witness identifies a mixture at which the upper confidence envelope reaches the threshold; it does not prove that the true loss has crossed it.

Proposition 3 (conservative stability). On $\mathcal{E}$, if $\bar{\Delta}_{j\ell}(\rho) < t$, then $\Delta_j(q) < t$ for every $q \in \mathcal{Q}_\rho(p)$. Consequently the certified stability radius cannot exceed the first radius at which the true robust contrast reaches t , under the usual infimum convention. This implication follows from Proposition 2 rather than a new robustness theorem.

6 Cohort-level calibration and model eligibility

6.1 One calibration unit per cohort

Simulation uncertainty and predictive calibration answer different questions. Monte Carlo intervals concern functionals of a declared simulator. A predictive interval concerns a future cohort outcome and requires calibration data with a justified relationship to that future cohort. A large number of pilots in a few cohorts does not provide the same evidence as many independent cohorts.

Let a frozen predictor return intervals $[\ell_{ci}, u_{ci}]$ for every pilot in calibration cohort c . For a fully observed target Y_{ci} , define one score

$$S_c = \max_{i \in I(c)} \max\{0, \ell_{ci} - Y_{ci}, Y_{ci} - u_{ci}\}.$$

For n calibration cohorts, set $k = \lceil (n+1)(1-\gamma) \rceil$. Let q_γ be the k th calibration score if $k \leq n$, and $+\infty$ otherwise. Expand every interval by that common amount. Point predictions are the special case $\ell = u$.

Proposition 4 (whole-cohort marginal coverage). If calibration cohorts and one future cohort are exchangeable, the predictor and selection rule are frozen independently of the calibration outcomes, and each score is computed from fully observed targets, then

$$\Pr\{Y_{*i} \in [\ell_{*i} - q_\gamma, u_{*i} + q_\gamma] \text{ for every } i \in I(*)\} \geq 1 - \gamma.$$

The benchmark freezes the simulator forecaster's 48 seeds and uses the same resulting forecasting map for every cohort; conditional on those seeds it is a deterministic function of the cutoff context. This preserves exchangeability without requiring independent forecast randomness across cohorts. Clipping intervals to a known target range preserves the result. For the benchmark's target

$R_i \in [0, H]$, the restricted outcome is observed even when the actual graduation time exceeds H , so this theorem applies to R_i . For an actual graduation target T_i that is right-censored, the implementation must not silently substitute H as a known graduation time. Its conservative calibration interface assigns an infinite score to an unresolved latent completion. Finite guarantees for T_i require additional survival or censoring assumptions not supplied here.

The result is marginal over a new exchangeable cohort, not conditional coverage for each pilot, each syllabus type, or each shifted resource regime. Larger classes, a new curriculum, changes in maintenance practice, and new patterns of weather can violate the required exchangeability. Such changes are explicit stress tests; they do not inherit a distribution-free guarantee by terminology alone.

6.2 Validation-first configuration selection

The forecasting study compares six configurations: an untrained workload heuristic, an untrained resource-fluid heuristic, Ridge regression, random forest regression, histogram gradient boosting, and a procedural mean forecast using 48 simulated worlds. Learned configurations use only features available at the forecast cutoff. Hyperparameters are fixed before final testing. The validation rule first retains configurations with mean absolute restricted-time error no greater than eight days, then selects the one with lowest median CPU prediction latency. If no configuration passes, selection returns no eligible configuration rather than relaxing the gate after seeing the results.

Selection is frozen before calibration and final testing. The calibrated intervals for the selected configuration consequently retain the split-conformal argument under exchangeable cohorts. Validation eligibility is not itself a confidence-certified population-risk statement: it is an operational selection rule whose generalization is assessed on a separate test set. Reported calibration results for nonselected configurations are descriptive comparisons, not permission to choose the most favorable interval after inspecting calibration or test outcomes.

7 Experimental protocol

7.1 Forecasting experiment

The benchmark generates disjoint streams of 60 training, 20 validation, 30 calibration, 40 matched-mixture test, and 20 shifted test cohorts, using root seed 20261022. A cohort’s family is drawn independently and uniformly from ordinary, instructor-constrained, weather-constrained, and interleaved-syllabus families. Each cohort receives one independently generated future path shared by its pilots. Forecast labels and the 48-world simulation forecast are conditional on the earliest-deadline scheduling policy. Training rows may contain several pilots, but cohort identity defines the split and the predictive coverage unit. Labels retain both restricted completion and censoring indicators. No future weather realization, realized maintenance outcome, or subsequent accomplishment enters the cutoff features.

The 27 cutoff features include remaining flight and simulator events, target slack, pilot priority, cohort workload, resource capacities, nominal weather and availability parameters, persistence, ordered-stage counts, and the pilot’s position fraction. Ridge uses standardized features and penalty 10. The random forest uses 120 trees, maximum depth 14, minimum leaf size four, and one CPU worker. Histogram gradient boosting uses 160 iterations, 15 leaves, learning rate 0.07, and L2 penalty three. These are fitted models, not labels attached to handwritten responses.

We report restricted-time MAE and RMSE, whole-cohort joint coverage, pilot-level coverage, median interval width, CPU latency, and model size. MAE and RMSE give each pilot equal weight; joint coverage gives each cohort equal weight. This changes the estimand rather than turning pilots within a cohort into independent calibration observations. With 30 calibration cohorts and nominal joint coverage 0.9, the conformal rank is $\lceil 31 \times 0.9 \rceil = 28$. A naive plus-or-minus-two-day interval is included as a deliberately simple baseline for interval usefulness. Accuracy, interval width, latency, and censoring must be interpreted together.

7.2 Policy comparison and allocation experiment

The policy experiment uses the first four held-out cohorts in the frozen model-output artifact, containing 24, 23, 18, and 19 pilots, without filtering on future outcomes. It compares FIFO, a frozen initial-workload priority, and a random-forest predicted-slack priority. The workload priority is minus the number of remaining events at the cutoff; its implementation identifier is `shortest_remaining`. It is not an online shortest-remaining-time oracle. The learned priority is predicted restricted completion minus target day, with larger scores served first. All scores are frozen before the evaluation futures are drawn, and the downstream simulator enforces resource and stage constraints. The random forest was fitted to forecast under earliest-deadline scheduling; using its predicted slack to intervene is a separate, explicitly tested adapter, not a model trained to optimize scheduling reward. This experiment evaluates the consequences of a fitted model’s recommendation, rather than assuming that forecast accuracy implies a good schedule.

Every unordered policy pair and three prespecified losses are retained: cohort mean restricted completion divided by H , fraction missing each pilot’s own declared target day, and fraction not complete by day $\lfloor 0.75H \rfloor$. All own target days in the tested cohorts are observable within the horizon; the executable adapter otherwise replaces that metric with a declared within-horizon completion fraction rather than silently labeling an unobserved deadline. There are $S = 4$ strata and $J = 9$ pair-metric contrasts. All policies in a comparison receive the same future disturbance path. Training, tuning, cheap integration, and fresh evaluation streams are separated.

The three control-variates methods each use 16 training and 12 tuning worlds per stratum, totaling 112 pilot worlds; the target-proportional baseline incurs no pilot. Each nonzero control integrates 2,048 independent feature worlds per stratum. The nominal mixture is $(0.70, 0.15, 0.10, 0.05)$, the TV radius is 0.05, and the total interval error level is 0.05 for each cohort-method evaluation across its three declared stages. Aggregating the four cohorts is descriptive; a post-hoc choice among methods does not inherit a joint 95 percent guarantee across all experiments. The allocation uses exploration weight $\epsilon = 0.2$, decision tolerance floor $\tau = 0.02$, and residual variance floor $\eta = 10^{-7}$. Root seed 20261022 and recorded cohort-specific stream derivations separate phases and match samples within a stratum across allocation methods.

Total expensive-world budgets are 512, 2,048, and 8,192. The no-control baseline spends the full budget on fresh outcomes; each control method spends 112 on pilot outcomes and the remainder on fresh evaluation. One world evaluates all three policies, so the logical policy-outcome charge is three times the world budget for every method. Cheap world generation and predictor calls are recorded separately. An execution cache reuses identical sampled worlds and their policy outcomes across ablations, but it does not reduce the logical per-method charge. Method-order-dependent cached wall times are not evidence of a runtime speedup.

This cost comparison concerns evaluation after the three candidate policies are fixed. The shared cost of fitting the forecasting models belongs to the separate model benchmark and is not removed

by the evaluator’s budget accounting.

The actual policy experiment has no exact or high-budget truth reference. It therefore reports interval width and decision resolution, without claiming empirical estimator RMSE or coverage against an unknown policy mean. Exact tiny-schedule examples and independently derived probabilities check feasibility and simulator behavior. The TV implementation is checked against both a simplex grid and an independent linear-program solver. A full path-space enumeration is infeasible for the persistent, multi-resource simulator, and there is no fixed finite table that can be exhausted more cheaply than the largest sampling budget.

The primary comparisons distinguish precision from decision power. A smaller variance does not imply that simultaneous bounds certify a useful improvement. Conversely, coverage with intervals spanning the entire feasible loss range is not evidence of a useful release procedure. We report nominal and robust uncertainty, unresolved comparisons, and possible-frontier size. A secondary noninferiority contract permits at most 0.05 worse normalized loss than FIFO on every metric: six restricted mean days at $H = 120$, and five percentage points for each miss fraction. These are predeclared demonstration tolerances, not Air Force-approved operational limits.

7.3 Independent stochastic-law validation

A separate experiment checks the statistical implementation against analytic truth, rather than treating the graduation simulator as its own truth oracle. It fixes 100 independent replications per method and compares frontier with target-proportional sampling at total budgets 256, 1,024, and 4,096. Four strata have masses (0.5, 0.25, 0.15, 0.1), three abstract arms have two bounded losses, and TV radius is 0.08. The continuous stochastic law is $Y_{a,k} = w_k \mathbf{1}\{U_k < r_{h,a,k}\} + (1 - w_k)V_{a,k}$, with shared uniform U_k , independent uniform $V_{a,k}$, and weights (0.7, 0.6). Its exact means are $w_k r_{h,a,k} + (1 - w_k)/2$; the recorded probability matrix defines all contrasts. Controls receive noisy versions $U_k + 0.35N_k$, stratum indicators, and an independent auxiliary feature. They do not receive exact means or robust extrema.

Frontier uses eight training and six tuning worlds per stratum and 512 cheap feature worlds per stratum; the pilot-free baseline uses its saved budget for fresh outcomes. Root seed is 2026102200. Each replication checks the joint event across all strata, pair-metrics, and three declared stages at $\alpha = 0.05$, and independently solves the truth-level TV programs. Closed-form truth is used only to score coverage, false dominance, contrast RMSE, and recovered true dominance edges. This is a distinct validation law, not another pilot-training data set.

8 Results and interpretation

8.1 Held-out forecasting and configuration choice

The 170 generated cohorts contain 3,341 pilot records. The training, validation, calibration, matched test, and shifted test splits contain 1,129, 416, 597, 805, and 394 pilots respectively. The corresponding numbers of right-censored actual completion times are 165, 65, 105, 132, and 211. These are pilot labels within 170 dependence units, not 3,341 independent graduation histories.

Only the 48-world Monte Carlo configuration meets the prespecified validation MAE limit of eight days. The validation rule therefore selects it without using calibration or test performance. Table 1 reports the complete matched-test comparison. MAE and RMSE concern $R = \min(T, 120)$ in calendar days. Joint coverage is the fraction of 40 test cohorts for which all pilots’ restricted targets

fall inside their intervals. Reported timing is median wall-clock prediction latency on the CPU per cohort in this execution environment; it is not a portable service-level latency claim.

Forecasting configuration	Validation MAE (days)	Test MAE (days)	Test RMSE (days)	Joint cohort coverage	Median width (days)	CPU latency (ms)
Remaining-workload heuristic	37.336	33.567	44.440	0.975	120.000	0.005
Resource-fluid heuristic	23.656	22.351	30.189	0.850	120.000	0.012
Ridge regression	12.031	11.614	14.937	0.925	79.331	0.178
Random forest	10.382	9.243	12.698	0.950	83.774	8.915
Histogram gradient boosting	10.307	9.356	12.908	0.925	79.198	2.486
Monte Carlo, 48 worlds (selected)	3.824	3.942	6.244	0.925	35.792	806.869

Table 1: Frozen forecasting configurations on the matched-mixture test set. Intervals use one maximum residual per calibration cohort and are clipped to the restricted target range $[0, 120]$. Nonselected models’ calibration results are descriptive. Median widths are across pilot intervals; coverage is computed across whole cohorts.

The selected configuration covers all pilots in 37 of 40 matched test cohorts, with conformal radius 17.896 days and median bounded interval width 35.792 days. The naive workload prediction with a fixed two-day interval covers all pilots in none of the 40 test cohorts; its pilot-level coverage is 0.119. This large discrepancy illustrates why a narrow interval alone is not useful evidence. Conversely, the workload heuristic achieves high joint coverage largely by returning almost the full 120-day target range. Coverage must be accompanied by width and a stated target.

The selected Monte Carlo forecast has access to the same correctly specified simulator and context probabilities used to generate the labels. Its lower error than regressors trained on 60 synthetic cohorts is an expected sanity result under that advantage. It establishes neither a general statistical superiority of Monte Carlo over machine learning nor accuracy on actual Air Force records. The learned models supply executable and much faster alternatives; under the declared error gate their speed does not make them eligible in this run. Changing the gate after seeing these results would answer a different selection question.

8.2 Shift stress and the limits of pooled calibration

The shifted test set adds workload and more interleaved prerequisite stages, reduces weather and aircraft availability, and increases weather persistence outside the calibration mixture. The selected configuration’s restricted-time MAE is 5.309 days and its whole-cohort coverage is 8 of 20 cohorts, or 0.400. Pilot-level coverage is 0.909, while median restricted interval width is 23.188 days. The narrower observed width partly reflects clipping near the horizon in a set with 211 of 394 actual graduation times censored; it does not indicate stronger knowledge of those actual graduation dates.

The difference between 0.909 pilot-level coverage and 0.400 whole-cohort coverage is operationally consequential. A planner who needs an interval containing every graduation outcome in a new

class cannot substitute an average per-pilot coverage statistic. More importantly, the pooled split-conformal theorem does not apply under this constructed shift. Even a correct conditional simulator forecast can have a different distribution of maximum residuals when cohort size, workload, and persistence change. The stress result therefore records a failed transfer of the matched calibration, rather than interpreting the nominal 90 percent target as a guarantee under shift.

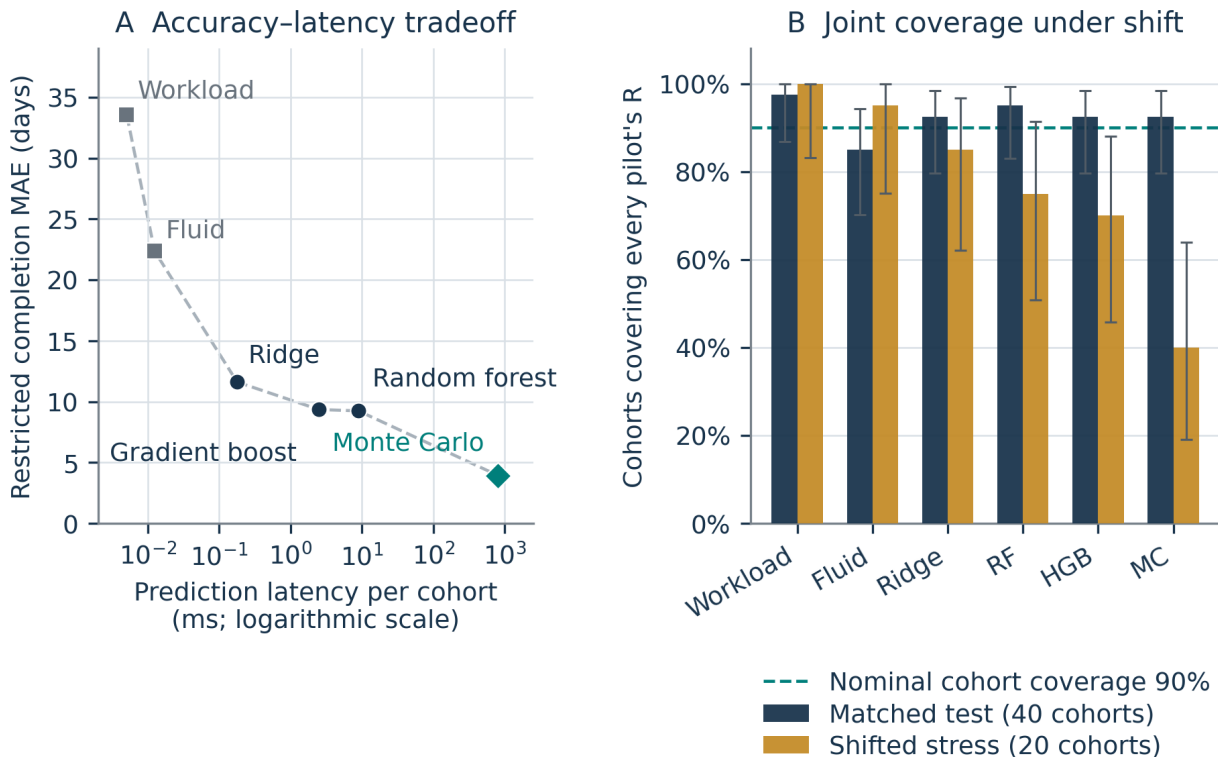


Figure 2: Canonical forecasting benchmark. Panel A compares restricted-time test MAE with measured per-cohort prediction latency on a logarithmic scale; the dashed curve is the observed accuracy-latency Pareto frontier. Selection used validation outcomes before calibration and testing. Panel B shows whole-cohort restricted-target coverage for matched and shifted cohorts, with exact 95 percent binomial intervals for the descriptive proportions. The nominal conformal statement requires exchangeability and does not extend to shift. The Monte Carlo configuration uses the same simulator family that generates labels.

8.3 Allocation ablation: the simpler baseline wins

Table 2 aggregates the four policy contexts with equal context weight. Each context contributes nine policy-pair-by-metric contrasts, giving 36 comparisons per method and stage. A resolved contrast has a robust interval excluding zero; it does not by itself imply that one policy dominates another on all three losses.

Total worlds per context	Evaluation design	Mean robust width	Resolved contrasts / 36	Mean possible-frontier size
512	Target, no control	0.7418	0	3.00
512	Target with control	1.0011	0	3.00
512	Residual variance	1.0098	0	3.00
512	Frontier	1.0189	0	3.00
2,048	Target, no control	0.2331	8	2.50
2,048	Target with control	0.3652	4	3.00
2,048	Residual variance	0.3340	5	2.75
2,048	Frontier	0.3305	6	2.75
8,192	Target, no control	0.0619	34	2.00
8,192	Target with control	0.1051	27	2.00
8,192	Residual variance	0.0978	27	2.00
8,192	Frontier	0.0969	27	2.00

Table 2: Cost-matched policy evaluation at TV radius 0.05. Width averages the nine robust contrast intervals and then the four contexts; all losses are normalized to $[0, 1]$. Counts aggregate descriptive per-context results and are not a new joint confidence statement across methods and contexts. Pilot worlds are included in the total. A possible frontier retains policies that have not been certified dominated.

The proposed frontier allocation does not beat the simpler pilot-free target-proportional design in this experiment. At 8,192 worlds its mean robust width is 0.0969 versus 0.0619, approximately 57 percent wider. It resolves 27 of 36 contrasts, compared with 34 for the no-control baseline. Frontier allocation modestly improves average width relative to the control-based target allocation, but neither that comparison nor the residual-variance ablation establishes a benefit over ordinary stratified paired simulation. At 512 worlds no method resolves any contrast. These are finite-sample precision and resolution findings; absent exact policy means, they are not empirical coverage or mean-squared-error measurements.

This negative result identifies a mismatch between the tuning objective and the release certificate. A covariance-based coefficient targets variance, whereas the implemented empirical-Bernstein radius also depends on a conservative deterministic residual support, cheap integration uncertainty, multiplicity, and the pilot budget. A smaller residual variance need not give a narrower valid interval. Appendix A.8 gives a concrete counterexample for these generic support bounds, even with a perfect control. This does not establish that control variates are intrinsically inferior. A coefficient selected to reduce the full certificate width, with an honestly evaluated zero-control option, is a plausible

future design; testing it on the same four observed contexts would be exploratory.

At the largest stage the 16 cohort-method evaluations logically charge 131,072 expensive worlds and 393,216 policy outcomes. Shared execution actually computes 35,157 unique expensive worlds, and the predictor arms separately charge 98,304 cheap feature worlds. The smaller unique count reflects reuse across ablations, not access to a finite exhaustive population or an algorithmic runtime improvement. Stage samples are nested, so summing the three stage budgets would double-count earlier observations.

8.4 A concrete conflict between graduation metrics

Table 3 reports the no-control method’s largest-stage estimates for the actual random-forest priority relative to FIFO. Positive differences indicate harm. The same fitted forecasting model produces different planning consequences in different contexts; it is never certified noninferior to FIFO under the three-metric tolerance contract in this experiment.

Held-out context	Family	Restricted mean difference (days)	Own-target miss difference (percentage points)	Day-90 miss difference (percentage points)
test-000	Ordinary	+11.648	+4.349	+5.527
test-001	Interleaved	+8.414	+8.988	-5.159
	syllabus			
test-002	Ordinary	+7.087	+8.409	-5.607
test-003	Instructor	+8.323	+4.432	+4.893
	constrained			

Table 3: Nominal estimated contrast, random-forest priority minus FIFO, on the four prespecified policy contexts. Restricted mean days multiply the normalized contrast by 120. Miss differences are percentage points. These are policy consequences under the simulator, not the forecasting errors in Table 1.

For test-001, the robust restricted-mean contrast interval is $[5.177, 12.132]$ days and the own-target miss interval is $[5.584, 12.355]$ percentage points, while the day-90 miss contrast is $[-8.473, -0.987]$ percentage points. Thus an evaluation rewarding only the fraction finished by day 90 can favor the learned priority while missing worsened mean completion and individual target commitments. This is the resource-coupled evaluation problem the framework is designed to expose. The intervals have the stated per-context, per-method simultaneous simulator-conditional interpretation; they do not validate the simulator’s operational law.

At 8,192 worlds the frozen initial-workload priority is certified to dominate both alternatives in test-000 and test-003, leaving a possible frontier of size one. All three policies remain possible-frontier members in test-001 and test-002, where the metrics exhibit tradeoffs. Under the secondary 0.05-loss tolerance contract, the workload policy is certified noninferior to FIFO in test-000 and test-003. Remaining candidates and unresolved comparisons are preserved instead of being converted into an artificial universal ranking.

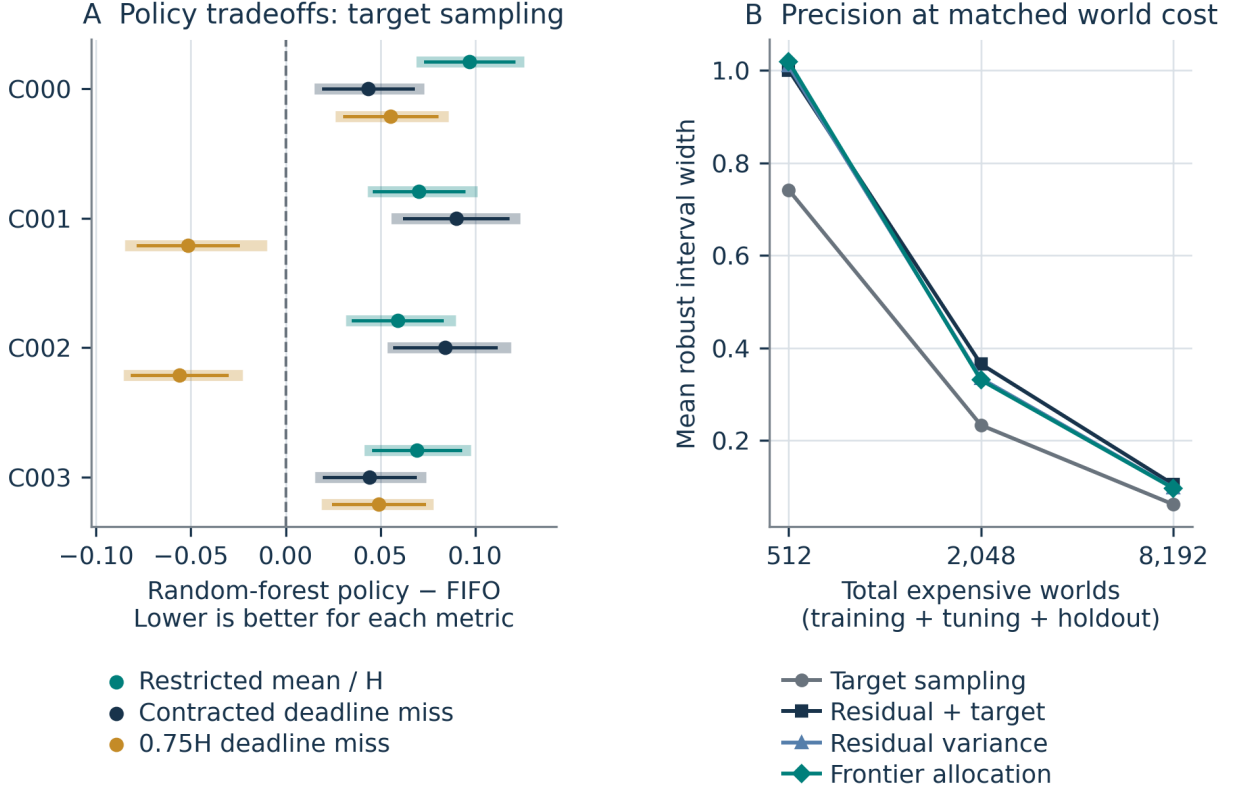


Figure 3: Policy tradeoffs and cost-matched precision on four prespecified synthetic contexts. Panel A shows random-forest priority minus FIFO under the pilot-free target method at 8,192 total worlds. Thin intervals use nominal regime masses; pale thick intervals include the TV radius 0.05 envelope. Dots show nominal estimates and smaller losses are better. Panel B averages robust widths across the four contexts at each total expensive-world budget, including training, tuning, and fresh outcomes. Cheap integration and cached execution are separate costs. Method comparisons are descriptive across contexts.

8.5 Coverage, false dominance, and power with known truth

On the separate analytic law, both methods cover every declared stratum contrast at all three stages in 100 of 100 independent replications. The exact 95 percent binomial interval for this law-specific coverage probability is $[0.9638, 1]$. Both also cover all truth-level robust extrema in 100 runs. Neither makes a false dominance assertion at any declared stage in any replication; the corresponding false-assertion probability interval is $[0, 0.0362]$. These finite checks support implementation correctness on the specified law and do not replace the propositions or demonstrate universal calibration.

Method, 4,096 total worlds	Runs with a dominance decision / 100	Mean fraction of true edges recovered	Mean robust width	Contrast RMSE
Target, no control	100	1.000	0.2019	0.004999
Frontier	80	0.545	0.3182	0.005380

Table 4: Decision power and precision on the independently specified analytic validation law. Unlike the graduation-policy experiment, this experiment has exact mean truth and a separate LP oracle. At the two earlier stages neither method makes a dominance decision. The larger width and lower decision power again contradict an efficiency advantage for the proposed allocation in the measured setting.

9 Discussion

9.1 Why cohort externalities change evaluation

The simulator exposes a failure mode that independent pilot-date regressions can miss. Shared instructor capacity couples flights and simulators, and a student’s prerequisite order determines which type of attempt can use an available slot. Priority changes therefore redistribute waiting times across pilots. A model that advances its focal student can worsen total restricted completion or the deadline-miss fraction. This motivates comparing the induced cohort distribution rather than scoring a single isolated recommendation.

The framework can be used for choosing forecasting models without asking them to schedule. It can also evaluate frozen priority policies or imported model-generated proposals, provided their actions are interpreted under a trusted resource contract. These uses share the completion outcomes and provenance conventions, but they require different evaluation samples. The present model benchmark validates predictive configurations; it does not establish the performance of an autonomous scheduling agent.

9.2 Practical scope and evidence needed for operational use

The immediate demonstration is a decision-support simulator with a defined import boundary. A local deployment can replace synthetic context fields with authorized records, retain the cutoff, and produce reproducible conditional forecasts. Before any operational accuracy claim, the simulator must reproduce syllabus accounting, calibration must be checked on genuinely held-out historical

cohorts, and disruption dynamics must be estimated without leakage from future records. Adjacent days from the same cohort should not be randomly split across training and testing.

Three distinct questions remain: does the simulator correctly implement the declared rules; does its distribution describe later operational outcomes; and does the evaluation select useful configurations at an acceptable cost? Unit tests address the first only in their covered cases. Statistical statements conditional on the simulator do not settle the second. The benchmark and independent selection protocol address the third within the synthetic experiment and cannot substitute for external validation.

9.3 Limitations

The cohorts, resources, event outcomes, and disruption frequencies are synthetic. There are no actual Air Force student records, real operational sortie outcomes, or demonstrated graduation improvements. The procedural simulation forecast uses the same correctly specified simulator that generates the labels, with the configured probabilities supplied. Its accuracy is therefore a well-specified sanity comparison, not evidence that Monte Carlo generally outperforms machine learning on operational data. Independently generated cohorts still come from one configured generator family mixture, rather than independently validated real syllabi. The simulator is deliberately compact and omits many features of real training, including detailed instructor qualifications, airspace allocation, medical holds, aircraft-specific maintenance histories, and complex syllabus branches unless provided through an extension. A predicted date is conditional on the chosen policy and these modeled dynamics.

The statistical design is adaptive only through independent pilot learning. It does not claim arbitrary online adaptivity or unrestricted anytime monitoring. Conservative simultaneous bounds can remain wide when many pairwise metrics are declared. The control coefficient is clipped and may sacrifice an optimal negative coefficient when the predictor is negatively correlated. Its quality can be assessed only on independent outcomes; the existence of a tuning step supplies no universal efficiency guarantee. Pilot, cheap integration, and computation costs can dominate the savings in expensive evaluations.

The allocation experiment covers four contexts and one recorded independent-pilot design per context. Its ablations compare four designs within the implemented estimator family; they are not a head-to-head reproduction of PPAT, R-AutoEval+, or the full range of modern sequential allocation methods. The fitted random-forest priority is one simple model-to-action adapter, not an optimized scheduling system. Broader independent contexts, repeated pilot designs, stronger scheduling comparators, and direct adaptations of the closest evaluation methods are required before any state-of-the-art efficiency claim.

The total-variation ambiguity set changes regime masses only. No mixture witness certifies real-world failure or repairs an incorrect within-regime law. The possible frontier can contain suboptimal configurations when evidence is insufficient. The fragility radius concerns the confidence envelope, not the unknown exact performance surface. Cohort-level conformal coverage requires exchangeability and may fail under structural or temporal shifts. Censoring restricts the target, and an interval for restricted completion must not be marketed as a guaranteed actual graduation date.

Finally, the work does not evaluate a large language model, a superintelligent system, or a general-purpose measure of intelligence. The term superintelligence names the broader research direction. A future advanced agent could be evaluated through the same bounded interfaces, but that extension

requires actual agent outputs, independent task construction, and a capability contract appropriate to its proposed use. Targeted literature review cannot guarantee priority; the proposed originality concerns the operational formulation, integrated evidence record, and reproducible benchmark rather than established statistical identities.

10 Conclusion

CLARK Frontier connects resource-coupled completion forecasting to evidence-based configuration choice. Its executable contract distinguishes feasibility, simulator sampling uncertainty, mixture sensitivity, predictive calibration, and selection. Independent pilot learning permits transparent simultaneous inference, while whole-cohort calibration respects shared resource dependence. The study also identifies limits: matched calibration fails under shift, a learned priority trades one graduation metric against others, and the proposed adaptive allocation loses to a simpler cost-matched baseline. These findings support evaluating increasingly capable systems by feasible operational consequences, preserving unresolved outcomes, and testing efficiency claims against strong alternatives. Operational deployment claims require independently validated training records and simulator dynamics.

11 Reproducibility and research disclosure

The accompanying source includes the daily simulator, cutoff-feature forecasting models, stratified evaluation code, TV optimizer, cohort calibration, unit and oracle tests, command-line interfaces, machine-readable scenarios, and recorded results. The root seed, software versions, split streams, source hashes, model hyperparameters, and timing conventions are retained with the experiment outputs. Re-running stochastic experiments may reproduce numerical results while CPU timings differ.

AI assistance was used for literature retrieval, mathematical exposition, implementation, and manuscript preparation. The author is responsible for independently reviewing the claims, references, code, and interpretation before public submission. No institutional affiliation with MIT and no Air Force endorsement is claimed. This is a research preprint; repository hosting and arXiv posting are separate from peer review.

12 Appendix A: proofs and technical details

12.1 A.1 Conditional unbiasedness of residual correction

Let $\mathcal{F}_0$ contain all independent training and tuning data. Conditional on $\mathcal{F}_0$, the predictor, β_j , the allocation, and all stage counts are deterministic. Write $\mu_{hj}^M = \mathbb{E}_h[M_{hj} \mid \mathcal{F}_0]$ and $\delta_{hj} = \mathbb{E}_h[D_{hj}]$. The expensive and cheap samples have the same declared within-stratum feature law, so

$$\mathbb{E}[\hat{\delta}_{hj\ell} \mid \mathcal{F}_0] = \mu_{hj}^M + \mathbb{E}_h[D_{hj} - M_{hj} \mid \mathcal{F}_0] = \delta_{hj}.$$

This argument does not require an accurate predictor. It requires that its cheap feature distribution matches that used in the expensive residual calculation. An outcome-dependent choice of stage can change the unbiasedness of the selected point estimate, even though the simultaneous interval event below still supports selection among declared stages.

If the cheap and expensive pools are independent, their conditional variance decomposes as

$$\text{Var}(\widehat{\delta}_{hj\ell} \mid \mathcal{F}_0) = \frac{\text{Var}_h M_{hj}}{N_h^C} + \frac{\text{Var}_h Z_{hj}}{n_{h\ell}}.$$

For a fixed scalar mixture and fixed predictor, minimizing the expensive-sample variance over real-valued counts with a fixed total gives counts proportional to $q_h \sqrt{\text{Var}_h Z_{hj}}$. The proposed multi-contrast frontier allocation replaces unknown quantities with pilot estimates and pools several objectives. It is therefore a heuristic extension of a classical variance-allocation calculation, not a proof of globally optimal experiment design.

12.2 A.2 Proof of Proposition 1

The empirical-Bernstein inequality for observations in a width- B interval is obtained from its unit-interval form by affine rescaling (Maurer and Pontil 2009). Applying the one-sided result to X and $-X$ with error $\delta/2$ each yields the two-sided radius used in the paper. The sample variance scales by B^2 and the linear term by B , giving exactly the stated formula. If $B = 0$, the random variable is known to be constant and its uncertainty radius is zero.

For each h, j , apply the bound to the cheap mean with failure probability δ_C . There are SJ cheap events, and their total failure probability is at most $SJ\delta_C = \alpha/2$. For each h, j, ℓ , apply the bound to the expensive residual mean with failure probability δ_E . The SJL expensive events have total failure probability at most $SJL\delta_E = \alpha/2$. Thus all cheap and expensive mean inequalities hold together with probability at least $1 - \alpha$, conditional on $\mathcal{F}_0$.

On that event, the triangle inequality gives

$$|\widehat{\delta}_{hj\ell} - \delta_{hj}| \leq |\overline{M}_{hj}^C - \mu_{hj}^M| + |\overline{Z}_{hj\ell}^E - \mathbb{E}_h Z_{hj}| \leq r_{hj\ell}.$$

Since $D_{hj} \in [-1, 1]$, intersecting with $[-1, 1]$ preserves coverage. No independence between these individual coverage events is used. Integrating the conditional failure bound over $\mathcal{F}_0$ gives the unconditional claim. This proof covers exactly the declared stages and functions; it does not license adding comparisons after inspecting the outcomes.

12.3 A.3 Proof of Proposition 2

On $\mathcal{E}$, $L_{hj\ell} \leq \delta_{hj} \leq U_{hj\ell}$ for each stratum. For any simplex vector q , nonnegativity of its coordinates implies

$$q^T L_{j\ell} \leq q^T \delta_j \leq q^T U_{j\ell}.$$

The actual q is an element of $\mathcal{Q}_\rho(p)$, so minimizing the left expression and maximizing the right expression over that set gives the stated lower and upper bounds. The derivation holds for every simplex vector on the same event. It therefore holds simultaneously for every radius without a union bound over radii. This conclusion does not extend to changing the within-stratum conditional laws, since those determine δ_j itself.

12.4 A.4 Greedy solution of the TV linear program

Consider maximizing $q^T v$ with q in the simplex and $\text{TV}(q, p) \leq \rho$. Every feasible change can be decomposed into transfers from coordinates with $q_h < p_h$ to coordinates with $q_g > p_g$. The total transferred mass is exactly $\text{TV}(q, p)$. A transfer of amount a from donor h to receiver g changes the objective by $a(v_g - v_h)$.

If a feasible solution transfers positive mass from a higher-valued donor while leaving lower-valued donor mass unused, swapping the donor cannot decrease the objective. Similarly, transferring to a lower-valued receiver while a higher-valued receiver has capacity can be exchanged without decreasing the objective. Repeated exchange yields a solution that exhausts the lowest available donors and highest available receivers first. Transfers stop when the radius is exhausted, the receiver reaches one, the donor reaches zero, or no positive objective gain remains. This is exactly the implemented sorted transfer procedure. Minimization follows by replacing v with $-v$.

The optimizer’s returned vector is a witness for the interval endpoint objective. Exactness of the LP does not turn an upper confidence endpoint into the true loss. The independent grid and general-solver comparisons test this optimization implementation on their covered cases, while the statistical argument is supplied by Proposition 1.

12.5 A.5 Proof of Proposition 3

Define $F(\rho) = \max_{q \in \mathcal{Q}_\rho(p)} q^T \delta_j$ and $U(\rho) = \max_{q \in \mathcal{Q}_\rho(p)} q^T U_{j\ell}$. On $\mathcal{E}$, Proposition 2 gives $F(\rho) \leq U(\rho)$ for every ρ . Both functions are nondecreasing because the feasible sets expand with the radius. Therefore $U(\rho) < t$ implies $F(\rho) < t$. The set of radii at which F reaches t is contained in the set at which U reaches t , proving that the confidence-envelope crossing cannot occur later than the true crossing. The all-mixtures and already-unresolved cases follow from the same inequality with the stated infimum conventions.

For multiple metrics, a single scalar crossing is not a summary of all possible tradeoffs. Report the metric, direction, threshold, confidence level, stage, nominal mixture, and witness alongside every radius. A change of loss normalization or decision tolerance changes its meaning.

12.6 A.6 Proof of Proposition 4

Condition on the independent fitting and validation-selection information. By assumption, the n calibration cohorts and the future cohort are exchangeable. Applying the same frozen prediction and maximum-score map to each preserves exchangeability of $S_1, \dots, S_n, S_*$. The rank argument for split conformal prediction gives $\Pr(S_* \leq q_\gamma) \geq 1 - \gamma$, with the infinite-threshold convention when $k > n$. Ties make this nonrandomized construction conservative.

The event $S_* \leq q_\gamma$ is equivalent to every pilot target in the future cohort lying in its expanded interval. Dependence among pilots within a cohort is unrestricted by the rank argument, because the cohort contributes one score. Clipping to a known target range cannot remove a target that was in the original interval. The result is marginal over the future cohort law; exchangeability does not imply conditional validity for every cohort subtype or validity after arbitrary distribution shift.

For $R = \min(T, H)$, the target is known even if $T > H$. For latent T under right censoring, its nonconformity score cannot generally be computed from the horizon alone. Assigning an infinite score is a conservative unresolved representation, whereas substituting H as if it were a completion date would change the target and invalidate the stated interpretation.

12.7 A.7 Reproduction and audit table

Item	Recorded object	Interpretation
Forecast cutoff	Scenario start and accomplishment snapshot	Later records must not enter features
Training and tuning	Separate random streams and outcome counts	Excluded from final interval samples
Cheap integration	Feature counts and bounded predictor artifact	Uncertainty is included, not treated as exact
Evaluation stages	Declared budgets and deterministic stratum counts	Finite simultaneous stage guarantee
Model selection	Validation error gate and latency rule	Frozen before calibration and testing
Calibration	One maximum score per whole cohort	Joint cohort marginal coverage
Censoring	Completion null, restricted target, censor flag	Restricted time is not actual graduation
Robustness	Nominal masses, TV radius, conditional regime definitions	Only regime mixture probabilities vary
Model identity	Fitted artifact, hyperparameters, software versions	No unrecorded model substitution
Compute	Pilot outcomes, fresh outcomes, cheap calls, CPU timings	Separate cold and amortized costs
Reference checks	Exact tiny cases and independent TV solver	No exact policy-performance truth is assumed

12.8 A.8 Why a perfect control can widen this certificate

This diagnostic concerns the generic support bounds used here, not an inherent defect of control variates. Consider one contrast $D \in \{-\varepsilon, +\varepsilon\}$ with $0 < \varepsilon < 1$. It can be represented by bounded losses $Y_a = (1 + D)/2$ and $Y_b = (1 - D)/2$. Suppose a hypothetical perfect control has $M = D$, $\beta = 1$, and hence $Z = 0$. This construction is not used as an empirical predictor in the experiments.

At a common fresh sample size n , write $C = \log(4/\delta_E)$. The implemented generic residual width is four for $\beta = 1$. Its residual radius alone is therefore $28C/[3(n - 1)]$, even though the observed residual variance vanishes. Cheap-mean uncertainty adds a nonnegative term. The no-control width is two, and the unbiased sample variance satisfies $s_D^2 \leq n\varepsilon^2/(n - 1)$, so its radius is at most

$$\varepsilon \sqrt{\frac{2C}{n - 1}} + \frac{14C}{3(n - 1)}.$$

Consequently, when $\varepsilon < \frac{14}{3\sqrt{2}}\sqrt{C/(n - 1)}$, the perfect-control radius exceeds the no-control radius. Choose sample sizes and ε so that interval truncation at $[-1, 1]$ is inactive. This supplies a finite-sample counterexample to a universal certificate-width improvement for the implemented generic bounds, even before charging pilot outcomes. If the global identity $M = D$ were independently

known, a justified residual support width of zero would remove that penalty. The issue is the declared bound and its tuning objective, not zero residual variance itself.

References

- Ahn, Dohyun, Huiyi Chen, and Lewen Zheng. 2026. “Wasserstein Distributionally Robust Rare-Event Simulation.” <https://arxiv.org/abs/2601.01642v1>.
- Alshammari, Shaden, Kevin Wen, Abrar Zainal, Mark Hamilton, Navid Safaei, Sultan Albarakati, William T. Freeman, and Antonio Torralba. 2026. “MathNet: A Global Multimodal Benchmark for Mathematical Reasoning and Retrieval.” In *International Conference on Learning Representations*. <https://arxiv.org/abs/2604.18584v2>.
- Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnica. 2023. “Prediction-Powered Inference.” *Science* 382: 669–74. <https://doi.org/10.1126/science.ad16000>.
- Angelopoulos, Anastasios N., John C. Duchi, and Tijana Zrnica. 2024. “PPI++: Efficient Prediction-Powered Inference.” <https://arxiv.org/abs/2311.01453v2>.
- Ashouritaklimi, Kianoosh, Valentin Kilian, Daolang Huang, Tom Rainforth, and Francois Caron. 2026. “Prediction-Powered Active Testing.” <https://arxiv.org/abs/2607.08347v1>.
- Balkir, Esma, Alice Pernthaller, Marco Basaldella, Jose Hernandez-Orallo, and Nigel Collier. 2026. “Confident Rankings with Fewer Items: Adaptive LLM Evaluation with Continuous Scores.” <https://arxiv.org/abs/2601.13885v2>.
- 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>.
- Berrada, Gabrielle, Jannik Kossen, Freddie Bickford Smith, Muhammed Razzak, Yarin Gal, and Tom Rainforth. 2025. “Scaling up Active Testing to Large Language Models.” <https://arxiv.org/abs/2508.09093v2>.
- 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/>.
- Center for AI Safety, Scale AI, and HLE Contributors Consortium. 2026. “A Benchmark of Expert-Level Academic Questions to Assess AI Capabilities.” *Nature* 649: 1139–46. <https://doi.org/10.1038/s41586-025-09962-4>.
- Dekoninck, Jasper, Nikola Jovanovic, Tim Gehrunger, Kari Rognvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. 2026. “Beyond Benchmarks: MathArena as an Evaluation Platform for Mathematics with LLMs.” <https://arxiv.org/abs/2605.00674v2>.
- Durmazkeser, Yavuz, Patrik Okanovic, Andreas Kirsch, Torsten Hoefer, and Nezihe Merve Gurel. 2026. “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>.
- 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/>.
- Glazer, Elliot, Ege Erdil, Tamay Besiroglu, et al. 2025. “FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI.” <https://arxiv.org/abs/2411.04872v7>.
- 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>.
- Kilian, Valentin, Stefano Cortinovis, and Francois Caron. 2025. “Anytime-Valid, Bayes-Assisted, Prediction-Powered Inference.” <https://arxiv.org/abs/2505.18000v2>.
- . 2026. “Asymptotically Log-Optimal Bayes-Assisted Confidence Sequences for Bounded Means.” <https://arxiv.org/abs/2605.07964v2>.
- 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>.
- Lee, Yonghoon, and Zhimei Ren. 2025. “Selection from Hierarchical Data with Conformal e-Values.” <https://arxiv.org/abs/2501.02514v1>.
- 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>.
- 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>.
- Mallick, Soham, Eric Tchetgen Tchetgen, Edgar Dobriban, and Yonghoon Lee. 2026. “Generalized Hierarchical Conformal Prediction.” <https://arxiv.org/abs/2608.15500v1>.
- Maurer, Andreas, and Massimiliano Pontil. 2009. “Empirical Bernstein Bounds and Sample Variance Penalization.” In *Proceedings of the 22nd Annual Conference on Learning Theory*. <https://arxiv.org/abs/0907.3740v1>.
- 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/>.
- Park, Sangwoo, Matteo Zecchin, and Osvaldo Simeone. 2025. “Adaptive Prediction-Powered AutoEval with Reliability and Efficiency Guarantees.” <https://arxiv.org/abs/2505.18659v2>.
- Sfyraki, Maria-Eleni, and Jun-Kun Wang. 2026. “Revisiting Active Sequential Prediction-Powered Mean Estimation.” <https://arxiv.org/abs/2604.18569v1>.
- Tu, Daohong, and Kay Giesecke. 2026. “Online Conformal Prediction for Non-Exchangeable Panel Data.” <https://arxiv.org/abs/2605.17705v2>.
- Xu, Shirong, and Will Wei Sun. 2026. “Do More Predictions Improve Statistical Inference? Filtered Prediction-Powered Inference.” <https://arxiv.org/abs/2602.10464v1>.
- Zhang, Chunxu, Tiantian Mao, and Ruodu Wang. 2026. “Wasserstein Distributionally Robust Quantile Regression.” <https://arxiv.org/abs/2603.14991v1>.
- Zhang, Haoting, Jinghai He, Rhonda Righter, and Zeyu Zheng. 2024. “Language Model Prompt Selection via Simulation Optimization.” <https://arxiv.org/abs/2404.08164v2>.